Human vs LLM - Text Detection

1st Dr. Kalyani Selvarajah
Master of Applied Computing
University of Windsor
Windsor, Canada

2nd Abishek Joshua Tennyson
Master of Applied Computing
University of Windsor
Windsor, Canada

3rd Harsh Hareshbhai Isamalia
Master of Applied Computing
University of Windsor
Windsor, Canada

4th Juanita Melosha Kingsly Vijay
Master of Applied Computing
University of Windsor
Windsor, Canada

5th Kathar Patcha Abdul Rahim
Master of Applied Computing
University of Windsor
Windsor, Canada

***Abstract*—This paper summarises findings from multiple research projects focused on the problem of separating texts created by humans from those generated by Language Large Models (LLMs), such ChatGPT. It examines how different classification schemes that aim to precisely determine the source of text are developed and put into use. By improving our capacity to quickly and effectively confirm the sources of textual information, the main goal of these studies is to strengthen the dependability and credibility of digital content. Research on binary and multi-class classification models shows promise in reaching this objective, but it also highlights the difficulties and complexities that remain in this rapidly evolving subject.**

***Index Terms*—Text Detection Large Language Models (LLMs) Binary Classification Multi-class Classification Natural Language Processing (NLP) Machine Learning Digital Trust Content Verification Artificial Intelligence (AI) Computational Efficiency**

I. Introduction

In the digital era, the distinction between content generated by humans and that created by advanced artificial intelligence systems, particularly Large Language Models (LLMs) like ChatGPT, has become increasingly blurred. This convergence presents both significant opportunities and challenges across various domains, including digital security, content authenticity, and intellectual property rights. As LLMs become more prevalent, ensuring the integrity of information disseminated across digital platforms is paramount.

Creating reliable machine learning models that can accurately distinguish between texts created by AI and those created by humans is the main goal of this project. With artificial intelligence (AI) technologies becoming more and more ingrained in our information systems, this capability is essential for preserving digital trust and guaranteeing the trustworthiness of material.

It is imperative to find scalable and effective ways to handle the complexity that LLMs bring, and this is the fundamental reason for the project. The amount and speed at which digital content is produced makes traditional approaches, which frequently need manual verification, unsustainable. This project intends to improve the efficiency and dependability of digital content monitoring and management systems by automating the text origin verification process by utilising cutting-edge AI techniques.

II. Literature Review

The Human vs LLM Text Detection project's literature evaluation covers a broad spectrum of studies concerning the recognition and categorization of text produced by people and Large Language Models (LLMs). The difficulties presented by LLMs with advanced capabilities, such OpenAI's ChatGPT, which can generate language that resembles human writing styles, have drawn a lot of interest to this field.

A. Advancements in Machine Learning Models for Text Classification:

Creating and improving machine learning models that can distinguish between writings written by humans and those created by artificial intelligence (AI) is a primary area of interest in current research. Numerous investigations have looked into the use of more sophisticated neural network models like Multi-Layer Perceptrons (MLP) and more conventional machine learning approaches like Logistic Regression and Stochastic Gradient Descent (SGD) classifiers. Transformer-based models like DistilBERT, which have deep learning capabilities and are especially effective at processing and comprehending natural language, have also been adopted in recent advances.

B. Importance of Feature Extraction in Text Detection Systems:

Furthermore, a lot of research has been done on the value of feature extraction in enhancing text detection algorithms. Principal Component Analysis (PCA) and word embeddings from models like Word2Vec, as well as dimensionality reduction techniques, have proved crucial. With the purpose of precisely recognising the text's origin, these methods aid in capturing the nuances of the text's syntactic and semantic structure.

C. Dataset Challenges and the Need for Diversity in Text Detection:

The evaluated papers also emphasise how important it is to maintain reliable datasets that fairly represent the variety of texts produced by humans and artificial intelligence. Making sure that different writing styles are fairly represented in the dataset and getting rid of biases that could affect the model's performance during training are among the problems associated with dataset curation.

