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 Daneshkhah [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, and Centre for Computational Sciences and Mathematical Modelling, Coventry University; alireza.daneshkhan@coventry.ac.uk
* Correspondence: xiaorui.jiang@coventry.ac.uk
\+ Equal contributions.

Abstract: Early identification of acute gout is crucial, enabling healthcare professionals to implement targeted interventions for rapid pain relief and preventing disease progression, ensuring improved long-term joint function. In this study, we comprehensively explored the potential of gout flare early detection based on nurse chief complaint notes in the Emergency Department (ED). 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. We investigate gout flare detection algorithms using both sparse representations by traditional NLP methods and dense encodings by medical domain-specific Large Language Models (LLMs), distinguishing between generative and discriminative models. Three methods are used to alleviate the issue of severe data imbalance, including oversampling, class weights, and focal loss. Extensive empirical studies are done on the Gout Emergency Department Chief Complaint Corpora. Sparse text representations like tf-idf proved to produce strong performance, achieving higher than 0.75 F1 Score. The best deep learning models are Roberta-large-PM-M3 and BioElectra, with a 0.8 and a 0.76 F1 Score respectively, on the 2019 dataset. On the 2020 dataset, the best performance achieved was 0.77 F1 by support vector machine trained with *n-gram* features. We concluded that although discriminative LLMs performed better for this classification task compared to generative LLMs, a combination of using generative models as feature extractors and employing SVM for classification yields promising results comparable to those obtained with discriminative models.

Keywords: Gout Flare; Natural Language Processing; Deep Learning; Large Language Models

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

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

Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Gout affects over 9 million Americans [1] 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 (ED) visits and resulting in annual billable charges exceeding $280 million [2]. It is important 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 [3].

While retrospective studies have leveraged NLP for gout flare detection, the prospective identification of patients in real-time ED settings presents a unique challenge, especially within the constraints of Emergency Department (ED) environments.

Despite of the success of natural language processing (NLP) techniques in healthcare [4], NLP-based Gout Flare Early Detection (DFED) is in severe lack of study. Only a few were identified, like Zheng et al [5], which however worked on Electronic Medical Records. The problem of early warning of acute gout flare becomes more challenging in the ED setting where only chief complaints of patients are taken by nurses in an extremely succinct format. It is of paramount challenge to develop an effective DFED algorithm using such limited amount of information. The current study tries to address this critical gap by advancing the methodologies proposed by Osborne et al [3]. Our study builds upon the groundwork laid by Osborne et al., who annotated two corpora of ED chief complaint notes for gout flares and paves the way for our exploration of effective text representation methods and state-of-the-art medical/clinical Large Language Models (LLM).

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

2.1 Data Collection

We utilized the dataset of ED chief complaint notes which were annotated by Osborne et al. for the presence of gout flares [3]. 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 (GF) status. Following this, a manual chart review was conducted by a rheumatologist and a post-doctoral fellow to ascertain the GF status for a small portion of the ED counters. These were served as the gold standard annotations of the real GF status. The corpora contain two datasets for the year 2019 and 202, namely GOUT-CC-2019-CORPUS and GOUT-CC-2020-CORPUS respectively. Table 1 shows the annotation statistics of the two datasets (from Osborne et al. [3]), while Table 2 illustrates some examples. In out experiments, we used the human-annotated samples using Chart Review, as what Osborne et al. did.

Table 1: Annotation Statistics of the Gout Flare Chief Complaint Datasets (Osborne et al. [3])

Dataset Name	GF-POS (Positive)	GF-NEG (Negative)	GF-UNK (Unknown)	Review	Agreement	Cohen's κ
GOUT-CC-2019-CORPUS	93	194	13	CC	0.883	0.825
GOUT-CC-2019-CORPUS*	70	118	9	Chart	0.849	0.774
GOUT-CC-2020-CORPUS	14	7992	129	CC	0.977	0.965
GOUT-CC-2020-CORPUS*	25	232	7	Chart	0.904	0.856

* Used for experiments as Osborne et al. [3]

Table 2: Examples of Chief Complaint Notes for Gout Flare (Osborne et al. [3])

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.

2.2 Feature Extraction

In the feature engineering approach, we extracted the *n*-grams ($n = 1, 2, 3$) and tested different combinations of *n*-grams and different feature sizes. CC texts were converted into sparse representations using *tf-idf* (Term Frequency-Inverse Document Frequency) [12] as initial feature values. A linear support vector classifier (Linear SVC) was trained. All implementations were done using the scikit-learn library[1].

