CIC-NLP at GenAI Detection Task 1: Advancing Multilingual Machine-Generated Text Detection

Abiola T. O[1], Tewodros A. B[1], Fatima Uroosa[1], Nida Hafeez[1], Ojo O. E.[1], Sidorov G.[1], Kolesnikova O.[1]

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico.

kolesnikova@cic.ipn.mx

Abstract

Machine-written texts are gradually becoming indistinguishable from human-generated texts, leading to the need to use sophisticated methods to detect them. Team CIC-NLP presents work in the Gen-AI Content Detection Task 1 at COLING 2025 Workshop: the focus of our work is on Subtask B of Task 1, which is the classification of text written by machines and human authors, with particular attention paid to identifying multilingual binary classification problem. Usng mBERT, we addressed the binary classification task using the dataset provided by the GenAI Detection Task team. mBERT acchieved a macro-average F1-score of 0.72 as well as an accuracy score of 0.73.

1 Introduction

Several researchers have worked on various binary classification tasks using ML models and LLMs in NLP, focusing on different areas such as hate speech detection (Zamir et al., 2024a; Ahani et al., 2024; Tonja et al., 2022; Ojo et al., 2022), sentiment analysis (Zhang et al., 2023; Hadi et al., 2024), fake news detection (Zamir et al., 2024b; Kanta and Sidorov, 2023), and hope speech identification (Tash et al., 2024a). These efforts aim to discern the nuanced aspects of human communication. Some of these classification tasks have been conducted on non-English and multilingual texts (Kanta and Sidorov, 2023; Ojo et al., 2023; Kolesnikova et al., 2024).

With the advancements in Large Language Models (LLMs), machine-generated content across various platforms, including news outlets, social media, educational, and academic publications (He et al., 2023) has reached an outstanding quality. Recent models like ChatGPT, GPT-4 (OpenAI, 2023), LLaMA 2 (Touvron et al., 2023), and Jais (Sengupta et al., 2023) generated remarkable coherence in responding to diverse user queries. This rapid advancement has raised concerns about the potential misuse of machine-generated text in different fields such as journalism, education, and academia (Uchendu et al., 2023; Crothers et al., 2023b), in addition to the influence on operations (Goldstein et al., 2023), disinformation (Buchanan et al., 2021), spam, or unethical authorship (Crothers et al., 2023a). Moreover, it poses challenges to maintain information integrity and ensure the accuracy of shared information. Consequently, the ability to effectively distinguish between human-generated and machine-generated content has become crucial for detecting possible instances of misuse (Jawahar et al., 2020; Stiff and Johansson, 2022; Macko et al., 2023). While significant progress has been made in detecting machine-generated text in English, we still need to improve it in multilingual settings.

In response to this gap, COLING Workshop organizers launched Gen-AI Content Detection Task 1: This shared Gen-AI Content Detection Task 1 introduces a new Binary Multilingual MGT Detection challenge to accelerate research in this area and improve cross-lingual detection capabilities (Wang et al., 2025) (Chowdhury et al., 2025) (Dugan et al., 2025). Being a shared task, it brings together researchers and practitioners interested in detecting machine-generated content reliably in many languages, reflecting the collaborative spirit and multidisciplinary innovation of shared tasks. At the broader level, the Gen-AI Content Detection Task 1 also highlights the importance of machine-generated text (MGT) detection. Also, it addresses the problem of keeping content authentic, fighting misinformation, and driving ethical use cases of AI in the multilingual realm. As CIC-NLP team, we used mBERT to detect and classify MGT as distinguished from human-generated text (HWT), the method used and results obtained are extensively highlighted in other sections of this report.

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 262–270
January 19, 2025.

2 Literature Review

Over the last few years there has been a great focus in the use of language models which in turn has created the need for keen classification of authentic and fake texts; this was historically stated mostly as a binary problem. The GenAI Detection Task 1 includes distinguishing between text written by a human and text written by a computer. There are two key approaches that have been broadly applied to text classification : classification with supervised methods (Kolesnikova and Gelbukh, 2019; Gelbukh and Kolesnikova, 2010; Kolesnikova and Gelbukh, 2010; Adebanji et al., 2022; Ojo et al., 2020; Gutiérrez-Hinojosa et al., 2023) and unsupervised (zero-shot) methods (Ojo et al., 2024a,b; Calvo and Gelbukh, 2004). Supervised methods normally do better in terms of accuracy but are more likely to overfit, particularly when new language structures are used (Su et al., 2023). On the other hand, unsupervised methods offer flexibility due to the absence of label information, however, they might call for impractical white-box access to the generating model.

