International Journal of
Environmental Research and Public Health

MDPI

Type of the Paper (Article, Review, Communication, etc.)

Effective Natural Language Processing Algorithms for Gout Flare Early Alert from Chief Complaints

Lucas Lopes Oliveira [1+], Aryalakshmi Nellippillipathil Babu [2+], Xiaorui Jiang [3*], Poonam Karajagi [4], and Alireza Daneshkhan [5]

[1] School of Computing, Mathematics and Data Sciences, Coventry University; lopesoll@uni.coventry.ac.uk
[2] School of Computing, Mathematics and Data Sciences, Coventry University; nellippila@uni.coventry.ac.uk
[3] Centre for Computational Sciences and Mathematical Modelling , Coventry University; xiaorui.jiang@coventry.ac.uk
[4] School of Computing, Mathematics and Data Sciences, Coventry University; karajagip@uni.coventry.ac.uk
[5] School of Computing, Mathematics and Data Sciences, Coventry University; alireza.daneshkhan@coventry.ac.uk
* Correspondence: xiaorui.jiang@coventry.ac.uk
\+ Equal contributions.

Abstract: In this study, we extend the exploration of gout flare detection initiated by Osborne, J. D. et al, through the utilization of their dataset of Emergency Department (ED) triage nurse chief complaint notes. Addressing the challenge of identifying gout flares prospectively during an ED visit, where documentation is typically minimal, our research focuses on employing alternative Natural Language Processing (NLP) techniques to enhance the detection accuracy. This study investigates the application of medical domain-specific Large Language Models (LLMs), distinguishing between generative and discriminative models. Models such as BioGPT, RoBERTa-large-PubMed-M3, and BioElectra were implemented to compare their efficacy with the original implementation by Osborne, J. D. et al. The best model was Roberta-large-PM-M3 with a 0.8 F1 Score on the Gout-CC-2019 dataset followed by BioElectra with 0.76 F1 Score. We concluded that discriminative LLMs performed better for this classification task compared to generative LLMs. However, a combination of using generative models as feature extractors and employing SVM for the classification of embeddings yielded promising results comparable to those obtained with discriminative models. Nevertheless, all our implementations surpassed the results obtained in the original publication.

Keywords: keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article yet reasonably common within the subject discipline.)

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

Gout affects over 9 million Americans[1,2] and is the most common form of inflammatory arthritis in men with a prevalence rate over 5%. The U.S. National Emergency Department Sample (NEDS) documents more than 200,000 annual visits where gout is identified as the primary diagnosis, constituting 0.2% of all Emergency Department visits and resulting in annual billable charges exceeding $280 million [3].

Despite strides in natural language processing (NLP) techniques for detecting gout flares from textual data, the prospective identification of such instances remains a complex task, especially within the constraints of Emergency Department (ED) environments. This study addresses this critical gap by advancing the methodologies proposed by Osborne, J. D. et al [1].

The importance of this research lies in the need to improve the continuity of care for gout patients, especially after an ED visit. Often, gout flares treated in the ED lack optimal follow-up care, necessitating the development of methods for identifying and referring patients with gout flares during an ED visit [1]. While retrospective studies have leveraged NLP for gout flare detection, the prospective identification of patients in real-time ED settings presents a unique challenge. The study builds upon the groundwork laid by Osborne, J. D. et al [1], who annotated a corpus of ED triage nurse chief complaint notes for gout flares, paving the way for our exploration.

1.1 Rationale for Using Large Language Models

Large language models, such as BERT (Bidirectional Encoder Representations from Transformers), GPT-3 (Generative Pre-trained Transformer 3), and their variants, have demonstrated remarkable success in a wide range of natural language processing tasks. The use of large language models in text classification offers several compelling reasons:

Contextual Understanding: Large language models leverage deep learning techniques to encode contextual information and relationships between words in a sentence. This contextual understanding allows them to capture subtle nuances and semantics, which is especially relevant in the medical domain where precise interpretation of clinical text is vital.

Transfer Learning: Pre-training on vast corpora of textual data enables large language models to learn general language patterns. This pre-trained knowledge can be fine-tuned on domain-specific datasets, making them adaptable and effective for text classification tasks in the medical field with relatively limited labelled data.

