

**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.

UOM Verified Signature

Signature of the supervisor: .

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

DECLARATION	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
LIST OF ABBREVIATIONS	viii
CHAPTER 1. INTRODUCTION	9
1.1. Overview	9
1.2. Problem and Motivation	10
1.3. Aim	12
1.4. Objectives	12
CHAPTER 2. LITERATURE REVIEW	13
2.1. Overview	13
2.2. Profile-Based Features	15
2.3. Content-Based Features	16
2.4. Graph-Based Features	21
2.5. Automation-Based Features	23
2.6. Neighbor-Based Features	24
2.7. Timing-Based Features	24
CHAPTER 3. METHODOLOGY	26
3.1. Overview	26
3.2. Data Collection	27
3.3. Data Preprocessing	30
3.4. Feature Extraction	34
3.4.1. Calculate Tweet Similarity	41
3.4.2. Calculate Mention Ratio	43
3.4.3. Calculate Hashtag Ratio	43
3.4.4. Calculate URL Ratio	43
3.4.5. Calculate Retweet Ratio	43
	iv

3.4.6.	Calculate F Ratio	44
3.4.7.	Spam Classification	44
CHAPTER 4.	EXPERIMENT	46
4.1.	Overview	46
4.2.	Data Analysis and Collection	46
4.3.	Experiments with Tweets	48
4.4.	Experiment with TF-IDF	50
4.5.	Experiment with BERT	50
4.6.	Summary	50
CHAPTER 5.	RESULTS AND ANALYSIS	51
5.1.	Overview	51
5.2.	Analysis	51
5.3.	Results	53
5.4.	Summary	56
CHAPTER 6.	DISCUSSION	57
6.1.	Overview	57
6.2.	A Deep Bidirectional Transformer Based Twitter Spam Detection And Profiling	57
CHAPTER 7.	CONCLUSION AND FUTURE WORK	58
7.1.	Conclusion	58
7.2.	Future work	59
REFERENCES		61
APPENDIX A:	CLASSIFIED TWEETS	66

LIST OF FIGURES

Figure 1-1. BBC journalist Louis Theroux hacked account	11
Figure 2-1. Finite state machine based spam template [2].....	17
Figure 2-2. Posting Heterogeneous Tweets [2].....	18
Figure 2-3. Spin Bot, ‘Spammers Are Becoming Smarter on Twitter - 2016’ [2]	19
Figure 2-4. Posting More Tweets, ‘Spammers Are Becoming Smarter on Twitter - 2016’ [2]....	19
Figure 2-5. Mixing tweets between tweets containing links to scam websites [2].....	20
Figure 2-6. Twitter Social Graph [11]	21
Figure 2-7. Centrality.....	22
Figure 3-1. A Deep Bidirectional Transformer Based Twitter Spam Detection And Profiling ...	26
Figure 3-2. Data storage architecture.....	27
Figure 3-3. Commonly used words in spam tweets.....	31
Figure 3-4. Commonly used words in genuine tweets.....	31
Figure 3-5. Retweet distribution	32
Figure 3-6. URL in tweets distribution.....	33
Figure 3-7. Mention in tweets distribution	33
Figure 3-8. Hash-tag in tweets distribution.....	34
Figure 3-9. Tweet vectorization.....	39
Figure 3-10. BERT model with pre-processed tweets	40
Figure 3-11. BERT model with not pre-processed tweets	40
Figure 4-1. Elon Musk Tweets about going to moon	49

LIST OF TABLES

Table 2-1. Research and detection methods	13
Table 3-1. Collected data representation	28
Table 3-2. Tweets and tokens	37
Table 3-3. Accuracy comparison of our BERT models.....	41
Table 3-4. Spam tweet prediction with our BERT model	45
Table 4-1. Result of data crawling.....	47
Table 4-2. Spam and genuine tweets distribution in gathered data	48
Table 4-3. Spam and genuine user distribution in gathered data.....	48
Table 5-1. Spam Tweets predictions evaluation.....	53
Table 5-2. Evaluation for not pre-processed tweets.....	54
Table 5-3. Evaluation for pre-processed tweets.....	55

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