Huge advancements in LLMs are currently driven by various platforms such as ChatGPT powered by GPT-3.5, GPT-4 (OpenAI, 2023), OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), LaMDA (y Arcas, 2022), and BLOOM, and emergent models like Vicuna (Zheng et al., 2023) and Alpaca (Taori et al., 2023). These models containing millions to billions of parameters are trained on huge amounts of data, have shown extraordinary results across multiple fields including finance, customer support, and the educational sector. Some of their most impressive features include their ability to write text that references human-generated text so closely that most people will initially find it hard to distinguish between the two. Also, it is possible to note that their multilingual skills enable them to generate clear and high-quality text in more than fifty languages (Workshop et al., 2022), thus making them more and more appropriate in the global business environment, but at the same time posing even higher problems to MGT detection.

To the best of our knowledge, several benchmarks have been proposed to assess multilingual MGT detection models in different languages (Wang et al., 2024). For example, the Human ChatGPT Comparison Corpus (HC3) (Guo et al., 2023) compares ChatGPT-generated text and human-written text, with authentications of English and Chinese languages using logistic regression models and RoBERTa-based classifiers built from features of Giant Language Model Test Room (GLTR). Others have replicated such approaches by testing other detectors including RoBERTa, XLM-R (Conneau, 2019), logistic regression based on features from NELA and other stylometric classifiers (Li et al., 2014; Horne et al., 2019). MULTITuDE has also been introduced by researchers within the news domain for 11 languages that offers a strong test bed for multilingual detection baselines (Macko et al., 2023). To detect MGT, researchers released benchmark environment (Uchendu et al., 2021) (Jawahar et al., 2020) to compare machine-generated text detection across multiple languages using monolingual and multilingual BERT models, which is consistent. As a result of comparison, it was found that multilingual-specific models tend to perform better than others. (Ruder et al., 2021) discussed challenges in multilingual NLP tasks and strategies for model adaptation across languages. While their work on sentiment analysis is almost exclusively concerned with model adaptation, their observations about the problem of improving machine-generated sentences are relevant for our work. Also, according to current literature, transformer models, such as LoRA-RoBERTa and XLM-RoBERTa, are found to be more accurate compared to classical machine learning techniques in multilingual MGT detection tasks, see for example (Xiong et al., 2024).

To summarize, researchers have been able to refine their methods of distinguishing human writing from computer scripts by integrating statistical analysis with other language models. The further development of these approaches proves that there are still challenges to differentiating between the advanced results produced by LLMs and works created by humans. Prior work has mainly considered the classification of synthetic text in few languages, certain LLMs, or certain domains like news (Zellers et al., 2019). Our work extends this scope to multiple languages and include a range of diverse and popular LLMs across different domains. To sum up, the previous methods and works provided useful information regarding the efficiency of various approaches to identifying AI vs. human written text, but more works required.

3 Methodology

We deployed multiple NLP techniques for data preprocessing, detection, and sorting to assess the performance of our approach to the Binary Multilingual Machine-Generated Text Detection task in the context of transformer-based models. Next, we loaded and preprocessed a broad multilingual dataset to normalize input formats and then applied language detection to guarantee certain types of processing on specific languages. We tokenized text into language-appropriate segments, translated text between language pairs, and sorted operations adjusted to that language's unique characteristics in this Gen-AI Content Detection Task 1. Using the mBert language model, which we pretrained and fine-tuned on the provided training datasets, we enhanced the model output by carefully approaching various linguistic constructs. Focusing on efficient management of code-mixed and pure multilingual data, our methodology determined the tokenization method by polyglot such that each input is associated with a particular language.

3.1 Dataset Analysis

The dataset provided for the Binary Multilingual Machine-Generated Text Detection task includes text data across nine diverse languages. This linguistic diversity adds complexity to Gen-AI Content Detection Task 1 (binary multilingual classification problem), requiring models that can handle varying scripts, grammatical structures, and cultural nuances in text patterns. The diversity of the training dataset used in this Gen-AI Content Detection Task 1 is further highlighted in (Wang et al., 2025) (Chowdhury et al., 2025) (Dugan et al., 2025) .

