Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI241070

Utilizing RAG and GPT-4 for Extraction of Substance Use Information from Clinical Notes

Fatemeh SHAH-MOHAMMADI[a,1] and Joseph FINKELSTEIN[a]
[a] *Department of Biomedical Informatics, School of Medicine, University of Utah, USA*
ORCiD ID: Fatemeh Shah-Mohammadi https://orcid.org/0000-0002-9034-7803,
Joseph Finkelstein https://orcid.org/0000-0002-8084-7441

Abstract. This research investigates the application of a hybrid Retrieval-Augmented Generation (RAG) and Generative Pre-trained Transformer (GPT) pipeline for extracting and categorizing substance use information from unstructured clinical notes. The aim is to enhance the accuracy and efficiency of identifying substance use mentions and determining their status in patient documentation. By integrating RAG to pre-filter and focus the input for GPT, the pipeline strategically narrows the scope of analysis to the most relevant text segments, thereby improving the precision and recall of the extraction. Utilizing the Medical Information Mart for Intensive Care III dataset, the performance of the pipeline was evaluated through manual verification, assessing various metrics including recall, precision, F1-score, and accuracy. The results demonstrated high precision rates (up to 0.99 for drug and alcohol mentions), and substantial recall (0.88 across all substances for status of the usage).

Keywords. Retrieval-augmented generation (RAG), GPT, Substance use information

1. Introduction

Substance use disorder is an increasingly prevalent multidimensional health issue nationwide. The complexity of assessing and treating substance use disorder arises from various factors such as the diversity of symptom expression, changing patterns of substance use, coexisting conditions, and social determinants of health that affect both clinical presentation and treatment outcomes. The Centers for Medicare and Medicaid Services acknowledge the necessity for specialized screening, prevention, and intervention protocols within clinical environments. Accordingly, they have provided incentives for the implementation of systematic approaches aimed at the early detection of substance use, ongoing monitoring of substance use patterns, and the facilitation of appropriate treatment referrals as required [1,2,3]. However, electronic health records lack the granularity for indicating these additional factors, as they are often documented in clinical notes in an unstructured format. As such, it is difficult to automatically flag any patients who may present specific risk factors [4,5]. To date, many conventional natural language processing (NLP) methodologies have been explored and applied for

[1] Corresponding Author: Fatemeh Shah-Mohammadi; E-mail: fateme.sh.mohammadi@gmail.com.

identifying supplementary patient risk factors in unstructured clinical documentation, achieving mixed outcomes [6,7]. Nonetheless, the substantial diversity in linguistic expression within clinical notes presents considerable challenges to the precision of traditional approaches that depend on parsing rules for text pattern recognition. Factors such as clinician typographical errors, neologisms, abbreviations, and other linguistic variations compromise the efficacy of these techniques. Recently, Large Language Models (LLMs) have emerged as a promising solution to the aforementioned challenges, particularly due to their capacity to autonomously "learn" and adapt to diverse linguistic patterns without requiring further model training. In this study, we investigate the utilization of LLMs, in specific generative pre-trained transformer (GPT), to address the challenge of extracting substance use information, in specific tobacco, alcohol and drug use information, from clinical notes; thereby contributing to risk assessment, treatment planning, patient safety, recovery, and overall well-being. Our proposed workflow involves a retrieval-augmented generation (RAG) strategy, which directs LLMs toward more accurate responses. The specific aim of the study is to showcase the creation of a specialized GPT-based pipeline integrated with a RAG system, that is tailored to extract the patients' substance use information.

2. Methods

The main data source for analysis in this paper is Medical Information Mart for Intensive Care III (MIMIC-III) dataset. Among all notes in this dataset, we focused on discharge summaries. This dataset contains 59,652 discharge summaries of 46,146 patients. We initially selected a random sample of 100 discharge summaries and manually analyzed the text to discern distinct patterns that facilitated the segmentation of each note into discrete chunks. We observed that specific sections within the discharge summaries such as social history, medication on discharge, and past medical history, demonstrated consistent patterns. These patterns were subsequently utilized to develop regular expressions that aided in the chunking of each note. To normalize the text within these chunks, we employed basic NLP techniques, such as converting text to lowercase and removing special characters. Following these preprocessing steps, the RAG model was integrated into our pipeline to generate embeddings and accordingly implement semantic search.

2.1. Integration of RAG

To generate embedding for each chunk within the note, we used embedding model "text-embedding-3-small", a small and highly efficient embedding model from OpenAI. After generating embeddings for each chunk within the note through RAG architecture, we leveraged GPT, and in specific GPT-4 model made available through an API by OpenAI. This approach was chosen to ensure that GPT-4 is only invoked with relevant sections of the clinical notes. Utilizing RAG allows us to leverage the clinical notes as an external data source, filtering the content to include only contextually relevant information in response to the query (prompt) presented to GPT-4, rather than the entirety of the content.

