Dual-Task Learning for AI-Generated Medical Text Detection and Named Entity Recognition

Saja Al-Dabet[1], Ban Alomar[1], Sherzod Turaev[1], Abdelkader Nasreddine Belkacem[2]
[1] Department of Computer Science and Software Engineering,
[2] Department of Computer and Network Engineering,
United Arab Emirates University
{700039885, 700039223, sherzod, belkacem}@uaeu.ac.ae

Abstract

The integration of artificial intelligence (AI) into the medical field has revolutionized documentation and diagnosis. However, the detection of AI-generated text within medical records remains a crucial task. This paper describes a dual-task learning framework using the ELECTRA model for detecting AI-generated medical texts and performing named entity recognition (NER). The dual-task model includes a binary classification head for identifying AI-generated texts and an NER head for extracting medical entities. Experiments on radiology report and medical texts datasets show that the proposed approach achieves robust performance, with F1 scores of 0.985 and 0.996 for classification and 0.51 and 0.68 for NER. The model achieves a high accuracy of 0.996 for medical text classification and 0.985 for MiMic classification, enhancing automated medical text analysis and supporting clinical decision-making.

1 Introduction

The advent of large language models such as ChatGPT (Generative Pretrained Transformer) has revolutionized various sectors (Radford et al., 2018), including the medical field, by enabling the generation of coherent and human-like text (Hamad et al., 2024; Hireche and Belkacem, 2024; Hireche et al., 2023; Jamil et al., 2024). These advances have facilitated tasks such as automated report generation, clinical documentation, and medical information dissemination. However, the spread of artificial intelligence (AI)-generated text in medicine raises significant concerns regarding the accuracy, reliability, and authenticity of the information contained therein. Misleading AI-generated medical content can have severe consequences, potentially compromising patient care and medical research integrity. Human medical writers, with their depth of understanding and expertise in the medical field, cannot at present be fully replaced by ChatGPT (Homolak, 2023; Liao et al., 2023; Tan et al., 2024). Additionally, there are concerns regarding potential bias in AI-generated content and the necessity for transparency in AI usage. This makes it essential to ensure the integrity and accuracy of medical information, indicating the important role of human oversight in creating medical content (Sajid and ul Hassan, 2022).

Distinguishing between human-written and AI-generated medical texts is challenging and requires robust detection methods. There are several differences between medical texts authored by humans and those generated by AI agents. Human-written texts have a larger vocabulary, greater diversity, and include specific information and numbers, making them detailed and contextually rich. In contrast, AI-generated texts use more common terminology, emphasizing fluency and logical structure, and are generally more neutral and positive in sentiment. In terms of parts-of-speech, AI-generated texts contain more nouns, determiners, plural nouns, and coordinating conjunctions, indicating a structured style, whereas humans use more cardinal digits and adverbs, reflecting greater specificity. Similarly, dependency parsing in AI-generated texts includes more determiners and conjuncts, with human texts having more numeric and adverbial modifiers. Furthermore, text perplexity is lower for AI-generated texts due to the replication of common patterns, whereas human texts display a greater degree of variation (Liao et al., 2023). Existing approaches, such as linguistic feature analysis and machine learning models, have shown promise, but often fall short in handling the complexities of medical language. To address these limitations, we propose a multitask model that leverages the capabilities of the ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020) language model. ELECTRA achieves superior performance in various natural language processing (NLP) tasks due to its efficiency in text encoding and understanding. Our proposed model utilizes ELECTRA for two primary tasks: differentiating between human-written and AI-generated texts and enhancing text comprehension through named entity recognition (NER). By integrating these tasks, the proposed model not only identifies AI-generated content, but also improves the understanding of medical texts, thereby increasing the accuracy of detection.

The integration of NER into the detection framework enables the model to identify and classify essential medical entities, thereby offering deeper insights into the context and content of the text. This dual-task approach ensures comprehensive analysis, capturing subtle differences between human- and AI-generated medical texts that may be overlooked by single-task models. Moreover, the enhanced text understanding provided by NER aids in the detection of inconsistencies and anomalies indicative of AI-generated content. This approach enhances parameter efficiency by sharing model parameters across tasks and leverages transfer learning, thereby allowing knowledge from one task to benefit the other. In addition, our model produces consistent predictions while simplifying deployment by reducing the need for separate models.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the previous literature, before Section 3 describes the proposed methodology used to develop the model. Section 4 covers the experiments and results. Finally, Section 5 concludes the paper.

2 Literature Review

In this section, recent papers on both AI-text detection and medical NER tasks are summarized.

2.1 AI-text detection task

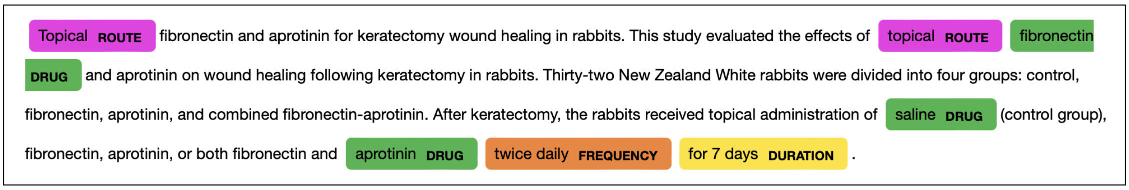
Guo et al. (2023) evaluated ChatGPT's performance in mimicking human expert responses using the Human ChatGPT Comparison Corpus (HC3), which includes around 40,000 questions and answers from both human experts and ChatGPT across various domains. The study utilized RoBERTa and GLTR models to analyze the text, revealing that RoBERTa significantly outperformed GLTR. Specifically, RoBERTa achieved F1 scores of 99.82% in full-text detection and 87.17% in sentence-level detection, compared with GLTR's 98.31% and 70.91%, respectively.

