CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Desidentificação de narrativas clínicas com modelos generativos de código aberto

De-identification of clinical narratives with open source generative models

Desidentificación de narrativas clínicas con modelos generativos de código abierto

Elisa Terumi Rubel Schneider[1], Fernando Henrique Schneider[2] , Yohan Bonescki Gumiel[1], Lilian Mie Mukai Cintho[3], Adriana Pagano[4], Emerson Cabrera Paraiso[5] , Marina de Sa Rebelo[1], Marco Antonio Gutierrez[1], Jose Eduardo Krieger[1] , Claudia Moro[5]

1 PhD, Instituto do Coração - InCor/HC FMUSP, São Paulo (SP), Brazil
2 BSc, Instituto do Coração - InCor/HC FMUSP, São Paulo (SP), Brazil
3 PhD, Universidade Estadual de Ponta Grossa (UEPG), Ponta Grossa (PR), Brazil
4 PhD, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte (MG), Brazil
5 PhD, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba (PR), Brazil

Autor correspondente: Prof. Dr. Claudia Moro
E-mail: c.moro@pucpr.br

Resumo

Objetivos: A desidentificação de narrativas clínicas é essencial para proteger a privacidade dos pacientes e garantir a conformidade com as regulamentações. No entanto, é uma tarefa complexa devido aos distintos tipos de entidades a serem desidentificadas e à necessidade de processar os textos localmente, por questões de segurança e privacidade. **Métodos:** Este artigo apresenta um estudo experimental sobre desidentificação de narrativas clínicas utilizando modelos generativos de código aberto, que podem ser executados localmente. **Resultados:** Avaliamos a eficácia de cinco modelos de linguagem, comparando-os ao GPT-4, um modelo proprietário. Os modelos foram avaliados com base na precisão, *recall* e *F-score*. Nossos resultados preliminares indicam que, embora o GPT-4 tenha atingido o melhor desempenho, o modelo aberto Llama3, da Meta, demonstrou robustez e eficácia nesta tarefa. **Conclusão:** O estudo contribui para o campo ao fornecer *insights* sobre o desempenho de diferentes modelos na anonimização de narrativas clínicas.

Descritores: Prontuários Médicos; Processamento de Linguagem Natural; Inteligência Artificial

Abstract

Objectives: De-identifying clinical narratives is essential to protect patient privacy and ensure regulatory compliance. However, this is a complex task due to the various types of entities to be de-identified and the need to process texts locally for security and privacy reasons. **Methods:** This article presents an experimental study on the de-identification of clinical narratives using open-source generative models that can be run locally. **Results:** We evaluated the effectiveness of five language models, comparing them to GPT-4, a proprietary model. The models were assessed based on precision, recall, and F-score. Our preliminary results indicate that while GPT-4 achieved the best performance, the open-source model Llama3 by Meta demonstrated robustness and effectiveness in this task. **Conclusion:** This study contributes to the field by providing insights into the performance of different models in anonymizing clinical narratives.

Keywords: Medical Records; Natural Language Processing; Artificial Intelligence

Resumen

Metas: La desidentificación de narrativas clínicas es esencial para proteger la privacidad de los pacientes y garantizar el cumplimiento de las normativas. Sin embargo, es una tarea compleja debido a los distintos tipos de entidades que requieren desidentificación y a la necesidad de procesar los textos localmente por razones de seguridad y privacidad. **Métodos:** Presentamos un estudio experimental sobre la desidentificación de narrativas clínicas utilizando modelos generativos de código abierto que pueden ejecutarse localmente. **Resultados:** Evaluamos la eficacia de cinco modelos de lenguaje, comparándolos con GPT-4, un modelo propietario. Los modelos fueron evaluados por la precisión, el recall y el F-score. Nuestros resultados preliminares indican que, aunque GPT-4 logró el mejor rendimiento, el modelo de código abierto Llama3 de Meta demostró robustez y eficacia en esta tarea. **Conclusión:** Este estudio contribuye al campo proporcionando información sobre el rendimiento de diferentes modelos en la anonimización de narrativas clínicas.

