National College of Ireland

Deep Learning-Based Approaches for Detecting AI-Generated Fake Reviews in Online Platforms

MSc Research Project
Data Analytics

Ashutosh Bangari
Student ID: 23412992

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ashutosh Bangari.....
Student ID:	23412992.....
Programme:	Msc Data Analytics..... **Year:** 2025-2026.....
Module:	Research Practicum.....
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	11 December 2025.....
Project Title:	Deep Learning-Based Approaches for Detecting AI-Generated Fake Reviews in Online Platforms .....
Word Count:	8724..... **Page Count** 22.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ashutosh Bangari.....

Date: 10 December 2025.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning-Based Approaches for Detecting AI-Generated Fake Reviews in Online Platforms

Forename Surname
Student ID

Abstract

The exponential growth of Artificial Intelligence and Natural Language Processing has blurred the boundaries between human-written and machine-generated text, creating new challenges in detecting synthetic content, ensuring authenticity, and maintaining academic integrity. While these technologies have advanced digital communication and sentiment analysis, they have also intensified the difficulty of distinguishing between real and AI-produced language. Existing approaches, including traditional machine learning and deep learning models, face limitations in efficiency, scalability, and contextual understanding. Classical ML algorithms are lightweight but lack semantic depth, whereas deep learning and transformer-based architectures capture rich linguistic features at the expense of high computational cost and long training times. This research leverages a hybrid pre-trained Transformer-GRU model to address these limitations. The model uses a pre-trained Transformer solely for feature extraction, capturing deep contextual embeddings without additional fine-tuning. These extracted representations are then fed into a Gated Recurrent Unit network that performs temporal sequence modelling and final classification. By freezing the transformer layers and training only the GRU, the hybrid design combines transformer-based semantic strength with computational efficiency. This structure enables faster convergence, lower resource consumption, and adaptability across multilingual and domain-specific applications. Experimental results show that the DistilBERT + GRU model outperformed baseline ML and DL models, achieving the highest accuracy of 86.12%. The framework aims to deliver an efficient, scalable, and context-aware system suitable for detecting AI-generated text and performing advanced sentiment analysis in real-world NLP environments.

1 Introduction

The massive development of the artificial intelligence and natural language processing has transformed digital communication, content creation, and sentiment analysis. As large-scale language models produce text that sounds human, it has become more challenging to draw the line between truly human text and AI-generated text, creating the problems of misinformation, academic dishonesty, and trust in online platforms. At the same time, the blistering development of online reviews, opinion and social media discussion requires strong text analysis systems that can process large, unstructured data. This has resulted in the use of fake reviews, sentiment classification and distinguishing between synthetic and human texts

to be identified as some of the most important areas of research in computational linguistics and machine learning (Alamleh et al., 2023).

With the ongoing development of AI-based text generation systems that become more and more fluent and coherent, the classical detection and classification systems have difficulties keeping the level of accuracy and efficiency. The current methods are based on shallow machine learning strategies with low contextual information or use deep learning models that are both computationally intensive and intractable. It, therefore, urgently needs effective hybrid systems to combine the semantic capacity of transformers with the lightweight efficiency of recurrent networks like Gated Recurrent Units (GRU). Such frameworks can be developed to advance real-time applications in academic integrity verification, review authenticity assessment and automated sentiment analysis in multilingual and domain-specific settings.

There is a great deal of literature on machine learning (ML) and deep learning (DL) approaches to text analytics. The classical ML models, including Logistic Regression, Support Vector Machines, Random Forest, and XGBoost, have proven to be quite accurate in text classification, sentiment detection, and fake reviews identification (Iqbal and Kashem, 2025; Alamleh et al., 2023). Accuracy is further increased by ensemble strategies such as ERF-XGB model which optimizes the feature selection (Alghazzawi et al., 2023). Conversely, the deep learning models like the LSTM, CNN and GRU networks have demonstrated to perform better in contextual semantics and sequential dependencies (Iqbal et al., 2022; Zamir et al., 2024; Kayabas et al., 2025). Transformer-based systems, such as BERT and XLM-RoBERTa, have been used in multilingual and domain-specific applications (Schaaff et al., 2024; De and Vats, 2025), whereas hybrid ensemble models have achieved impressive improvements by combining the power of ML and DL (Alhayan and Himdi, 2024; Boutadjine et al., 2025). Nevertheless, these methods tend to be computationally slow, inefficient in scaling, and do not have adaptability especially with low-resource languages and real-time systems.

Research Question

How can a light weight hybrid pre-trained Transformer GRU model improve the efficiency and accuracy of AI-generated text detection compared to conventional ML and DL approaches?

The necessity to come up with models that retain the same semantic depth of transformer-based embeddings with a small computational cost justifies this question and thus the trade-off between accuracy and efficiency that has been noted in existing systems.

The leveraging research will combine a pre-trained Transformer to extract the contextual features and then a lightweight GRU network to make the final classification. The leverage model will help (1) minimize calculation cost and training time by extracting features with frozen Transformer layers, (2) improve the accuracy of the model with the help of the ability of GRU to extract the temporal dependencies. Using the advantages of both paradigms, the

research presents a scalable, resource-saving framework, which can be used in various NLP applications like text identification with AI-generated texts, a gap that is essential in the modern environment of text analytics.

2 Literature Review

The AI and NLP have evolved at a fast rate, changing the trend of text generation, sentiment analysis, and content identification in digital media. Detecting a fake and an original text written by humans or AI, determining fake reviews, and understanding sentiments have become highly important tasks that demand sophisticated computational methods. The literature review will discuss the current studies of machine learning and deep learning methods, the opportunities of hybrid and ensemble models, and the most relevant gaps in research in order to use an effective Transformer-GRU-based model to conduct better text analysis and classification.

2.1 The Role of Machine Learning Models in AI Generated Review detection

The use of machine learning in text classification has been widely used in such tasks as identifying AI-generated content, sentiment analysis, and detecting fake reviews. The traditional ML models are used because they are efficient and can process high-dimensional text information. To illustrate, Iqbal and Kashem (2025) employed such feature extraction algorithms as Count Vector, Bag of Words, and Hashing Vectorization to convert text data into numerical forms, which can be used in the work of ML algorithms. They further increased the accuracy of the classification using the dimensionality reduction methods such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), with the accuracy of 99.4 percent with the combination of HV, LDA, and XGBoost. Similarly, Alamleh et al. (2023) have demonstrated that the ML classifiers can be used to distinguish between texts written by AI and human writers, and SVM and RF are the most precise and least computationally intensive. These findings underscore the fact that classical ML applications to the resolution of the existing text classification challenges continue to be relevant even in the face of the growing role of deep learning.