1.2 Natural Language Processing and Large Language Models in Healthcare

In recent years, the domain of healthcare has witnessed a revolutionary transformation due to the rapid advancements in Natural Language Processing (NLP) and the emergence of Large Language Models (LLMs). These technologies have the potential to revolutionize the healthcare industry by enhancing medical decision-making, patient care, and biomedical research.

Some tasks in NLP could be automated using LLM such as text classification [4, 5, 6], keyword Extraction [7, 8, 9], machine translation [10], and text summarization [11]. Furthermore, NLP and LLM can assist in the early detection and diagnosis of diseases by sifting through vast datasets to identify patterns, symptoms, and risk factors.

1.3 Gaps and Limitations of Current Literature

Insufficient Comparative Studies Between Domain-Specific Generative LLMs and Discriminative LLMs

While some studies have compared a single generative LLM (GPT) with discriminative LLMs, a comprehensive comparison between multiple domain-specific generative LLMs and discriminative LLMs for medical intent classification is lacking. Such comparisons are essential to determine the performance disparities between different LLM types and guide the selection of the most suitable model for our specific medical intent classification task.

In light of these gaps, our research aims to bridge these deficiencies in the current literature. We specifically focus on intent classification of medical letters by leveraging domain-specific generative LLMs as feature extractors. Additionally, our study includes comparative analyses of multiple domain specific generative LLMs and discriminative LLMs to gain comprehensive insights into their performance on this particular medical classification task. By addressing these gaps, we hope to contribute novel findings and enrich the existing literature.

In the current research landscape, the use of Large Language Models (LLMs) in the medical domain has demonstrated remarkable success. LLMs, such as Roberta-large-Pm-m3-voc, BioElectra, and BioBart, have shown promise in their ability to comprehend and process medical text […]. The integration of these advanced models in gout flare detection within chief nurse complaints presents an exciting avenue for exploration. Furthermore, the study introduces a novel approach of using LLMs for feature extraction, followed by classification with a Support Vector Machine (SVM), contributing to the evolving methodologies in this field […].

2. Materials and Methods

Data Collection

We utilized the dataset curated by Osborne, J. D. et al, consisting of Emergency Department (ED) triage nurse chief complaint notes. This dataset, annotated for the presence of gout flares, served as the foundation for our investigation. Each Chief Complaint (CC) in the dataset was annotated to determine its indication of a gout flare, non-indication of a gout flare, or remained unknown in terms of gout flare status. Following this, a manual chart review was conducted by a rheumatologist (MID) and a post-doctoral fellow (GR) to ascertain the gout flare status for 197 out of the 300 Emergency Department (ED) encounters. The following table, extracted from the publication by Osborne, J. D. et al., illustrates the data structure.

Table 1: GOUT-CC-2019-CORPUS Examples (Osborne, J. D. et al,)

Chief Complaint Text	Predicted*	Actual**
AMS, lethargy, increasing generalized weakness over 2 weeks. Hx: ESRD on hemodialysis at home, HTN, DM, gout, neuropathy	No	No
I started breathing hard" hx-htn, gout, anxiety,	No	No
R knee pain x 8 years. pmh: gout, arthritis	Unknown	No
Doc N Box DX pt w/ R hip FX on sat. Pt states no falls or injuries. PMH: gout	Unknown	No
out of gout medicine	Yes	Yes
sent from boarding home for increase BP and bilateral knee pain for 1 week. Hx of HTN, gout.	Yes	Yes

*Consensus predicted gout flare status determined by annotator examination of CC
**Gout flare status determined by chart review.

Large Language Models

In this study, we harnessed the power of Large Language Models (LLMs) and transfer learning for the task of gout flare detection within Emergency Department (ED) triage nurse chief complaint notes. LLMs are state-of-the-art natural language processing models, designed to comprehend and generate human-like text trained on vast amounts of pre-existing linguistic data.

Model Selection

We employed several LLMs tailored for the medical domain, for their ability to capture intricate patterns within medical text, making them well-suited for discerning nuances in chief complaints related to gout flares.

Discriminative models

In the domain of discriminative Large Language Models (LLMs), we strategically incorporated robust models renowned for their discriminative prowess—Roberta-PM-M3-Voc and BioElectra.