Scheibe and Mandl (2023) explored the effectiveness of models in distinguishing between human-written and machine-generated texts. Their study was framed within the AuTexTification 2023 shared task, focusing on automated text identification. The methodology uses the pre-trained DeBERTaV2 model (He et al., 2020), selected for its capabilities in handling text classification tasks, and a dataset that comprises a balanced mix of human and machine-generated texts, resulting in a robust training environment. In terms of results, the DeBERTaV2 model achieved a macro-F1 score of 67.2%, ranking 15th out of 76 submissions for subtask 1.

Verma et al. (2023) introduced Ghostbuster, developed by UC Berkeley researchers to detect AI-generated text. Ghostbuster uses the GPT-3 Davinci configuration to extract probabilistic features, and employs a linear classifier to identify machine-generated text. Token probabilities from the text-generating AI are not required, making Ghostbuster effective even with complex models. Tested on three datasets covering student essays, creative writing, and news articles, Ghostbuster achieved a 99% F1 score, outperforming models including DetectGPT and GPTZero.

Alamleh et al. (2023) explored machine learning-based approaches to detect ChatGPT-generated text. The authors evaluated their models on a Kaggle dataset of 10,000 instances, half from human sources and half generated by GPT-3.5. They employed a variety of machine learning and deep learning algorithms, including random forests, logistic regression, decision trees, support vector machines, AdaBoost, bagging classifiers, multilayer perceptrons, and long short-term memory (LSTM) networks, with a special focus on the extremely randomized trees classifier for its robustness in handling random data points. Their methodology involves sentence vectorization using the term frequency–inverse document frequency (TF-IDF) followed by classification. The highest achieved accuracy for distinguishing between human- and ChatGPT-generated texts was 77%.

Mitrović et al. (2023) investigated the ability of machine learning to detect AI-generated short online reviews, comparing a Transformer-based model with a perplexity-based approach. Two datasets were created: one with ChatGPT-generated texts from custom prompts and another with rephrased human-written reviews. The Shap-



Topical ROUTE fibronectin and aprotinin for keratectomy wound healing in rabbits. This study evaluated the effects of topical ROUTE fibronectin DRUG and aprotinin on wound healing following keratectomy in rabbits. Thirty-two New Zealand White rabbits were divided into four groups: control, fibronectin, aprotinin, and combined fibronectin-aprotinin. After keratectomy, the rabbits received topical administration of saline DRUG (control group), fibronectin, aprotinin, or both fibronectin and aprotinin DRUG twice daily FREQUENCY for 7 days DURATION .

Figure 1: Medical NER tagging example using en-core-med7-lg pretrained model.

ley additive explanations were used to identify influential features. The Transformer-based model proved more effective, achieving up to 98% accuracy for straightforward AI-generated texts and 79% for rephrased texts.

Liao et al. (2023) highlighted the risks of AI-generated content in medical contexts. The authors constructed datasets of both human-written and ChatGPT-generated medical texts, before analyzing the linguistic properties and employing machine learning to identify AI-generated content. The key differences were found to be the more detailed and varied human texts versus more general and fluent AI texts. A BERT-based model achieved an F1 score of over 95% in identifying AI-generated texts.

2.2 Medical NER tasks

For medical NER tasks, several studies have targeted different languages using machine and deep learning approaches. Review articles have specifically addressed medical and clinical NER approaches (Ahmad et al., 2023; Pagad and Pradeep, 2022).

Gaschi et al. (2023) evaluated cross-lingual transfer (CLT) and translation-based methods for medical NER in English, French, and German. They used the N2C2, MedNERF, and GERNERMED datasets, and applied fine-tuned multilingual models (XLM-R, mBERT) to N2C2 for CLT, as well as translating N2C2 for training (translate-train) and testing (translate-test). CLT with the XLM-R base yielded F1 scores of 79.1% for French and 72.2% for German. The translate-train method achieved F1 scores of 78.6% for French and 74.8% for German, while DrBERT PubMed scored 78.8% for French and 75% for German.

Park et al. (2023) developed a web service using BioBERT to integrate NER and relation extraction (RE) in the biomedical domain. The BioBERT base was fine-tuned using the NCBI Disease Corpus and BC2GM Corpus (for NER) and the Genetic Association Database (for RE). The NER system demonstrated high performance, achieving a precision of 85.16%, recall of 83.65%, and an F1 score of 84.4% for gene/protein recognition, and 89.04%, 89.69%, and 89.36%, respectively, for disease recognition. The Django-based web service allows users to input PubMed IDs, retrieve abstracts, and view color-coded NER results and interactive RE graphs.

