VerbaNexAI at MEDIQA-CORR 2024: Efficacy of GRU with BioWordVec and ClinicalBERT in Error Correction in Clinical Notes

David Villate[1,*], Laura Tinjaca[2,*], Laura Estrada[3,*], Edwin Puertas[4,+], Juan Pajaro[5,*]
[*]Pontificia Universidad Javeriana
[+]Universidad Tecnologica de Bolívar
[1]juand.villate@javeriana.edu.co, [2]tinjacac.l@javeriana.edu.co
[3]l-estrada@javeriana.edu.co, [4]epuerta@utb.edu.co, [5]juanpajaro@javeriana.edu.co

Abstract

The automatic identification of medical errors in clinical notes is crucial for improving the quality of healthcare services.LLMs emerge as a powerful artificial intelligence tool for automating this task. However, LLMs present vulnerabilities, high costs, and sometimes a lack of transparency. This article addresses the detection of medical errors through the fine-tuning approach, conducting a comprehensive comparison between various models and exploring in depth the components of the machine learning pipeline. The results obtained with the fine-tuned ClinicalBert and Gated recurrent units (Gru) models show an accuracy of 0.56 and 0.55, respectively. This approach not only mitigates the problems associated with the use of LLMs but also demonstrates how exhaustive iteration in critical phases of the pipeline, especially in feature selection, can facilitate the automation of clinical record analysis.

1 Introduction

Large language models (LLMs) demonstrate promise in tackling unseen tasks with notable competencies. However, these models exhibit a fundamental vulnerability. LLMs are costly to train and utilize: their cost has increased 10 to 100-fold since 2015 and must be run on giant compute clusters. The training data used for corporate models is a closely guarded secret that lacks transparency [3]. Additionally, the success of LLMs has led to certain online content being generated entirely by these models, which are susceptible to producing non-factual information. In specialized domains, online information can be unreliable, detrimental, and contain logical inconsistencies that impede the models' reasoning ability. Nevertheless, most prior research on common sense detection has concentrated on the general domain. [1].

In this context, our study focuses on the challenge of identifying common sense errors in clinical notes. Unlike correcting these errors, which requires a deep understanding and specific knowledge of the medical field, identification is a crucial first step that demands the models' ability to recognize inaccuracies and anomalies in the text. This work explores how advanced natural language processing (NLP) technologies, such as GRU with BioWord-Vec, and especially ClinicalBERT[5], can be useful for analyzing unstructured medical texts. Our methodological approach involves a comprehensive comparative analysis among these models, highlighting their efficiency in identifying errors in clinical notes, underscoring the relevance of adapting model training to the peculiarities of medical data. We seek to demonstrate that, through specialization and fine-tuning of these LLMs models, it is possible to significantly improve their ability to detect erroneous or missing information, crucial for diagnosis and treatment in the clinical setting. This study not only aims to demonstrate the capabilities and limitations of advanced models in specialized medical contexts but also to emphasize the importance of integrating specialized knowledge within LLMs to optimize the reliability and usefulness of clinical notes in medical practice.

This document is described as part of our participation in the Shared Task Medical Error Detection and Correction of the Association for Computational Linguistics [1].

2 Related Work

In recent advances in the field of NLP, the ability to identify common sense errors in clinical notes poses a significant challenge and represents an opportunity to improve the quality of healthcare. The relevance of this study lies in exploring the applicability of advanced NLP models for the accurate detection of inaccuracies in medical records. These models constitute a promising advance over the inherent limitations off LLMs especially those arising from the quality and diversity of their training

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 461–469
June 21, 2024 ©2024 Association for Computational Linguistics

datasets [4]. LLMs often require domain-specific adaptations to effectively handle specialized tasks due to these limitations [4]. Moreover, models like ClinicalBERT have been shown to significantly improve their performance in interpreting clinical language by adapting to specific contexts [11].

NLP has the capacity to derive meaningful insights from unstructured data, specifically in the domain of categorizing incident reports and adverse events. Understanding the nature and reasons behind these incidents is crucial for analyzing adverse events. If NLP can enable the extraction of these insights from larger datasets, it has the potential to enhance learning from adverse events in the healthcare field. [13].