Model	Roberta-PM-M3-Voc	BioElectra	BioBart
Model Size	355M Parameters	---	139M Parameters

Hidden Size	1024	768	768
Model Size	24 Layers, 16 attention heads	12 Layers, 12 attention heads	12 Layers, 12 attention heads
Base Model	RoBERTa-large	Electra Base	Bart Base
Training Data	PubMed and MIMIC-III corpora	PubMed and PubMed central (millions of articles)	PubMed abstracts, PMC articles

Generative models

In the realm of generative Large Language Models (LLMs), we strategically chose BioGPT, BioMedLM, and PMC_LLaMA_7B for their renowned scale and exceptional performance in natural language processing tasks. These models represent the forefront of generative language understanding, and their comprehensive specifications, training data, and architectural features are elucidated below.

Table 2: Description of Generative LLMs implemented

Model	**BioGPT**	**BioMedLM**	**PMC_LLaMA_7B**
Model Size	347M Parameters	2.7B Parameters	7B Parameters
Hidden Size	1024	2560	4096
Model Size	24 Layers, 16 attention heads	32 Layers, 20 attention heads	32 Layers, 32 attention heads
Base Model	GPT2-medium	GPT2	LLaMA_7B
Training Data	15M PubMed abstracts from scratch	All the PubMed abstracts and full documents from The Pile.	4.8 million Biomedical academic papers from the S2ORC dataset.

Benchmark methods

To facilitate a comprehensive benchmarking analysis, we incorporated benchmark methods for comparison with our Large Language Models (LLMs). The benchmark methods involved the transformation of textual data into numerical vectors, a crucial step for machine learning algorithms that inherently require numerical input.

Textual Data Transformation:

Given that machine learning algorithms cannot interpret textual data directly, we employed Sklearn's 'TfidfVectorizer' algorithm to translate textual information into numerical vectors. This algorithm transforms documents into a matrix of tf-idf (term frequency-inverse document frequency) characteristics, capturing the significance of words within the corpus.

N-gram Exploration:

The tf-idf vectorizer, in its default setting, considers single-word tokens (unigrams) from sentences. In our research, we expanded this exploration by incorporating and evaluating various n-gram combinations of words. N-grams represent consecutive sequences of n words in a sentence. After experimentation, we opted for the (1, 2) ngram setting, utilizing both unigrams and bigrams to capture a more comprehensive contextual understanding of chief nurse complaints.

Performance Evaluation

The performance of each model was evaluated using standard metrics, including precision, recall, and Macro F1-score. We compared our results with the original algorithm proposed by Osborne et al., ensuring a comprehensive assessment of the advancements achieved.

3. Results

In this section, we meticulously analyze and compare the performance of three distinct models—Roberta-large-Pm-m3-voc, BioElectra, and BioBart—on two separate datasets: Gout-cc-2019 and Gout-cc-2020. The comprehensive assessment involves a thorough examination of overall recall and F1-score metrics, providing insights into the models' respective capabilities in capturing and identifying instances of gout flares within chief nurse complaints.

3.1. *Direct LLMs Classification*

This subcategory encompasses results obtained by directly employing Large Language Models (LLMs) for the classification of Chief Complaints (CCs). Analyze and present the performance metrics, such as recall and F1-score, achieved by each LLM (Roberta-large-Pm-m3-voc, BioElectra, and BioBart, BioGPT, BioMedLM) when used independently for gout flare prediction within CCs.

Table 3: Direct LLM Classification

	Gout-CC-2019			**Gout-CC-2020**		
Model	Precision	Recall	F1-score	Precision	Recall	F1-score
Roberta-large-PM-M3	0.80	0.79	0.80	0.62	0.72	0.63
BioElectra	0.76	0.76	0.76	0.63	0.68	0.65
BioBart	0.74	0.73	0.73	0.65	0.70	0.67
BioGPT	0.62	0.59	0.60	0.45	0.50	0.48
BioMedLM	0.49	0.49	0.47	0.51	0.52	0.52

3.2. LLMs *Embedding Extraction and Classification with SVM*

In this subcategory, we explore the outcomes derived from using LLMs to extract embeddings from Chief Complaints, followed by a secondary classification using a Support Vector Machine (SVM).

Table 4: LLMs Embedding Extraction and Classification with SVM

	Gout-CC-2019			**Gout-CC-2020**		
Algorithm	Precision	Recall	F1-score	Precision	Recall	F1-score
SVM with BioGPT Embeddings	0.79	0.79	0.75	0.69	0.73	0.71*
SVM with BioMedLM Embeddings	0.70	0.72	0.70	0.59	0.70	0.61

SVM with PMC_LLaMA_7B Embeddings	0.64	0.64	0.64	0.60	0.60	0.60

3.3. *Benchmark Methods*

This subcategory involves benchmarking the performance of traditional methods for textual data transformation, specifically focusing on the Tf-idf vectorizer with different n-gram settings. Contrast and compare the results obtained with these benchmark methods against the outcomes achieved by the LLMs, providing valuable insights into the effectiveness of each approach for gout flare prediction. In this section we have also included the results from the original publication (shaded), the results will be discussed further in the discussion section.

