PRE-TRAINED LANGUAGE MODEL-BASED SEMI-SUPERVISED LEARNING APPROACH FOR CONTENT-BASED EMAIL CATEGORIZATION

Nuwan Dinusha Kankanamge

189383 U

Degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

October 2020

PRETRAINED LANGUAGE MODEL-BASED SEMI-SUPERVISED LEARNING APPROACH FOR CONTENT-BASED EMAIL CATEGORIZATION

Nuwan Dinusha Kankanamge

189383 U

This thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

October 2020

Declaration

I declare that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organizations.

UOM Verified Signature

Name of Student Signature of Student

N.D. Kankanamge Date: 22/05/2021

Supervised by

Name of Supervisor(s) Signature of Supervisor(s) *UOM Verified Signature*

Dr. (Ms.) A.T.P. Silva Date: 24/05/2021

Dedication

I dedicate this thesis to my parents for the support they have given me to fulfil my dreams. I dedicate this work to people at the University of Moratuwa for their help and guidance throughout this research.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. A.T.P. Silva for guiding me throughout the project. She has given his fullest cooperation to me whenever I sought advice.

I must thank Prof. Asoka Karunananda for the advice given, which had a significant impact on making my mind to think differently and his advice and techniques helped me a lot in the process of developing this solution.

I must also thank all members of the lecture panel. During one year and three months, these valuable lectures helped me to think differently.

For this research, I had to refer to many books and research papers as reference. I would like to thank all the authors of those publications.

Also, my batch mates helped me a lot in many ways. I would like to pay my gratitude for them. They made my life enjoyable during the course period.

I would like to place my gratitude to my loving parents and wife for always encouraging me on higher studies.

Last but not least, I would like to thank all my colleagues and others who are not mentioned, for all the support extended to me. Without their dedication, the project would not have been successful.

Abstract

Recent developments in neuroscience have revolutionized modern trends in artificial intelligence. Artificial neural networks (ANN), which is the artificial model of the human brain, have started to dominate in the field of artificial intelligence. The major usage of ANN is for data classification and prediction. There are numerous applications of ANN, ranging from health, education, entertainment, and business.

Email classification has been an issue for many of the large organizations as it needs human interaction. There are many artificial intelligence-based solutions have been proposed. When it comes to content-based email filtering, many recent researchers have identified that the use of ANN-based approaches are much more useful than conventional natural language modelling methods, as the volume of data increased. One reason for this is ANN has been able to capture some of the hidden styles of writing which have not been captured by conventional natural language processing. However conventional ANN has been suffering from lack of labeled data for training. This has been the major drawback of conventional ANN approach as generating labeled data needs human interaction and therefore making it a costly process. This has limited ANN solutions from providing a generic approach for email classification in any domain since to succeed, it needs large a number of labeled data from each of these domains to train the particular ANN.

This thesis report on our research on content-based email classification using semi-supervised learning which will address the issues with conventional ANN. Semi-supervised learning was introduced around 15 years back but came to play an important role in the field of artificial intelligence recently. Semi-supervised learning provides a solution to this issue as it needs a minimum amount of labeled data for training and it can use unlabeled data to increase its' accuracy. Proposed solution is a multi-view core-training approach that takes labeled emails, unlabeled emails and the names of the different categories as inputs. Output of the project is a trained model that can classify emails to given categories. We have tested our solution with 10000 training samples where only 10% to 20% were given to the system as labeled data and others were used as unlabeled data. We managed to achieve around 0.888 accuracy which is more than 5% accuracy improvement from the total system.

Contents

List of Figures

List of tables

Abbreviations

ANN	-	Artificial Neural Network
SVM	-	Support Vector Machine
NLP	-	Natural Language Processing
LSTM	-	Long Short-Term Memory