Another important field in which ML models have demonstrated a great success is sentiment analysis. Yang, J.L. (2025) suggested the ML framework to classify sentiments and detect fake reviews with the help of such algorithms as SVM, and NB. They efficiently identified positive, negative, and neutral reviews and fake content using their system and showed that optimally trained ML models can be used to compete well without using complex deep learning architectures. Al-Qudah et al. (2025) took it to the next level and applied ordinal regression with evolutionary XGBoost (PSO-XGB) to the sentiment analysis performed by machine learning, which demonstrated better performance than SVM and other optimization-based methods, especially in Arabic-language data. Additionally, Alghazzawi et al. (2023) proposed the XGBoost (ERF-XGB) model that made use of the Harris Hawk Optimization (HHO) algorithm to select features, which contributed to improvement of various problems,

including polysemy and word disambiguation. This model was able to reach 98.7% accuracy on the ChnSentiCorp dataset and 98.2% on IMDB data which shows the significance of ensemble and optimization methods in optimizing the performance of the ML model.

Machine learning has played a central role in detection of fake news and misinformation. Zamir et al. (2024) showed the use of ML models such as SVM, LR, KNN, DT, and RF to classify the news articles on the topic of COVID-19 as real or fabricated in order to enhance media literacy and resilience of society to misinformation. As Kansal and Diwaker (2025) highlighted, sentiment analysis systems based on machine learning on social media platforms can assist in business intelligence applications and reputation management in situations where a large volume of unstructured text data is involved. Moreover, Uzun (2023) has talked about different ML-based tools including, but not limited to, copy leaks or Turnitin, stylometric analysis or metadata inspection to identify AI-generated content and maintain academic integrity. Taken together, these analyses depict that machine learning is a sound, versatile, and solution to various text analysis problems, such as sentiment classification and fake review detection, or misinformation filter and AI-content detection, as a solid methodological base of more sophisticated deep learning and hybrid models.

2.2 Deep Learning based AI generated review detection

Deep learning systems have transformed natural language processing tasks by learning complex hierarchical representations directly on data, and are much more effective than traditional machine learning techniques in most text classification tasks. Contrary to traditional models, which are highly sensitive to manual feature engineering, DL models like Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Gated Recurrent Units and Transformer-based models are able to discover contextual and semantic patterns in large amounts of unstructured text automatically. Iqbal et al. (2022) showed that three LSTM variations are effective in sequential dependencies of consumer reviews and result in better sentiment classification with less complexity. In the same manner, Zamir et al. (2024) used BiGRU networks to detect fake news in the COVID-19 pandemic and achieved high results with an accuracy of 0.91 and F1-score of 0.92. These findings indicate that recurrent neural networks (RNNs) and its variants, in particular the ones that are most effective on sequential data, can aid the process of constructing subtle information in the text that the traditional models may overlook.

Deep learning has contributed to the distinction of artificial and human writing, as well as the fight against the increasing problem of fake news, plagiarism, and online trust. LSTM networks were also used by Kayabas et al. (2025) to identify AI generated Turkish texts compared to human written texts with an accuracy of 97.28 and an F1 score of 0.97. In the paper, the strength of LSTM in low resources language sequential classification has been presented, indicating its flexibility and effectiveness. The authors also were able to conclude that LSTM-based models can identify AI-based essays with high accuracy, precision, recall, and F1-score, which is why they can be useful in the context of academic integrity (Chimata et al., 2024). Boutadjine et al. (2025) have employed a transfer learning model that is based

on the state-of-the-art transformer-based LLMs, and they achieved a 94-99 percent accuracy to recognize synthetic texts. In this paper, it was found that the gap in capabilities between the human brain and machine models to detect AI-generated content is large, which justifies the application of DL-based technologies in order to detect Deepfake text with high precision.

Also, deep learning can be used in more complicated tasks than binary classification like text analysis in multilingual and domain-specific applications. Schaaff et al. (2024) created classifiers based on the perplexity, semantic similarity, readability, and AI feedback feature to identify AI-generated and AI-rephrased texts of various languages, which was superior compared to current tools, including GPTZero and ZeroGPT. In their study, De and Vats (2025) used transformer-based architectures such as Indic-BERT, IndicSBERT, MuRIL, and XLM-RoBERTa to identify AI-generated product reviews in Dravidian languages (Tamil and Malayalam), which showed that advanced DL models can be useful in the low-resource language environment. Al Bataineh et al. (2025) used deep learning algorithms including LSTM, CNN, and CNN-LSTM with different text vectorization methods and obtained a high accuracy of more than 90 percent and allowed explainable predictions in SHAP. Overall, the above studies show that deep learning models not only provide better results concerning the accuracy of text classification, but also the scalability, and cross-lingual applicability of the NLP systems, which puts them at the centre of future improvements in AI-generated text detection and sentiment analysis.

2.3 Hybrid and Ensemble Techniques in Text Analysis and Content Detection

The recent progress in the field of natural language processing demonstrated that the performance of text classification can be greatly enhanced by applying several models, either machine learning or deep learning, or both. These hybrid and ensemble methods combine the advantages of the separate models to minimize their defects which makes them more accurate, strong and able to generalize. Alghazzawi et al. (2023) suggested the Ensemble Random Forest-based XGBoost (ERF-XGB) sentiment polarity classification model of e-commerce product reviews. Their model, with predictive capabilities of Random Forest and XGBoost as well as feature selection with the Harris Hawk Optimization (HHO) algorithm, had impressive accuracy rates of 98.7% and 98.2% on two benchmark datasets. This shows how ensemble techniques can alleviate the problems such as polysemy and word disambiguation, which can in most cases be a problem with standalone models. On the same note, Alhayan and Himdi (2024) examined the use of ensemble learning in the process of ranking human and AI-generated Arabic reviews, which revealed that soft voting between Logistic Regression and Convolutional Neural Networks had a high accuracy rate of 89.7 percent, which is close to that of AraBERT. These results suggest that ensemble methods can be more accurate predictors, as well as provide better robustness to across languages and text types.

Multilingual and domain-specific problems are also important issues that are best addressed by hybrid solutions that fail to work with the traditional single-model approach. De and Vats

(2025) showed that hybrid models of the traditional methods of ML and sophisticated transformer-based models such as Indic-BERT, IndicSBERT, and XLM-RoBERTa were effective in detecting AI based product reviews in Dravidian languages. Their findings emphasized the benefits of hybrid models in low-resource language environments, where data sparseness can usually restrict the effectiveness of individual models. Also, Schaaff et al. (2024) integrated different feature-based systems, including perplexity, semantic similarity, and readability, and deep learning classifiers to distinguish AI-generated and AI-rephrased text in diverse languages and fields. Their ensemble models outperformed the current detectors such as GPTZero and ZeroGPT by a large margin, demonstrating the applicability and flexibility of hybrid solutions in multilingual NLP tasks. The studies highlight that hybridization is not just concerned with stacking models but also incorporation of various feature engineering, optimization and vectorization methods to tackle the complicated problems in the real world.

In addition, hybrid and ensemble methods also help in interpretation and practical implementation, closing the gap between theory and practice. Iqbal and Kashem (2025) used explainability to combine Hashing Vectorization with LDA and XGBoost to reach a 99.4% accuracy as well as explain model predictions, using explainable AI (XAI) tools such as SHAP and LIME. Al Bataineh et al. (2025) took this to a new level by incorporating both ML and DL models into a user-friendly web application that uses Stream lit to enable the interaction of non-experts with AI detection models in real time. Another study by Boutadjine et al. (2025) highlighted the usefulness of ensemble and transfer learning techniques to identify Deepfake text where the accuracy was 94-99% in far exceeding the ability of humans to detect Deepfake text. These demonstrate that hybrid methods can not only be able to boost predictive performance, but also be more usable, and scalable, which makes them essential to fight fake reviews, misinformation, and AI-generated content in practice. With the ongoing evolution of NLP, integrative approaches of this nature are expected to be at the heart of designing the next generation text analysis systems.