3.1.1 Language Distribution

The dataset is balanced concerning languages so that models trained on it can generalize to multilingual text. Each language presents unique challenges: Arabic and Urdu are right-to-left languages, their grammar is more than complex in script; German and Russian have more intricate grammar and syntactical structure.

3.1.2 Content Sources

The dataset contains text samples that are extracted from online sources such as social media posts, articles from web pages, and other digital content. This variety corresponds to the broad range of text that models may encounter in actual deployment, from informal posts to more structured forms of article-style content.

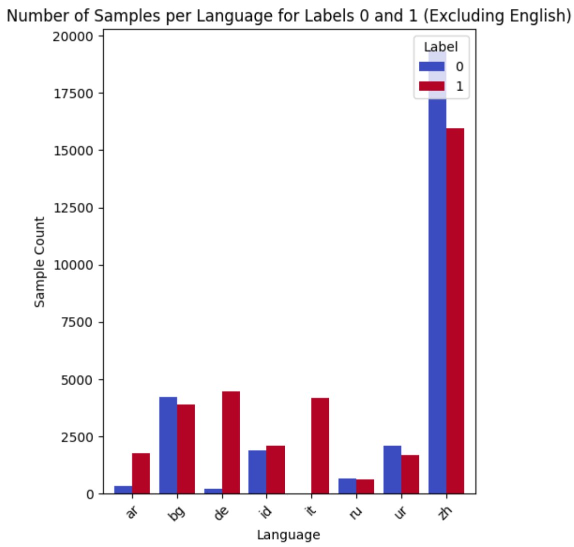

Figure 1: Languages in the training dataset (English excluded)

3.1.3 Class Labels

The dataset is labeled with each entry as human-generated (1) or machine-generated (0). Due to these binary labels, this is a simple classification task for models to learn to distinguish fine-grained features associated with machine generation (such as repetitive phrasing and lower variation in tone).

3.1.4 Tokenization and Script Variability

The dataset is multilingual; we adopted a polyglot multilingual tokenizer to segment the texts. This tokenizer was found reliable even with Chinese texts that have no spacing and also with some Cyrillic systems. Arabic and Urdu are abjad-based scripts, meaning they provide mostly consonants and have fewer vowels, and all these disparities in the multilingual text data were accounted for to enhance the model in classification.

3.1.5 Potential Language-Specific Features

Machine-generated text may have different features that show the distinct characteristics of each language. For example, Arabic and Urdu have complex morphology, which can show up in stylistic differences in machine-generated text compared to English and German, so detecting human vs. machine-generated text for these languages would be subtler.

3.1.6 Text Length and Complexity

The text lengths and simplicity levels of the dataset are probably very different from those of the online sources. Short, informal texts (e.g., social media snippets) and longer, structured articles offer diverse linguistic challenges. The variety supports the ability to train models that can process variations of text lengths and learn the stylistic characteristics inherent to machine generation for each language.

3.1.7 Class Imbalance

One of the main factors of the dataset's design is the balance among languages and classes. When evaluating the model, a macro-average F1-score will handle minor imbalances and ensure the model's robust performance in all languages involved.

The dataset has a multilingual, balanced structure to capture languages and to train a model for working well with many linguistic backgrounds. Our analysis of languages and classes in this section serves as a basis for understanding the diversity within the dataset and for building preprocessing pipelines and model architectures suited for varied language patterns and scripts.

3.2 Shared Task Description

The Binary Multilingual Machine-Generated Text Detection Task, part of Gen-AI Content Detection Task 1, gave the participants a rich multilingual dataset for distinguishing human- and machine-generated text. The dataset contains content from various languages across domains, including social media, news articles, and educational materials. We had machine-generated or human-authored labels for each text entry; we carefully labeled them in the binary classification tasks.

Finally, we participated in developing models to reliably detect machine-generated text across different languages, evidence of the need for cross-lingual detection abilities. The macro-average F1-score was evaluated as a metric based on precision and recall while covering multiple languages and text types. The purpose of the shared Gen-AI Content Detection Task 1 was to develop multilingual capabilities for machine-generated text detection with the growing demand for authenticity in multilingual digital content and for innovations in reliably detecting AI-generated content within different linguistic contexts.

