A Word Embeddings and Stylistic Features based Approach for Generative AI Authorship Verification

Thota Kesava Rao
Associate Professor, Dept of CSE
Swarnandhra College of Engineering and Technology,
Narasapur, AP, India
thota.kesavarao@gmail.com

K V Peddi Raju
Assistant Professor, Dept of CSE
Swarnandhra College of Engineering and Technology,
Narasapur, AP, India
kvpraju37@gmail.com

V S Rama Krishna
Assistant Professor, Dept of CSE
Swarnandhra College of Engineering and Technology,
Narasapur, AP, India
vsramakrishna2k@gmail.com

Sabbithi Chanti
Assistant Professor, Dept of CSE
Swarnandhra College of Engineering and Technology,
Narasapur, AP, India
chantisabbithi@gmail.com

Karunakar Kavuri
Associate Professor, Dept of CSE
Swarnandhra College of Engineering and Technology,
Narasapur, AP, India
karunakar.mtech@gmail.com

2025 International Conference on Electronics, AI and Computing (EAIC) | 979-8-3315-5358-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/EAIC66483.2025.11101533

***Abstract*—The remarkable capacity of generative large language models such as GPTs, to generate high-quality content in a variety of fields that nearly mimics human writing has garnered attention in recent years, posing significant problems in differentiation. Consequently, it is now more important than ever to detect machine-generated text. PAN 2024 organizes a competition for differentiating the text written by human or machine. They provided the dataset of human writings and machine generated texts. In this work, we developed a hybrid model for predicting whether the text was generated by the human or a machine. The proposed model integrates the stylistic features and embeddings from two sophisticated transformer models, such as RoBERTa and T5 models. The PAN 2024 Generative AI Authorship Verification dataset is used in this experiment. The dataset contains files of concepts that are written by humans and different large language models. In the proposed model, the files are represented as vectors by combining the stylistic features based vector representation and word embedding techniques based vector representation. These vectors are forwarded to LightGBM algorithm for training. The performance of proposed model represented in terms of numerous evaluation metrics such as ROC-AUC, C@1, Brier, F0.5u, F1, and mean. The proposed hybrid model attained scores of 0.996 for ROC-AUC, 0.970 for Brier, 0.989 for C@1, 0.972 for F1, 0.971 for F0.5u, and 0.979 for mean on the training dataset. These results demonstrate the importance of including transformer embeddings along with stylistic features to improve the performance of authorship verification.**

Keywords—Authorship Verification, Large Language Models, T5 Model, RoBERTa Model, Stylistic Features

I. Introduction

Authorship Verification (AV) is a crucial application in the area of Natural Language Processing (NLP), which is based on text classification. The purpose of this concept is to evaluate whether two texts were written by the single person or not. AI Authorship Verification (AIAV) is vital for authenticate the validity of documents, recognizing the origins of publications, and detecting plagiarism, thereby safeguarding the integrity of written text across diverse sectors. In association with the ELOQUENT Lab (Voight-Kampff Task), the PAN competition [1, 2] introduced a task of Generative AI Authorship Verification (GAIAV) [3] that builds on earlier challenges and specifically aims to differentiate among texts that are authored by humans and those that are generated by machines. To aid humans in discriminating among text created by humans and text created by machines, numerous classification techniques have been developed. Conventional approaches depend on surface-level features like stylistic components, syntactic patterns, and word frequency, while advanced Large Language Models (LLMs) may readily imitate these features [4, 5]. As a result, authorship verification in human vs. machine language context is still a difficult and crucial issue.

Since LLMs are now broadly accessible and extensively available, there is a rise in machine-generated content on a variety of environments, such as educational resources, social media, academic settings, Q&A forums, news writing, advertising, storytelling, and code production. The ability of recent developments in LLM technology, such as GPT-4, ChatGPT, Qwen [6], and ChatGLM [7], to generate logical answers to the majority of user queries makes these models more appealing for substituting human labour in a variety of applications. This accessibility has, however, sparked worries about possible misuse, including the creation of fake news, effects on the financial services sector, disturbances in educational environments, and effects on the legal domain. In order to reduce the hazards involved, automated systems that can recognise machine-generated content must be developed immediately, as humans have difficulty in telling the difference between text that has been created by machines and text that has been produced by humans.

With millions parameters to billions parameters, LLMs detect the tokens probability distributions based on their observed context. The deep learning architecture of transformer, which debuted in 2017 [8], serves as the foundation for the majority of LLMs. Text generation through prompting has been made possible by the more latest LLMs (e.g., Llama, GPT-4, GPT-3.5) development with huge number of parameters (billions), but LLMs (GPT-2) with more number of parameters (millions) have long been able to produce texts with human- level efficiency. Because of this, practically anyone can now rapidly and simply create extremely high-quality machine-written documents.

There are two primary types of detection techniques for text produced by LLMs. The first technique, known as zero-shot detection, uses the source model that produced the text to immediately recognize the AI-generated text. This approach determines if the text is machine-generated by

using the loss values or output logits of the source model rather than pre-trained datasets. Zero-shot detection has the benefit of being able to be used on new text and requiring little in size of training data. However, its dependence on the performance of the proxy model or source model is a major drawback. The efficacy of the detection may be low if the source model and proxy model differ significantly. Deep neural network (DNN) classifiers, which use supervised training models to identify both human-generated and AI-produced text, are the foundation of the second technique. This method's benefit is that it can enhance performance of detection by using a lot of training data. However, the learned classifiers are susceptible to attacks like backdoor [9] and adversarial [10], and DNN-based classifiers have low generalisation capabilities and high data requirements [11].