Descriptores: Registros Médicos; Procesamiento de Lenguaje Natural; Inteligencia Artificial

Introduction

Every time a patient interacts with a healthcare system, medical documentation is produced, including both structured and unstructured data. Given this scenario, there are billions of electronic medical records (EMRs) in existence [1]. These records are crucial for data-driven medical research and the advancement of healthcare practices[1]. For instance, narrative clinical texts contain detailed clinical information like disease treatment and medication, being increasingly recognized as vital for clinical studies and medical applications[2]. Using EMRs information offers tremendous potential, but also raises significant concerns regarding patient confidentiality and privacy breaches[3].

Advancing medical knowledge often relies on clinical studies where patient volunteers provide informed consent before any procedures or data collection begins[4]. However, retrospective studies leveraging the vast information in EMRs also offer significant potential for breakthroughs. These studies, designed after patient care has been documented, still require informed consent, which can be logistically challenging to obtain from each patient or their families[4]. In the absence of informed consent, the use of EMRs containing protected health information (PHI) is not allowed until all PHI is de-identified[4]. Therefore, de-identification (or anonymization), which involves identifying and removing PHI, is a critical step in making clinical data accessible for research [1]. This process removes or replaces sensitive information while keeping the records otherwise unchanged, ensuring that patient privacy is protected without compromising the utility of the data for research purposes [1].

In the United States, the guidelines for de-identification are outlined by HIPAA[5], which mandates the protection of personal health information. Meanwhile, in Brazil, similar protections are enforced under the so-called LGPD (General Data Protection Act [1]), which not only protects privacy rights but also categorizes sensitive data, including health-related information, to prevent its misuse.

[1] Act Nº 13.709, 14th August, 2018. Available at https://www.gov.br/cidadania/pt-br/acesso-a-informacao/lgpd/classificacao- dos-dados

Manually de-identifying clinical texts is highly impractical and costly in terms of time, effort, and expense. Consequently, there is a significant need for an automated de-identification system[6]. Clinical notes vary depending on their purpose and institutional conventions, and they incorporate PHI in ways that are difficult to identify and redact [7].

In this context, natural language processing (NLP) and machine learning (ML) techniques are employed to identify and redact PHI from clinical texts, treating this process as a named entity recognition (NER) task. Initially, many approaches to this task were rule-based, relied on dictionary matching, or were hybrid methods that combined machine learning algorithms with rules. For instance, Prado et al. [8] utilized regular expressions and lists to anonymize cardiology texts in Brazilian Portuguese within a rule-based framework. In the realm of hybrid approaches, Deleger et al.[9] implemented Conditional Random Fields (CRFs) alongside rules, while Grouin and Névéol[4] applied CRFs and rules to texts in French across various specialties, document types, and hospitals. Additionally, Yang and Garibaldi [2] employed CRFs combined with rule-based methods.

With the advancement of deep learning technologies, several new methodologies have gained prominence. In terms of deep learning-based models: (1) Obeid et al.[10] compared Convolutional Neural Networks (CNNs) with traditional ML techniques; (2) Ahmed, Alziz, and Mohammed[11] explored the use of Recurrent Neural Networks (RNNs) and self-attention mechanisms; (3) Liu et al.[1] developed an ensemble method that integrates CRFs, Long Short-Term Memory networks (LSTMs), and rules; (4) Shweta focused on various RNN architectures; (5) Hartman et al.[7] applied Bidirectional LSTM (BI-LSTM) in conjunction with CRFs; (6) Catelli et al.[12] evaluated BI-LSTMs and CRFs with different embeddings against BERT models in Italian medical records concerning COVID-19, (7) Khin, Burckhardt, and Padman [13] utilized BI-LSTM with contextualized word embeddings and (8) Santos et al. [14] used Bi-LSTM-CRF and contextualized word embeddings for the-identification of patient names. These developments reflect the dynamic evolution of machine learning in the field of medical document de-identification.

Recently, large language models (LLMs) in the field of natural language processing (NLP) have attracted significant attention from the general public,

particularly with the rise of ChatGPT [15]. The development of advanced LLMs such as GPT-3 and GPT-4, which are pre-trained on vast datasets, has facilitated zero-shot and few-shot in-context learning, broadening their application in real-world scenarios [15]. Notably, a study by Liu et al.[15] is one of the few identified that explores the potential of GPT-4 to automatically detect and redact named entities (PHI) from clinical texts. However, this study did not address the use of open-source models, which are crucial for reducing de-identification costs. In terms of the Brazilian Portuguese language, previous research by Prado et al.[8] and Santos et al. [14] did not utilize LLMs. Therefore, our study aims to bridge this gap by investigating the efficacy of open-source LLMs in identifying PHI within cardiology clinical texts written in Brazilian Portuguese.