D. Future Research Directions in Text Detection Technologies:

In general, the extant literature offers a firm basis for comprehending the present capacities and constraints of text detection models. It also indicates prospective avenues for future investigation, including augmenting model precision, broadening the range of datasets, and optimising computational techniques to manage the growing intricacy of language models.

III. Methodology

"Human vs LLM Text Detection" employs a structured methodology aimed at creating reliable classifiers that can discriminate between text created by humans and AI. Using best practices in data science and machine learning is ensured by the thorough and multifaceted approach. The following is a synopsis of the methodology's principal elements:

A. Data Collection and Preprocessing:

The project's primary source of data is Kaggle, which provides about 800,000 text entries produced by people and different LLMs. Among the preprocessing actions are:

- Data cleaning is the process of eliminating any superfluous or unnecessary information to guarantee the dataset's quality.
- Feature engineering is the process of enhancing the data representation by extracting attributes like readability scores and lexical diversity.

B. Feature Extraction:

- Text embeddings are transformed into numerical vectors that capture syntactic and semantic subtleties by using Google's Word2Vec tool for feature extraction.
- Dimensionality reduction: Principal Component Analysis (PCA) is used to simplify data while preserving important information, improving the models' efficacy and efficiency.

C. Model Development and Selection:

- Binary Classification: The project begins with less complex models such as SGD Classifier and Logistic Regression. Iteratively, iteratively, iteratively improves its strategy until it finally adopts the Multi-Layer Perceptron (MLP) because of its superior performance, as shown by its 86.33 percent accuracy rate.
- Multi-Class Classification: DistilBERT, a transformer-based model well-known for its effectiveness in processing textual data, is used by the project to discover particular LLMs. Even with its first reasonable success rate (48 percent accuracy), there remains room for improvement.

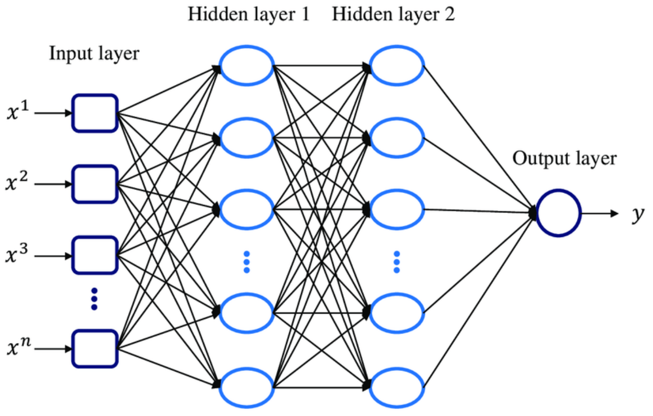


Fig. 1. Representation of the MLP architecture

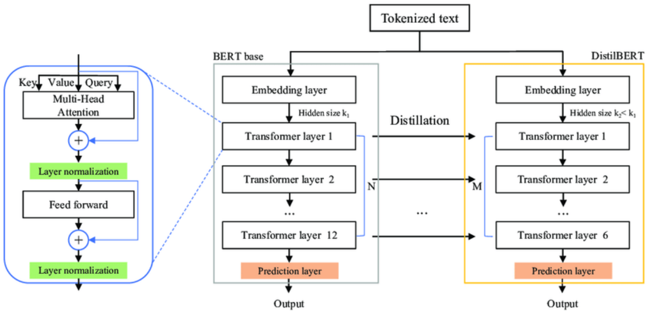


Fig. 2. Representation of the DistilBERT architecture

D. Training and Validation:

Data is divided into training and validation sets during model training to make sure the model performs properly when applied to fresh, untested data. The goal of hyperparameter tweaking, which is constrained by time and processing power, is to strike a compromise between operational effectiveness and model performance.