Due to the small distinctions among machine-generated and human written text, existing methods frequently perform badly in this task. This is especially true when generative models continue to develop, making it harder to spot these differences. To solve this issue, we developed a hybrid model for authorship verification task. In the proposed model, extract the most significant stylistic features from the dataset to differentiate the writing styles of human-written texts and machine produced texts. Two transformer based models like RoBERTa and T5 models are utilized to produce word embeddings. The hybrid model represents files as vectors by combining stylistic features based representation and word embedding techniques based representation. The LightGBM algorithm is used for generating classification model by training these vectors.

This paper is planned in 6 sections. Section 2 discuss about different research works proposed for GAIAV task. The characteristics of the PAN 2024 AV dataset is presented and discussed in section 3. The methodology of proposed hybrid model and the components used in the model are explained in section 4. The performance metrics and experimental outcomes of proposed hybrid model are explained in section 5. The section 6 specifies the conclusions of this article with possible future enhancements to this work.

II. Survey on Existing works of Generative AIAV

The task of GAIAV is distinguishing among texts produced by humans and those created by machines. To solve this task, Ye Zhu et al., explored [12] the utilization of the Deberta which is a pre-trained language model. The proposed method entails fine-tuning of the Deberta model using a curated dataset that includes both machine-generated and human-produced text. They used random sampling technique to control the imbalance in their dataset and guarantee that both kinds of texts were represented balanced way during training. Initial experiments indicate that although their methodology performs similarly to other existing methods, there is an excessive potential for further enhancement and optimisation in the identification of texts written by humans.

Ye et al., presented [13] a method for Voight-Kampff GAIAV task using Next Token Prediction as Implicit Classification. This approach is justified by the fact that text classification tasks can be successfully completed through token prediction. In order to address the gap among downstream and pre-training tasks, they employed the token prediction technique to directly determine whether the given text was written by a human or by a particular AI model. They used PAN's Generative AI Authorship Verification datasets to assess the efficiency of their approach. The developed method reached different performance scores of 0.947 for max, 0.926 for 75-th Quantile, 0.922 for Median, 0.896 for 25-th Quantile, and 0.527 for minimum on the test dataset. These results confirm the efficiency of their developed approach in performing the task of GAIAV.

It is now more complex to tell the difference among texts that were generated by humans and those that were created by machines due to the large language models' (LLMs) widespread adoption and rapid development. Even though a number of classification techniques have been developed to assist in determining the origins of texts, they frequently failed to solve the inherent challenges and fundamental feasibility of the task. Based on their vast experience in authorship verification, Zepeng Wu et al., presented [14] BertT, which is a novel hybrid approach that combines Transformer and BERT technologies. It was developed especially for the GAIAV, which was organized in association with ELOQUENT Labs and PAN. The proposed BertT model exhibited strong performance across multiple metrics, utilising the transformer's efficient sequence processing power and the BERT's deep semantic understanding capabilities. Their model accomplished a Brier score of 0.903, ROC-AUC score of 0.967, and accuracies that ranged from a low score of 0.354 to a high score of 0.980. This demonstrates its capacity to efficiently handle complex and diverse textual contexts.

Humans now find it more difficult to distinguish among human produced and machine produced text due to the proliferation of machine-generated content across several platforms produced by Large Language Models (LLMs). Automated solutions that can recognise machine-generated text and reduce associated risks are desperately needed to address this problem. In response, the task of GAIAV was presented by the PAN competition in partnership with the ELOQUENT Lab. Andric Valdez-Valenzuela et al., proposed [15] a novel model architecture that combines pre-trained Language Models with stylometric features and Graph Neural Networks (GNNs) in order to categorise text documents as either machine-generated or human-generated. The two-path structure is used in the proposed method. The first path used data augmentation to convert the text documents into co-occurrence graphs with GNN for processing, and the second path used a BERT-BASE model and stylometric features to extract and fine-tuning embeddings. By combining these embeddings, classification robustness and accuracy are enhanced. In order to balance the dataset, they described the data stratification technique, which includes adding of human-generated text. The efficiency of the proposed model is verified through several experimental results, especially final-run7-gnnllm_llmft_stylofeat-fullpartitionA and final-run4-gnnllm_llmft_stylofeat-partitionB received high scores consistently across a variety of evaluation measures, considerably outperforming all standard methods.

It is getting difficult for people to tell if a specific text was produced by a machine or a human as LLMs continue to develop at startling speeds and are progressively implemented by more people. Verifying authorship has grown to be a crucial and difficult task. Guihong Sun et al., developed [16] a hybrid model by combining BERT with

Convolutional Neural Networks (CNNs) to improve performance of text classification. This model makes use of effective local feature extraction skills of CNN and strong contextual understanding capabilities of BERT. CNN successfully makes up for limitations of BERT in phrase-level feature extraction, especially when it comes to recognizing local features available in the text, including n-gram based features. According to experimental outcomes, the suggested BCAV hybrid model performs remarkably better than all baseline models, improving by up to 6% in the score of ROC-AUC measure and by over 3% in the score of Mean measure.