It was hard to extract more advanced syntactic or semantic features due to the noisiness of CC texts. As can be observed from Table 2, CC texts are extremely succinct, often containing a sequence of medical terms or abbreviations, which record the facts reported by patients. Such CCs are not meaningful sentences for us to extract features from the syntactic analysis results. Semantic analysis tools are either immature or nonexistent in this particular area. However, we could still observe quite good performances from fine-tuning a machine learning model using the right sparse feature representation of CC texts.

2.3 Large Language Models

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 GF. All LLMs belong to the Transformers family [13] because we hoped that the multi-headed self-attention mechanism of the Transformers architecture could be able to learn the meaningful association between certain words of CC texts to indicate the existence of GF.

2.3.1 Discriminative models

We strategically incorporated three robust discriminative LLMs renowned for their discriminative prowess—Roberta-PM-M3-Vo[2], BioElectra[3] [14], and BioBART[4] [15]. These are the domain-specific versions of the RoBERTa [16], Electra [17] and BART [18] models respectively. Although BART was a language model pretrained in a sequence-to-sequence fashion, it can be used equally well and in the same way as a discriminative model [18]. So, we treated it as one representative of the discriminative category. The details of the discriminative LLMs are shown in Table 3.

2.3.2 Generative models

In the realm of generative LLMs, we strategically chose BioGPT[5] [20], BioMedLM[6], and PMC_LLaMA_7B[7] [21] for their renowned scale and exceptional performance in natural language processing tasks. BioGPT and PMC_LLaMA_7B are the domain-specific versions of the GPT-2 [22] and LLaMA [23-24] models respectively, while BioMedLM is a bespoke LLM pretrained for medical applications. These models represent the forefront of generative language understanding, and their comprehensive specifications, training data, and architectural features are elucidated in Table 4.

[1] https://scikit-learn.org/

[2] https://huggingface.co/Sedigh/RoBERTa-large-PM-M3-Voc

[3] https://github.com/kamalkraj/BioELECTRA

[4] https://github.com/GanjinZero/BioBART

[5] https://huggingface.co/docs/transformers/model_doc/biogpt

[6] https://github.com/stanford-crfm/BioMedLM

[7] https://github.com/chaoyi-wu/PMC-LLaMA

Table 3: Description of Discriminative LLMs Implemented

Model	Roberta-PM-M3-Voc	BioElectra	BioBART
Model Size	355M Parameters	---	139M Parameters
Hidden Size	1024	768	768
Model Size	24 Layers, 16 heads	12 Layers, 12 heads	12 Layers, 12 heads
Base Model	RoBERTa-large	Electra Base	BART Base
Training Data	PubMed articles and MIMIC-III corpora[8] [19]	PubMed articles	PubMed abstracts and articles

Table 4: 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 heads	32 Layers, 20 heads	32 Layers, 32 heads
Base Model	GPT2-medium	GPT2	LLaMA_7B
Training Data	15M PubMed abstracts from scratch	All PubMed abstracts and full texts from The Pile benchmark [25].	4.8 million Biomedical publications from the S2ORC dataset [26].

2.4 Fine-tuning

Fine tuning was implemented to improve the models' ability to understand and capture the nuances in the texts. For the discriminative models full fine tuning was implemented, but for the generative models due to the size of the models and hardware constraints full fine tuning was not possible.

2.4.1 Fine-tuning of Discriminative LLMs

All three discriminative LLMs use a bidirectional encoder as BERT [27]. The encoder part of these models was used to encode each CC text, and the "[CLS]" token was used as the dense representation. For Roberta-PM-M3-Voc and BioElectra, a further feature transformation was applied. Essentially, the classification head was a Multiple Layer Perceptron (MLP), the hidden layer of which made a nonlinear transformation (of the same size). On the contrary, BioBART used a linear classification head following the tradition of BART usage.

2.4.2 Fine-tuning of Generative LLMs

Similarly, generative LLMs were used for encoding CC texts, and the "Extract" token (for all three models as they all belong to the GPT family) were used to extract the dense representation, which was then sent to a linear classification head. Due to their large sizes, the generative LLMs were not fully fine-tuned. Instead, we used LoRA (Low Rank Adaptation) to efficiently adapt LLMs to specific tasks by only modifying a small portion of the whole parameter space.

The main idea behind LoRA is to exploit the low-rank structure of the model's weight matrices during task adaptation, resulting in reduced memory usage and computational complexity [28]. The idea was inspired by Aghajanyan et al.'s finding that pre-trained language models have a low "intrinsic dimension" meaning that they can still lean efficiently even when their weight matrices are randomly projected to a smaller subspace [29].