	Gout-CC-2019			**Gout-CC-2020**		
Algorithm	Precision	Recall	F1-score	Precision	Recall	F1-score
SVM with TF-IDF	0.75	0.75	0.75	0.82	0.74	0.77
NAIVE-GF	0.23	1.00	0.38	0.28	0.56	0.37
SIMPLE-GF	0.44	0.84	0.58	0.37	0.40	0.38
BERT-GF	0.71	0.48	0.56	0.79	0.47	0.57

4. Discussion

4.1. *General Analysis*

The results on the GOUT-CC-2019-CORPUS and GOUT-CC-2020-CORPUS datasets were unsatisfactory in relation to machine learning standards. The highest performance on these datasets was the SVM with BioGPT embeddings and oversampled data on the merge of the datasets with 70% accuracy but after further analysis of the results its clear the model is not able to predict the positive label as well as the negative label, and the high results of the negative class indicate a bias of the model towards the negative class, even after oversampling.

None of the models employed in this study were able to accurately make predictions of the GOUT-CC2019-CORPUS and GOUT-CC-2020-CORPUS datasets. The unsatisfactory results are related to the nature of the dataset. All the chief nurse complaints contain the keyword “gout” and most of the nurse complaints did not contain any clear indicator of gout flare. This is proven by the analysis of the predict column, where the professional annotators attempted to predict the presence of GOUT flare bases solely on the complaint. In the test set used more than half were miss classified by the professional rheumatologists.

4.2. *Comparative Analysis*

The following table compares the results acquired from this study, with the results obtained from the paper by Osborne et al.

	Gout-CC-2019			**Gout-CC-2020**		
Algorithm	Precision	Recall	F1-score	Precision	Recall	F1-score
Roberta-large-PM-M3	0.80	0.79	0.80*	0.62	0.72	0.63
BioElectra	0.76	0.76	0.76	0.63	0.68	0.65
BioBart	0.74	0.73	0.73	0.65	0.70	0.67
BioGPT	0.62	0.59	0.60	0.45	0.50	0.48

BioMedLM	0.49	0.49	0.47	0.51	0.52	0.52
SVM with BioGPT Embeddings	0.79	0.79	0.75	0.69	0.73	0.71*
SVM with Bio-MedLM Embeddings	0.70	0.72	0.70	0.59	0.70	0.61
SVM with PMC_LLaMA_7B Em-beddings	0.64	0.64	0.64	0.60	0.60	0.60
SVM with TF-IDF	0.75	0.75	0.75	0.82	0.74	0.77*
NAIVE-GF	0.23	1.00	0.38	0.28	0.56	0.37
SIMPLE-GF	0.44	0.84	0.58	0.37	0.40	0.38
BERT-GF	0.71	0.48	0.56	0.79	0.47	0.57

As shown in the above table Roberta was the best performing model on the GOUT-CC-2019-CORPUS dataset followed by BioElectra, showcasing the superiority of discriminative LLMs in classification tasks. The SVM with BioGPT embedding and TF-IDF also performed well in relation to the other models. In the GOUT-CC-2020-CORPUS dataset the best LLM was SVM with BioGPT embeddings which outperformed all the discriminative LLMs due to the use of oversampling, which was not possible using the discriminative LLMs. This result was still outperformed by SVM with TF-IDF features. All of our models outperformed the models used in the study by Osborne et al.(in grey) in both datasets.

4.3. *Future Directions*

Some improvements can be done to enhance the results obtained in this research:

Full Fine-Tuning and Distributed Computing: While parameter-efficient fine-tuning, specifically LoRA, was applied in this study due to hardware constraints and the models' size, pursuing full fine-tuning would enhance the results of the models. Implementing distributed computing is necessary to apply full fine tuning, due to the very large size of the models this process requires distributing the model load across different GPUs to perform the calculations. This strategy would enable more comprehensive fine-tuning, potentially leading to an increase in model performance.