Given the complexity of clinical notes and the necessity for a high degree of precision in their analysis, this study is grounded in the review of previous research that has addressed similar issues in the domain of medical text classification. A relevant study focused on clinical text classification using rule-based features and knowledge-guided convolutional neural networks, leveraging trigger phrases and Unique Medical Concepts (CUIs) from the unified Medical Language System (UMLS) to enhance classification accuracy in class-imbalanced situations [12]. This approach underscores the effectiveness of integrating deep learning with explicit medical knowledge, emphasizing the importance of adapting model training to the specificities of clinical data.

Additionally, a comparative investigation evaluated various deep learning models, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), GRU, Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and a Transformer encoder, in their ability to handle unstructured medical note texts affected by different levels of class imbalance [6]. This analysis provides a critical perspective on the variability in model performance in the face of the unique challenges posed by medical data, highlighting the need for more specialized and adaptive approaches.

These studies and similar efforts outline the current state of using advanced NLP technologies in medical text classification. The present study draws inspiration from these research endeavors to advance understanding of the application of specific NLP models in error identification in clinical notes. In doing so, we aim to contribute to the field by providing valuable insights for future research and practices in this essential domain.

3 System Description

In the system description of our study, we address the implementation of an advanced predictive model specifically designed for detecting errors in clinical notes. This model relies on two fundamental pillars of NLP: GRU and the ClinicalBERT architecture [5]. The formulation of our central hypothesis questions the effectiveness of lexical and contextual features obtained through these NLP technologies to identify inaccuracies within clinical texts.

We propose two main methodological strategies. The first strategy implements GRU to extract lexical features, leveraging its ability to process complex temporal dependencies in the data [8]. This aspect is reinforced by the use of BioWordVec, which provides detailed vector representations of medical terms, thereby facilitating the capture of the semantic complexities of clinical language. The adaptability of GRU models to variable-length sequences proves particularly useful for analyzing medical texts, where critical information may be irregularly distributed throughout the document [9].

The second strategy focuses on harnessing ClinicalBERT, a model known for its ability to weigh the relevance of words through attention mechanisms, thereby enabling a deep understanding of the context in which medical terms are embedded. This approach significantly benefits from transfer learning, adapting previously acquired knowledge from extensive medical text corpora to fine-tune the model for our specific task. The synergy between GRU and ClinicalBERT enables a comprehensive analysis of the texts, evaluating not only coherence but also the semantic accuracy of the clinical content [6].

ClinicalBERT exhibits superior performance in identifying significant connections between medical concepts, a validation corroborated by medical experts [6].. This model has surpassed several benchmarks in predicting 30-day hospital readmissions, using discharge summaries and notes from early intensive care units, covering multiple clinically relevant metrics [6]. The attention weights generated by ClinicalBERT facilitate the interpretation of predictions, providing a deeper understanding of the context in which medical terms are embedded. We have released the model parameters and training scripts to encourage further research in this field. Thanks to its flexible structure, ClinicalBERT can be easily adapted to other predictive tasks with minimal engineering effort, making it ideal for studies requiring detailed analysis of clinical language [6].

Based on the outlined strategies, we configure a detailed Training System as depicted in Figures 1 and 2 of the study. This system unfolds through a sequence of well-defined stages: data ingestion and preliminary cleaning, generation of training instances, extraction of both lexical and contextual features, followed by the classification phase, and finally, model evaluation. This process ensures comprehensive treatment of clinical notes, optimizing error detection through the joint evaluation of long temporal dependencies and detailed contextual analysis.

This approach highlights not only the relevance of incorporating advanced NLP tools in the assessment of clinical texts but also the potential of these technologies to progress towards a higher degree of accuracy and reliability in medical documentation.

4 Data Description

The dataset provided by MEDIQA-CORR @ NAACL-ClinicalNLP 2024 [2] offers a comprehensive collection of medical texts, each corresponding to a clinical case report. This dataset stands out for its structured and detailed content, tailored for facilitating the analysis and identification of medical errors. Below are the key features of this dataset:

This dataset represents a valuable tool for research in the field of NLP applied to medicine, especially in tasks related to the identification and correction of errors in clinical texts. The richness and specificity of the data facilitate the development and evaluation of advanced NLP models, as addressed in this study, providing a solid foundation for detailed analysis and improvement of the quality of clinical notes.

5 Embeddings

In the process of generating embeddings for our analysis, we applied meticulous preprocessing to the provided data. This preprocessing consisted of a series of essential steps to ensure the quality and uniformity of the text, including correcting encoding errors and normalizing medical terms and units of measurement. This preliminary treatment of the texts is crucial to mitigate variations and ensure the integrity of the analyzed data.

