A DEEP BIDIRECTIONAL TRANSFORMER BASED TWITTER SPAM DETECTION AND PROFILING

Thivaharan.V

179353U

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

A DEEP BIDIRECTIONAL TRANSFORMER BASED TWITTER SPAM DETECTION AND PROFILING

Thivaharan.V

179353U

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science specializing in Data Science Engineering and Analytics

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

UOM Verified Signature

Signature: …………………………. Date: 31/May/2021

Name: Thivaharan.V

The supervisor/s should certify the thesis/dissertation with the following declaration. I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the Master of Science.

Signature of the supervisor: UOM Verified Signature Date: 31/May/2021

Name: Dr. Uthayasanker Thayasivam

ABSTRACT

Online social networks are becoming extremely popular among Internet users as they spend a significant amount of time on popular social networking sites like Facebook, Twitter, and Google+. These sites are turning out to be fundamentally pervasive and are developing a communication channel for billions of users.

Twitter has become a target platform on which spammers spread large amounts of harmful information. These malicious spamming activities have seriously threatened normal users' personal privacy and information security. An effective method for detecting spammers is to learn about user features and social network information.

However, social spammers often change their spamming strategies for evading the detection system. To tackle this challenge, in this research we determine various features to capture the consistency of users' behavior.

In this research, we investigate additional criteria – spam patterns, to measure the similarity across accounts on Twitter. We propose a method to define the relation among accounts by investigating their tweeting patterns and content. Our real data evaluation reveals that, given some initially labelled spam tweets, this approach can detect additional spam tweets and spam accounts that are correlated to the initially labelled spam tweets.

Keywords: Classification, Word embedding, Vectors, Cosine similarity, Crawler

ACKNOWLEDGEMENTS

I would like to express profound gratitude to my advisor, Dr. Uthayasanker Thayasivam, for his invaluable support by providing relevant knowledge, materials, advice, supervision and useful suggestions throughout this research work.

His expertise and continuous guidance enabled me to complete my work successfully. I thank him for the patient guidance, enthusiastic encouragement and useful critiques of this research work. His continuous supervision greatly helped me in keeping the correct phase in research work.

I especially appreciate the frequent feedback, which helped me to correct and fine-tune it to this level. I would like to thank all staff from the Department of Computer Science and Engineering for various support rendered by them throughout this effort.

I am in much debt to my friends who lent a helping hand through various means. Also, I would like to extend thanks to everyone who helped me behind the scene to pursue this degree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
BERT	Bidirectional Encoder Representations
OSN	Online Social Network
CEO	Chief Executive Officer
US	United States of America
URL	Uniform Resource Locator
API	Application Programming Interface
DB	Database
PBF	Profile-Based Features
CBF	Content-Based Features
GBF	Graph-Based Features
NBF	Neighbor-Based Features
ABF	Automation-Based Features
TBF	Timing-Based Features
FFNN	Feed-Forward Neural Network