2.2. Semantic Search- Integration of GPT

The goal here is to implement a semantic search functionality utilizing sentence embeddings to retrieve chunks based on their semantic similarity to a specified query. Prompt engineering was employed to craft the optimal prompt for querying GPT. Our finalized prompt selected to be as follow: "*Extract mentions of tobacco use, illicit substance use, and alcohol use, along with the status of each (categorized as present, past, none, or non-mentioned) in bullets. Note that the status should be classified as non-mentioned if there is no information available regarding the patient's substance use profile.*" The embeddings generated in the previous stage will be used for semantic search. These embeddings capture the semantic essence of the text by representing the underlying meanings and relationships within the text rather than just the literal words. When a query is initiated (i.e., the crafted prompt is activated), the semantic search mechanism kicks in. It compares the semantic embeddings of the query prompt with those of the stored chunks from the clinical notes. This comparison is, in this study, based on measuring cosine similarity. Chunks whose embeddings show high similarity to the query embedding are considered relevant. This means they likely contain information that is semantically related to the query's focus which is extraction of patient's substance use information. The most similar chunk is then fed into GPT as context to the prompt. GPT, informed by both the specific prompt and the contextually relevant chunk, generates a response. This response aims to accurately address the query based on the semantically relevant information extracted from the clinical note. Table 1 shows example of RAG output (most relevant chunk to the prompt) and GPT output. This table, that showcase the interaction between the RAG model and GPT, highlights the efficacy of RAG in isolating pertinent sections of clinical notes for query processing. Considering first row of this table, the RAG model retrieves a segment under "Social History," which mentions the patient's use of tobacco and alcohol. RAG is capable of identifying content that has some relevance to the queried substances. GPT then processes this content, generating outputs that accurately reflect the patient's substance use containing the mentions of the use in the note and devising the status of the usage (whether it is past, present, none or not-mentioned). This ability of GPT to categorize substance use accurately is enhanced by the precision with which RAG selects the relevant text.

Table 1. Screenshot of pipeline output

RAG output	GPT output
relatedness=0.64 Social History: Retired, never used tobacco, occasional ETOH.	**- Tobacco use**: None (the patient never used tobacco) **- Illicit substance use:** Non-mentioned (there is no information about illicit substance use) **- Alcohol use:** Present (the patient occasionally uses alcohol)
relatedness=0.35 Past Medical History: Throat CA s/p radiation in [**2090**], Hypothyroidism, Tobacco, ETOH, COPD, Mild Dementia	**- Tobacco use:** Present **- Illicit substance use:** Non-mentioned **- Alcohol Use:** Present (implied by ETOH)

The second row of the table demonstrates the RAG model selecting the "Past Medical History" section of a discharge summary as the most relevant chunk for a query regarding substance use. Upon manual inspection, it was discovered that the associated discharge summary lacked a "Social History" section, typically where substance use is explicitly documented. Consequently, the "Past Medical History" section became the focus, as it contains indirect references to substance use, such as mentions of "Tobacco" and "ETOH" along with medical conditions like throat cancer and COPD, which are

often associated with these substances. The entire dataset processing was completed in approximately 20 minutes, with the bulk of this time attributed to retrieving and analyzing results from the GPT.

3. Results

In the evaluation of the performance of the system, assessments were conducted using classification metrics, including accuracy, precision, recall and F1 score metrics. For this project, Python version 3.9 was used. To calculate the aforementioned metrics, the output of the system was manually reviewed. Table 2 summarizes the results. In terms of extracting part of the text that mentions the substance use, i.e. "Mention of the Use", the recall scores are quite strong, indicating that the pipeline effectively captures true instances of substance use mentioned within the notes, with tobacco at 0.75, drugs at 0.78, and alcohol at 0.69. Precision is exceptionally high for all substances, with tobacco scoring 0.97, and both drugs and alcohol at 0.99, suggesting a high likelihood of correctness in predictions made by the pipeline. The F1-scores, which balance precision and recall, reflect the overall efficacy of the pipeline in identifying these mentions, recording scores of 0.85 for tobacco, 0.87 for drugs, and 0.81 for alcohol. However, the accuracy, which assesses the overall correctness including both true positives and true negatives, is somewhat lower, though still robust, with tobacco at 0.74, drugs at 0.77, and alcohol at 0.69.

