Development of an Extended Biomedical Named Entity Recognition and Relation Extraction Model for Malaria using BioBERT

AJAO, ADEMOLU DANIEL
(22PCG02416)
BSc Computer Science, Mountain Top University, Ogun State

AUGUST, 2024

Development of an Extended Biomedical Named Entity Recognition and Relation Extraction Model for Malaria using BioBERT

BY

AJAO, ADEMOLU DANIEL
(22PCG02416)
BSc Computer Science, Mountain Top University, Ogun State.

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENT OF THE MSc DEGREE IN COMPUTER SCIENCE (SOFTWARE ENGINEERING) OF THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY, OTA, OGUN STATE, NIGERIA

AUGUST, 2024

ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the degree of Master of Sciences in Computer Science in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, Nigeria.

Miss Adefunke F. Oyinloye
(Secretary, School of Postgraduate Studies) **Signature and Date**

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) **Signature and Date**

DECLARATION

I declare that **AJAO, ADEMOLU DANIEL (22PCG02416)**, carried out this research entitled **"DEVELOPMENT OF AN EXTENDED BIOMEDICAL NAMED ENTITY RECOGNITION AND RELATION EXTRACTION MODEL FOR MALARIA USING BIOBERT".** It was carried out under the supervision of Dr. Oluranti Jonathan, Concepts of this research project are results of the research carried out by Ajao, Ademolu Daniel and ideas of other researchers have been fully recognized.

AJAO, ADEMOLU DANIEL

Signature and Date

CERTIFICATION

This is to certify that this dissertation titled "**DEVELOPMENT OF AN EXTENDED BIOMEDICAL NAMED ENTITY RECOGNITION AND RELATION EXTRACTION MODEL FOR MALARIA USING BIOBERT"** is original research carried out by **AJAO, ADEMOLU DANIEL (22PCG02416)** in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota. Ogun State, Nigeria, under the supervision of Dr. Oluranti Jonathan. We have examined and found this work acceptable as part of the requirements for the award of Master of Science (M.Sc.) in Computer Science.

Dr. Oluranti Jonathan

(Supervisor) **Signature and Date**

Prof. Olufunke O. Oladipupo
(Head of Department) **Signature and Date**

Prof. Zacchaeus O. Omogbadegun
(External Examiner) **Signature and Date**

Prof. Akan B. Williams
(Dean, School of Postgraduate Studies) **Signature and Date**

DEDICATION

I dedicate this project to God Almighty for His Help and ever-sufficient Grace, Wisdom and Knowledge given to me throughout my Master's Degree Programme. My Family and Friends for their words of encouragement and support.

ACKNOWLEDGEMENTS

I use this medium to appreciate the following people/organizations that have supported me in one way or another during my study.

First and foremost, I would like to thank my project supervisor, Dr. Oluranti Jonathan, for his invaluable guidance, insightful feedback, and constant encouragement throughout the project.

Secondly, I want to thank the Head of the Department of Computer and Information Sciences, Prof. Olufunke Oladipupo, for giving me the unique privilege to pursue my interests and for her guidance.

Also, I would like to thank the faculty members for their feedback and advice and for giving expert opinions that helped shape me in the right direction.

I also want to thank CApiC-ACE for sponsoring, supporting and giving me the privilege to further my studies for this degree.

I also want to appreciate my parents, Mr Adefemi and Mrs Florence Ajao and my siblings for their support and encouragement that gives me the courage to carry on towards the journey of excellence

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Named Entity Recognition	NER
Relation Extraction	RE
Large Language Model	LLM
Biomedical Bidirectional Encoding Representations from Transformers	BioBERT
Natural Language Processing	NLP
Convolutional Neural Network	CNN
Kernelized Support Vector Machine	KSVM
Biomedical Information Extraction	BioIE
Biomedical Natural Language Processing	BioNLP
Biomedical Named Entity Recognition	BioNER
Long Short-Term Memory	LSTM
Random Forest	RF
Support Vector Machine	SVM
Gradient Boosting Machines	GBM
Bidirectional Encoder Representations from Transformers	BERT

ABSTRACT

This study aims to enhance the biomedical Named Entity Recognition and Relation Extraction model for use in the malaria subdomain by fine-tuning the existing BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) model. The process of fine-tuning involves adjusting the parameters of the BioBERT model to suit the characteristics of the malaria subdomain better. The model is intended to improve the process of recognizing entities and their relationships in the context of malaria-associated publications. This solves an essential problem in connection with the inapplicability of the previously developed models in the biomedical field. The study uses complex and highly effective machine learning algorithms, such as Long Short-Term Memory, Random Forests, Support Vector Machine and Gradient Boosting Machine, to fine-tune the existing BioBERT model, leading to the FT-BioBERT model. The fine-tuned model is compared with other models, such as BioBERT and Multi-BioNER, over three datasets, namely BC5CDR-Disease, BioRED, and NCBI-Disease. The fine-tuned model achieved notable performance improvements: achieving 92.4% in accuracy, which is a 3.13% increase from BioBERT and 2.33% from Multi-BioNER and attaining 91.8% in precision, 92.7% in recall, and 92.2% F1-score which is a of 3.15% improvement over BioBERT, and 2.23% improvement over Multi-BioNER. Based on the results, we confirm that the proposed model can effectively identify and extract entities and their relationships when supplied with malaria literature and, therefore, is suitable for biomedical text mining. We hope that the study's findings will provide new avenues that will lead to the creation of domain-related NLP applications in malaria-related fields.

Keywords: Named Entity Recognition, Relation Extraction, BioBERT, Biomedical Language Model, Biomedical Natural Language Processing