3.3 Model Architecture

Our model architecture is built on fine-tuning mBERT for multilingual GenAI Detection Task 1, with a focus on the binary classification of MGT and HWT with the challenge of making it robust to efficiently classify different languages. For this purpose we chose the mBERT-cased version, a choice for dealing with more than one language, including less-resourced ones. This architecture integrates three primary modules: language detection and tokenization with polyglot, and training and prediction with mBERT which we optimized to capture different languages in the datasets and unseen ones that surfaced in the test dataset.

We trained the model with a few meticulously chosen hyperparameters for optimizing the training process with the ADAM optimizer and adjusting the learning rate to best suite the classification. The metrics we used are exclusively listed in the appendix section of this paper. The categorical cross-entropy was used as a loss function, and the batch size was well adjusted to maximize the computing resources available as well as prevent overfitting. We also adopted early stopping to prevent overfitting using validation performance-enabled training across the three epochs. Three epochs were used as a time factor and computational resources at our disposal were considered. The model was engineered to be computationally fast and memory efficient overall. Its design makes it scalable to the large datasets provided and maintains high performance.

3.4 Experimental Setup

Our experiments employed a training validation split on the multilingual set, configured language-specific preprocessing rules, and set up the model in a high-performance computing environment. The translation and detection models were initialized and then fine-tuned using the training dataset to capture multilingual patterns using weights from pre-trained models. We built a complete evaluation pipeline to monitor model performance in each language and used accuracy and F1-score as critical metrics. Unseen test data were used for model evaluation and generalization. Additionally, the model was evaluated in language-agnostic embeddings, using multiple languages and contexts to show robustness. The model hyper-parameters have also been experimentally optimized to trade precision with increased computational efficiency.

Epoch	Training Loss	Validation Loss	F1-score
1	0.200	0.241	0.916
2	0.093	0.286	0.925
3	0.046	0.155	0.953

Table 1: Metrics generated by the model during training

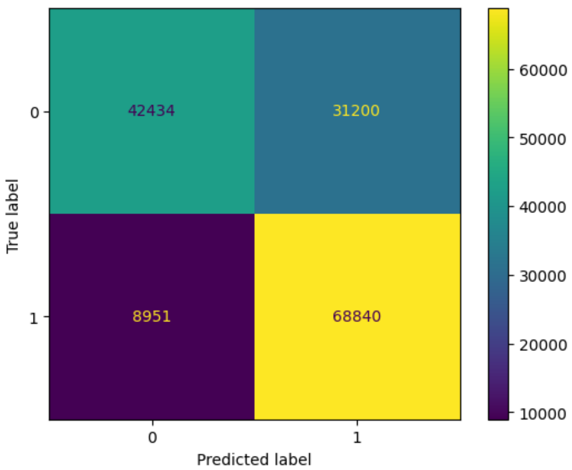

Figure 2: Confusion matrix

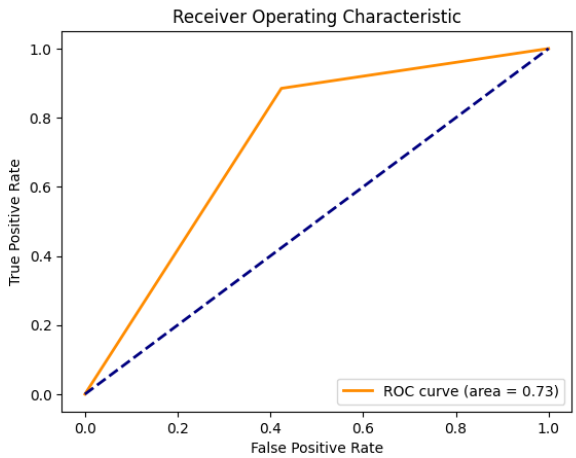

Figure 3: ROC curve

Model	Micro_F1	Macro_F1
mBERT	0.734	0.726

Table 2: Result obtained from the test set

3.5 Predictions on Unseen Data

We evaluated the model's generalizability on unseen data with text entries in multiple languages. The model generated predictions to see how it translated, sorted and identified machine-generated text. The model's output for translation with linguistic accuracy, language-specific sorting correctness, and detection precision were analyzed. We showed that the multilingual model preserves language nuances, sorts accurately, and identifies machine-generated text reliably on a diverse set of language pairs. We found that language performance differed slightly in low-resource languages, but the model met the multilingual detection benchmarks of the shared Gen-AI Content Detection Task 1.