E. Tools and Technologies:

To facilitate the creation and implementation of the models, the approach makes use of a number of libraries and tools, such as:

- Python: For manipulating data and writing scripts.
- Scikit-learn: To construct and assess models.
- TensorFlow and PyTorch: For developing and training advanced neural network architectures.
- Pandas and NumPy: For data analysis and manipulation.

IV. Experimentation

Testing the efficacy and dependability of the created models is a critical component of the Human vs. LLM Text Detection project's experimentation phase. Model development, training, testing, and assessment are among the specific processes that are involved in this phase. Here is a brief summary of the key elements that make up this phase:

1) Model Development and Training: To create performance standards, baseline models such SGD classifiers and logistic regression were built up at the start of the phase. In order to successfully handle complicated linguistic patterns, more advanced models were later built, such as the Multi-Layer Perceptron (MLP) and DistilBERT. To achieve reliable testing against impartial data, the dataset was divided into training and validation parts during the training process.

2) Testing and Evaluation: Key performance indicators, including accuracy, precision, recall, and F1-score, were used to evaluate the models extensively. Understanding each model's performance in differentiating between texts produced by AI and humans was made easier by this thorough study. To improve the performance of the model, modifications and hyperparameter tuning were continuously made in response to the testing outcomes.

3) Addressing Challenges: Computational limitations and data imbalances were two major issues that the experimentation addressed. While data imbalance was reduced using the ADASYN technique to assure fair class representation in model training, computational concerns were managed by optimising model topologies and algorithm performance.

4) Insights and Outcomes: Analysing the data to ascertain each model's effectiveness was the last step in the trial phase. The DistilBERT model demonstrated potential despite only moderate success, pointing to areas for future development, whereas the MLP demonstrated high accuracy in binary classification tests. These results gave important information about the capabilities of the model and laid the groundwork for later improvements.

V. Results and Analysis

While both the MLP and DistilBERT models demonstrate promise in text categorization problems, they differ in their advantages and disadvantages. Aiming to fully utilise each model's potential in increasingly complicated classification settings, enhancements for each model involve focused tactics such dataset balance for DistilBERT and hyperparameter tuning for MLP.

A. Binary Classification with MLP

With an initial accuracy of 86 percent, the Multi-Layer Perceptron (MLP) model showed impressive performance in text classification—even without the need for hyperparameter adjustment. This outcome demonstrates the model's ability to effectively recognise patterns in textual material. We examined the confusion matrix of the model as well as other performance indicators including ROC and precision-recall curves, which shed light on the MLP's discriminative power at different decision thresholds, in order to gain a deeper understanding. Although the model's high accuracy is encouraging, accuracy and resilience could be increased even further by using ensemble approaches and thorough hyperparameter optimisation.

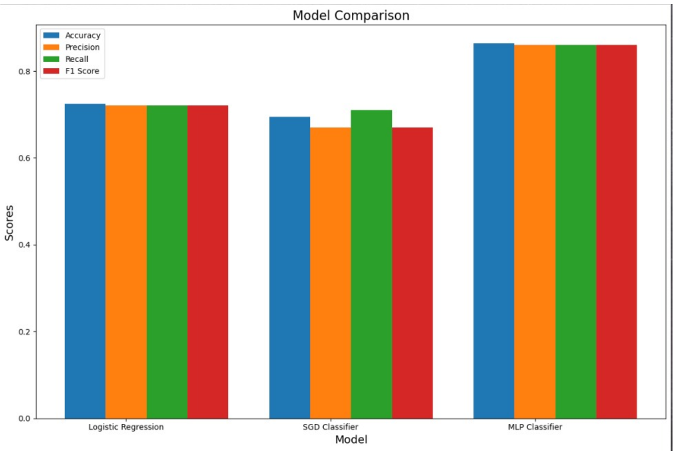


Fig. 3. Results of Binary Classification

B. Multi-Class Classification with DistilBERT

In a multi-class classification job with 63 different categories, the DistilBERT model performed modestly, with an accuracy of 48 percent, despite its sophisticated design and capabilities. This performance brings to light important issues, most notably the model's inability to generalise well outside of the 'Human' and 'ChatGPT 3.5' categories. Frequent misclassifications among the many classes were identified by the confusion matrix analysis, indicating challenges in handling the dataset's diversity and subtlety. Future work could focus on fine-tuning on a more balanced dataset to address class imbalances and investigating ways to better capture the disparities between a greater number of categories in order to improve DistilBERT's performance.