[8] https://www.nature.com/articles/sdata201635

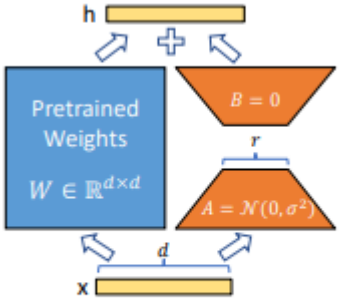

Figure 1: Parametrization of LoRA. Only A and B are trained. [12]

More precisely, LoRA hypothesizes that updates to model's weight matrix, W_0, can be represented by a low-rank decomposition, which is given by $W_0 + \Delta W = W_0 + BA$, where $B \in R^{d \times r}, A \in R^{r \times k}$, and $\Delta W = BA$ represents weight updates. During training (i.e., fine-tuning), W_0 is frozen while A and B contain the trainable parameters.

In our fine-tuning process, we applied the following LoRA parameters:

- The rank (r) of A and B was set to 8.
- The LoRA regularization coefficient α was set to 16.
- To prevent overfitting and enhancing model generalisation, we applied a LoRA dropout rate of 0.1.
- A learning rate of 3e-4 was used, enabling efficient convergence during training.

2.5 Classification

In the feature engineering approach, a Linear SVC was trained. When finetuning discriminative LLMs, either an MLP or a linear classifier was applied. Similarly, a linear layer was used for classification with generative LLMs. In the experiments, we also tested using generative LLMs only as the feature extractor and trained a Linear SVC for classification. In this alternative approach, which required significantly less computational resources, generative LLMs were frozen, used to encode CC texts, and the hidden states of the "Extract" token were extracted as dense representation. A Linear SVC was then trained in the similar way as in the feature engineering approach. This was to demonstrate LLMs' native ability to understand and represent medical texts for the downstream task.

2.6 Optimisation

2.6.1 Model Selection

To report the optimized performance in a more robust way, for all methods, we used 5-fold cross-validation (Seed 42 was used across all experiments) for model selection, i.e., determining hyperparameters for both Linear SVC and deep learning methods. Note that for the purpose of increasing recall, we merged the GF-UNK class into GF-POS. The final performance was reported using the mean of the performances of the five folds.

2.6.2 Handling Data Imbalance

We also observed severe data balance in the corpora. The data imbalance ratio of GOUT-CC-2019 is (70 + 9) / 118 = 0.6695, while the imbalance ratio of GOUT-CC-2020 is (25 + 7) / 232 = 0.1379. One method used to handle data imbalance was class weights, which were set according to the relative sizes of each class as in Eq. (1),

$$w_j = N/(K \times N_j), \quad (1)$$

where w_j is the weight for the j-th class, K is the total number of classes, N is the total number of samples, and N_j is the number of samples of the j-th class [30].

However, class weighting in Eq. (1) did not help improve the performances on GOUT-CC-2020 much, which is 5 times more imbalanced than GOUT-CC-2019. Although the discriminative LLMs performed strongly in our experiments, they were extremely sensitive to this severe data imbalance. Therefore, we performed random upper sampling on GOUT-CC-2020. More precisely, in each fold of the cross-validation, the positive samples in the training splits, including GF-POS and GF-UNK combined, were randomly duplicated to match the size of GF-NEG.

2.6.3 Focal Loss

In the context of our classification tasks, the choice of a suitable loss function plays a pivotal role in training and optimizing our models. We employed two distinct loss functions as per dataset and model requirement, namely cross-entropy loss and focal loss [12], to effectively guide the training process and address specific challenges posed by our datasets.

In instances where class imbalance persisted even after oversampling the training data, such as in the case of GOUT-CC-2020, we employed focal loss as an alternative to cross-entropy to combat class imbalance in classification tasks, as in Eq. (2).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{2}$$

where p_t is the posterior probability of each target t (here $t = 0\ or\ 1$), $\alpha_t \in [0,1]$ is the scaling parameter, γ is the focusing parameter and $(1-p_t)^{\gamma}$ is the modulating factor of the original cross-entropy loss [31].

3. Results

In this section, we meticulously analyze and compare the performances of all methods. Appendix A shows the hyperparameters optimized for performance reporting here. 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. [3], ensuring a comprehensive assessment of the advancements achieved.

3.1 Fine-tuned LLM

This subcategory encompasses results obtained by directly employing LLMs for CC classification. Table 5 shows the results.