4 Results

On the test set, the model predicted the classes with an accuracy of 0.7348 and a macro-average F1-score of 0.7265, which indicates balanced test performance across the languages. Our results also demonstrate that the model is capable of handling multiple languages without much performance degradation. We present figure 2 showing the confusion matrix for better analysis of the model predictions as it revealed the model strength towards accurately predicting MGT with accuracy of 0.88 and the model got weak results by confusing some HWT for MGT with accuracy of 0.58.

5 Conclusion

This paper shows how a multilingual transformer-based model detects machine-generated text in various languages. Our results confirm the model's adaptability and scalability and evidence to its promising performance in high-resource languages and its potential for improvement in low-resource scenarios. We show that with appropriate data preprocessing, machine-generated text detection can be successfully extended to multilingual applications using fine-tuning and balanced datasets. This work will be continued to improve the performance for low-resource languages and deploy the model to handle more complex linguistic features such as code-switching and mixed scripts.

Ethics Statement

This paper is fully committed to transparency and ethical AI utilization, especially in multilingual digital content authentication. Ethical responsibility must be first prioritized for machine-generated text detection, as wrong classifications may impact individuals and organizations. However, we take the responsible use of our model seriously and want feedback on minimizing any negative impacts. A primary goal is to add value to online digital content verification, combatting misinformation while paying due respect to the plurality of the linguistic scopes in online media.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

O.O. Adebanji, I. Gelbukh, H. Calvo, and O.E. Ojo. 2022. Sequential models for sentiment analysis: A comparative study. In O. Pichardo Lagunas, J. Martínez-Miranda, and B. Martínez Seis, editors, *Advances in Computational Intelligence. MICAI 2022. Lecture Notes in Computer Science*, volume 13613, pages 220–231. Springer, Cham.

Z. Ahani, M. Tash, M. Zamir, and I. Gelbukh. 2024. Zavira@ dravidianlangtech 2024: Telugu hate speech detection using lstm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 107–112.

NiCole T. Buchanan, Marisol Perez, Mitchell J. Prinstein, and Idia B. Thurston. 2021. Upending racism in psychological science: Strategies to change how science is conducted, reported, reviewed, and disseminated. *American Psychologist*, 76(7):1097–1112.

Hiram Calvo and Alexander Gelbukh. 2004. Unsupervised learning of ontology-linked selectional preferences. In *Progress in Pattern Recognition, Image Analysis and Applications: 9th Iberoamerican Congress on Pattern Recognition, CIARP 2004, Puebla, Mexico, October 26-29, 2004. Proceedings 9*, pages 418–424. Springer.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Shammur Absar Chowdhury, Hind Al-Merekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, Georgios Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

A. Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

E. Crothers, N. Japkowicz, and H. Viktor. 2023a. Machine-generated text: A comprehensive survey of threat models and detection methods. *arXiv preprint arXiv:2210.07321*.

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

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Callison-Burch Chris. 2025. Genai content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Alexander Gelbukh and Olga Kolesnikova. 2010. Supervised learning for semantic classification of spanish collocations. In J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, and J. Kittler, editors, *Advances in Pattern Recognition. MCPR 2010. Lecture Notes in Computer Science*, volume 6256, pages 311–320. Springer, Berlin, Heidelberg.

J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Sandra J. Gutiérrez-Hinojosa, Hiram Calvo, and Marco A. Moreno-Armendáriz. 2023. Automatic extractive summaries using supervised learning approach. *Open Review Journal*.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1):1–23.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Selam Kanta and Grigori Sidorov. 2023. Selam@ dravidianlangtech: Sentiment analysis of code-mixed dravidian texts using svm classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

Olga Kolesnikova and Alexander Gelbukh. 2010. Supervised machine learning for predicting the meaning of verb-noun combinations in spanish. In G. Sidorov, A. Hernández Aguirre, and C.A. Reyes García, editors, *Advances in Soft Computing. MICAI 2010. Lecture Notes in Computer Science*, volume 6438, pages 260–269. Springer, Berlin, Heidelberg.

Olga Kolesnikova and Alexander Gelbukh. 2019. A study of lexical function detection with word2vec and supervised machine learning. *Special Section: Selected Papers of LKE 2019*.

Olga Kolesnikova, Mesay Gemeda Yigezu, Alexander Gelbukh, Selam Abitte, and Grigori Sidorov. 2024. Detecting multilingual hate speech targeting immigrants and women on twitter. *Journal of Intelligent and Fuzzy Systems*, 10.3233(IFS-219350).