2.4 Emerging Perspectives on AI-Generated Reviews and Detection Techniques

However, recent studies have shifted a significant focus towards the comprehension of the essence of differences between the so-called fake review generated by AI, the fake one generated by the human, and the one generated by the human. The research of reviews was carried out on a massive scale with Zhao et al. (2025) discovering that AI-written fake reviews are distinctively different compared to classic deceitful human reviews. In comparison with fakes that are produced by humans and tend to include signs of psychological manipulation (like exaggeration and emotional bias), AI-generated fake reviews are seen to be more mechanicalized, empathetic, and specific. Such results do not match the current fake review detection paradigm which is mainly formulated to detect fake intent that may belong to humans. The paper emphasizes that it is important to come up with detection models in which algorithmic patterns of writing are explicitly considered and not

just the traditional indicators of linguistic deception thus redefining the theoretical underpinnings of fake review detection.

In addition to textual analysis, the recent experience in the field of deep learning architectures addresses the idea of combining spatial and temporal representation to identify AI-generated content in terms of various modalities. A CNN -LSTM -based system that combines convolutional neural networks to extract spatial features and LSTM layers to learn time dependencies were proposed by Longwei et al. (2025) as a method of identifying AI-generated deep fake videos. The principle behind hybrid architectures though involves video data is not very different, even though contextual and sequential dependencies are paramount in text-based fake review detection. On the same note, Blake et al. (2025) developed a Bi-LSTM and attention-based AI-generated text detection model with the help of convolutional layers, bidirectional recurrent network, and attention-based mechanisms to identify salient linguistic aspects. As they convincingly prove, hybrid deep learning frameworks can be highly accurate yet highly efficient in terms of computational cost, which makes the combination of sequential modeling and sophisticated feature extraction methods to be effective.

Alongside the detection of the algorithmic one, the concept of consumer perception and features of consumer behavior have received recent literatures. Tuomi et al. (2025) analyzed the potentiality of AI-generated restaurant reviews and revealed that users had difficulties differentiating AI-generated reviews and those written by human beings, with the accuracy of the analysis being parallel to random guesses. This vulnerability highlights the threat of AI-generated content, which is increasing and the necessity to have automated detection technologies. To support this view, Mughal et al. (2025) put more weight on user behavioral features when detecting fake reviews, especially with the low-resource language like Roman Urdu. They have found that behavioral features in most cases perform better than textual features and that gradient boosting models performed better using such features. Together, these works indicate that trying to identify the presence of fake reviews in the future by combining linguistic patterns and sequential learning with behavioral information is necessary to deal with the changing nature of AI-generated reviews.

2.5 Research Gaps

Although machine learning and deep learning methods have made a major advancement in text classification, sentiment analysis, and AI-generated content detection, there are still a number of limitations and disadvantages in the current research. More efficient yet models of traditional machine learning do not always work with the complex contextual and semantic relationship found in textual data (Iqbal and Kashem, 2025.). They have a restricted scalability and adaptability due to manual feature engineering that are required by multilingual or domain-specific applications (Al-Qudah et al., 2025). Conversely, deep learning models, and large-scale transformer models, such as BERT, XLNet, and ERNIE, are more effective but are computationally inefficient, require a lot of memory, and are slow to

train (Boutadjine et al., 2025; Schaaff et al., 2024). The computational resources used in these models are also very high and this limits their use in actual systems particularly in low resource settings. Moreover, purely deep learning-based methods are typically not and the tendency to overfit limited data also indicates the importance of more efficient, scalable as well as explicable methods.

One such possible solution to these shortcomings Is the creation of hybrid architectures that use pre-trained transformer models to extract features and lightweight recurrent networks such as GRU to predict. In this method, the models are transformer-based models that have been trained solely to extract features from the text in their capacity to extract deep contextual representations of the text without being re-trained entirely. These dense feature representations are subsequently passed through a GRU layer to do final classification. As the transformer component is already trained and frozen in training, the model is much less expensive in terms of computational cost and training time than end-to-end transformer fine-tuning. Furthermore, simplified in structure and having fewer parameters than LSTM, GRU can acquire the knowledge of temporal dependencies and classify objects with a high degree of accuracy. This separation of labour, transformer-generated feature extraction and GRU-based prediction, allows the hybrid model to retain the semantic richness of transformer embeddings and significantly reduces computational time, memory, and training time.

The potential of such a hybrid pre-trained Transformer-GRU model is that it can be more successful than current ML, DL, and even transformer-based models due to the integration of their respective advantages and the reduction of their disadvantages. This model uses transformers as feature extractors only, which means that it does not require the large resource requirements and prolonged fine-tuning periods that transformer training has. Simultaneously, GRU guarantees rapid convergence and efficient learning of downstream tasks, which allows the model to scale on languages, domains, and resource limitations. Not only is this approach more memory-efficient than training large transformer networks end-to-end, but it is also more accurate than traditional ML approaches. Consequently, this type of hybrid architectures may be the future solution to sentiment analysis, AI-generated text detector, and fake content classifier, being high-performance, efficient, and applicable to the real-life use of academic, commercial, and multilingual NLP applications.

3 Methodology

3.1 Research Frameworks Using CRISP-DM Methodology

Research framework is based on CRISP-DM (Cross-Industry Standard Process for Data Mining) to make the approach systematic and structured. The process is initiated with Business Understanding, in which the necessity to identify AI-generated content with the high degree of efficiency is established. The Data Understanding step will be based on gathering and exploring various text data, both AI and human-written texts. The Data Preparation stage is concerned with cleaning, tokenization and encoding of text to be fed into the transformers. The Modelling phase is the one when the frozen transformer is combined

with the GRU architecture, and then the Evaluation, during which the model performance is evaluated by the help of the corresponding metrics, including precision, recall, and F1-score. Lastly, the Deployment stage defines possible implementation plans of real-time detection system and sentiment analysis tools. CRISP-DM alignment with this hybrid model guarantees the reproducibility, scalability and flexibility of the research model.

3.2 Dataset

The dataset used in this study is the reviews made by customers on the different category, and each record has four major attributes namely the category, rating, label, and text. The category column determines the product domain and the rating column gives numerical feedback of 1 to 5 which is the level of satisfaction of the user. The label field is used to note the status of the review as either CG (computer-generated) or OR (human-written) to allow binary classification of AI-generated text detection. The text column is where the review text is stored and it is the main input to the model training and analysis. This data is a good representation of both real and fake reviews, with a variety of different linguistic patterns, tones, and degrees of sentiment. It is also especially applicable to testing the hybrid Transformer-GRU model, which can extract semantic features of complex sentences and learn temporal dependencies in sequential textual data, which will enable the effective assessment of the model to perform the task of authenticity and sentiment analysis of reviews in the real world.