Xu et al. (2018) presented a combined deep learning approach for medical NER. Utilizing datasets from the 2010 i2b2/VA NLP Challenges, their study implemented an attention-based LSTM architecture combined with a conditional random field to target document-level global information. This method employs pretrained word embeddings and bidirectional language models trained on the MIMIC-III corpus, and addresses the limitations of sentence-level NER by incorporating global context through neural attention mechanisms. The model achieved an impressive micro-F1 score of 85.71%.

Naseem et al. (2021) constructed BioALBERT, a domain-specific language model optimized for biomedical NER. The model was trained on large-scale biomedical corpora from PubMed and PMC, addressing the limitations of existing models through techniques such as factorized embedding parameterization, cross-layer parameter sharing, and sentence-order prediction. BioALBERT demonstrated significant performance improvements across various datasets: 7.47% for NCBI Disease, 10.63% for BC5CDR-Disease, 4.61% for BC5CDR-Chem, 3.89% for BC4CHEMD, 12.25% for BC2GM, 6.42% for JNLPBA, 6.19% for LINNAEUS, and 23.71% for Species-800.

Košprdić et al. (2023) proposed a biomedical NER approach using zero- and few-shot learning with six public corpora: CDR, CHEMDNER, BioRED, NCBI Disease, JNLPBA, and N2C2. They fine-tuned the BioBERT and PubMedBERT models, converting multiclass token classification into binary token classification to recognize unseen entity classes through semantic similarities from pretraining. The method achieved average F1 scores of 35.44% for zero-shot NER, 50.10% for one-shot NER, 69.94% for 10-shot NER, and 79.51% for 100-shot NER.

3 Proposed Model

This section describes the proposed model. Multitask learning is used to train a single model on multiple tasks simultaneously, improving both generalization and performance through shared representations. This approach enhances parameter efficiency by sharing model parameters across tasks, which is beneficial when there are limited computational resources or datasets. Transfer learning leverages knowledge from one task to enhance another. Multitask models yield consistent and coherent predictions, simplifying deployment by reducing the need for separate models. In a dual-task model for text classification and NER tagging, shared linguistic and entity recognition capabilities enhance the overall performance. The following subsections detail the proposed architecture. Figure 2 provides an overview.

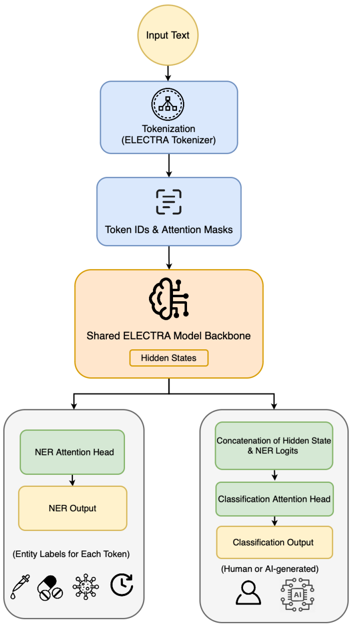

Input Text
Tokenization
(ELECTRA Tokenizer)
Token IDs & Attention Masks
Shared ELECTRA Model Backbone
Hidden States
NER Attention Head
NER Output
(Entity Labels for Each Token)
Concatenation of Hidden State & NER Logits
Classification Attention Head
Classification Output
(Human or AI-generated)
AI

Figure 2: Proposed model architecture.

3.1 Data preprocessing

Data preprocessing makes a significant contribution to model performance. The following preprocessing steps are implemented:

- *Identification of null entries*: an initial assessment is performed to identify columns containing missing values.
- *Tokenization*: text data are tokenized to convert sentences into single tokens. This process is essential for subsequent text processing and model input preparation.
- *Remove special characters*: special characters that do not add to the semantic meaning of the text are removed. This step helps in cleaning the data and minimizing noise.
- *NER data annotation*: the data are annotated with medical NER tags using the SpaCy pretrained model (“en-core-med7-lg” version) (AI, 2024). This model is specifically designed for medical NER, identifying entities such as DRUG (names of medications), DOSAGE (dosage information and units), DURATION (duration of medication use or treatment), FORM (medication forms, i.e., tablets or injections), FREQUENCY (how often a medication is taken), ROUTE (route of administration, i.e., oral or intravenous), and STRENGTH (strength of the medication). The use of these NER tags ensures the precise identification and categorization of relevant medical entities within the text. Figure 1 illustrates an example of an annotated medical report tagged using the “en-core-med7-lg” pretrained model.
- *Encoding extraction*: encoding vectors and labels are extracted from the tokenized datasets for both classification and NER tasks. This involves generating numerical representations of the text data that are suitable for model training.
- *Label padding and conversion for NER*: NER labels are padded and converted from string tuples to integer labels using a label map. This ensures that the labels have a consistent format and are aligned with the input sequences, which is necessary for effective model training.