Using a model fusion approach, Rui Qin et al., proposed [17] a method to differentiate among text produced by generative AI models and human-generated language. There are three steps in the proposed method. First, add an external dataset from the well-known machine learning and data science competition platform of Kaggle to the PAN competition dataset of CLEF 2024. Then, misspelled words are corrected by using the Levenshtein distance technique. Following that, text pairs of datasets are created using a common theme and divided into training, testing, and validation datasets. In second step, train a refined BERT as the basic model and R-Drop technique is used by the BERT model to solve the problem of overfitting. The final phase involves combining the two models using a voting strategy and an ensemble learning method. According to the experimental findings, the fusion model outperformed the baseline model Fast-DetectGPT (Mistral) by 5.6%, achieved a ROC-AUC metric score of 0.932.

New and more powerful iterations of pre-existing models, or even entirely new models, are continuously being generated in the fast changing area of artificial intelligence. Both industry and humanity are investing more and more in these developments. Humans have utilised these models in several real-world situations, either to help themselves or to deceive. Two major events that occur frequently are the generation of fake news, and scholars and students refusing to probe deeply into knowledge. Therefore, a classifier that can identify and differentiate between text created by AI and text written by humans must be developed. There have been a number of excellent approaches, but they need to keep changing as LLMs change. This year's PAN shared work at CLEF clarifies the previously noted necessity. To accomplish the task, Panagiotis Petropoulos et al., developed [18] an architecture that combines RoBERTa and Bi-LSTM on top. According to the results, the proposed method and architecture can distinguish between texts produced by humans and artificial intelligence (AI) with a high accuracy degree, with an F1-score and accuracy of over 90%. The model's ability to consistently distinguish between texts produced by AI and those produced by humans is demonstrated by the strong C@1 and AUC values.

In order to detect AI produced content, Benjamin Ostrower et al., suggested [19] an ensembled method that combines three techniques including a factual coherence graph, a neural dependency graph, and a fine-tuned RoBERTa. These techniques produce their prediction logits, which are subsequently concatenated and forwarded to an XGboost classifier. Using a PAN competition's bootstrapped dataset of Bard and human produced text, the ensembled approach accomplished an accuracy of 61%. In order to crack the Generative AI AV task, Jiajun Lv et al., proposed [20] a technique that integrates contrastive and meta-learning. Their goal is to maximise the relationships among samples by using supervised contrastive learning to improve the discriminative power of the model. They also improved the generalisation ability on out-of-domain data by using the meta-learning method Reptile. The model that performed the best on the dataset of validation was finally selected by the authors. On the leaderboard, the proposed approach attained scores of 0.98 for roc-auc, 0.945 for brier, 0.954 or c@1, 0.93 for F1, 0.935 for F0.5u, and 0.949 for Mean. These outcomes demonstrate how well the suggested approach performs in the task of GAIAV.

Zhaojian Lin et al., developed [21] a technique for optimising the T5 pre-trained language model for the task of GAIAV. Training samples and explicit instructions make up the input sequence during the training phase, and the output sequence displays the results of classification as positive or negative. The vocabulary of model is limited to positive and negative words during inference, and the word that has more probability was selected as the outcome of classification. In conclusion, their performance metrics score for the maximum, 75th percentile, median, 25th percentile, and minimum values on the test set were 0.877, 0.874, 0.744, 0.529, and 0.138 respectively.

Haotian Lei et al., optimised [22] the large language model ChatGLM using the LoRA approach. The LoRA technique can expand the model's linguistic representation, making it a more adaptive and professional. They changed the class labels and forwarded to a multi-labelled task of classification in order to balance the data distribution of the dataset. This makes it possible for the LLMs to learn the writing styles of both machines and humans by better understanding the variations in expression between various authors on the same topic. In order to decide whether the content was written by a machine or a human, the author changed the inference process's final outcome by assigning it to a binary classification task. The objective of their method is to complete the task of GAIAV. With a mean score higher than 0.7, the evaluation findings on the test dataset of PAN corpus show that this approach is successful.

By using pre-trained language models, domain adaptation, and contrastive learning, Kaicheng Huang et al., [23] successfully completed the task of Generative AI Authorship Verification, which involves identifying which of two texts, one is created by AI and the other by a human, is a human text. In contrast to traditional machine learning techniques, they employed unsupervised domain adaptation and self-supervised contrastive learning methods to efficiently use unlabelled target domain and labelled source domain data, extract features in both AI and human texts, and utilize these features to categorise the text. Their experiments show that the average score of their model in terms of metrics such as F1, ROC-AUC, c@1, Brier, and F0.5U reached to 0.994 on the validation dataset, while the average score in the test dataset of PAN reached to 0.480.

The application of Authorial Language Models (ALMs) for AIAV was presented by Weihang Huang et al., [24]. AIAV is the task of identifying which of two texts, one authored by a machine and the other by a human, was created by the machine or by the human. The proposed method resolved this task by utilizing Support Vector Machine (SVM) to detect whether the text was produced by a machine. For independent testing on the primary dataset of

testing and its versions that are complicated against detection, they submitted their approach as a docker-contained software. They have been notified that their approach outperformed all baseline approaches on the main dataset, attained a score of roughly 0.979 on all suggested evaluation metrics. Additionally, they outperformed all baselines with a 0.935 median score on the main dataset variants. They credit the effectiveness of ALMs in this situation to the utilisation of numerous refined authorial language models, which they feel strengthens the approach's resilience by exploiting the potentially differentiating information quantity extracted from the fundamental textual data.