Subsequently, we focused on transforming these normalized texts into vector representations using the BioWordVec model. This model, specifically trained on extensive medical corpora, was selected for its ability to accurately capture the semantics and clinical context of the terms used in the notes. By converting the texts into 200-dimensional vectors, representations of unrecognized words were adjusted using the <OOV> token, following a standardized approach for sequence length. This text-to-embeddings transformation procedure is essential for subsequent analysis using NLP techniques.

We used BioWordVec based a previous study, which findings across five models utilizing various word embeddings indicate that BioWordVec embeddings marginally enhanced the Bi-LSTM model's performance for certain datasets. Overall, models incorporating BioWordVec embeddings exhibited slightly superior performance compared to those utilizing GloVe embeddings[9] .

Through tokenization and sequence adjustment, we prepared the data for processing by advanced models such as GRU and ClinicalBERT. These models require structured and coherent inputs to effectively interpret the information contained in the clinical notes and thus identify possible errors. The meticulousness of this approach highlights the importance of preprocessing in NLP-supported clinical research. By transforming clinical notes into contextualized embeddings, we facilitate deep and accurate analysis by NLP models, enhancing error detection. This process not only enhances the analytical capability of the models but also underscores the value of rigorous data preparation in the field of artificial intelligence applied to medicine.

6 Data Transformation

After normalizing the data, we proceeded with its segmentation into training and test sets, adjusting this split according to the specific model to be used and experimenting with different partitions to always seek optimal accuracy. For the analysis with GRU, we selected an 80-20 split for training and testing, respectively, while for the evaluation using BioWordVec and ClinicalBERT, the distribution was adjusted to 70-30. This differentiation allowed us to adapt the learning and validation process to the peculiarities of each model, optimizing their ability to analyze and understand complex clinical texts.