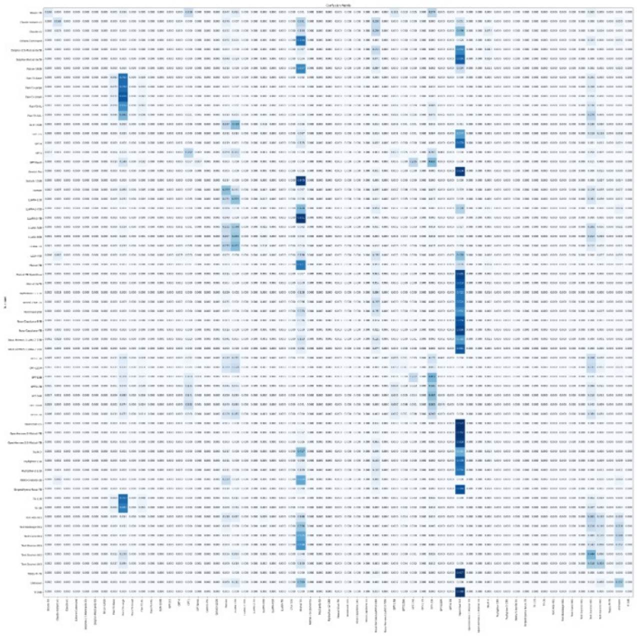
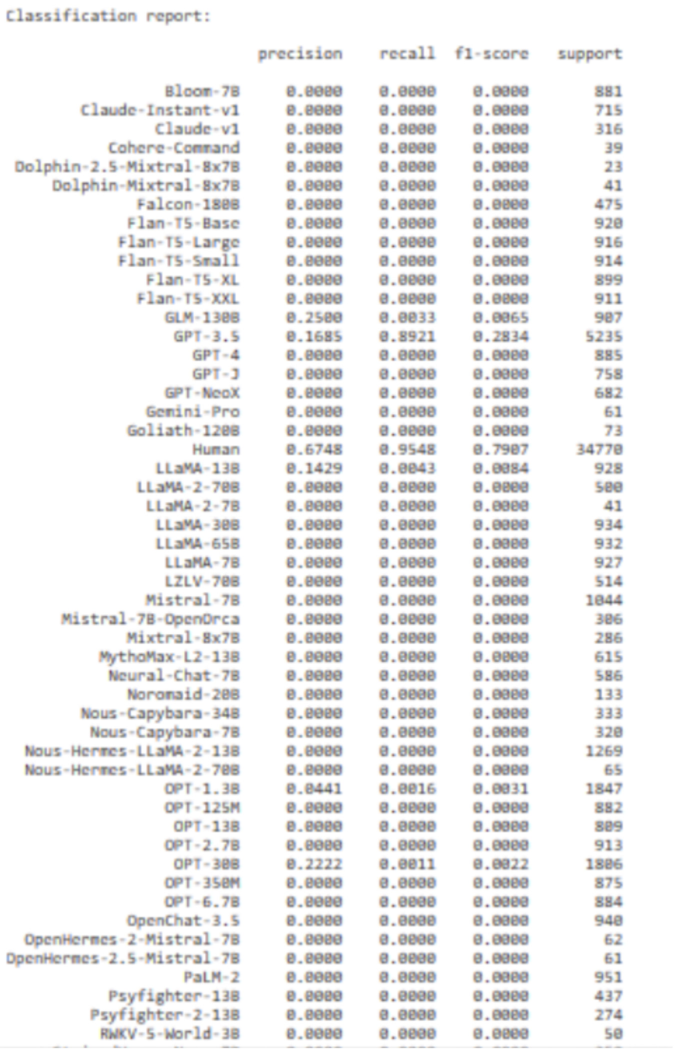