The dataset used in this research is collected from e-commerce platforms. It contains product reviews across multiple categories such as Home and Kitchen, Sports and Outdoors, Electronics, Movies and TV, Tools and Home Improvement, Pet Supplies, Kindle Store, Books, Toys and Games, and Clothing, Shoes and Jewelry. These categories represent typical review sections commonly found on large online shopping platforms. The goal of the study is to predict whether a given review in this dataset is computer-generated or written by a human, making the approach applicable to real-world e-commerce review systems.

3.3 Data Preprocessing

Preprocessing of data is a critical process to ascertain that the textual data is clean, consistent and prepared to be extracted into features and model trained. The data sample of this study originally had four columns category, rating, label and text of 40, 432 entries. Nonetheless, the label and text columns were kept in order to classify the texts as the two fields directly reflect the type of each review and the corresponding review text. The data was checked on missing data and duplication. No records were lost, and 20 duplicate entries were detected and then eliminated, and a clean dataset of 40,412 samples was obtained. This guaranteed elimination of redundancy, which enhanced reliability of data and avoided biased training outcomes.

Following the primary filtering process, the text-cleaning process was conducted to make the textual data analytical. The reviews were all transformed to lower-case to ensure consistency and also avoid the possibility of case sensitivity to influence the representation of words. Noise like HTML tags, URLs, digits and punctuation were eliminated in a systematic way so as to avoid interference with the interpretation of textual patterns by the model. Another aspect that was eliminated was excess whitespaces to standardize the composition of any sentence. The text was then cleaned and tokenized into separate words, which was then finer controlled during lemmatization. Common words that do not add much to the semantic meaning were removed and the rest of tokens were lemmatized with WordNet Lemmatize to simplify them to their base forms. This move guaranteed that the similar meaning words were dealt with equally and the model could better capture meaningful linguistic patterns.

After cleaning the text, the label column was coded as numbers with the help of a label encoder to convert the categorical labels (e.g. fake or real) into numbers that are compatible with machine learning algorithms. To extract features, a Term Frequency- Inverse Document Frequency (TF-IDF) vectorizer was used to transform the text processed into numerical vectors. TF-IDF method shows relevance of words in relation to the entire corpus, and the method reflects the weightings of words in a relevant context. The vectorizer was set to take into account the unigrams and the bigrams with a feature limit of 5,000 and default stop word removal which is inherent in the English language to further streamline the computational performance. Lastly, the data set was split into training and testing data sets by the 80-20 split to allow balanced evaluation and model generalization. All these preprocessing stages helped to pre-process the data, making it easier to learn, and the hybrid Transformer-GRU model operated on structured, meaningful, and noise-free textual data.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted in order to get a rough idea about the structure and distribution of the dataset prior to model training. A number of visualizations were created, such as count plots to see how the labels of reviews were distributed, so that both real and generated ones were fairly represented. A histogram of text lengths was used to investigate differences in the size of reviews and the writing patterns by samples. Moreover, a word cloud diagram was produced that pointed to the most used words to gain an idea of the most common linguistic patterns and words that are prevalent in the corpus of the reviews.

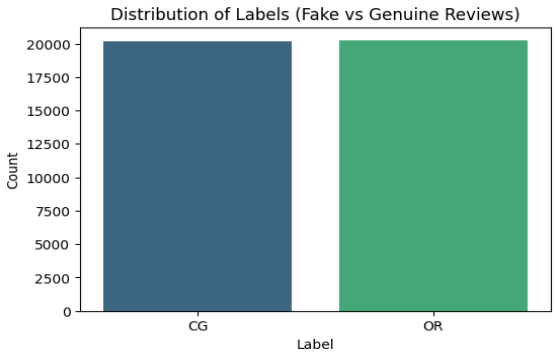

Figure 1. Review Classification Labels

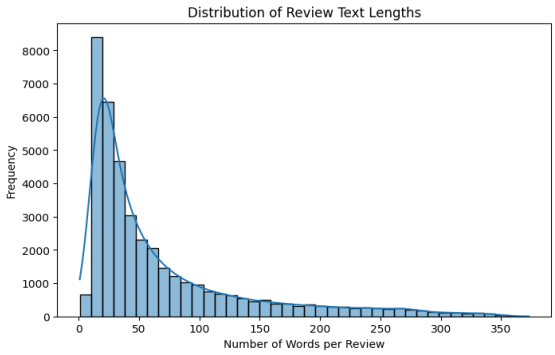

Figure 2. Text Length Distribution

In figure 1, is the distribution of CG and OR review labels. Both groups have almost equal numbers of samples and this means that they have a well-balanced dataset which can be used in classification. The consistency of the bar heights proves that there is no stereotyping of the classes, which is in favor of fair and impartial training of the model. Figure 2 below demonstrates the length of review texts. Reviews are mostly less than 50 words in length, which is a skewed right distribution, with the short and terse reviews being the most frequent, with longer and elaborate ones being relatively fewer.

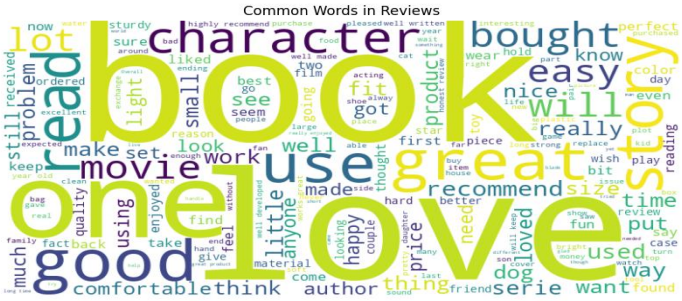

Figure 3. Most Frequently Used Words in Customer Reviews

Figure 3 represents a visualization of the most common words used in all the reviews within the dataset in a plot word cloud. Words like book, love, great, good, and use are the larger words and thus can be seen to have greater occurrence and capture general sentiments and subjects that are articulated by users. It offers a summary of the most common themes in the reviews, focusing on the positive experiences and commonly spoken about things.

3.5 Hybrid Pre-Trained Transformer-GRU Model

To develop an effective and context-sensitive text classification model, this paper uses a hybrid pre-trained DistilBERT-GRU model. Another element of Transformer is a lightweight version of BERT, which is called DistilBERT and is a powerful feature extractor. It compresses input text into dense contextual representations that encode linguistic, syntactic and semantic data and is both computationally and memory efficient as compared to the original BERT. As DistilBERT is trained in the frozen state, without further fine-tuning it

retains its general knowledge of language in large-scale pre-training, and minimizes the computational cost at training.

GRU network represents the second level of the architecture in which the output of DistilBERT is inputted in the form of sequential input. GRU conducts the representations over time steps to learn about temporal and contextual dependencies of the text. GRUs are an effective long term dependency modeling using update and reset gates with minimal training complexity as an alternative to LSTMs. This architecture is only trained by the GRU, which allows achieving a faster convergence, lesser memory consumption, and better scalability.