The table shows that Roberta-large-PM-M3 outperforms the other four models in terms of precision, recall, and F1-score for both datasets. This suggests that this model is more effective at detecting gout flares from clinical notes. Table 5 also shows that BioBERT and BioElectra have similar performance, while BioGPT and BioMedLM have the lowest performance among the five models.

Table 5: Performances of Gout Flare Detection using Fine-Tuned LLMs

	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 Frozen LLMs as Feature Extractors

In this subcategory, we used LLMs to embed CC texts to dense feature vectors and use Linear SVC for classification. Table 6 shows the results.

The table shows that SVM with BigGPT Embeddings has the best performance among the four algorithms on both datasets. It achieves an F1-score of 0.67 on Gout-CC-2019 and 0.71 on Gout-CC-2020. This indicates that this algorithm can effectively extract the relevant features from CC texts and classify them accurately.

The table also shows that SVM with BioMedLM Embeddings and SVM with PMC_Llama_7B Embeddings have similar performance, but lower than SVM with BigGPT Embeddings. They both have an F1-score of 0.66 on Gout-CC-2019 and 0.61 on Gout-CC-2020. This suggests that these algorithms are less robust and consistent in handling the variability and complexity of CC texts.

Table 6: Performances of Gout Flare Detection using LLM Embeddings

	Gout-CC-2019			**Gout-CC-2020**		
Algorithm	Precision	Recall	F1-score	Precision	Recall	F1-score
SVM with BioGPT Embeddings	0.68	**0.67**	**0.67**	0.69	**0.73**	**0.71**
SVM with BioMedLM Embeddings	0.69	0.66	0.66	0.59	0.70	0.61
SVM with PMC_LLaMA_7B Embeddings	0.66	0.66	0.66	0.60	0.60	0.60

3.3 Sparse Text Representation

This subcategory involves performance of the traditional feature engineering approach, which generated sparse text representations using tf-idf of *n*-gram features. Contrast and compare these results against the outcomes achieved by the LLMs, providing valuable insights into the effectiveness of each approach for GF prediction. In this section we have also included the results from the original publication of Osborne et al. [3], which are shaded. All results will be discussed further in the discussion section. Table 7 shows the results.

Table 7: Performances of Gout Flare Detection using Sparse Text Representations

	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.

Table 8: Comparing the Performances of All Gout Flare Detection Methods.

	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	0.68	**0.67**	**0.67**	0.69	0.73	0.71*
SVM with BioMedLM Embeddings	0.69	0.66	0.66	0.59	0.70	0.61
SVM with PMC_LLaMA_7B Embeddings	0.66	0.66	0.66	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. 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.
2. 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.
3. 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.
4. B. Zhou, G. Yang, Z. Shi and S. Ma, "Natural Language Processing for Smart Healthcare," in IEEE Reviews in Biomedical Engineering, 2022, doi: 10.1109/RBME.2022.3210270.
5. Chengyi Zheng, Nazia Rashid, Yi-Lin Wu, River Koblick, Lin Antony T, Levy Gerald D, Cheetham T Craig. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. Arthritis care & research. 2014;66(11):1740–1748.
6. 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

7. 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
8. 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
9. 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
10. 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
11. 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
12. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
13. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is All you Need. In *Proceedings of the 30-th Conference* on *Neural Information Processing Systems* (*NIPS'2017*).
14. Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA: Pretrained Biomedical text Encoder using Discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (*BioNLP'2021*), pages 143–154, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.bionlp-1.16
15. Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proceedings of the 21st Workshop on Biomedical Language Processing* (*BioNLP'2022*), pages 97–109, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bionlp-1.9
16. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692
17. Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proc. International Conference on Learning Representations* (*ICLR'2019*).
18. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (*ACL'2020*), pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703
19. Johnson, A., Pollard, T., Shen, L. et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035. https://doi.org/10.1038/sdata.2016.35
20. Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in Bioinformatics, Volume 23, Issue 6, November 2022, bbac409, https://doi.org/10.1093/bib/bbac409
21. Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. https://arxiv.org/abs/2304.14454
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
23. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971
24. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://arxiv.org/abs/2307.09288
25. Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. https://arxiv.org/abs/2101.00027
26. Lo, K., Wang, L.L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (*ACL'2020*), pages 4969–4983, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.447

27. 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
28. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/10.48550/arxiv.2106.09685
29. Aghajanyan, A., Zettlemoyer, L., & Gupta, S. (2020). Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. https://arxiv.org/abs/2012.13255
30. Singh, K. (2020, Oct 6). How to Improve Class Imbalance using Class Weights in Machine Learning? Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights.
31. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar. (2017). Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV' 2017*), pp. 2980-2988.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.