Methods

In this section, we will present the methodology of this work, describing our workflow in detail, from data collection and preprocessing to evaluation, as can be seen in Figure 1.

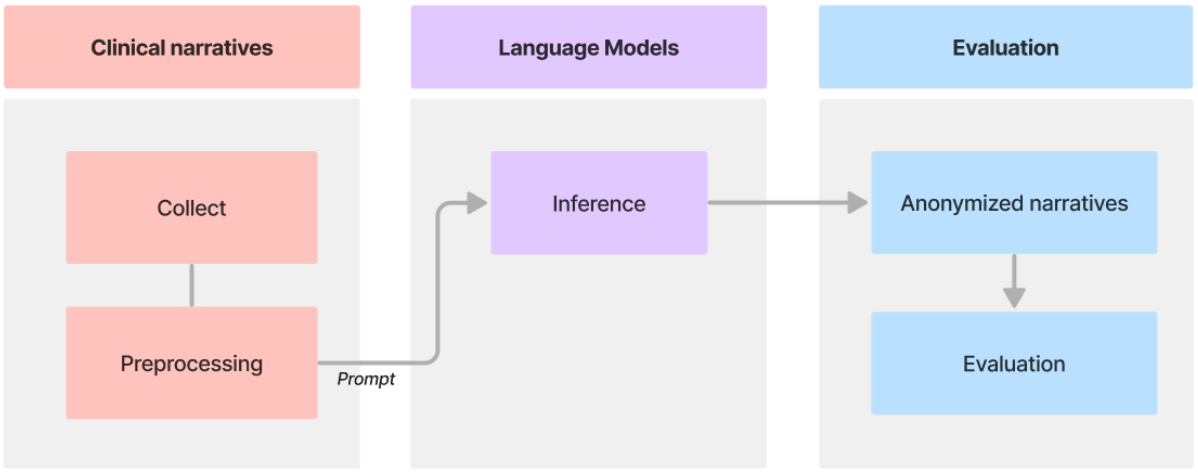

Figure 1 - Overview of the Clinical Narratives De-Identification Process

Data Acquisition

The data used in our experiment are from a tertiary hospital specialized in cardiopulmonary diseases in São Paulo, Brazil, collected in the work of Prado et al. [8]. The complete database contains approximately 160,000 clinical narratives, where the personal and sensitive data were replaced by a specific token composed of

special characters (#). We randomly selected some narratives from this collection and pre-processed them by re-identifying the data that had been anonymized with fictitious data. In some cases, we merged segments from two or more narratives to expand the pool of personal data examples for anonymization. As a result of this work, we obtained 10 entirely fictitious clinical narratives, but with a structure similar to real-world narratives.

In Figure 2, we present an example of a segment of a narrative used in the experiment alongside its gold standard, where personal information has been anonymized.