Lastly, a classification layer is used to map the output of the GRU to desired categories (e.g. human written vs. AI written reviews). Optimization of the model is achieved with the help of algorithms such as Adam with the corresponding loss functions (binary or categorical cross-entropy). Regularization methods like dropout are used to avert overfitting and aid in generalization. This hybrid architecture strikes a productive compromise between the depth of contextual understanding and the efficiency of sequential learning with the rich semantic embeddings of DistilBERT and the sequential learning efficiency of the GRU, which makes it an effective architecture in multilingual classification and AI-generated content detection in specific domains.

3.6 DistilBERT-GRU Architecture

The DistilBERT-GRU architecture adopted in this research paper incorporates the advantages of the Transformer-based contextual understanding and the recurrent sequential learning, which is quite effective when it comes to differentiating between AI-generated and human-written reviews. DistilBERT is the initial part of the model and it is a pre-trained Transformer encoder that transforms every review to dense and context-rich embeddings. The embeddings are able to memorize profound semantic data including word sense, tone, and language structure without further fine-tuning. This allows the model to store strong language representations trained on large-scale training corpora, allowing it to be sensitive to small textual variations between machine and authentic human generated writing.

The contextual embeddings generated by DistilBERT are then fed to the GRU (Gated Recurrent Unit) network that takes them as a sequence to learn temporal and structural relationships in the text. GRUs are especially effective with sequential tasks as they are capable of storing important information across time and are less complicated to compute than LSTMs. Regarding the present study, the GRU determines the stylistic and context-related patterns, including repetitive phrases, unnatural transitions, or degree of coherence, which tend to vary between AI and human-generated reviews. Training the GRU alone with DistilBERT frozen makes the architecture lightweight, faster to train, and less memory-intensive and is a perfect fit where real-time or resource-constrained applications are required.

The last classification layer converts the representations learned by the GRU to the output of the labels that allow the correct prediction of whether a review is AI-generated (CG) or human-written (OR). The hybrid model has a number of benefits: it is less expensive to compute, it does not overfit as fully fine-tuned transformers, and it can perform well with different text lengths and writing styles. DistilBERT semantic richness and the sequential dependency learning of GRU guarantee the presence of minor linguistic hints, which is why the architecture is well-fit to address the purpose of the study to detect fake or AI-written reviews in real-world data.

The transformer layers are frozen, which simplifies the model training process and decreases the number of parameters that are updated. This minimizes the cost of computation and memory and comes in handy when such resources are scarce. It also implies, however, that the model cannot adapt to new domains or languages completely as the frozen layers do not acquire new patterns. In this project, this trade-off was reasonable since DistilBERT already has a good general language understanding, and the GRU layer is concerned with task-specific learning.

3.7 Model Evaluation Metrics

All the model performance measures are evaluated in terms of standard classification measures such as Accuracy, Precision, Recall, and F1-Score in order to get reliable detection of AI-generated and human-written reviews. Accuracy is a general performance measure, which is the ratio of total reviews correctly classified. Nevertheless, as accuracy alone can be deceiving when there is an imbalance in the classes, other measures are applied. The precision measures the percentage of reviews estimated as AI or human-written that are actually accurate and, as such, it reduces false positive rates. Recall is used to determine the capacity of the model to identify all the relevant instances correctly, which assists in determining the effectiveness of the system to capture AI-generated reviews without omitting them. The F1-Score which is the harmonic mean of the precision and recall is a more balanced measure of performance particularly when the cost of misclassification is high. Combined, these measures give a holistic picture of the model to identify how well it is differentiating between AI-generated and authentic reviews.

4 Design Specification

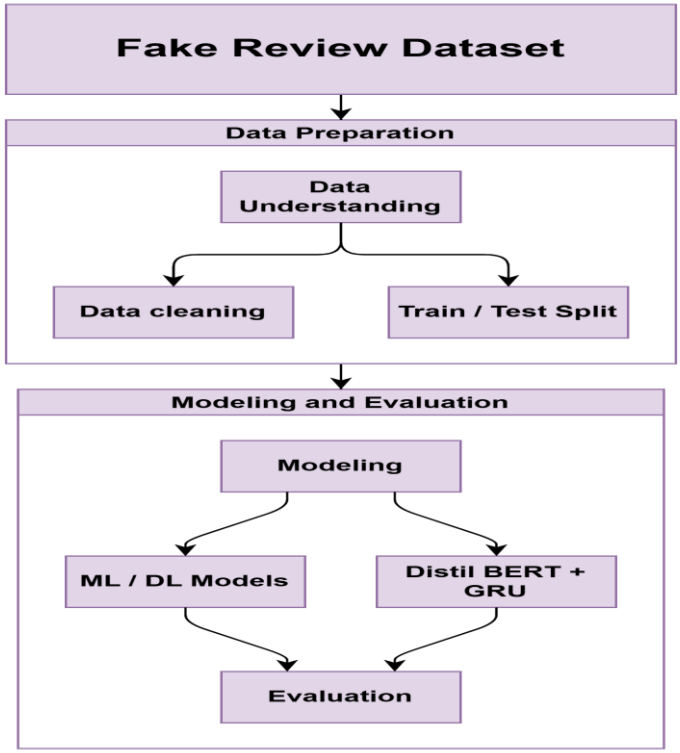

Figure 4. Architecture Overview

4.1 System Architecture Overview

The architecture of the system depicted in figure 4, in the diagram, is that of a pipeline that converts raw review data into correct fake-review forecast. It starts with Data Preparation whereby the dataset is understood, cleaned, and a train-test split to provide high-quality and well-structured input. The obtained processed data then proceeds to the Modeling and Evaluation phase which consists of two parallel modeling paths: traditional Machine Learning and Deep Learning models and the leverage DistilBERT + GRU hybrid model. Traditional models provide their foundations whereas the hybrid architecture relies on DistilBERT to obtain deep contextual embeddings and GRU to acquire sequential patterns to classify them. The two directions culminate in a single Evaluation stage, in which the model performance can be evaluated in terms of accuracy, precision, recall, and F1-score. This architecture will guarantee an efficient, systematic flow of raw data to healthy fake-review detection.

4.2 Workflow and Pipeline Implementation

The pipeline and workflow implementation of the fake review detection system is designed in a structured and modular way such that it ensures an efficient data flow in terms of ingestion and prediction in machine learning, deep learning, and the DistilBERT + GRU hybrid model. The system starts with a general design specification whereby every component data ingestion, preprocessing, feature engineering, modeling, and evaluation are structured into independent but interdependent modules to improve scalability, maintainability and

reproducibility. These modules are incorporated into a layered system in the architecture where raw text is initially received by the data ingestion layer, cleaned and normalized in the preprocessing layer and then sent to distinct model training layers based on the algorithmic method of choice. The workflow transforms the dataset by loading the raw review text, eliminating duplicates and noise, tokenizing and lemmatizing, encoding labels, creating TF-IDF vectors to be used in ML models, scaling numerical features when necessary, and transforming sequences to neural network models; the pipeline used to transform text to token IDs and attention masks and extract contextual embeddings is DistilBERT + GRU. The model training pipeline provides three branches: traditional ML models using TF-IDF matrices including Decision Tree, Random Forest and XGBoost deep learning models using padded word-index sequences including GRU; and the hybrid model of DistilBERT + GRU where the former uses transformer-based tokenization and attention encoding to generate dense embeddings which are used as inputs to a GRU layer to learn sequentially and be finally classified. The evaluation of all trained models is based on the accuracy, precision, recall, and F1-score to guarantee reliable and comparative evaluation of the architectures to enable a holistic and consistent pipeline of detecting AI-generated and human-written reviews.