3.2 Framework architecture

To overcome the challenges of distinguishing between human-written and AI-generated medical texts, a multitask framework leveraging the ELECTRA language model is proposed. This framework is designed to perform two primary tasks simultaneously: differentiating between human- and AI-generated texts and enhancing text comprehension through NER. By integrating these tasks, the model not only enhances the accuracy with which AI-generated content is detected, but also provides a deeper understanding of the context and content of the medical texts. ELECTRA (Clark et al., 2020) represents a pretraining approach for text encoders, diverging significantly from traditional masked language modeling methods such as BERT (Devlin et al., 2018). Rather than masking random tokens and predicting their original forms, ELECTRA modifies the input by replacing specific tokens with plausible alternatives produced by a smaller auxiliary network, known as the generator.

The primary model (discriminator) is then tasked with identifying whether each token in the modified input is original or has been replaced. This replaced token detection strategy leverages the entire input sequence, thereby enhancing both sample efficiency and computational effectiveness. The discriminator plays a critical role, as it learns to differentiate between authentic tokens and those introduced by the generator using the full context of the input data. This discriminative task not only improves the training efficiency, but also enhances performance on downstream tasks. The architecture of ELECTRA integrates both a generator and a discriminator, resulting in superior results with fewer computational resources (Hao et al., 2021; Ozyurt, 2020). Algorithm 1 in the appendix describes the proposed dual-task learning process for both classification and NER tasks.

4 Experiments and Results

4.1 Dataset

The medical dataset utilized in this study comprises two primary components, as described by Liao et al. (2023). The medical abstract dataset is sourced from a publicly available Kaggle dataset (Kamath, 2023) and includes texts related to five medical conditions: digestive system diseases, cardiovascular diseases, neoplasms, nervous system diseases, and general pathological conditions. The radiology report dataset, which is based on the work of Johnson et al. (2016), includes selected radiology reports. A total of 4400 samples were obtained from both the radiology report and medical abstract datasets as human-written medical texts. To create corresponding ChatGPT-generated texts, a text continuation method was applied, resulting in datasets containing 8800 samples each for the medical abstracts and radiology reports. Both datasets were then divided into 70% for training, 10% for validation, and 20% for testing subsets, yielding 3080 samples for training, 440 for validation, and 880 for testing in each dataset.

4.2 Evaluation metrics

To evaluate the performance of the proposed model, a comprehensive set of evaluation metrics was employed. The precision, recall, and F1 score are essential metrics in the context of distinguishing between AI-generated and human-written medical texts.

4.3 Experimental settings

The experiments were conducted on the Kaggle platform using the GPU-enabled feature. The applied model based on the "electra-small-discriminator" checkpoint and tokenization was handled by the ElectraTokenizer layer. For classification tasks involving both the MiMic and medical datasets, the batch size for classification tasks was set to 16, whereas for NER tasks, it was set to 8. NER tasks utilize seven labels, while classification tasks are binary, involving two labels. The AdamW optimizer was used (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-5}. The training process involved separate head optimization with five epochs for both the classification and NER heads, followed by joint optimization epochs.

4.4 Results and discussion

4.4.1 Evaluating dual-task performance

Table 1 compares the proposed model with other models from the literature. The proposed model performs robustly across all metrics for both NER and classification tasks, outperforming ELECTRA, RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2020), and XLNet (Yang et al., 2019), and surpassing the baseline model of Liao et al. (2023). Four main models were used: Perplexity-CLS, CART, XGBoost, and BERT. For Perplexity-CLS, BioGPT calculates the text perplexity, with the optimal threshold identified using the validation set. The CART model uses TF-IDF for vectorization, a decision tree with a maximum depth of four, and the Gini impurity for feature division. The XGBoost model also uses TF-IDF and sets the maximum depth for base learners to four. The BERT model achieves the best performance due to its advanced text processing capabilities.

Table 1: Performance evaluation for models.

Classification Task								
Model	Medical text				MiMic			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Perplexity-CLS (Liao et al., 2023)	-	0.728	0.724	0.723	-	0.831	0.828	0.828
CART (Liao et al., 2023)	-	0.777	0.745	0.738	-	0.829	0.825	0.824
XGBoost (Liao et al., 2023)	-	0.898	0.893	0.893	-	0.899	0.898	0.898
BERT (Liao et al., 2023)	-	0.958	0.958	0.958	-	0.968	0.967	0.967
BioBERT	0.948	0.943	0.942	0.944	0.970	0.968	0.968	0.969
RoBERTa	0.999	0.999	0.999	0.999	0.989	0.990	0.989	0.989
XLNet	0.998	0.998	0.998	0.998	0.988	0.988	0.988	0.988
ELECTRA	0.996	0.996	0.996	0.996	0.985	0.985	0.986	0.985
Named Entity Recognition Task								
RoBERTa	0.54	0.41	0.47	0.58	0.75	0.72	0.72	0.73
XLNet	0.72	0.42	0.48	0.51	0.45	0.31	0.36	0.38
ELECTRA	0.68	0.45	0.51	0.56	0.93	0.91	0.91	0.92

In comparison with the other models considered in this study, ELECTRA demonstrates faster and more robust performance. The ELECTRA model utilizes a pretraining method that is more computationally efficient than the traditional masked language modeling employed by models such as BERT. Instead of masking and predicting random tokens, ELECTRA modifies the input by replacing some tokens with plausible alternatives generated by a small auxiliary network, and then trains a discriminator to determine whether each token is original text or substituted text. For the classification task, ELECTRA demonstrates robust performance on both the medical text and MiMic datasets. Specifically, ELECTRA achieves an accuracy of 0.985 for the MiMic dataset, with precision and recall scores of 0.985 and 0.986, respectively, resulting in an F1 score of 0.985. For the medical texts, ELECTRA achieves an accuracy of 0.996, with precision, recall, and F1 scores all at 0.996. This performance is comparable to, and in some cases exceeds, that of other transformer-based models such as BioBERT. The high F1 scores indicate that ELECTRA is highly effective at differentiating between AI-generated and human-written texts, making it a strong candidate for this classification task.