This meticulous preparation and segmentation

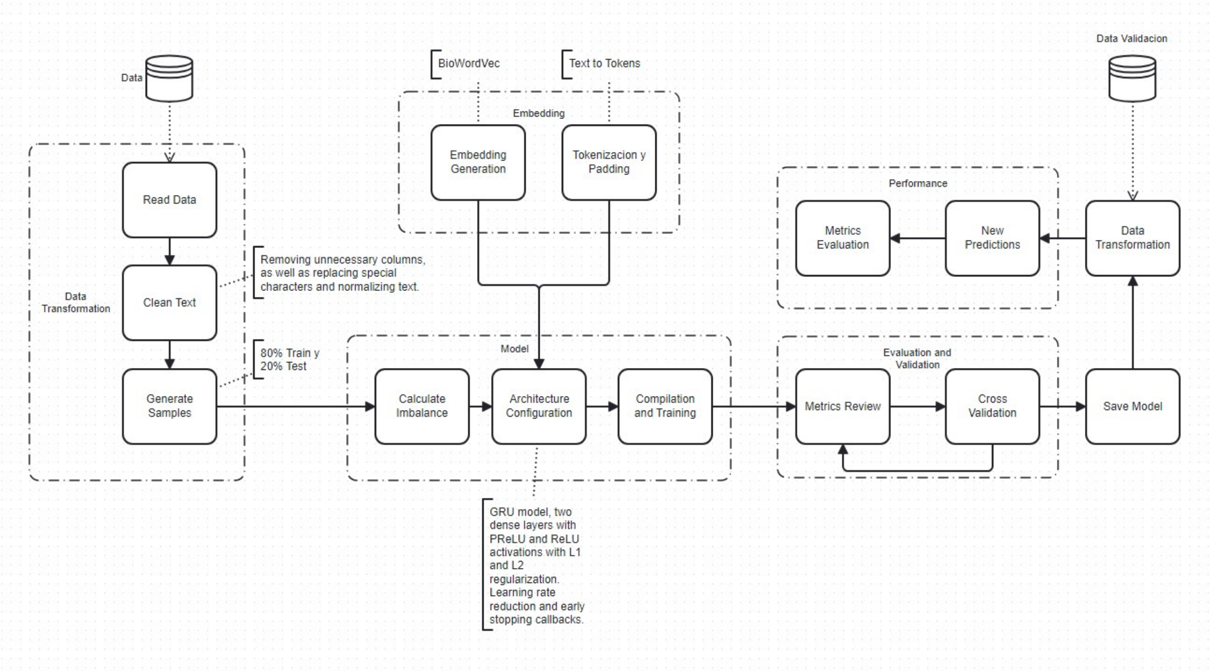

Figure 1: Model GRU

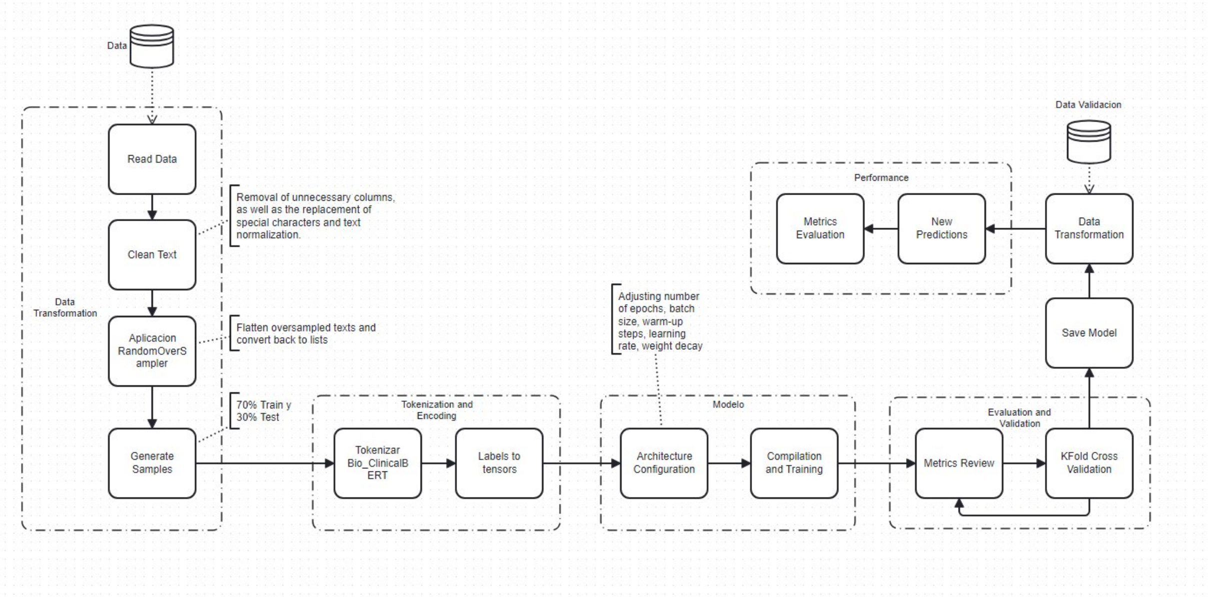
Data
Data Transformation
Read Data
Clean Text
Removal of unnecessary columns, as well as the replacement of special characters and text normalization.
Aplicacion RandomOverSampler
Flatten oversampled texts and convert back to lists
Generate Samples
70% Train y 30% Test
Tokenization and Encoding
Tokenizar Bio_ClinicalBERT
Labels to tensors
Adjusting number of epochs, batch size, warm-up steps, learning rate, weight decay
Modelo
Architecture Configuration
Compilation and Training
Evaluation and Validation
Metrics Review
KFold Cross Validation
Save Model
Data Validacion
Performance
Metrics Evaluation
New Predictions
Data Transformation

Figure 2: Model Bio Clinical

of the data reflect the rigor with which we approach the implementation of advanced NLP techniques. By establishing solid foundations for the training and evaluation of models such as GRU, BioWordVec, and ClinicalBERT, our goal is to maximize their effectiveness in the precise identification of errors in medical documentation. This commitment to a detailed and adaptive methodology underscores our objective to advance the application of artificial intelligence to improve the accuracy and reliability of clinical documentation.

7 Feature Extraction

The process of extracting lexicographic features involved analyzing fundamental aspects of the text, such as the use of specific terms and the overall semantic structure of the clinical notes. This included evaluating polarity and the frequent use of parts of speech, which are indicative of the tone and intention of the medical text. Through this analysis, we sought to better understand how lexicographic features can influence the presence of errors within the notes.

For the GRU-based model, we adjusted the class weights to address the imbalance in our data, using the number of unique classes derived from the training set. This adjustment was crucial for training a balanced model capable of effectively classifying texts based on the presence or absence of medical errors. The GRU model was configured with layers specifically designed to capture and analyze complex temporal dependencies within clinical texts, including regularization layers to prevent overfitting and optimize overall performance.