5 Implementation

5.1 Machine Learning Models Implementation

The machine learning models used in this work Decision Tree, Random Forest, and XGBoost were trained with feature vectors based on TF-IDF on the preprocessed reviews text that allows handling high-dimensional sparse data and enables efficient work with the high-dimensional feature space. To do the hyperparameter tuning of each model, GridSearchCV was applied with fold cross-validation to generate a systematic evaluation of several settings and determine the best settings to achieve precise fake-review classification.

The Decision Tree classifier was adjusted with the max depth parameter by using a range of numerical values which comprised of 10, 20, 30, 40, 50, 70, 100 and 200 to ensure that the model tested a wide range of structural complexity as indicated in figure 5. Following the comparison of the performance of these configurations, it was found that the depth of 100 was the best that gave the most effective balance between the learning capacity and preventing overfitting. Using this optimal depth, the model was trained and evaluated performance that provided an overall review of its ability to differentiate between AI-written and human-written reviews.

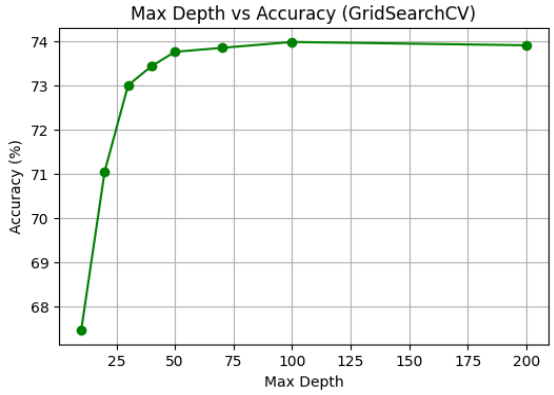

Figure 5. Accuracy Decision Tree Accuracy vs Max Depth

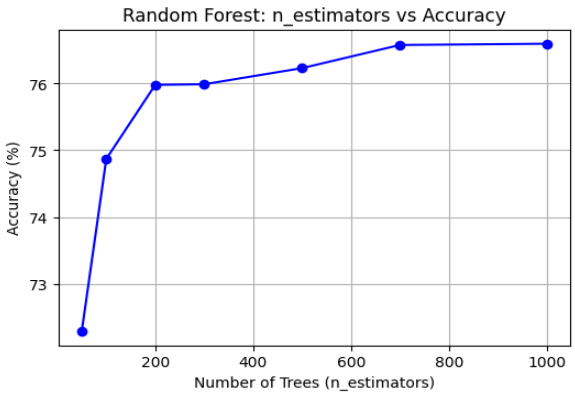

Figure 6. Accuracy of the random Forest VS estimator's number.

Random Forest model was tuned using a single parameter that is the n-estimators, which is used to regulate the number of decision trees in the ensemble. A number of numerical setups were experimented namely 50, 100, 200, 300, 500, 700 and 1000 trees to investigate the influence of increasing the size of the ensemble on predictive accuracy and the stability of the model as illustrated in figure 6. Of these values, number of estimators 1000 gave the highest scores showing that the greater the amount of collected trees, the better the generalization and the greater the reliability in differentiating between AI-generated and human-written reviews. Once this optimum parameter has been chosen, the model was trained and tested through the use of different measures in order to test its performance.

In the case of the XGBoost classifier, the most important parameter to tune was the number of estimators, which dictates the number of boosting rounds to use in training the model. There were four numerical settings 30, 50, 100, and 150 that were tested to determine the impact of deeper boosting sequences on learning effectiveness. The number of estimators configuration was considered the best performing and, thus, it can be concluded that more boosting rounds improved the capacity of the model to detect minor differences between AI-generated reviews and original human-written reviews without overfitting. The last model was then measured through different measures as well as a classification report and confusion matrix to fully gauge the classification ability of the model.

5.2 Deep Learning Model Implementation

In this study, a neural architecture based on a GRU was used to implement the deep learning model that was used to identify sequential patterns in the review text. In order to ready the input data, all the review labels were initially coded into numerical values using a label encoder, and the text was converted into fixed-length sequences by using a tokenizer with a vocabulary size of 20,000 and converting each review to a sequence of integer indices. Such sequences were then padded or truncated to a constant length of 100 to have similar input sizes in the neural network. The model was a two-stage GRU architecture, where the embedding dimension was 64 to map input tokens to dense vectors, and the first GRU layer had 64 units and output complete sequences, and the second GRU layer had 32 units to learn

to learn compressed time dependencies. Dropout layers were used, along with batch normalization, were integrated to prevent overfitting and stabilize training. A fully connected layer with 32 neurons and ReLU activation further refined the learned features, and the final output layer used a sigmoid activation for binary classification of AI-generated versus human-written reviews. The model was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 512, incorporating early stopping with a patience to prevent unnecessary training once validation performance stopped improving.

5.3 DistilBERT–GRU Hybrid Model Implementation

DistilBERT-GRU hybrid model was introduced in order to utilize the strength of the pre-trained transformer in terms of contextual embedding and the ability of the recurrent neural network in terms of sequential learning. This was initiated by loading the DistilBERT tokenizer and the model so that contextual embeddings of each of the reviews were generated with the transformer using the evaluation mode to make sure that its weights were not changed. The textual data was tokenized and converted to fixed length sequences of 128 token and the DistilBERT generated sentence embeddings of 768 dimensions by averaging the final hidden states. Such embeddings were used as input to the GRU classifier, and there was no longer a need to use text vectorization. GRU-based architecture was an input layer, which received the 768-dimensional embeddings, a reshape operation, which created a sequence-like structure, a GRU layer with dropout rates as a regularization method, batch normalization to ensure the stability of training, and a dense layer of 64 neurons with ReLU as a final output layer to generate a binary classification. The Adam optimizer was used to train the model with a learning rate, binary cross-entropy loss, and a batch size of 64 up to 15 epochs and early stopping in conjunction with adaptive learning rate reduction used to improve convergence. The predictions were made and tested after the training to show how the hybrid model was able to utilize transformer-level semantic knowledge, as well as GRU-based temporal pattern learning, to achieve very high levels of fake review detection.