Enhanced Dataset Quality and Size: with such a limited number of samples the model cannot be properly trained, validated and tested. To address this more samples must be acquired or whole new datasets to test the models effectively.

Exploring Embeddings and Discriminative LLMs: A new direction to follow would be a similar approach to the one employed in this study where the embeddings of the discriminative LLMs are extracted and used for classification using a separate classifier, in order to test the different embeddings side by side in a similar setting.

Ensemble Learning for Enhanced Embeddings: A promising route is the utilization of deep learning models to create an ensemble that enhances embeddings before their application in text classification. This strategy could potentially enhance the information captured by the embeddings, thereby leading to improved classification outcomes.

5. Conclusions

Overall this study highlighted the potential of generative LLMs for classification tasks, achieving results comparable to the discriminative models. Additionally the models also have shown potential as feature extractors for classification tasks even without fine tuning, due to their ability to understand contextual information and produce contextual rich embeddings. Despite the results between the two types of models being comparable, the computational requirements to perform the same task is much greater using the generative LLMs employed in this study. Similar or superior results can be obtained using much smaller discriminative models.

Still, this research highlights the importance of using the domain specific variants of the models when the text contains specialized and out of word vocabulary.

Given the considerations mentioned above, the following conclusions can be drawn. The integrations of Large language models trained on medical publications holds potential to reshape classification tasks in the medical domain for the future.

References

1. Osborne JD, Booth JS, O'Leary T, Mudano A, Rosas G, Foster PJ, Saag KG, Danila MI. Identification of Gout Flares in Chief Complaint Text Using Natural Language Processing. AMIA Annu Symp Proc. 2021 Jan 25;2020:973-982. PMID: 33936473; PMCID: PMC8075438.
2. Michael Chen-Xu, Chio Yokose, Rai Sharan K, Pillinger Michael H, Choi Hyon K. Contemporary prevalence of gout and hyperuricemia in the united states and decadal trends: the national health and nutrition examination survey, 2007–2016. Arthritis & Rheumatology. 2019;71(6):991–999.
3. Singh Jasvinder A, Shaohua Yu. Time trends, predictors, and outcome of emergency department use for gout: a nationwide us study. The Journal of rheumatology. 2016;43(8):1581–1588.
4. Xu, B., Gil-Jardiné, C., Thiessard, F., Tellier, E., Avalos, M., & Lagarde, E. (2019). Pre-training A Neural Language Model Improves The Sample Efficiency of an Emergency Room Classification Model.10.48550/arxiv.1909.01136
5. Veladas, R. et al. (2021). Aiding Clinical Triage with Text Classification. In: Marreiros, G., Melo, F.S., Lau, N., Lopes Cardoso, H., Reis, L.P. (eds) Progress in Artificial Intelligence. EPIA 2021. Lecture Notes in Computer Science(), vol 12981. Springer, Cham. https://doi.org/10.1007/978-3-030-86230-5_7
6. Noor, K., Smith, K., Bennett, J., OConnell, J., Fisk, J., Hunt, M., Philippo, G., Xu, T., Knight, S., Romao, L., Dobson, R. J., & Wong, W. K. (2022). Predicting Clinical Intent from Free Text Electronic Health Records.10.48550/arxiv.2204.09594
7. Ding,L.,Zhang,Z.,Liu,H.,Li,J. & Yu,G.(3921).Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling. Journal of Data and Information Science,6(3) 35-57. https://doi.org/10.2478/jdis-2021-0013
8. Ke, H., Lee, C. S., & Sugiyama, K. (2021). Bert-Based Chinese Medical Keyphrase Extraction Model Enhanced with External Features. Towards Open and Trustworthy Digital Societies (pp. 167-176). Springer International Publishing AG. 10.1007/978-3-030-91669-5_14
9. Tang, M., Gandhi, P., Kabir, M. A., Zou, C., Blakey, J., & Luo, X. (2019). Progress Notes Classification and Keyword Extraction using Attention-based Deep Learning Models with BERT.10.48550/arxiv.1910.05786
10. Han, L., Erofeev, G., Sorokina, I., Gladkoff, S., & Nenadic, G. (2022). Investigating Massive Multilingual Pre-Trained Machine Translation Models for Clinical Domain via Transfer Learning.10.48550/arxiv.2210.06068
11. Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., ... & Peng, Y. (2023). Evaluating Large Language Models on Medical Evidence Summarization. medRxiv, 2023-04.