In the NER task, ELECTRA produces balanced performance across the datasets. On the MiMic dataset, ELECTRA achieves an accuracy of 0.93, precision and recall of 0.91, and an F1 score of 0.92. The medical reports dataset, however, presents a more challenging environment for the model due to the nature of the written text. ELECTRA achieves an accuracy of 0.68 and an F1 score of 0.56 on this dataset, with precision at 0.45 and recall at 0.51.

In the field of medical AI, the development of a stable architecture capable of both classification and NER tasks is essential. ELECTRA demonstrates efficient classification, achieving high F1 scores on both the medical text and MiMic datasets, thereby ensuring precise differentiation between AI-generated and human-written texts. Although there is potential for improvement in terms of NER performance, the ability of ELECTRA to identify and classify medical entities remains significant. This stability across multiple tasks enhances the reliability of automated medical text analysis, facilitating more accurate clinical decision-making and efficient information processing.

To evaluate the effect of using a dual-task model instead of a single classification model, the ELECTRA classification model was tested alone and achieved an accuracy of 0.967, precision of 0.968, recall of 0.967, and an F1 score of 0.967. Using the dual-task ELECTRA model, which integrates NER parameters, enhances the results over those given by the ELECTRA model alone. The integration of NER allows the model to better understand and classify complex medical texts by recognizing and categorizing relevant entities within the text, thus improving the overall accuracy and reliability of the classification.

The receiver operating characteristic (ROC) curves are shown in Fig. 3. These curves evaluate the performance of the classification and NER tasks on the medical and MiMic datasets. The top-left plot shows the overall ROC curves, with an area under the curve (AUC) of 1.0 for both datasets, indicating significant classification performance in distinguishing AI-generated from human-written text. The top-right plot displays the ROC curves for the NER task, with slightly better performance on the MiMic dataset than the medical dataset.

The bottom-left plot presents multiple ROC curves for the medical dataset's NER performance across different classes (0–6, representing form, dosage, route, frequency, drug, strength, and duration), with Class 4 (Drug) and Class 0 (Form) having the highest and lowest AUCs, respectively. Similarly, the bottom-right plot shows the MiMic dataset's NER performance, with Class 1 (Dosage) and Class 6 (Duration) being the most challenging and easiest classes, respectively. The micro-average curves in the bottom plots indicate good overall NER performance.

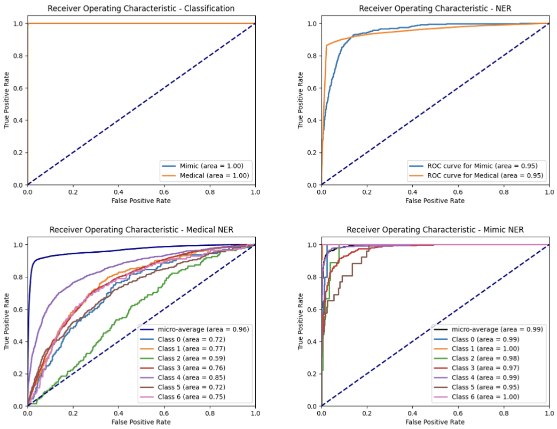


Figure 3: ROC curves for both experimented tasks.