Simultaneously, for the implementation based on ClinicalBERT, we proceeded with data tokenization and preparation using the AutoTokenizer from 'emilyalsentzer/BioClinicalBERT'. This preparation was essential to adapt our clinical notes to the format required by ClinicalBERT, allowing the model to process and classify the texts efficiently. The training of the model focused on binary classification of texts, training on contextualized representations generated to identify the presence of errors with high precision.

The training of the GRU and ClinicalBERT models was conducted under carefully selected configurations to optimize their learning and evaluation on the dataset. These configurations included defining the number of epochs, batch size, and learning rate, which are fundamental elements for the success of the training.

8 Settings

In the setup of the study, specific adjustments were made to the hardware and software parameters to optimize the analysis of the GRU and ClinicalBERT models. These adjustments included the optimization of processors and the allocation of execution threads, essential for the efficient processing of the clinical dataset.

Additionally, differentiated configurations were implemented in the software environment to adapt to the peculiarities of each model. This involved optimizing data loading, preprocessing, and embedding generation, ensuring that both GRU and ClinicalBERT operated under optimal conditions for text analysis. Adapting the computational environment allowed for maximizing the capabilities of each model, facilitating a thorough and precise analysis of clinical texts.

The computational infrastructure was also configured to log and store the highest performing features and classifiers during the experimental phase. This systematic approach allowed for continuous monitoring of model performance, providing a solid foundation for iteration and enhancement of analysis strategies.

This detailed setup reflects the methodical and rigorous approach adopted for the preparation and execution of the NLP models. By optimizing computational resources and adapting the software, the necessary conditions were established for an effective evaluation of the models' ability to identify errors in medical documentation.

9 Experiments and Analysis of Results

Comprehensive evaluations of multiple natural language processing models were conducted using the dataset provided by MEDIQA-CORR @ NAACL-ClinicalNLP 2024, with the goal of identifying those with the best performance in detecting errors in clinical notes . These experiments not only allowed for the adjustment of model configurations but also served to identify optimal techniques that significantly contribute to the analysis of medical texts. Among the evaluated models, GRU and BioClinicalBERT proved to be the most effective across various metrics and scenarios, which is why they were selected for further detailed analysis.

During the initial evaluations, models such as RF, RoBERTa, BERT, and CNN were also tested.

Hyperparameters for these models were adjusted to obtain better results, revealing their potential when dealing with larger datasets [7]. The implementation of RF and CNN models highlighted the importance of feature identification and automatic feature extraction, respectively [10]. Moreover, the use of BERT models leveraged the transformer architecture to pre-train language representations, enhancing the understanding of context and semantics in clinical terms [7]. This extensive evaluation facilitated the refinement of strategies and parameters for each model, aiming to maximize their accuracy in classifying texts based on the presence of errors.

Throughout multiple iterations in the pre-evaluation phase, strategies and parameters for each of these selected models were refined with the goal of maximizing their ability to classify texts accurately based on the presence of errors. Standard competition metrics, with a special emphasis on accuracy (ACC), were employed to measure the performance and effectiveness of the developed systems.

The experiments revealed notable differences in the efficacy of the GRU and BioClinicalBERT models for analyzing the medical corpus. While GRU excelled in its ability to process text sequences and capture temporal dependencies, BioClinicalBERT proved to be particularly effective in understanding the context and specific semantics of clinical terms. This distinction underscores the complementarity of the models in handling complex medical texts.

The results, summarized in Table 1, provide a clear view of the performance of the models under study. Compared to other traditional classification algorithms, GRU and BioClinicalBERT provided a deeper and more nuanced analysis of clinical notes, demonstrating their superiority in identifying inaccuracies and textual anomalies.

This detailed analysis reinforces the importance of adopting advanced NLP approaches in the realm of clinical documentation. The findings not only demonstrate the viability of these models to improve error detection in medical texts but also open new avenues for future research in the field of NLP applied to health, marking a step forward in the goal of elevating the quality and reliability of medical information through technology.

10 Result Test

Table 2 summarizes the performance of various classifiers in terms of accuracy during the training and testing phases, showing both the absolute accuracy (Training Accuracy, Testing Accuracy) as well as the accuracy differences between these phases for each evaluated model. This initial evaluation allowed us to identify models with promising performance.