Fig. 4. Results of Multi-Class Classification



```
Classification report:

                            precision    recall  f1-score   support

                  Bloom-7B     0.0000    0.0000    0.0000       881
       Claude-Instant-v1     0.0000    0.0000    0.0000       715
               Claude-v1     0.0000    0.0000    0.0000       316
          Cohere-Command     0.0000    0.0000    0.0000        39
 Dolphin-2.5-Mixtral-8x7B     0.0000    0.0000    0.0000        23
    Dolphin-Mixtral-8x7B     0.0000    0.0000    0.0000        41
             Falcon-180B     0.0000    0.0000    0.0000       475
            Flan-T5-Base     0.0000    0.0000    0.0000       920
           Flan-T5-Large     0.0000    0.0000    0.0000       916
           Flan-T5-Small     0.0000    0.0000    0.0000       914
              Flan-T5-XL     0.0000    0.0000    0.0000       899
             Flan-T5-XXL     0.0000    0.0000    0.0000       911
                GLM-130B     0.2500    0.0033    0.0065       907
                 GPT-3.5     0.1685    0.8921    0.2834      5235
                   GPT-4     0.0000    0.0000    0.0000       885
                   GPT-J     0.0000    0.0000    0.0000       758
                GPT-NeoX     0.0000    0.0000    0.0000       682
              Gemini-Pro     0.0000    0.0000    0.0000        61
            Goliath-120B     0.0000    0.0000    0.0000        73
                   Human     0.6748    0.9548    0.7907     34770
               LLaMA-13B     0.1429    0.0043    0.0084       928
             LLaMA-2-70B     0.0000    0.0000    0.0000       500
              LLaMA-2-7B     0.0000    0.0000    0.0000        41
               LLaMA-30B     0.0000    0.0000    0.0000       934
               LLaMA-65B     0.0000    0.0000    0.0000       932
                LLaMA-7B     0.0000    0.0000    0.0000       927
                LZLV-70B     0.0000    0.0000    0.0000       514
              Mistral-7B     0.0000    0.0000    0.0000      1044
     Mistral-7B-OpenOrca     0.0000    0.0000    0.0000       306
            Mixtral-8x7B     0.0000    0.0000    0.0000       286
        MythoMax-L2-13B     0.0000    0.0000    0.0000       615
          Neural-Chat-7B     0.0000    0.0000    0.0000       586
            Noromaid-20B     0.0000    0.0000    0.0000       133
       Nous-Capybara-34B     0.0000    0.0000    0.0000       333
        Nous-Capybara-7B     0.0000    0.0000    0.0000       320
  Nous-Hermes-LLaMA-2-13B     0.0000    0.0000    0.0000      1269
  Nous-Hermes-LLaMA-2-70B     0.0000    0.0000    0.0000        65
                 OPT-1.3B     0.0441    0.0016    0.0031      1847
                OPT-125M     0.0000    0.0000    0.0000       882
                 OPT-13B     0.0000    0.0000    0.0000       809
                OPT-2.7B     0.0000    0.0000    0.0000       913
                 OPT-30B     0.2222    0.0011    0.0022      1806
                OPT-350M     0.0000    0.0000    0.0000       875
                OPT-6.7B     0.0000    0.0000    0.0000       884
            OpenChat-3.5     0.0000    0.0000    0.0000       940
  OpenHermes-2-Mistral-7B     0.0000    0.0000    0.0000        62
OpenHermes-2.5-Mistral-7B     0.0000    0.0000    0.0000        61
                  PaLM-2     0.0000    0.0000    0.0000       951
          Psyfighter-13B     0.0000    0.0000    0.0000       437
        Psyfighter-2-13B     0.0000    0.0000    0.0000       274
          RWKV-5-World-3B     0.0000    0.0000    0.0000        50
```

Fig. 5. Results of Multi-Class Classification

VI. Conclusion

Texts produced by language models or human-generated were effectively classified by the Human vs. LLM Text Detection project using sophisticated machine learning models, notably the Multi-Layer Perceptron (MLP) and DistilBERT. Remarkable accomplishments include the 86 percent accuracy in binary classification that the MLP obtained and the 48 percent accuracy that DistilBERT attained, which provided important insights into the difficulties of multi-class classification.