Jijie Huang et al., treated [25] generative AI authorship verification task as a problem of binary classification and introduced Tri-Sentence Analysis (TSA) approach for solving this task. TSA improves the model's capacity to determine the source of text by capturing detailed contextual information. It increases the resilience of model while dealing with lengthy texts by better comprehending the consistency and semantic relationships among sentences. In order to enhance the model's differentiation and efficiency for brief texts, they also included the MPU method. Lastly, they integrated these techniques into a BERT model that had already been trained. Their relative performance metrics on the test set are 0.999, 0.989, 0.976, 0.936, and 0.883 and 0.999 for the 25-th Maximum, 75th Quantile, Median, 25th Quantile, and Minimum scores, respectively.

NLP has advanced significantly in latest years because of Large Language Models (LLMs) like GPT-4, BERT, and GPT-3, which have improved tasks like question answering, language translation, and document summarisation. Despite these advantages, societal issues like plagiarism and misinformation have been brought up by the credibility and authenticity of texts produced by these models. The PAN organisation has started a number of tasks to distinguish between writings that are produced by humans and those that are generated by machines in order to address these problems. Linjiu Guo et al., proposed [26] a model by merging multi-text feature approaches with transformer encoders, which is based on BiLSTM and BERT, improves discrimination capabilities of text. In addition to using a pre-trained BERT for extraction of deep features, the model also uses transformer encoders for classification and supplementary text features computed by the spaCy that are then further handled by BiLSTM. According to experimental results, the model outperformed all baseline models with a 0.971 of mean score on the validation dataset of PAN. This method is important for preserving the reliability and authenticity of information in the age of automatic generation of content since it not only increases accuracy of detection but also improves adaptation to different text types.

III. Dataset Characteristics

The Voight-Kampff GAIAV 2024 PAN competition [1] offers a training dataset made up of many JSONL files, which covers both fraudulent and fraudulent news articles from several US news headlines of 2021 year. Thirteen JSONL files with writings are produced by well-known LLMs, such as alpaca, Mistral, Llama [27], Gemini pro [28], and GPT-4 [29]. One JSONL file with news items are authored by humans. The dataset includes both machine-generated and human-authored text, with a notable imbalance in the ratio among the two categories (1:13). Table 1 describes the dataset's structure in detail.

TABLE I. The Characteristics of Dataset

Data Source	Number of Topics	Range of Tokens
Human	1087	25 - 7989
gemini-pro	1087	1205-5881
bigscience-bloomz-7b1	1087	96 - 3557
alpaca-7b	1087	0 - 3141
chavinlo-alpaca-13b	1087	0 - 5505
gpt-4-turbo-preview	1087	1189-6931
gpt-3.5-turbo-0125	1087	75-5961
meta-llama-llama-2-7b-chat-hf	1087	367-5865
meta-llama-llama-2-70b-chat-hf	1087	1209-5957
mistralai-mixtral-8x7b-instruct-v0.1	1087	811-6928
mistralai-mistral-7b-instruct-v0.2	1087	1446-6274
text-bison-002 1087	1087	0-5613
qwen-qwen1.5-72b-chat-8bit	1087	1404-3917
vicgalle-gpt2-open-instruct-v1	1087	52-3653

There are 1087 different concepts in the dataset. For every concept, it contains AI-generated text files of 13 and human-written text file of one. As listed in Table 1, these 13 files were produced by a distinct LLM model. The same topic is covered in both machine-produced and human-generated texts in the train dataset. The PAN Dataset was combined and categorised into a new dataset called "combine". There are two columns in this dataset such as "text" and "label". The text's content is represented by the "text" column, and its source is indicated by the "label" column, where machine-produced texts are labelled as 0 and human-generated texts as 1. There are 15,218 articles in all in the combined dataset. We considered 12,174 training samples and 3,044 validation samples by using 80% of the dataset with the labels 1 and 0 for training and 20% of the dataset for validation. The dataset spans multiple domains such as science, history, general knowledge, and current events. This diversity is beneficial for training generalized detectors but may also introduce domain-specific biases if some categories are overrepresented. The dataset is balanced number of instances in each category.

IV. Proposed Methodology

The Figure 1 shows the framework of the proposed hybrid model. In this model, first we collected the PAN 2024 GAIAV dataset. Then, apply suitable pre-processing techniques like stopword removal and lemmatization to remove irrelevant data not required for further analysis. Extract all tokens from the cleaned dataset and a set of stylistic features are extracted that are useful for discriminating the writing styles of text generated by human or machine. The RoBERTa and T5 models are used to generate word embedding vectors for all the tokens extracted from the dataset. Merge these two varieties of embedding vectors (RoBERTa represents every token as 768 dimensional vector and T5 represent every token as 768 dimensional vectors) for all tokens. Represent each file in the dataset with the averaging of combined word embedding vectors (Each token represented with 1536 dimensional vector) of all tokens that are present in that file. Simultaneously the files are represented as vectors with

stylistic features. The files are finally represented with vector representation of statistical features and embeddings based file representation. These files vectors are used to train the machine learning algorithm of LightGBM algorithm [30]. The trained model predicts the performance of hybrid model for Generative AI Authorship Verification. The next subsections explains the components used in the implementation of proposed hybrid model.