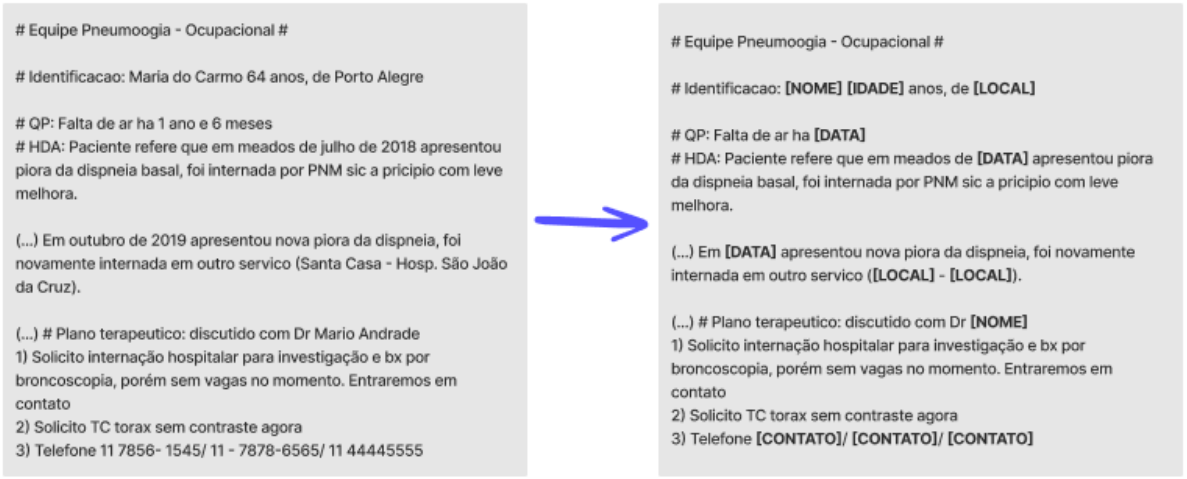

Figure 2 - A clinical narrative (left) and its de-identified gold standard (right) used in our experiment.

Prompt

In this work, as in the work of Liu et al.[15], we adopted a zero-shot approach, relying solely on prompt engineering techniques. This method enables the achievement of meaningful results without additional training on specific tasks or datasets, showcasing the effectiveness and efficiency of prompt-based learning.

To create the prompt with instructions for the models, we based our approach on the work of Liu et al.[15], which requires the de-identification of 18 HIPAA identifiers that can be used to identify, locate, or contact individuals. This is crucial for processes involving data-sharing and transmission of clinical text documents. The work of Liu et al.[15] maps HIPAA identifiers to the i2b2/UTHealth benchmark,

providing a template for optimized prompts to redact sensitive information in line with HIPAA guidelines.

We adapted the prompt to explicitly indicate the type of information being anonymized. This adjustment was made to facilitate evaluation and, if necessary, to enable re-identification using fictitious data, to train machine learning models for instance. We opted to keep the prompt in English for improved performance, as the models used in our experiments were pre-trained on multilingual data, with a higher prevalence of English texts.

Also, our prompt ensures compliance with the Brazilian General Data Protection Law, guaranteeing the protection of personal health information. By doing so, it not only meets the strict requirements of the LGPD but also ensures the secure handling and confidentiality of sensitive data.

Figure 3 illustrates our experiment's prompt, containing the task, the command, and the rules for the model to follow.

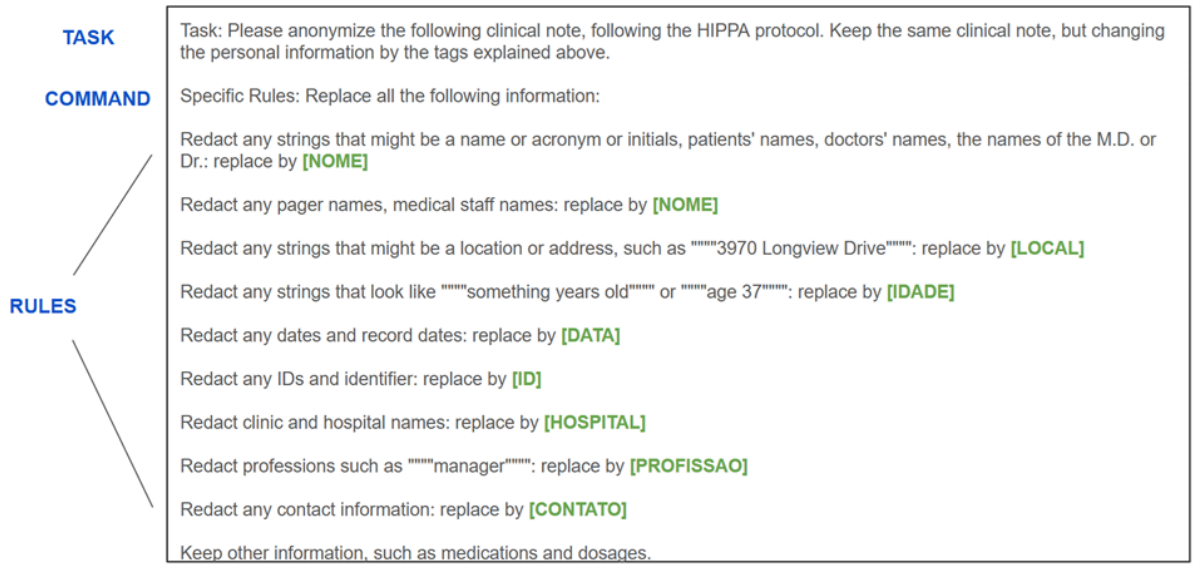

Figure 3 - Illustration of the prompt used in our experiment.