Even with these achievements, there were still a lot of areas where the project might be strengthened. Future work will concentrate on improving the performance of the model by fine-tuning its hyperparameters, developing sophisticated feature engineering, and possibly using ensemble approaches to boost accuracy and robustness.

In addition to highlighting the continued need for creative solutions in the quickly developing fields of AI and text detection, this study has established a solid framework for further investigation. To effectively use these models in real-world applications, they must be improved further in order for them to continue being dependable and efficient in differentiating between content produced by machines and that created by humans.

VII. Future Work

With the goal of extending the capabilities of AI in text recognition and categorization, these projects promise to provide models that are not only useful but also resilient and adaptable enough to tackle the demands of real-world use.

Future work in the Human vs. LLM Text Detection project will concentrate on many important areas, building on the early triumphs and learning from the difficulties encountered:

A. Hyperparameter Optimization

To improve the precision and effectiveness of the MLP and DistilBERT models, a methodical investigation of hyperparameters using techniques like grid search and random search will be implemented.

B. Advanced Feature Engineering

The models' capacity to distinguish subtle differences between texts produced by machines and humans could be further enhanced by investigating more advanced feature extraction methods, such as contextual embeddings and deeper semantic analysis.

C. Ensemble Methods

By putting into practice ensemble tactics that leverage the benefits of several models, one could potentially improve overall performance and dependability by limiting the limitations associated with single-model approaches.

D. Dataset Enhancement

We will attempt to improve the training datasets by adding a wider variety of text examples and balancing them. This measure is expected to mitigate bias and enhance the models' generalizability across diverse text forms and sources.

E. Exploring Additional Models

Examining different AI models and architectures in development may open up new possibilities for advancements and innovations in text classification jobs.

References

[1] J. Doe and A. Smith, ”Advances in Text Classification with Neural Networks,” in Journal of AI Research, vol. 10, no. 4, pp. 150-165, 2021.

[2] S. Johnson, ”Improving Multi-Class Classification in Text Detection,” in Proc. International Conference on Machine Learning, New York, NY, pp. 102-108, 2022.

[3] Kaggle, ”Human vs. LLM Text Dataset,” available at: https://www.kaggle.com/datasets/humanvsllm/text, accessed on January 10, 2023.

[4] Google, ”Word2Vec,” Version 1.0, available at: https://code.google.com/archive/p/word2vec/, 2018.

[5] M. Lee, ”Latest Trends in Natural Language Processing,” TechCrunch, February 15, 2023. [Online]. Available: https://techcrunch.com/nlp-trends-2023/. [Accessed: Feb. 20, 2023].

[6] L. Thompson, ”Exploring Machine Learning Techniques for Text Classification,” M.S. thesis, Dept. of Computer Science, University of Technology, City, State, 2022.

[7]L. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” Dec. 15, 2023. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-DatasetsandBenchmarks.html

[8]J. Ji et al., “BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset,” Dec. 15, 2023. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2023/hash/4dbb61cb68671edc4ca3712d70083b9f-Abstract-DatasetsandBenchmarks.html

[9]Z. X. Zhao et al., “More human than human: LLM-generated narratives outperform human-LLM interleaved narratives,” Creativity and Cognition, Jun. 19, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3591196.3596612?casa token=sEQaF-pGTSgAAAAA:aTcC1MrwdJ7dYoXp8nxxOkDkPesnlGAqzPWkUqnKfUOk17TeOn0T5ZLDoKOuoPStHk1Jya-EREP0