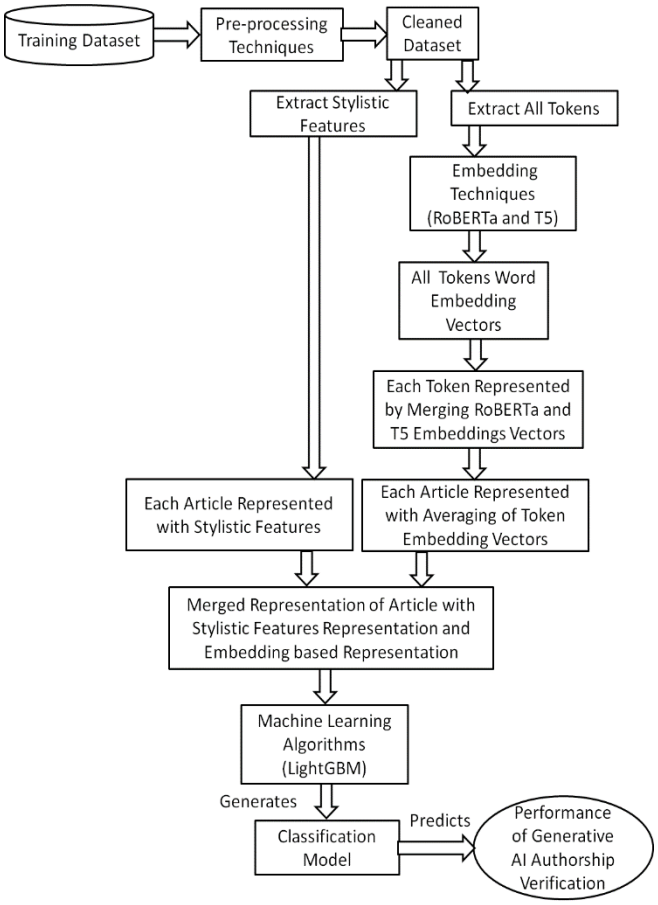

Fig. 1. The Architecture of Proposed Hybrid Model

A. *RoBERTa Embeddings*

Researchers at Facebook AI and Washington University created the RoBERTa ("Robustly Optimised BERT Approach") [31] model which is a variation of BERT (Bidirectional Encoder Representations from Transformers) model. RoBERTa is a language model which is based on transformer, similar to BERT that processes sequences of input and creates contextualised words representations in a sentence by using self-attention. The fact that RoBERTa was learnt using a more effective process of training and a far larger dataset is one of the primary distinctions among it and BERT. Specifically, 160GB of text was utilized to train RoBERTa, which is ten times more than the dataset size utilized for BERT training.

RoBERTa shows higher performance when compared with BERT and other most popular models across numerous NLP tasks, including question answering, text classification, and language translation. It has gained popularity in both academic and commercial applications and has served as the foundation for numerous successful NLP models. Overall, RoBERTa is a powerful and efficient language model that has significantly contributed to advancements in NLP and its diverse applications.

B. *T5 model Embeddings*

Google AI created the T5 (Text-to-Text Transfer Transformer) [30] family of LLMs, which were first released in 2019. T5 models are encoder-decoder Transformers, just as the original Transformer model, in which the decoder creates the output text after processing the input text by the encoder. After being pretrained on a sizable dataset of code and text, T5 models are able to execute text-based tasks that are comparable to the pretrained tasks. They can be fine-tuned to carry out additional tasks as well. Machine translation, chatbots, text summarisation, robotics, and code generation are just a few of the applications that have used T5 models. The Colossal Clean Crawled Corpus (C4) contains a code and text that has been scraped from the internet, and is used to pre-train the original T5 models. This process of pre-training allows the models to learn understanding and generation capabilities of general language. After then, fine-tune T5 models to implement well on a variety of downstream tasks by modifying their knowledge.

The encoder-decoder Transformers of the T5 series come in a variety of capabilities and sizes, with the encoder processing the given text and the decoder producing the output text. These models are frequently identified by the number of parameters they contain, which reveals the model's potential and complexity. Five models, including T5 small, base, large, 3B, and 11B, were reported in the original work [30]. In this work, we generated word embeddings using the T5 base model. The decoder and encoder networks in the T5 base model consist of 12 layers, with 12 attention heads in every block of attention. The embedding vectors have 768 dimensions, the feed-forward network has 3072 dimensions within each decoder and encoder layer, and the value and key vectors used in the mechanism of self-attention have 64 dimensions.

By utilizing its pre-trained layers of transformer, the T5 model, which was initially created for text comprehension and generation tasks, can also be utilized to generate word embeddings. T5 uses its architecture of transformer to translate text into a meaningful latent space. The hidden states of its encoder or decoder can be utilized as high-quality embeddings for input text.

C. *Stylistic Features*

Differentiating between LLM (Large Language Model) generated text and human-written text often relies on identifying subtle stylistic and structural differences. In this work, we identified various stylistic features to differentiate the text generated by machine or human.