Among the evaluated models, ClinicalBERT and GRU stood out for their robust performance across various metrics and were selected for further detailed analysis. After rigorous validation, which included reviewing performance and learning curves, Table 3 details the accuracy of these models on the validation set, confirming their efficacy.

The selection of ClinicalBERT and GRU was based on a rigorous analysis of their capacity to process and analyze complex clinical texts, showing notable superiority in identifying errors in medical documentation. The validation of these models confirms the effectiveness of our selection strategy and highlights the importance of exploring in depth how these models can contribute to improving the analysis of clinical notes in the future.

11 Discussion and Conclusion

The meticulous selection of embeddings and NLP models, specifically GRU and ClinicalBERT, is crucial for the accurate analysis of clinical texts, as evidenced in our findings. These decisions are vital for optimizing error detection in clinical notes. However, there is a need to expand experimentation with a broader spectrum of models and embeddings to validate their effectiveness in specific clinical contexts. The analysis of the results, presented in Table 3, compares models from the most basic to the more complex ones (excluding large language models), revealing a progression and the importance of a detailed methodology and the adaptation of models to clinical textual peculiarities. This approach underscores the urgency of increasing experimentation to enhance precision and applicability in improving clinical documentation.

Despite considering the use of advanced LLMs like Gemini or ChatGPT-4, this study highlights the efficacy of alternative models such as ClinicalBERT and GRU. This preference is due not only to their competent performance but also to their specific adaptability to the demands of clinical texts. This approach is crucial in environments

Modelos	Train		Test	
	F1	**Acc.**	**F1**	**Acc.**
Roberta	0.71	0.55	0.71	0.56
Roberta_Sobremuestreo	0.64	0.53	0.61	0.47
Roberta_96_warmup_steps_9_epochs	0.64	0.56	0.55	0.45
Roberta_48_warmup_steps_15_epochs	0.86	0.86	0.5	0.46
Roberta_15_epochs	0.88	0.87	0.52	0.48
Roberta_20_epochs	0.66	0.49	0.67	0.5
Roberta_25_epochs	0.99	0.99	0.51	0.48
Roberta_30_epochs	0.99	0.99	0.51	0.48
Roberta_35_epochs	0.99	0.99	0.44	0.5
Roberta_40_epochs	0.99	0.99	0.48	0.51
Roberta_sobremuestro_steps_45_epochs	1	1	0.41	0.43
Bio_medical_sobremuestro_5_epochs	0.67	0.67	0.53	0.51
Bio_medical_96_warmup_steps_5_epochs_8_batch	0.69	0.7	0.48	0.48
Bio_medical_96_warmup_steps_10_epochs_8_batch	0.95	0.95	0.45	0.45
Bio_medical_sobremuestro_10_epochs_16_batch	0.94	0.94	0.48	0.47
Bio_medical_sobremuestro_7_epochs_16_batch	0.8	0.78	0.49	0.45
Gpt2-medium_1_batch	0.66	0.5	0.7	0.54
Longformer-base-4096	0.66	0.5	0.73	0.58
Random Forest_split_vectorizar	0.36	0.56	0.35	0.54
Random Forest_vectorizar_split	0.36	0.56	0.35	0.53
Random Forest_vectorizar_split_10_leaf	0.35	0.56	0.35	0.53
Random Forest_80_train_20_test	0.35	0.56	0.33	0.51
Stacking RL, SVC y RF	0.3609	0.3597	0.7514	0.7511
Stacking RL, SVC, RF, GB y DT	0.023	0.021	0.822	0.8219
GRU_No_Embbeding	0.665	0.5	0.69	0.53
GRU_glove-wiki-gigaword-200	0.47	0.52	0.33	0.42
GRU_glove-wiki-gigaword-200_DropOut	0.98	0.97	0.54	0.51
GRU_BioWordVec_PubMed_MIMICIII_d200	0.67	0.56	0.6	0.49
GRU_BioWordVec_MIMICIII_d200_desbalanceo	0.71	0.56	0.69	0.53
GRU_BioWordVec_MIMICIII_d200_L2	0.72	0.56	0.7	0.54
GRU_BioWordVec_MIMICIII_d200_L1_L2	0.53	0.57	0.52	0.53
LR	0.2836	0.2841	0.2836	0.2841
CNN	0.4043	0.5619	0.3746	0.5365
RNN	0.2668	0.4380	0.2935	0.4634
LSTM	0.4043	0.5619	0.3746	0.5365