Experiments

In our research, we evaluated five open-source language models in the task of de-identification clinical narratives using a zero-shot prompting approach.

We focused our experiments on open-source models due to the importance of being able to be executed locally, as the narratives contain sensitive and personal data. We also ran experiments for comparison using the proprietary model GPT-4.0 from OpenAI. The model was accessed through its API, suitable for our experiment, but not recommended for real-world use, as it can expose sensitive data. Table 1 provides further details of the models used in the experiment.

Table 1 – Overview of language models used in our experiment. In the de-identification process, the prompt with instructions was concatenated with the clinical note and sent to each model for inference, one at a time.

Alias (used in this work)	Model	Number of parameters	Size of context (tokens)	License
Llama3 (16)	meta-llama/Meta-Llama-3-70B-Instruct	70B	8K	Llama3
Mistral (17)	mistralai/Mixtral-8x7B-Instruct-v0.1	46.7B	32K	Apache 2.0
Zephyr (18)	HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1	141B	2K	Apache 2.0
Command (19)	CohereForAI/c4ai-command-r-plus	104B	128K	CC-BY-NC-4.0
Gemma (20)	google/gemma-1.1-7b-it	7B	8K	Gemma License
GPT4 (21)	GPT 4.0	1T	8K	Proprietary

We leveraged the user-friendly Hugging Face Chat interface as a platform for conducting our experiments, enabling seamless interactions with the models and facilitating the inference process. HuggingChat [2] interface is an AI-powered chatbot developed by Hugging Face. A specialized assistant for the de-identification task was created, as shown in Figure 4. For each clinical note processed, a new chat session was created to ensure that previous results did not interfere with the current inference.

As the assistant may be useful for other research purposes, it was made available at: https://hf.co/chat/assistant/663a986ce556dc50f6d9e5a1.

