**NAME: ADEYEMI VICTOR GBADAMOSI**

**STUDENT ID:699795**

**SUPERVISOR: DR. CLAIRE CASHMORE, DR. EVA SOUSA**

**PROJECT PROPOSAL FOR THE FINETUNING OF PRETRAINED LANGUAGE MODELS FOR CLINICAL TAG EXTRACTION IN EVIDENCE BASED MEDICINE**

**ABSTRACT**

*The goal of this proposal is to expedite information retrieval from medical journals by investigating the use of pretrained language models (PLM) in evidence-based medicine. Using PLMs speeds up this process and shows how they can be applied to a range of medical-related tasks. Detailed steps, proposed timelines, and anticipated outcomes of the project are outlined, emphasizing the transformative potential of PLMs in streamlining information access within the medical field.*

***BACKGROUND AND OBJECTIVES***

***What is Evidence based medicine?***

Evidence-based medicine (EBM) stands as a systematic approach to transforming clinical challenges into well-structured questions, followed by the systematic identification, evaluation, and application of contemporary research findings as the foundation for informed clinical decisions(Rosenberg & Donald, 1995). Recently, EBM has gained a lot of popularity amongst clinicians and is considered the favourable approach to finding answers using clinical trials. Practicing EBM requires medical professionals to maintain an up-to-date awareness of pertinent developments and research within their field. Traditionally, this involves perusing medical journals and textbooks, a process that, while valuable, can be laborious and time intensive. So, many clinicians rely on other methods of making clinical decisions(Isaacs & Fitzgerald, 1999).

*Pretrained Language Models*

In recent years, pretrained language models has been the major building block for many natural language tasks which is well suited to aid clinical practitioner with EBM. A downstream use of NLP which is Named entity recognition involves using machine learning to identify pre-informed tags in a group of text. These tags could then be grouped into named entities or classified or extracted for other use. Specifically, PLM which has be trained on medical corpora which has shown acceptable accuracy in the field of Named Entity Recognition would be instrumental in extracting key words the clinical practitioner is interested in from a medical journal or textbooks.

*Our Approach*

Building on the work of researchers such as (Lei et al., 2014) and (Tian et al., 2021) who utilized PLMs for specific NER tasks, our research distinguishes itself through the creation of a bespoke dataset. We aim to fine-tune the model to recognize and categorize specific tags within the medical domain, deviating from existing studies in terms of both dataset composition and the selection of pretrained models. Our approach involves crafting a specialized dataset and fine-tuning the model to align with these tags. Subsequently, we will assess the performance of four biomedical pretrained models on this unique dataset, with the overarching objective of enhancing the tagging process for documents. This, in turn, aims to facilitate medical practitioners in efficiently accessing relevant information for their clinical tasks.

**RESEARCH DESCRIPTION**

Our primary objective is to develop an advanced model for medical tag extraction based on a biomedical pretrained language model. Presently, we employ an existing tool, SWT, which effectively extracts text from medical articles and journals. The tool, leveraging a database of predefined keywords, demonstrates a certain level of accuracy in extracting crucial information from text documents. However, its rule-based nature constrains its ability to generalize, lacking the adaptability and efficiency characteristic of pretrained large language models. To overcome these limitations, our project aims to create a refined model by leveraging a biomedical pretrained language model. This model will undergo fine-tuning, incorporating rule tags previously employed by SWT, to enhance its proficiency in tagging words within medical documents.

*Preview relevant materials*

Initiating our project, a critical phase involves an in-depth review of pertinent literature to enrich our understanding and elaborate on our proposed methodology. Our preliminary efforts have involved working with the Evidence-Based Medicine Patient, Intervention, Comparison, and Outcome (EBM PICO) dataset and a bespoke dataset utilizing BioElectra PLM (raj Kanakarajan et al., 2021). Furthermore, we plan to explore existing research on biomedical pretrained language models, evaluating their performance in downstream tasks. To ensure the selection of the most effective models for Named Entity Recognition (NER) tasks, a rigorous evaluation process will be employed. Models with the highest metrics in NER tasks will be identified and subsequently trained and fine-tuned in our project. This strategic approach not only aligns with best practices in model selection but also underscores our commitment to implementing state-of-the-art techniques in the development of an efficient and accurate medical tag extraction model.

*Data formatting and cleaning*

In the initial stages of our project, once the project methodology is established, our approach involves the utilization of the SWT tool to extract pertinent textual information from a corpus of medical articles. Employing a rule-based methodology inherent to the SWT tool, we then proceed to annotate the extracted data. This initial annotation serves as a crucial step in establishing a foundation for subsequent processing.