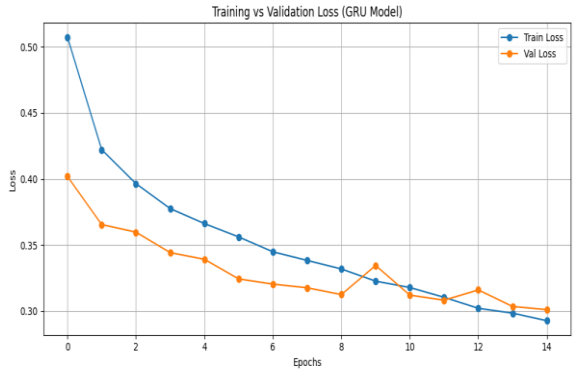

Figure 7. Training vs Validation Loss of Distil Bert +GRU

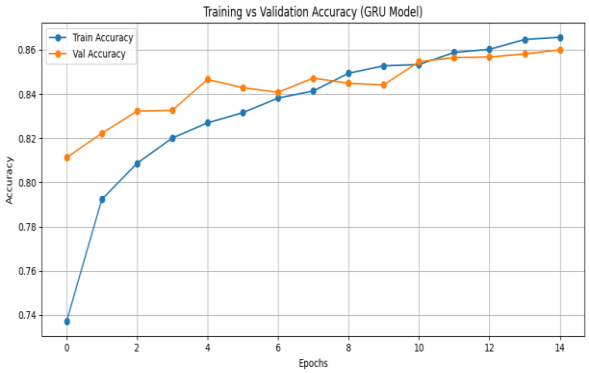

Figure 8. Training vs Validation accuracy Distil Bert +GRU

The training and validation loss curves in figure 7 demonstrate a consistent decrease in loss throughout the epochs, which implies that the GRU model is learning successfully and continuously decreasing error in prediction. The proximity of the two curves indicates consistent learning patterns and consistency in analyzing new information. Both in figure 8, Training and validation accuracy gain steadily with time indicating that as the model is trained it becomes more accurate. The accuracy of the validation is nearly the same as the training accuracy, which validates the consistent learning and good performance of the unseen data.

5.4 Libraries Used

Data manipulation and numerical operations were done using NumPy and Pandas, preprocessing, label encoding and machine learning models were done using scikit-learn, deep learning and GRU-based architectures were built and trained using TensorFlow and Keras. The loading, tokenization, and embedding generation of DistilBERT became possible with the help of PyTorch and the Hugging Face Transformers library. The visualization and performance plots were created with the help of Matplotlib and Seaborn. The combination of these libraries allowed developing and testing models efficiently throughout the workflow.

Table 1. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	74.18%	74.19%	74.18%	74.18%
Random Forest	77.32%	77.42%	77.32%	77.30%
XGBoost	82.15%	82.21%	82.15%	82.14%
GRU	83.36%	84.93%	83.36%	83.17%
DistilBERT + GRU	86.12%	86.13%	86.12%	86.12%

6 Evaluation Results

In a bid to determine the efficiency of the strategies, all models were well-tested based on different metrics as the main performance metrics. The findings of this paper give an idea of the ability to distinguish AI-generated and human-written reviews by each model, which can be used to draw a clear picture of machine learning, deep learning, and transformer-based models. The strengths and weaknesses of each method are also brought out in the evaluation as the various methods are shown to work or not when they are used on actual review data. The results below are the specific results of each type of model as displayed in table 1.

6.1 Machine Learning Model Results

The Decision Tree model was able to reach an accuracy of 74.18, which implies that the model was able to classify more than three-quarters of the reviews as human or AI generated. It had a precision of 74.19 which means that when the model made a prediction that a particular review would fall into a specific class, it was right seventy four percent of the time. The model had a recall value of 74.18% indicating that the model had the capacity to identify approximately three-quarters of all instances of each class that occurred. The fact that the model works with an F1-score of 74.18, a number that represents a compromise between the accuracy of predictions and detection completeness, though with an average classification ability overall, indicates that the model is consistently accurate and successful at making predictions and detecting instances of fraudulent activity, respectively.

The Random Forest model has provided a better result with a 77.32 accuracy, so it gave more accurate results than the Decision Tree. Its accuracy of 77.42% is an indicator of a higher capacity to make positive predictions right, and the recall of 77.32 is an indication of the higher capacity to identify the true cases of an AI-generated and human-written reviews. F1-score of 77.30 percent demonstrates a more balanced and stronger performance which proves that the ensemble methodology was successful in minimizing errors and enhancing the stability of the model in all the measures of evaluation.

Among the machine learning models, XGBoost was the best with an accuracy of 82.15 as it demonstrated that it performed the best in correctly classifying the types of reviews. Its accuracy of 82.21% shows that it makes accurate predictions with minimum falses, whereas a recall of 82.15% shows that it correctly recognized most of the true cases of classes. The F1-score of 82.14% indicates a great compromise of the precision and the recall, which proves that XGBoost follows complex patterns in the data better than the other ML models and has the best overall classification.

6.2 Deep Learning Model Results

The deep learning model based on the GRU demonstrated an accuracy of 83.36, i.e. the percentage of accurate determination of the generated review as an AI-generated or a human-

written one. Its accuracy of 84.93 percent indicates that the positive predictions of the model were very accurate, whereas its recall of 83.36 percent has shown that the model was effective in identifying most of the true values in each category. The F1-score of 83.17, which is a trade-off between precision and recall is the indicator of the stable and efficient model performance. All in all, these findings indicate that the GRU model has better classification power than the conventional machine learning models.

6.3 DistilBERT + GRU Results

The DistilBERT + GRU hybrid model scored 86.12, which is a very efficient feature of the model to differentiate between AI-generated and human-written reviews with a high level of reliability. This good precision is due to an integrated benefit of the rich contextual embeddings of DistilBERT and the sequentially capabilities of a GRU network, which enable the model to learn semantic meaning as well as style differences that simple models can fail to recognize. The model has a precision score of 86.13% which means that the model makes very few false positive predictions, that is, it has a lot of confidence in its prediction and it correctly classifies the reviews it classifies in each group. Equally, the recall score of 86.12% indicates that the model is quite efficient at identifying almost all the pertinent cases, and the true cases of both AI-generated and human-written reviews are not likely to be missed. F1-score, 86.12, also, proves that the model has a very good balance between precision and recall and has the same or reliable performance in both measures.

All these findings prove that the architecture of DistilBERT +GRU promotes the classification process significantly using transformer-based language comprehension and recurrent processing of text features. The sensitivity of the model to contextual information, syntactic information, and sequential dependencies allow it to perform better than a traditional machine learning model, and even out of the box GRU deep learning model. The high performance of the hybrid model underscores the efficiency of combining the pre-trained transformer embeddings with the lightweight sequential neural networks; hence this hybrid network model is the most powerful and accurate in detecting fake reviews as per the scope of this study.

6.4 Findings of the Study and Superiority

The findings clearly show that the DistilBERT + GRU hybrid model was the best of all models as it was able to blend deep contextual information with efficient sequential pattern learning. DistilBERT offers high-quality semantic embeddings that reflect subtle language aspects, style of writing, and subtle distinctions between AI and human-written text, whereas the GRU layer improves the performance by learning time and structural as well as dependencies within the embeddings. This combination enables the model to comprehend the meaning and flow of the sentences in a better manner compared to the traditional machine learning models and even the standalone GRU network. Consequently, the hybrid architecture enjoys the advantages of transformer-level representation power and recurrent

learning efficiency, which result in a more accurate prediction and an overall high-quality fake review detection.