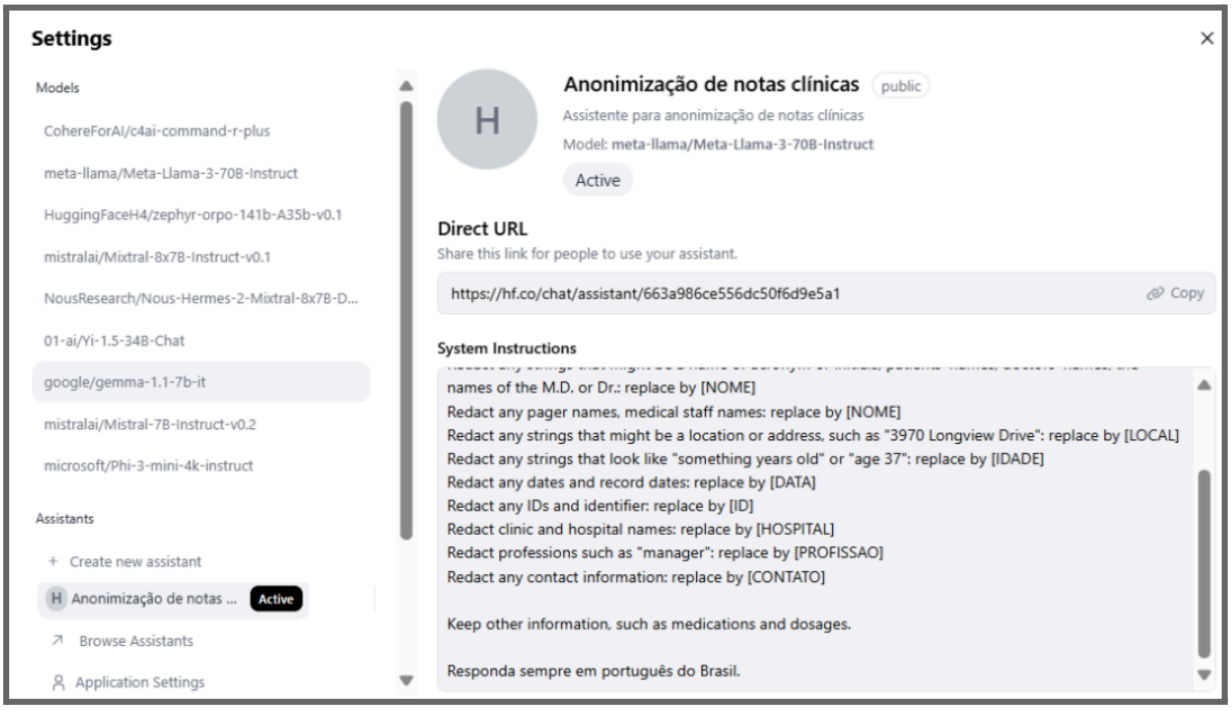

Figure 4 - Screenshot of the Hugging Chat assistant created for our task.

[2] https://huggingface.co/chat/

Evaluation

The performance of the models was evaluated in an entity-level way, without considering partial matches. We calculated the metrics of precision, recall, and F-score for each model, through equations (1), (2), and (3):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F-score} = 2 * \frac{Precision*Recall}{(Precision+Recall)} \quad (3)$$

where True Positives (TP) indicate the correct instances that were correctly identified, False Negatives (FN) indicate the instances that should have been identified but were not, and False Positives (FP) indicate the instances that were incorrectly identified as positive.

Precision measures the accuracy of the positive predictions made by the model, while recall, also known as sensitivity, measures the model's ability to identify all relevant instances. F-score (or F1-score) is the harmonic mean of precision and recall, a single metric that balances both precision and recall, which gives a more comprehensive view of the model's performance.

Results and discussion

Table 2 shows the precision, recall, and F-score values for all evaluated models, ordered by F-score, where we can see that the proprietary model GPT4 achieves the best performance, as expected since it has more parameters and extensive training data.

Among the open-source models, Llama3 had the best performance, demonstrating its robustness and effectiveness compared to other open-source alternatives.

Table 2 – Model results in terms of precision, recall, and F-score, where scores in bold indicate the highest values.

Model	Precision	Recall	F-score

Open-source			
Llama3	0,753	**0,884**	0,814
Command	0,524	0,867	0,653
Gemma	0,847	0,418	0,560
Zephyr	0,377	0,455	0,412
Mistral	**1,000**	0,238	0,385
Proprietary			
GPT4	**1,000**	0,828	**0,906**

These preliminary results can provide valuable insights into the strengths and weaknesses of each model in this task. GPT-4 stands out with a perfect precision of 1.000 and a high recall of 0.828, resulting in an F-score of 0.906. This indicates that GPT-4 was very accurate in its predictions and robust in identifying relevant instances. Among the open-source models, Llama3 performs notably well, with a balanced precision of 0.753 and recall of 0.884, leading to a strong F-score of 0.814. This demonstrates its effectiveness in managing the trade-off between precision and recall. In contrast, Command and Gemma exhibit significant imbalances, with Command achieving high recall but low precision and Gemma showing high precision but low recall. Zephyr and Mistral exhibit low F-scores, highlighting their need for substantial improvements to be competitive.

Models with a large number of parameters tend to perform better, suggesting that greater representational capacity, enabled by more parameters, can result in better generalization and performance in complex natural language processing tasks. On the other hand, models with fewer parameters, such as Gemma and Mistral, may struggle to capture the nuances and complexities of the data, leading to inferior performance.

Overall, while GPT-4 and Llama3 show robustness and reliability, the other models display varying degrees of performance challenges. Since de-identification is a critical task where false negatives can lead to the exposure of sensitive data, further studies must be conducted to ensure the effectiveness of this process and to comply with the LGPD and the guidelines of the HIPAA. This study represents an initial step toward future research and advancements in de-identifying clinical narratives in Brazilian Portuguese.

Conclusions

This study on de-identification of clinical narratives has highlighted the effectiveness and promise of open-source generative models in protecting the privacy of sensitive data. The experimental results demonstrated that, while there are challenges such as minimizing false negatives, the investigated models, such as Llama3, have shown the ability to achieve promising results, with a good balance between precision and recall. This F-score demonstrates that Llama3 is a strong open-source alternative for this task.