Subsequently, the acquired dataset undergoes a meticulous formatting process to align with the requirements of our named entity recognition (NER) task. For this purpose, we employ the Hugging Face framework in Python, leveraging the capabilities of PyTorch as a foundational element (Wolf et al., 2019). This framework proves instrumental in seamlessly integrating NER tags and tokens into the dataset, ensuring a standardized and machine-readable structure conducive to subsequent analysis.

Additionally, we adopt the IOB (Inside, Outside, Beginning) tagging format, a convention recommended by (Ramshaw & Marcus, 1999). This tagging schema facilitates the identification and categorization of entities within the dataset, enhancing the precision and interpretability of subsequent model training and evaluation processes.

This meticulous approach to data formatting and cleaning not only establishes a robust foundation for subsequent phases of our project but also aligns with best practices and established methodologies in the field of natural language processing and named entity recognition.

*Training, Evaluation and Finetuning*

The identified pretrained language models (PLMs) will be acquired from the Hugging Face model repository and subsequently fine-tuned using the designated dataset. In the pursuit of optimal model performance, we advocate for the adoption of an 80:10:10 ratio for partitioning the dataset into training, testing, and evaluation sets, respectively. The default hyperparameters provided by the Hugging Face training method will serve as our starting point. However, we plan to experiment with variations in learning rates and weight decay to achieve the most favorable convergence.

Informed by the work of (raj Kanakarajan et al., 2021) who have elucidated the best hyperparameters yielding optimal results for the BioElectra model, we intend to explore relevant literature for each selected PLM to discern the hyperparameter configurations that have proven efficacious in prior studies.

It is essential to acknowledge the computational demands associated with PLM training, necessitating access to robust computing resources. To address this, we propose to conduct our training procedures on the VIPER supercomputer, leveraging its high-end GPU capabilities for efficient model training.

The anticipated outcomes will be benchmarked against the base model of Bioelectra, serving as a comparative reference point to assess the performance enhancements achieved through our fine-tuning efforts. This comprehensive approach, encompassing both hyperparameter exploration and computational optimization, aims to yield insights into the nuanced dynamics of PLM performance within the specific context of medical document information extraction.

**TIMESCALES**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 18th october | 25th october | 10th november | 15th december | 10th january |
| Material Review and model verification |  |  |  |  |  |
| data preprocessing |  |  |  |  |  |
| training and finetuning |  |  |  |  |  |
| hyperparameter Optimization |  |  |  |  |  |
| Documentation and round up |  |  |  |  |  |

Table 1. Gantt Chart for the timeline of the project

**EXPECTED OUTCOMES**

This project holds the promise of equipping the developed named entity recognition models with the capability to discern specialized medical tags, facilitating seamless integration into document-based tools such as SWT. This integration empowers medical practitioners with a streamlined mechanism for sifting through extensive volumes of medical articles, enabling swift access to crucial information. Consequently, this optimization in information retrieval serves to liberate valuable time for medical professionals, redirecting their attention to more critical tasks, notably the imperative mission of saving lives.

Beyond the realm of document-based tools, the applications of medical named entity recognition models extend to diverse use cases. For instance, in the domain of clinical decision systems, these models offer a potent tool for efficiently filtering and organizing patients' information based on pre-identified tags. This functionality not only enhances the speed and accuracy of information processing but also contributes to the overall efficacy of clinical decision-making processes. In essence, the outcomes of this project not only pave the way for enhanced document analysis and information retrieval but also open avenues for transformative applications in healthcare, aligning with the broader objective of advancing medical practices and improving patient outcomes.

**BIBLIOGRAPHY**

Isaacs, D. & Fitzgerald, D. (1999) Seven alternatives to evidence based medicine. *Bmj*, 319(7225), 1618.

Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M. & Xu, H. (2014) A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5), 808-814.

raj Kanakarajan, K., Kundumani, B. & Sankarasubbu, M. (2021) BioELECTRA: pretrained biomedical text encoder using discriminators, *Proceedings of the 20th Workshop on Biomedical Language Processing*.

Ramshaw, L. A. & Marcus, M. P. (1999) Text chunking using transformation-based learning, *Natural language processing using very large corpora*Springer, 157-176.

Rosenberg, W. & Donald, A. (1995) Evidence based medicine: an approach to clinical problem-solving. *Bmj*, 310(6987), 1122-1126.

Tian, S., Erdengasileng, A., Yang, X., Guo, Y., Wu, Y., Zhang, J., Bian, J. & He, Z. (2021) Transformer-based named entity recognition for parsing clinical trial eligibility criteria, *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R. & Funtowicz, M. (2019) Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.