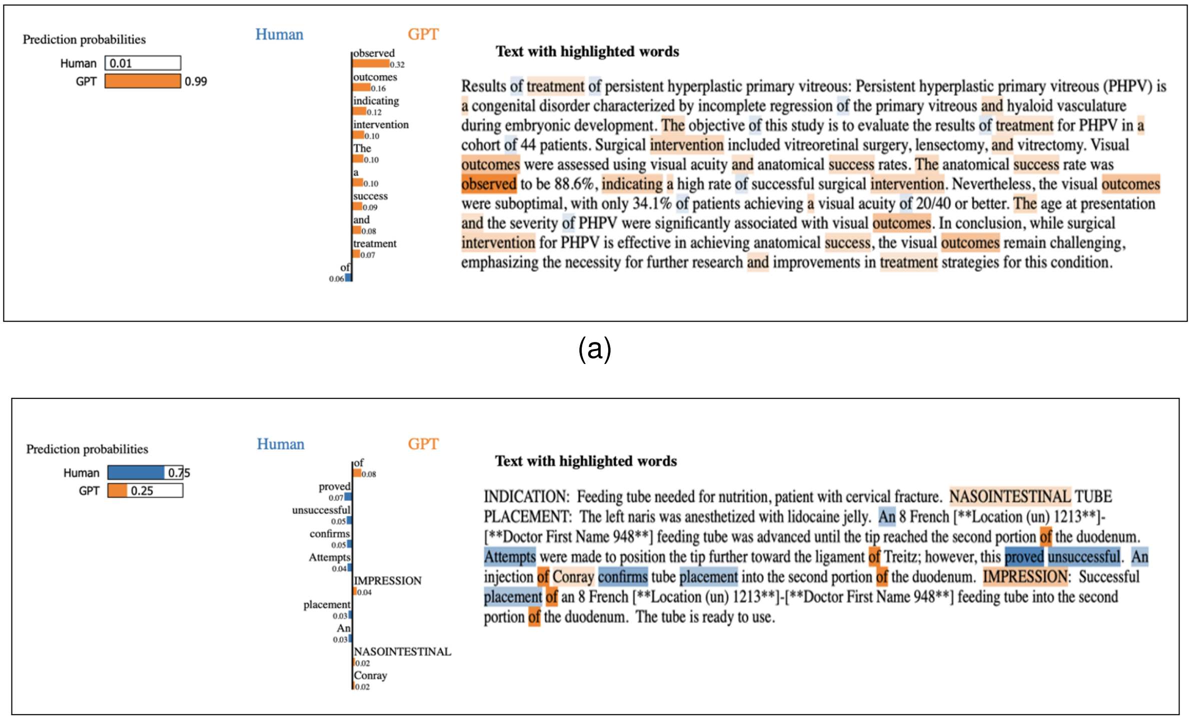
To add more explainability to the trained model, the Local Interpretable Model-agnostic Explanations (LIME) tool was utilized (Ribeiro et al., 2016). LIME increases interpretability by approximating the behavior of complex models around specific predictions. The trained weights of the ELECTRA model were used to plot these figures. In Fig. 4(a), the model assigns a high probability of 0.99 to the text being GPT-generated and only 0.01 to it being human-written, with key terms such as "treatment", "outcomes", "indicating", and "intervention" highlighted in orange, indicating their significant contribution to the model's classification decision. Figure 4(b) shows the prediction probability of 0.75 for the text being human-written, while the probability for GPT generation is 0.25, where key terms such as "proved", "unsuccessful", "confirms", "attempts", and "placement" are highlighted in blue, indicating their significant contribution to the human-written classification. In contrast, the terms "of" and "Conray" are highlighted in orange, showing their association with the GPT-generated classification. The resulting predictions are both correct.

To evaluate the effect of using Transformer models on the NER task alone, additional explorations were conducted. ClinicalBERT (Huang et al., 2019), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), and BioBERT are variants of the BERT model (Devlin et al., 2018) tailored for medical and clinical usage. ClinicalBERT is pretrained on clinical notes and medical records, enhancing its effectiveness in healthcare-related tasks. SciBERT is pretrained on scientific literature from Semantic Scholar, making it suitable for scientific and academic applications. BlueBERT is trained on a combination of biomedical and clinical texts, specifically PubMed abstracts and MIMIC-III clinical notes, allowing it to handle both domains proficiently. BioBERT is pretrained on extensive biomedical literature, including PubMed abstracts and full-text articles from PubMed Central, resulting in optimization for understanding biomedical texts. Other models such as BERT, RoBERT, and ALBERT (Lan et al., 2019) were also included in this experiment. Table 2 presents the results obtained using these Transformers for the NER task.

Table 2: NER task evaluation.

Model	**Dataset**	**Acc**	**Precision**	**Recall**	**F1-score**
ClinicalBERT	Medical	0.93	0.80	0.76	0.78
	MiMic	0.99	0.98	0.99	0.99
SciBERT	Medical	0.88	0.70	0.63	0.65
	MiMic	0.95	0.95	0.93	0.94
BlueBERT	Medical	0.91	0.77	0.69	0.72
	MiMic	0.97	0.98	0.96	0.97
BioBERT	Medical	0.91	0.78	0.71	0.74
	MiMic	0.97	0.97	0.96	0.96
ELECTRA	Medical	0.88	0.70	0.57	0.63
	MiMic	0.97	0.97	0.96	0.96
BERT	Medical	0.90	0.75	0.69	0.72
	MiMic	0.99	0.99	0.99	0.99
RoBERT	Medical	0.86	0.66	0.48	0.55
	MiMic	0.93	0.92	0.87	0.89
ALBERT	Medical	0.86	0.54	0.41	0.47
	MiMic	0.62	0.26	0.26	0.26

Comparative analysis of the pretrained Transformer-based models for NER tasks across the medical and MiMic datasets reveals significant performance variability. ClinicalBERT and BERT demonstrate exceptional proficiency, achieving the highest F1 scores of 0.99 on the MiMic dataset and 0.78 on the medical dataset. This outstanding performance can be attributed to their architecture, which enhances their ability to accurately identify and classify named entities. SciBERT and ELECTRA achieve moderately good performance on the medical dataset (F1 scores of 0.65 and 0.63), but perform strongly on the MiMic



Results of treatment of persistent hyperplastic primary vitreous: Persistent hyperplastic primary vitreous (PHPV) is a congenital disorder characterized by incomplete regression of the primary vitreous and hyaloid vasculature during embryonic development. The objective of this study is to evaluate the results of treatment for PHPV in a cohort of 44 patients. Surgical intervention included vitreoretinal surgery, lensectomy, and vitrectomy. Visual outcomes were assessed using visual acuity and anatomical success rates. The anatomical success rate was observed to be 88.6%, indicating a high rate of successful surgical intervention. Nevertheless, the visual outcomes were suboptimal, with only 34.1% of patients achieving a visual acuity of 20/40 or better. The age at presentation and the severity of PHPV were significantly associated with visual outcomes. In conclusion, while surgical intervention for PHPV is effective in achieving anatomical success, the visual outcomes remain challenging, emphasizing the necessity for further research and improvements in treatment strategies for this condition.