Table 2. Performance of the pipeline

	Mention of the use			**Status of the use**		
Metrics	Tobacco	Drug	Alcohol	Tobacco	Drug	Alcohol
Recall	0.75	0.78	0.69	0.88	0.88	0.88
Precision	0.97	0.99	0.99	0.92	0.92	0.85
F1-score	0.85	0.87	0.81	0.90	0.90	0.87
Accuracy	0.74	0.77	0.69	0.82	0.82	0.77

For devising the status of the use, i.e "Status of the Use", the recall is consistently high at 0.88 across all substances, demonstrating the pipeline's excellent ability to correctly identify the status of use when mentioned. Precision varies slightly with excellent scores for tobacco and drugs (0.92 each) but is slightly lower for alcohol at 0.85, which may indicate some challenges in accurately determining the status of alcohol use. The F1-scores are high across the board, showing a strong balance of precision and recall in categorizing use status, with scores of 0.90 for tobacco and drugs, and 0.87 for alcohol. Accuracy for use status is notably higher than for use mention, indicating that the pipeline is more effective at correctly determining the status of the usage once a mention is identified, with values of 0.82 for tobacco and drugs and 0.77 for alcohol.

4. Discussion

Presented results highlight that the pipeline is highly effective in detecting mentions of substance use and accurately categorizing the status of that use. By leveraging retrieval to focus on relevant text chunks before generation, RAG minimizes the impact of

common errors that might arise from broader text analysis done by GPT. For example, clinician typos, neologisms, and non-standard abbreviations, which are typical in unstructured clinical notes, are less likely to mislead the analysis when the model is focused on segments pre-identified as relevant. This specificity not only boosts precision but also recall, as the model is less likely to overlook correct instances due to noise. This pipeline has the potential to significantly enhance the doctor-patient relationship by allowing for quick and accurate extraction of information regarding a patient's substance use from their medical records. By automating the retrieval and categorization of such data, doctors can gain a more comprehensive understanding of a patient's medical and social history without spending extensive time reviewing documents. A comparative evaluation of the traditional NLP methods was performed previously [8]. We plan to compare GPT with different NLP approaches in future research.

5. Conclusions

This study underscores the potential of combining retrieval and generative models to address the challenges posed by the variability and complexity of language in clinical notes. By focusing on the most relevant parts of the notes, RAG not only streamlined the workflow but also improved the accuracy and relevance of the responses generated by GPT. This process exemplifies how RAG's selective retrieval of data sharpens the focus of queries, reducing computational load and refining the response quality in clinical assessments. The proposed pipeline not only improves the reliability of automated text analysis in medical contexts but also offers a scalable approach for enhancing electronic health record systems through seamlessly integrating into existing EHR systems, where it can function in the background, analyzing incoming patient notes and updating patient profiles with relevant information. Moreover, with the pipeline's ability to accurately identify and categorize substance use, clinicians can make more accurate diagnoses and treatment decisions. Understanding the extent of tobacco and alcohol use can influence the management of various conditions like cardiovascular diseases or pulmonary issues. This leads to more personalized care plans that are directly tailored to the patient's specific needs, ultimately improving health outcomes.

References

[1] Trenz RC, et al. Early onset of drug and polysubstance use as predictors of injection drug use among adult drug users. Addict Behav 2012; 37 (4): 367–72. doi: 10.1016/j.addbeh.2011.11.011
[2] Levy SJL, Williams JF; Committee on Substance use and Prevention. Substance use screening, brief intervention, and referral to treatment. Pediatrics 2016; 138 (1): e20161211.
[3] Ghitza UE, et al. Common data elements for substance use disorders in electronic health records: The NIDA clinical trials network experience. Addiction 2013; doi: 10.1111/j.1360-0443.2012.03876.x
[4] Wu LT, Payne EH, Roseman K, et al. Clinical workflow and substance use screening, brief intervention, and referral to treatment data in the electronic health records: a national drug abuse treatment clinical trials network study. EGEMS (Wash DC) 2019; 7 (1): 35. doi: 10.5334/egems.293
[5] Levy S, et al. Screening adolescents for alcohol use. J Addict Med 2017; 11 (6): 427–34.
[6] Poulsen MN, et al. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. Frontiers in Public Health. 2022;10:850619.
[7] Patra BG, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. JAMIA. 2021;28(12):2716-27. doi: 10.1093/jamia/ocab170
[8] Shah-Mohammadi F, Cui W, Finkelstein J. Comparison of ACM and CLAMP for Entity Extraction in Clinical Notes. Annu Int Conf IEEE Eng Med Biol Soc. 2021 Nov;2021:1989-1992.