6.5 Limitations

The model can have difficulty identifying AI-generated text that closely resembles human writing. Advanced AI text often follows natural grammar, tone, and structure, which reduces clear signals for detection. Since DistilBERT is a lightweight model, some fine-level semantic details may not be fully captured. Although the GRU helps learn sequence patterns, very polished AI-written content can still appear similar to genuine human reviews.

7 Conclusion and Future Work

7.1 Conclusion

The paper was able to establish the usefulness of the integration of conventional machine learning and deep learning as well as hybrid transformer-based methods to recognize AI-generated and human-written reviews. By systematic experimentation, it was found that although machine learning models like Decision Tree, the Random Forest, and the XGBoost offered a good baseline performance, the deep learning models especially the GRU network were found to be more accurate by learning more intricate sequential patterns of the text. The DistilBERT + GRU hybrid model was the most successful and the fact that transformer-based contextual embeddings are integrated with recurrent neural processing is beneficial. This method was particularly useful to grasp some nuanced elements of language and structure that can be unique to a real review and not an AI-written one. On the whole, the study proves that the hybrid architectures with the help of pre-trained transformers are an effective and strong solution to the contemporary text classification tasks, especially the fake review detection.

7.2 Future Work

There are a number of directions that can be investigated in future work to improve the accuracy, scalability, and practicality of fake review detection systems further. A possible avenue is to include more transformer models like BERT-large, RoBERTa, or GPT-based encoders, which might be able to pick up more semantic regularities but need more efficient training methods to be computationally efficient. It would also be beneficial to expand the dataset to encompass multilingual or domain specific reviews in order to test the generalizability of the model to a wide variety of real-world situations. Moreover, the explainable AI methods might be implemented to offer a better understanding of what makes the model label some of the reviews as AI-generated, which would enhance the perception of transparency and trustworthiness in practice. There can be real-time deployment strategies, model distillation or lightweight inference frameworks, which can be further used in industry settings. Lastly, it could be worthwhile to research ensemble techniques or hybrid

architectures consisting of multiple transformers or recurrent layers to achieve even greater performance and reliability when it comes to the detection of more and more advanced AI-generated content.

References

Al Bataineh, A., Sickler, R., Kurcz, K. and Pedersen, K., 2025. AI-generated vs. human text: Introducing a new dataset for benchmarking and analysis. IEEE Transactions on Artificial Intelligence.

Al-Qudah, D.A., Ala'M, A.Z., Cristea, A.I., Merelo-Guervós, J.J., Castillo, P.A. and Faris, H., 2025. Prediction of sentiment polarity in restaurant reviews using an ordinal regression approach based on evolutionary XGBoost. PeerJ Computer Science, 11, p.e2370.

Alamleh, H., AlQahtani, A.A.S. and ElSaid, A., 2023, April. Distinguishing human-written and ChatGPT-generated text using machine learning. In 2023 Systems and Information Engineering Design Symposium (SIEDS) (pp. 154-158). IEEE.

Alghazzawi, D.M., Alquraishee, A.G.A., Badri, S.K. and Hasan, S.H., 2023. ERF-XGB: Ensemble random forest-based XG boost for accurate prediction and classification of e-commerce product review. Sustainability, 15(9), p.7076.

Alhayan, F. and Himdi, H., 2024. Ensemble learning approach for distinguishing human and computer-generated Arabic reviews. PeerJ Computer Science, 10, p.e2345.

Blake, J., Miah, A.S.M., Kredens, K. and Shin, J., 2025. Detection of AI-generated texts: A Bi-LSTM and attention-based approach. *IEEE Access*.

Boutadjine, A., Harrag, F. and Shaalan, K., 2025. Human vs. machine: A comparative study on the detection of AI-generated content. ACM Transactions on Asian and Low-Resource Language Information Processing, 24(2), pp.1-26.

Chimata, S., Bollimuntha, A.R., Devagiri, D. and Puligadda, S., 2024, June. An Investigative Analysis on Generation of AI Text Using Deep Learning Models for Large Language Models. In 2024 International Conference on Smart Systems for Electrical, Electronics, Communication and Computer Engineering (ICSSEECC) (pp. 7-12). IEEE.

De, S. and Vats, A., 2025, May. AiMNLP@ DravidianLangTech 2025: Unmask It! AI-Generated Product Review Detection in Dravidian Languages. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 46-55).

Iqbal, A., Amin, R., Iqbal, J., Alroobaea, R., Binmahfoudh, A. and Hussain, M., 2022. Sentiment analysis of consumer reviews using deep learning. Sustainability, 14(17), p.10844.

Iqbal, M.S. and Kashem, M.A., 2025. A Machine Learning Framework for Identifying Sources of AI-Generated Text. Statistics, Optimization & Information Computing, 13(5), pp.2186-2204.

Kansal, R. and Diwaker, C., 2025. Efficiency Determination of Various Machine Learning Techniques for Sentiment Analysis on Social Media Platforms. Engineering, Technology & Applied Science Research, 15(4), pp.25584-25589.

Kayabas, A., Topcu, A.E., Alzoubi, Y.I. and Yıldız, M., 2025. A deep learning approach to classify AI-generated and human-written texts. Applied Sciences, 15(10), p.5541.

Longwei, X., Jhanjhi, N.Z., Saeed, S., Ashfaq, F., Wicaksana, F.A. and Sidik, A.D.W.M., 2025, April. Deep Fake Detection: Enhancing Content Verification Using CNN-LSTM for Mitigating AI-Generated Video Threats. In *2025 International Conference on Metaverse and Current Trends in Computing (ICMCTC)* (pp. 1-10). IEEE.

Mughal, N., Mujtaba, G., Mughal, M.H., Manaf, A. and Kamangar, Z., 2025. Fake Reviews Detection on E-Commerce Websites Using Novel User Behavioral Features: An Experimental Study. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(9), pp.1-44.

Schaaff, K., Schlippe, T. and Mindner, L., 2024. Classification of human-and AI-generated texts for different languages and domains. International Journal of Speech Technology, 27(4), pp.935-956.

Tuomi, A., Zainal Abidin, H., Tuominen, P. and Ascenção, M.P., 2025, February. Convincingness of AI-Generated Restaurant Reviews. In *ENTER e-Tourism Conference* (pp. 437-448). Cham: Springer Nature Switzerland.

Uzun, L., 2023. ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. Language Education and Technology, 3(1).

Yang, J.L., 2025. Classification of Artificial Intelligence-Generated Product Reviews on Amazon. *Engineering Proceedings*, *92*(1), p.17.

Zamir, M.T., Ullah, F., Tariq, R., Bangyal, W.H., Arif, M. and Gelbukh, A., 2024. Machine and deep learning algorithms for sentiment analysis during COVID-19: A vision to create fake news resistant society. PloS one, 19(12), p.e0315407.

Zhao, Y., Tang, S., Zhang, H. and Lyu, L., 2025. AI vs. human: A large-scale analysis of AI-generated fake reviews, human-generated fake reviews and authentic reviews. *Journal of Retailing and Consumer Services*, *87*, p.104400.