1) *Linguistic Features*

Sarcasm and Humour (Humans are better at employing and comprehending subtle context-based sarcasm and humour, while LLM difficulties with subtle humour or sarcasm), Use of Figurative Language and Idioms (Human uses idiomatic and figurative language more creatively and richly, while LLM uses idioms more literally and may misuse uncommon idiomatic terms), Sentence Length Distribution (While human sentence length fluctuates and frequently reflects natural speaking patterns, LLM tries to keep uniform sentence lengths), Sentence Complexity

(Human can have asides, interruptions, and incomplete sentences, while LLM frequently utilises well-formed, medium-length phrases), Diversity of Vocabulary (humans exhibit a diverse vocabulary linked to particular contexts or areas of competence, whereas LLM may employ a broad vocabulary but occasionally lack depth in topic-specific jargon), Repetition (particularly in lengthy passages, LLM tends to repeat specific phrases, ideas, or structures, but humans generally avoid excessive repetition unless intentional), and Frequency of part-of-speech (POS) tagging.

2) Structural Features

Personalisation (unless specifically instructed, LLM rarely personalises content, and human incorporates personal tales, viewpoints, or distinctive viewpoints), Relevance (Human often adheres more closely to the context or topic at hand, whereas LLM may introduce tangential or excessively generic stuff), Topic Depth (Human exhibits deeper investigation and individual understanding into particular areas, whereas LLM offers wide, general responses that can appear superficial on speciality themes), Dependency parsing to identify sentence structure patterns; logical flow (LLM shows great logical coherence, but may misuse connectors like "however" or "therefore", while human flow may be less rigid but more instinctive, with regular transitions).

3) Semantic Features

Specificity (unless trained differently, LLM responds with high-level and generic responses, whereas humans are more specific, providing examples and facts from the real world), Redundancy (human redundancy is less common until a point is being emphasised, while LLM repeats points in slightly varied wording), Factual Errors (Human errors are usually related to ignorance rather than fabrication, whereas LLM may "hallucinate" facts or give certain but inaccurate information).

4) Quantitative Features

Lexical Density (human lexical density varies according to the aim and writing style, while LLM exhibits higher lexical density with an emphasis on informational content), Phrase and word Frequency (Due to training biases, LLM exhibits regular patterns in word usage, whereas humans exhibit more flexibility and originality in word choice), vocabulary richness (type-token ratio), N-gram frequencies (unique bigram/trigram patterns), Readability scores (LLM is frequently optimised for a middle-grade readability level, while human readability varies greatly depending on the author's goal and audience).

5) Behavioural Features

Metadata usage (Human frequently incorporates metadata related to real-world context, whereas LLM might not contain authorial metadata such as references or timestamps), typographical errors (humans occasionally make typos, slang, or informal shorthand, whereas LLM is unlikely to produce typos, and errors are more semantic or structural), and Emotional Tone (Human exhibits a variety of emotions and tonal shifts depending on context, whereas LLM may seem neutral or too positive and lacks deep emotional complexity).

D. LightGBM (Light Gradient Boosting Machine)

Microsoft developed LightGBM [30], which is a distributed, open-source, and high-performance gradient boosting system designed for scalability, accuracy, and efficiency. LightGBM was built on decision trees, which enhances model performance while minimizing memory usage. To optimize time for training and consumption of resources, LightGBM employs innovative techniques including Gradient based One Side Sampling (GOSS), which retains instances selectively with large gradients in the time of training. Additionally, it controls histogram based methods for effective construction of trees. These strategies are combined with various optimizations including effective data storage structures and leaf-wise tree development, which give LightGBM a good edge over other frameworks of gradient boosting.

LightGBM's primary characteristics include handling of categorical features directly, requiring less processing resources and memory, and training huge datasets more quickly than XGBoost. In this work, LightGBM (Light Gradient Boosting Machine) is used to determine if a text was created by a large language model (LLM) or by a human. This method takes either stylistic traits or text embeddings as input. LightGBM is appropriate for this authorship verification task since it can efficiently handle high-dimensional and large datasets.

V. Experimental Results

In this work, we selected the RoBERTa and T5 model as our pre-trained base models for generating word embeddings. Our hyperparameters for these models are set the margin to 0.5, set batch size to 16, and set the sequence length to maximum 512 (sequences greater than this are terminated). Five epochs of training are conducted, with the initial learning rate at 2e-5. For optimisation, we employ AdamW optimizer in every training session. We train the model using the officially supplied labelled dataset during the training phase.

A. Evaluation Measures

The proposed model was evaluated by using a various standard evaluation measures such as ROC-AUC, C@1, F1, Brier score, and F0.5u, which are typically applied in the task of authorship verification, along with the arithmetic mean of these measures to offer a systematic performance overview.

ROC-AUC gives information about how well the model can distinguish among classes at all thresholds [32] by measuring the area of curve under the receiver operating characteristics. The curve of ROC plots the True Positive Rate (TPR) against False Positive Rate (FPR) at different threshold values. The formula for ROC-AUC was given by Equation (1).

$$ROC - AUC = \int_0^1 TPR(t)\, d\left(FPR(t)\right) \qquad (1)$$

The Brier Score indicates the probability predictions accuracy by evaluating the mean squared error of the assigned probabilities [33]. Since the Brier score indicates a nearer proximity to the actual outcome, the lower the score, and the better performance. Equation (2) is used in Brier Score computation.