(a)

INDICATION: Feeding tube needed for nutrition, patient with cervical fracture. NASOINTESTINAL TUBE PLACEMENT: The left naris was anesthetized with lidocaine jelly. An 8 French [**Location (un) 1213**]-[**Doctor First Name 948**] feeding tube was advanced until the tip reached the second portion of the duodenum. Attempts were made to position the tip further toward the ligament of Treitz; however, this proved unsuccessful. An injection of Conray confirms tube placement into the second portion of the duodenum. IMPRESSION: Successful placement of an 8 French [**Location (un) 1213**]-[**Doctor First Name 948**] feeding tube into the second portion of the duodenum. The tube is ready to use.

(b)

Figure 4: Explainable examples using LIME. (a) Medical text example. (b) MiMic text example.

dataset (F1 scores of 0.94 and 0.96). BlueBERT and BioBERT consistently perform well across both datasets, achieving F1 scores of 0.72 and 0.74 on the medical dataset and 0.97 and 0.96 on the MiMic dataset. RoBERTa and ALBERT display weak performance on the medical dataset (F1 scores of 0.55 and 0.47), with ALBERT underperforming on the MiMic dataset (F1 score of 0.26).

In multitask learning, the separate tasks can influence each other's outcomes. In a dual-task setup with text classification and NER, classification is often improved by joint training. This is due to shared representations capturing general features that are useful for both tasks, with NER enhancing the model's linguistic and semantic understanding. This positive transfer acts as regularization, reducing overfitting and boosting classification performance. However, NER might perform better alone due to task interference and complexity in balancing losses in a dual-task model. Thus, while multitask learning benefits classification, it poses challenges for optimizing both tasks.

5 Conclusion and Future Work

This study developed a dual-task learning framework using the ELECTRA model to detect AI-generated medical texts and perform NER. The integrated approach, combining a binary classification head and an NER head, showed robust performance across medical text and radiology report datasets. The framework effectively distinguishes human-written from AI-generated texts and extracts critical medical entities, enhancing detection accuracy and text comprehension. Experiments demonstrated that the ELECTRA model outperforms others in terms of inference speed and prediction robustness, achieving high F1 scores for both classification and NER tasks.

Future work will attempt to extend and refine the proposed framework by exploring additional datasets and domains to evaluate the model's generalizability and robustness across various types of medical texts. Moreover, incorporating more advanced techniques for handling complex medical terminology and context-specific nuances could further improve the framework's performance and applicability in real-world scenarios.

Acknowledgments

This research was partially funded by the United Arab Emirates University (UAEU) under Grant UPAR-12T055, with additional support from the UAEU Strategic Research Grant G00003676 through the Big Data Analytics Center, and the UAEU SURE+ Grant G00004748.

References

Pir Noman Ahmad, Adnan Muhammad Shah, and KangYoon Lee. 2023. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare*, volume 11, page 1268. MDPI.

Explosion AI. 2024. spacy models: en_core_web_md. Version 3.5.0.

Hosam Alamleh, Ali Abdullah S AlQahtani, and AbdEl-Rahman ElSaid. 2023. Distinguishing human-written and chatgpt-generated text using machine learning. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 154–158. IEEE.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical ner: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arxiv. *Preprint posted online on*, 18.

Zineb Touati Hamad, Nuraini Jamil, and Abdelkader Nasreddine Belkacem. 2024. Chatgpt's impact on education and healthcare: Insights, challenges, and ethical consideration. *IEEE Access*.

Yaru Hao, Li Dong, Hangbo Bao, Ke Xu, and Furu Wei. 2021. Learning to sample replacements for electra pre-training. *arXiv preprint arXiv:2106.13715*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Abdelhadi Hireche and Abdelkader Nasreddine Belkacem. 2024. Integrating pepper robot and gpt for neuromyth educational conversation. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–6. IEEE.

Abdelhadi Hireche, Abdelkader Nasreddine Belkacem, Sadia Jamil, and Chao Chen. 2023. Newsgpt: Chatgpt integration for robot-reporter. *arXiv preprint arXiv:2311.06640*.

Jan Homolak. 2023. Opportunities and risks of chatgpt in medicine, science, and academic publishing: a modern promethean dilemma. *Croatian Medical Journal*, 64(1):1.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Nuraini Jamil, Fahed Namir Saghir, Hassan Alshanqiti, Ali Khalifa Ali Almansoori, Abdulrahman Saeed, Ali Ahmad, and Abdelkader Nasreddine Belkacem. 2024. On combining the potential of social robots and chatgpt for enhanced learning. In *2024 12th International Conference on Information and Education Technology (ICIET)*, pages 226–231. IEEE.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Chaitanya Kamath. 2023. Medical text dataset. Accessed: 2024-06-04.

Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milosevic. 2023. From zero to hero: Harnessing transformers for biomedical named entity recognition in zero-and few-shot contexts. *Available at SSRN 4463335*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

W Liao, Z Liu, H Dai, S Xu, Z Wu, Y Zhang, X Huang, D Zhu, H Cai, T Liu, et al. 2023. Differentiate chatgpt-generated and human-written medical texts. arxiv 2023. *arXiv preprint arXiv:2304.11567*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. *bioRxiv*, pages 2020–05.

Naveen S Pagad and N Pradeep. 2022. Clinical named entity recognition methods: an overview. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2*, pages 151–165. Springer.

Yeon-Ji Park, Min-a Lee, Geun-Je Yang, Soo Jun Park, and Chae-Bong Sohn. 2023. Web interface of ner and re with bert for biomedical text mining. *Applied Sciences*, 13(8):5163.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ujala Sajid and Faheem ul Hassan. 2022. Chatgpt and its effect on shaping the future of medical writing. *Pakistan Journal of Ethics*, 2(2):38–43.

Tatjana Scheibe and Thomas Mandl. 2023. Univ. of hildesheim at autextification 2023: Detection of automatically generated texts.

Songtao Tan, Xin Xin, and Di Wu. 2024. Chatgpt in medicine: prospects and challenges: a review article. *International Journal of Surgery*, pages 10–1097.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Guohai Xu, Chengyu Wang, and Xiaofeng He. 2018. Improving clinical named entity recognition with global neural attention. In *Web and Big Data: Second International Joint Conference, APWeb-WAIM 2018, Macau, China, July 23-25, 2018, Proceedings, Part II 2*, pages 264–279. Springer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Appendix

Algorithm 1 Multitask learning for text classification and NER.

```
Input:
$\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{Y}_i^{\text{NER}})\}_{i=1}^{N}$: Dataset
$\mathcal{M}$: Pretrained ELECTRA model
$\mathcal{T}$: Tokenizer
$L_{\text{max}}$: Maximum sequence length
$\mathcal{L}$: Label set for NER
$E_c, E_n$: Epochs for classification and NER pretraining
$E_j$: Epochs for joint training
Output: Trained multitask model $\mathcal{F}$
                                        // Preprocessing
for $i = 1$ to $N$ do
    $\mathbf{x}_i^t \leftarrow \mathcal{T}(\mathbf{x}_i, L_{\text{max}})$
    $\mathbf{Y}_i^{\text{NER}} \leftarrow$ ConvertAndPad$(\mathbf{Y}_i^{\text{NER}}, \mathcal{L}, L_{\text{max}})$
end for
                                        // Model architecture
$\mathcal{F}_{\text{BERT}} \leftarrow \mathcal{M}$
$\mathcal{F}_{\text{NER}} \leftarrow$ LinearLayer$(d_{\text{BERT}}, |\mathcal{L}|)$
$\mathcal{F}_{\text{CLS}} \leftarrow$ LinearLayer$(d_{\text{BERT}} + |\mathcal{L}|, 2)$
                                        // Loss functions
$\mathcal{L}_{\text{CLS}} \leftarrow$ CrossEntropyLoss()
$\mathcal{L}_{\text{NER}} \leftarrow$ CrossEntropyLoss(ignore_index $= -1$)
                                        // Separate pretraining
for $e = 1$ to $\max(E_c, E_n)$ do
    if $e \leq E_c$ then
        Train $\mathcal{F}_{\text{BERT}}$ and $\mathcal{F}_{\text{CLS}}$ using $\mathcal{L}_{\text{CLS}}$
    end if
    if $e \leq E_n$ then
        Train $\mathcal{F}_{\text{BERT}}$ and $\mathcal{F}_{\text{NER}}$ using $\mathcal{L}_{\text{NER}}$
    end if
end for
                                        // Joint training
for $e = 1$ to $E_j$ do
    for $(\mathbf{x}_i^t, y_i, \mathbf{Y}_i^{\text{NER}})$ in $\mathcal{D}$ do
        $\mathbf{H}_i \leftarrow \mathcal{F}_{\text{BERT}}(\mathbf{x}_i^t)$
        $\mathbf{Z}_i^{\text{NER}} \leftarrow \mathcal{F}_{\text{NER}}(\mathbf{H}_i)$
        $\mathbf{h}_i^{\text{CLS}} \leftarrow \mathbf{H}_i[0, :]$
        $\mathbf{z}_i^{\text{CLS}} \leftarrow \mathcal{F}_{\text{CLS}}([\mathbf{h}_i^{\text{CLS}}; \mathbf{Z}_i^{\text{NER}}[0, :]])$
        $L_{\text{CLS}} \leftarrow \mathcal{L}_{\text{CLS}}(\mathbf{z}_i^{\text{CLS}}, y_i)$
        $L_{\text{NER}} \leftarrow \mathcal{L}_{\text{NER}}(\mathbf{Z}_i^{\text{NER}}, \mathbf{Y}_i^{\text{NER}})$
        Update $\mathcal{F}$ by minimizing $L_{\text{CLS}} + L_{\text{NER}}$
    end for
end for=0
```