Table 1: Detailed Model Results

Classifiers	Train Acc.	Diff. Train Acc.	Test Acc.	Diff. Test Acc.
ClinicalBERT	0.77	0.00	0.51	0.00
GRU	0.57	-0.20	0.53	0.02
Random Forest	0.56	-0.11	0.54	0.03
CNN	0.56	-0.11	0.53	0.02
LSTM	0.56	-0.11	0.53	0.02
RoBERTa	0.55	-0.12	0.56	0.05
GPT-2	0.50	-0.17	0.54	0.03
Longformer	0.50	-0.17	0.58	0.07
Stacking RL, SVC y RF	0.35	-0.32	0.75	0.24
RL	0.28	-0.39	0.28	-0.23
Stacking RL, SVC, RF, GB y DT	0.21	-0.46	0.82	0.31

Table 2: Model and Results Record

Classifiers	Acc Validation	Diff. Acc.
ClinicalBERT	0.56	0.00
GRU	0.55	-0.01

Table 3: Selected Models Validation Set Results

where data security, privacy, and time and resource constraints are primary considerations. In such contexts, the need for efficient yet less demanding models makes specialized alternatives surpass more generalist LLMs, aligning better with practical limitations and data protection imperatives in clinical research

Throughout the experiments conducted, it was observed that specific models such as GRU and ClinicalBERT demonstrate significant potential in processing medical text, emphasizing that, with proper data preparation and model tuning, it is possible to effectively manage the complexities inherent in clinical texts. Although the highest accuracy percentages obtained do not significantly exceed the random decision threshold, these results do not detract from the effectiveness of the models employed, but rather underline the importance of a meticulous selection and configuration of modeling features and parameters.

This study demonstrates that advancements in NLP can significantly contribute to the clinical field, although it also highlights the ongoing challenge of adapting these technologies to the specificities of medical language and data. NLP models, even in the face of accuracy challenges, prove to be valuable tools when carefully adjusted based on a deep understanding of the context and specific objectives of the task.

For future research, feature selection is highlighted as the primary strategy. It is suggested to focus on the development and application of advanced methodologies for feature extraction and selection, with the aim of refining the analytical capabilities of models for the precise processing and understanding of medical texts. This methodological approach not only anticipates an increase in the accuracy of models for anomaly detection and error identification in clinical documentation but also promises to deepen our understanding of the adaptation and optimization of NLP techniques for specific needs within the healthcare domain.

In conclusion, this study significantly contributes to the field of NLP applied to the medical domain, promoting the continuous innovation and optimization of models that, through meticulous choice and configuration of features, have vast potential to elevate the quality of clinical documentation. A notable finding is the moderate impact that pre-trained embeddings have on model performance, indicating that the integration and thorough exploration of these pre-trained tools can be crucial for amplifying the effectiveness of NLP in clinical contexts. This constant adaptation and improvement of technologies promise to advance towards optimizing the practical utility of NLP models, thereby contributing to improving the standards of care and documentation in the healthcare sector.

References

[1] Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024. Overview of the mediqa-corr 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

[2] Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin.

[3] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. BioMedLM: A 2.7b parameter language model trained on biomedical text. *Preprint*, arxiv:2403.18421 [cs].

[4] Matt Casey. 2023. Large language models: their history, capabilities and limitations.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

[6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint*. ArXiv:1904.05342 [cs].

[7] Jyoti Kumari and Abhinav Kumar. 2023. JA-NLP@LT-EDI-2023: Empowering Mental Health Assessment: A RoBERTa-Based Approach for Depression Detection. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 89–96, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

[8] Lishuang Li, Jia Wan, Jieqiong Zheng, and Jian Wang. 2018. Biomedical event extraction based on GRU integrating attention mechanism. *BMC Bioinformatics*, 19(9):285.

[9] Hongxia Lu, Louis Ehwerhemuepha, and Cyril Rakovski. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. 22(1):181.

[10] Yuanren Tong, Keming Lu, Yingyun Yang, Ji Li, Yucong Lin, Dong Wu, Aiming Yang, Yue Li, Sheng Yu, and Jiaming Qian. Can natural language processing help differentiate inflammatory intestinal diseases in china? models applying random forest and convolutional neural network approaches. 20(1):248.

[11] Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. *arXiv preprint*. ArXiv:2304.05368 [cs].

[12] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, 19(3):71.

[13] Ian James Bruce Young, Saturnino Luz, and Nazir Lone. 2019. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*, 132:103971.