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}\left(\text{predicted probability}_i - \text{actual outcome}_i\right)^2 \quad (2)$$

C@1 is an improved accuracy that penalises uncertainty by averaging the remaining cases accuracy for non-answers (predictions with a 0.5 confidence score) [34]. When it is better to make no prediction than to make a mistaken one, this measure is especially helpful. Equation (3) represents the formula for C@1.

$$C@1 = \frac{Number\ of\ correct\ answers}{Total\ number\ of\ cases - Number\ of\ non\ answers} + \frac{Number\ of\ nonanswers}{Total\ number\ of\ cases} \quad (3)$$

The classifier's precision and recall capabilities are balanced by the F1 Score, which is the harmonic mean of recall and precision [35]. It is especially helpful when achieving a balance among recall and precision is desired. Equation (4) is a representation of the F1 score formula.

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Precision is denoted in Equation (5).

$$\text{Pr}ecision = \frac{TP}{TP + FP} \quad (5)$$

Recall is represented in Equation (6).

$$\text{Re}call = \frac{TP}{TP + FN} \quad (6)$$

Where, TP is True Positives, FN is False Negatives, and FP is False Positives.

A variation of the F-measure called F0.5u, which prioritises precision over recall. F0.5u is appropriate in conditions where false negatives are less expensive than false positives [36]. Equation (7) is used in F0.5u computation.

$$F0.5u = \left(1 + 0.5^2\right) \cdot \frac{\text{Precision} \times \text{Recall}}{0.5^2 \cdot \text{Precision} + \text{Recall}} \quad (7)$$

B. Results of AI Authorship Verification

The Table 2 shows the performance of proposed hybrid model for the task of GAIAV.

TABLE II. THE PERFORMANCE OF PROPOSED HYBRID MODEL FOR GENERATIVE AI AUTHORSHIP VERIFICATION TASK

Evaluation Metric	Score
ROC-AUC	0.996
Brier	0.970
C@1	0.989
F1	0.972
F0.5u	0.971
Mean	0.979

According to Table 2, the proposed hybrid model attained scores of 0.996 for ROC-AUC, 0.970 for Brier, 0.989 for C@1, 0.972 for F1, 0.971 for F0.5u, and 0.979 for mean on the training dataset of GAIAV task of PAN 2024. The ROC-AUC score of 0.996 suggests that the model has near-perfect performance in differentiating classes. A high Brier score of 0.970 suggests very well-calibrated probabilities, which is less common since the metric usually lies between 0 and 1. The C@1 score of 0.989 indicates highly confident and consistent predictions with minimal uncertainty. F1 score of 0.972 demonstrates strong recall and precision, indicating effective and balanced classification. The F0.5u score of 0.971 shows the model performs well, particularly in minimizing false positives while maintaining acceptable recall. The mean score of 0.979 indicates consistently high performance across all evaluated metrics.

C. Limitations of Proposed Approach

The limitations of proposed approach are susceptibility to adversarial attacks (e.g., backdoor or adversarial examples) and the challenge of maintaining high performance as generative models continue to evolve. In the susceptibility to adversarial attacks, maliciously crafted inputs or subtle perturbations, including backdoor triggers and adversarial examples, may degrade the reliability of AV detection or manipulate outputs in unintended ways. The rapid evolution of generative models presents a challenge in maintaining consistent detection performance. As LLMs grow more sophisticated and contextually aware, AV may become subtler and harder to identify, potentially requiring continual adaptation of the detection methodology.

VI. CONCLUSION AND FUTURE SCOPE

In the field of NLP, LLMs like BERT, GPT-3, Llama2, ChatGPT, GPT-4, and PaLM2, have shown outstanding performance in recent years. They are commonly used for tasks including answering questions, translating languages, and summarising documents. But as these technologies have become more widely used, concerns about the reliability and authenticity of writings produced by these models have drawn more public attention. Important societal difficulties include the dissemination of false information, the creation of illogical or deceptive content, and the piracy of original works and intellectual property. To differentiate the text written by human or machine, we proposed a hybrid model in this work. The proposed model used stylistic features, RoBERTa model embeddings, T5 model embeddings for representing the files as vectors. The LightGBM classifier was utilized to assess the efficiency of the proposed model. The PAN 2024 GAIAV dataset was considered in this experiment. The proposed hybrid model attained scores of 0.996 for ROC-AUC, 0.970 for Brier, 0.989 for C@1, 0.972 for F1, 0.971 for F0.5u, and 0.979 for mean on the training dataset of GAIAV task of PAN 2024.

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 ThinkingAnalysis, 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 theFifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in ComputerScience, Springer, Berlin Heidelberg New York, 2024.

[2] Raghunadha Reddy, T., Vijaya Pal Reddy, P. (2024). A New Text Representation Technique-Based Approach for Authorship Verification. Accelerating Discoveries in Data Science and Artificial Intelligence I. ICDSAI 2023. Springer Proceedings in Mathematics & Statistics, vol 421, pp 705-714, May 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] D. Ippolito, D. Duckworth, C. Callison-Burch, et al., Automatic detection of generated text is easiest when humans are fooled, arXiv preprint arXiv:1911.00650 (2019).