For future work, de-identification of Portuguese clinical narratives could be enhanced by fine-tuning existing models on larger and more diverse datasets to improve performance. Additionally, developing domain-specific annotation guidelines and datasets could improve model performance on clinical narratives.

To the best of our knowledge, this is the first paper on de-identification in Brazilian Portuguese clinical narratives using open-source generative models. Additionally, we are releasing a Hugging Chat assistant for further experimentation and development in this area.

Acknowledgments

This study was supported by Zerbini Foundation and Foxconn Brazil, as part of the research project "Natural Language Processing in Cardiovascular Medicine".

References

[1] Liu, Zengjian et al. "De-identification of clinical notes via recurrent neural network and conditional random field." Journal of biomedical informatics vol. 75S (2017): S34-S42. doi:10.1016/j.jbi.2017.05.023

[2] Yang, Hui, and Jonathan M Garibaldi. "Automatic detection of protected health information from clinic narratives." Journal of biomedical informatics vol. 58 Suppl,Suppl (2015): S30-S38. doi:10.1016/j.jbi.2015.06.015

[3] Meystre, Stéphane M et al. "Text de-identification for privacy protection: a study of its impact on clinical text information content." Journal of biomedical informatics vol. 50 (2014): 142-50. doi:10.1016/j.jbi.2014.01.011

[4] Grouin, Cyril, and Aurélie Névéol. "De-identification of clinical notes in French: towards a protocol for reference corpus development." Journal of biomedical informatics 50 (2014): 151-161.

[5] Act, Accountability. "Health insurance portability and accountability act of 1996." Public law 104 (1996): 191.

[6] Yadav, Shweta, et al. "Deep learning architecture for patient data de-identification in clinical records." Proceedings of the clinical natural language processing workshop (ClinicalNLP). 2016.

[7] Hartman, Tzvika, et al. "Customization scenarios for de-identification of clinical notes." BMC medical informatics and decision making 20 (2020): 1-9.

[8] Prado, Carolina Braun, et al. "De-Identification Challenges in Real-World Portuguese Clinical Texts." Latin American Conference on Biomedical Engineering. Cham: Springer Nature Switzerland, 2022.

[9] Deleger, Louise, et al. "Large-scale evaluation of automated clinical note de-identification and its impact on information extraction." Journal of the American Medical Informatics Association 20.1 (2013): 84-94.

[10] Obeid, Jihad S., et al. "Impact of de-identification on clinical text classification using traditional and deep learning classifiers." Studies in health technology and informatics 264 (2019): 283.

[11] Ahmed, Tanbir, Md Momin Al Aziz, and Noman Mohammed. "De-identification of electronic health record using neural network." Scientific reports 10.1 (2020): 18600.

[12] Catelli, Rosario, et al. "A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records." Ieee Access 9 (2021): 19097-19110.

[13] Khin, Kaung, Philipp Burckhardt, and Rema Padman. "A deep learning architecture for de-identification of patient notes: Implementation and evaluation." arXiv preprint arXiv:1810.01570 (2018).

[14] Santos, Joaquim, et al. "De-identification of clinical notes using contextualized language models and a token classifier." Brazilian Conference on Intelligent Systems. Cham: Springer International Publishing, 2021.

[15] Liu, Zhengliang, et al. "Deid-gpt: Zero-shot medical text de-identification by gpt-4." arXiv preprint arXiv:2303.11032 (2023).

[16] AI@Meta, 2024. Llama 3 model card. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[17] Mistral AI Team, 2024. Model Card for Mixtral-8x7B. URL: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

[18] Hong, J., Lee, N., Thorne, J., 2024. Orpo: Monolithic preference optimization without reference model. arXiv:2403.07691.

[19] CohereForAI, 2024. Model Card for C4AI Command R+. URL: https://huggingface.co/CohereForAI/c4ai-command-r-plus.

[20] Google, 2024. Gemma Model Card. URL: https://huggingface.co/google/gemma-1.1-7b-it.

[21] OpenAI, 2024. Gpt-4 technical report. arXiv:2303.08774.