Jenny S. Li, John V. Monaco, Li-Chiou Chen, and Charles C. Tappert. 2014. Authorship authentication using short messages from social networking sites. In *Proceedings of the 2014 IEEE 11th International Conference on e-Business Engineering*, pages 314–319. IEEE.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, et al. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

O. E. Ojo, T.-H. Ta, A. Gelbukh, H. Calvo, G. Sidorov, and O. O. Adebanji. 2022. Automatic hate speech detection using deep neural networks and word embedding. *Computación y Sistemas*, 26(2):1007–1013.

O.E. Ojo, A. Gelbukh, H. Calvo, O.O. Adebanji, and G. Sidorov. 2020. Sentiment detection in economics texts. In L. Martínez-Villaseñor, O. Herrera-Alcántara, H. Ponce, and F.A. Castro-Espinoza, editors, *Advances in Computational Intelligence. MICAI 2020. Lecture Notes in Computer Science*, volume 12469, pages 318–329. Springer, Cham.

Olumide E. Ojo, Olaronke O. Adebanji, Hiram Calvo, Damian O. Dieke, Olumuyiwa E. Ojo, Seye E. Akinsanya, Tolulope O. Abiola, and Anna Feldman. 2023. Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*. Association for Computational Linguistics.

Olumide E. Ojo, Olaronke O. Adebanji, Alexander Gelbukh, Hiram Calvo, and Anna Feldman. 2024a. Medai dialog corpus (medic): Zero-shot classification of doctor and ai responses in health consultations. *Preprint*, arXiv:2310.12489.

Olumide Ebenezer Ojo, Olaronke Oluwayemisi Adebanji, Alexander Gelbukh, Hiram Calvo, and Anna Feldman. 2024b. Evaluating embeddings for one-shot classification of doctor-ai consultations. *Preprint*, arXiv:2402.04442.

R. OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

H. Stiff and F. Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

M. Tash, Z. Ahani, M. Zamir, O. Kolesnikova, and G. Sidorov. 2024a. Lidoma@ lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189.

A. L. Tonja, M. Arif, O. Kolesnikova, A. Gelbukh, and G. Sidorov. 2022. Detection of aggressive and violent incidents from social media in spanish using pre-trained language model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings*. CEUR-WS.org.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, et al. 2024. SemEval-2024 Task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Feng Xiong, Thanet Markchom, Ziwei Zheng, Subin Jung, Varun Ojha, and Huizhi Liang. 2024. NCL-UoR at SemEval-2024 Task 8: Fine-tuning large language models for multigenerator, multidomain, and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 163–169.

Blaise Agüera y Arcas. 2022. Do large language models understand us? *Daedalus*, 151(2):183–197.

M. Zamir, M. Tash, Z. Ahani, A. Gelbukh, and G. Sidorov. 2024a. Lidoma@ dravidianlangtech 2024: Identifying hate speech in telugu code-mixed: A bert multilingual. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 101–106.

M. Zamir, M. Tash, Z. Ahani, A. Gelbukh, and G. Sidorov. 2024b. Tayyab@ dravidianlangtech 2024: Detecting fake news in malayalam lstm approach and challenges. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 113–118.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in Neural Information Processing Systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al. 2023. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998*.

A Appendix

Training Arguments

```
training_args = TrainingArguments(
output_dir=output_dir,
evaluation_strategy="epoch",
save_strategy='epoch',
load_best_model_at_end=True,
learning_rate=2e-5,
per_device_train_batch_size=128,
per_device_eval_batch_size=128,
num_train_epochs=3,
weight_decay=0.01,
logging_dir='./logs',
logging_steps=10,
fp16=True, Enable mixed precision
gradient_accumulation_steps=2,
)
```

Tokenizer

```
tokenizer = AutoTokenizer.from_pretrained('bert-base-multilingual-cased')
```

```
def tokenize_function(examples):
encoding = tokenizer(
examples["tokens"],
padding="max_length",
truncation=True,
is_split_into_words=True,
max_length=512
)
encoding["labels"] = examples["label"]
encoding["id"] = examples["id"]
return encoding
tokenized_train n̄ew_ds['train'].map(tokenize_function,
batched=True, num_proc=8)
tokenized_dev n̄ew_ds['dev'].map(tokenize_function,
batched=True, num_proc=8)
tokenized_test t̄okenized_test.remove_columns(["tokens"])
```