[5] K. Kavuri, T. R. Reddy, A. Gelli and B. H. D. D. Priyanka, "A Hybrid Approach for Language Variety Prediction Using BERT and T5 Embeddings," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, Bengaluru, India, 2025, pp. 2075-2083, doi: 10.1109/IDCIOT64235.2025.10914742.

[6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).

[7] A. Zeng, X. Liu, Z. Du, Z.Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414 (2022).

[8] T. Raghunadha Reddy, Naveed Wasim, Mohd Muzzammil Hassan, An Approach for Writing Style Change Detection using Pre-trained BERT model with similarity measures, International Journal of Novel Research and Development, Volume 8, issue 5, pp 653-658, 2023.

[9] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, M. Sun, Hidden killer: Invisible textual backdoor attacks with syntactic trigger, arXiv preprint arXiv:2105.12400 (2021).

[10] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, Mgtbench: Benchmarking machine-generated text detection, arXiv preprint arXiv:2303.14822 (2023).

[11] Ramesh Adavelli, K Karunakar, B Yugandhar and Raghunadha Reddy T, "Influence of Similarity Measures in Authorship Attribution", IEEE International Conference on New Trends in Engineering & Technology, GRT Institute of Engineering and Technology, Tiruvallur Dist, Chennai, Tamil Nadu, 7-8 September, 2018.

[12] Ye Zhu, Leilei Kong, AI Authorship Verification Based On Deberta Model, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[13] Zhanhong Ye, Yutong Zhong, Zhen Huang and Leilei Kong, Token Prediction as Implicit Classification for Generative AI Authorship Verification, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[14] Zepeng Wu, Wenyin Yang, Li Ma and Zikai Zhao, BertT: A Hybrid Neural Network Model for Generative AI Authorship Verification, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[15] Andric Valdez-Valenzuela, Helena Gómez-Adorno, 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, September 09–12, 2024, Grenoble, France

[16] Guihong Sun, Wenyin Yang, Li Ma, BCAV: A Generative AI Author Verification Model Based on the Integration of Bert and CNN, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[17] Rui Qin, Haoliang Qi and Yusheng Yi, A Model Fusion Approach for Generative AI Authorship Verification, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[18] Panagiotis Petropoulos, Vasilis Petropoulos, RoBERTa and Bi-LSTM for Human vs AI Generated Text Detection, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[19] Benjamin Ostrower, Jacob Wessell, and Abhinav Bindal, AI Authorship Verification: An Ensembled Approach, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[20] Jiajun Lv, Yong Han and Leilei Kong, Meta-Contrastive Learning for Generative AI Authorship Verification, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[21] Zhaojian Lin, Fanzhi Zeng, Yan Zhou, Xiangyu Liu and Yuexia Zhou, Voight-Kampff Generative AI Authorship Verification Based on T5, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[22] Haotian Lei, Xiangyu Liu, Guo Niu, Yan Zhou and Yuexia Zhou, Generative AI Authorship Verification based on ChatGLM, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[23] Kaicheng Huang, Haoliang Q, and Kai Yan, Generative AI Authorship Verification based on Contrastive Learning and Domain Adaptation, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[24] Weihang Huang, Jack Grieve, Authorial Language Models For AI Authorship Verification, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[25] Jijie Huang, Yang Chen, Man Luo and Yonglan Li, Generative AI Authorship Verification Of Tri-Sentence Analysis Base On The Bert Model, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[26] Linjiu Guo, Wenyin Yang, Li Ma, Jinli Ruan, BLGAV : Generative AI Author Verification Model Based on BERT and BiLSTM, Notebook for the PAN Lab at CLEF 2024, September 09–12, 2024, Grenoble, France

[27] K. Kavuri and M. Kavitha, "A Novel Document Representation Method for Author Profiling using Auto-Encoders," *Journal of Information Systems Engineering and Management*, vol. 10, no. 11s, 2025. [Online]. URL: https://doi.org/10.52783/jisem.v10i11s.1648

[28] T. G, R. Anil, S. Borgeaud, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023)

[29] OpenAI, GPT-4 Technical Report, 2023. URL: http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs].

[30] T. Raghunadha Reddy, B. Madhubala, G. Varshini, S. K. Fayaz, A Deep Learning Approach for Author Profiling using Word Embeddings, International Journal for Research in Applied Science & Engineering Technology, Volume 11, pp 1553-1558, 2023.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Vaseline Stoyanov:

[32] RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019).

[33] A. M. Carrington, D. G. Manuel, P. W. Fieguth, et al. "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation", IEEE Transactions on Pattern Analysis and Machine Intelligence 45.1 (2022): 329-341.

[34] W. Yang, J. Jiang, E. M. Schnellinger, et al. "Modified Brier score for evaluating prediction accuracy for binary outcomes." Statistical methods in medical research 31.12 (2022): 2287-2296.

[35] Archana Gelli, Karunakar Kavuri, T Raghunadha Reddy, Lakshmi Narayana M, "Distance Measures based Approach for Hate Speech Spreaders Detection", Journal of Applied Science and Computations, Volume IX, Issue XII, December/2022, Pages: 227 – 233

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.