TWITTER SPAM DETECTION USING MACHINE LEARNING APPROACH

CHAN YAN SHIH

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: [**TWITTER SPAM DETECTION USING MACHINE LEARNING APPROACH**]

SESI PENGAJIAN: [2020/ 2021]

Saya: CHAN YAN SHIH

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
____✓_____	TIDAK TERHAD	

YAN SHIH

(TANDATANGAN PELAJAR)

Alamat tetap: 37, Lorong Air Putih 4
25300, Kuantan, Pahang

Tarikh: _____11/9/2021_____

fadzilah

(TANDATANGAN PENYELIA)

DR NUR FADZILAH BINTI OTHMAN
Nama Penyelia

Tarikh: ____14/9/2021_____

CATATAN: * Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

TWITTER SPAM DETECTION USING MACHINE LEARNING APPROACH

CHAN YAN SHIH

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Security) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I hereby declare that this project report entitled

TWITTER SPAM DETECTION USING MACHINE APPROACH

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT : _____YAN SHIH_____ Date : __11/9/2021___

([CHAN YAN SHIH])

I hereby declare that I have read this project report and found

this project report is sufficient in term of the scope and quality for the award of

Bachelor of [Computer Science (Computer Security)] with Honours.

SUPERVISOR : _____ fadzilah _____ Date : 14/9/2021

NUR FADZILAH BINTI OTHMAN

DEDICATION

I would wish to dedicate my work to my beloved parents whose motivation or determination rings in my ears and continuously provide their moral, spiritual, and funding.

I also dedicate this dissertation to my siblings, grandparents, mentor and friends who advised and motivated me all the time to finish this study.

Finally, I dedicate this book to the Great God for granting me power, courage, skills as well as healthy life to complete this project.

ACKNOWLEDGEMENTS

Firstly, my sincere gratefulness goes to Universiti Teknikal Malaysia Melaka (UTeM), for providing us; undergraduates from the Faculty of Information and Communication Technology (FICTS) a precious opportunity on the road to fulfill our final year project within the 28 weeks.

Following gratitude extends to my dearest supervisor, Dr. Nur Fadzilah binti Othman for all her mentoring and reassurance have been invaluable and her insightful ideas initiated the largest part of this paper.

Above ground, my appreciation also extends to my beloved parents, Chan Tsung Fond and Lee Sow Kheng for all the words of inspiration and financial support. Last but not least, to all my course mates and friends, thank you for assisting me to complete this study as a whole.

ABSTRACT

Nowadays, the application of social media has grown widely in our daily routine. People can freely post and share any contents on social media. With the growth of social media, people can now make use of it for building connections whether for business or personal gain. The popularity of Twitter has also noted to attract awareness of spammers who make use of Twitter for their own malevolent objectives such as conducting acts of phishing real Twitter users or spreading malicious software through URLs that are shared in tweets as well as hijack topics to attract users' attention. The Internet is a boundless platform for information and data sharing. Detecting spam contents from social media network is an intriguing research topic because it is important for cyber forensic agencies to detect the way of social media in broadcasting malicious activities or attacks before offenses are performed. This research attempts to detect spam in Twitter platform using three different machine learning classifier models which is Naïve Bayes, Support Vector Machine (SVM), and Random Forest in addition propose the model that produce the highest accuracy and precision in predicting spam by comparing each of the model's result. At the end of this study, the results of each model's analysis will be explained and compared to achieve the objective of this study. The dataset is categorized into Training and Testing and the samples for testing is divided into 5 categories such as 100, 200, 300, 500, and 1000 sample tweets. The reason of dividing the samples into different size is to analyses whether the size of samples affect the analysis results or not. After comparing the results, we can conclude that Naïve Bayes has the highest accuracy and precision value in predicting spam while Random Forest has the worst accuracy. Thus, this research includes all features from extracting contents from social media network such as Twitter, applying different machine learning classifiers based on specific keywords like URLs on social media network to finally classifying them as Spam or Ham as well as equating the accuracy differences between each of the machine learning classifiers.

ABSTRAK

Pada masa kini, penggunaan media sosial telah berkembang secara meluas dalam rutin harian kita. Orang ramai boleh menghantar dan berkongsi kandungan di media sosial dengan bebas. Dengan pertumbuhan media sosial, orang kini dapat menggunakannya untuk membina hubungan sama ada untuk perniagaan atau keuntungan peribadi. Popularity Twitter juga diperhatikan untuk menarik kesedaran spammer yang menggunakan Twitter untuk tujuan jahat mereka sendiri seperti melakukan tindakan memancing pengguna Twitter sebenar atau menyebarkan perisian jahat melalui URL yang dikongsi dalam tweet serta topik rampasan untuk menarik pengguna 'perhatian. Internet adalah platform tanpa batas untuk berkongsi maklumat dan data. Mengesan kandungan spam dari rangkaian media sosial adalah topik penyelidikan yang menarik kerana penting bagi agensi forensik siber untuk mengesan cara media sosial dalam menyiarkan aktiviti atau serangan jahat sebelum kesalahan dilakukan. Penyelidikan ini cuba mengesan spam di platform Twitter menggunakan tiga model pengkelasan pembelajaran mesin yang berbeza iaitu Naïve Bayes, Support Vector Machine (SVM), dan Random Forest di samping mencadangkan model yang menghasilkan ketepatan dan ketepatan tertinggi dalam meramalkan spam dengan membandingkan masing-masing hasil model. Pada akhir kajian ini, hasil analisis setiap model akan dijelaskan dan dibandingkan untuk mencapai objektif kajian ini. Set data dikategorikan ke dalam Latihan dan Pengujian dan sampel untuk ujian dibahagikan kepada 5 kategori seperti 100, 200, 300, 500, dan 1000 contoh tweet. Sebab membahagikan sampel ke dalam ukuran yang berbeza adalah dengan menganalisis sama ada ukuran sampel mempengaruhi hasil analisis atau tidak. Setelah membandingkan hasilnya, kita dapat menyimpulkan bahawa Naïve Bayes mempunyai nilai ketepatan dan ketepatan tertinggi dalam meramalkan spam sementara Random Forest mempunyai ketepatan terburuk. Oleh itu, penyelidikan ini merangkumi semua ciri dari mengekstrak kandungan dari rangkaian media sosial seperti Twitter, menerapkan pengelasan pembelajaran mesin yang berbeza berdasarkan kata kunci tertentu seperti URL di rangkaian media sosial untuk akhirnya mengklasifikasikannya sebagai Spam atau Ham serta menyamakan perbezaan ketepatan antara masing-masing pengelasan pembelajaran mesin.

Table of Contents

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

SVM – Support Vector Machine

2D – Two Dimensional

3D – Three Dimensional

PSM – Projek Sarjana Muda

LIST OF APPENDICES

Chapter 1: INTRODUCTION

1.1 Introduction

People's communication modes are no longer restricted and limited to only on-site features in this advanced era where technology has evolved tremendously. Social media networks, particularly social media platforms like Twitter, have grown in importance as a means of communication and news dissemination, attracting spammers from all over the world to divert users' attention. (Sepideh Bazzaz Abkenar, 2020). Twitter's popularity, among so many social media platforms, has made it an easy and appealing platform for spammers to spread spam to the point where it has become a serious problem. Twitter spam is often described as unwanted tweets that contains hostile links that redirect victims to third-party web site infested with risks such as phishing, terrorists, drug sales, malware downloads, scam, and lots of others (Susana Boniphace Maziku, 2020).

A series of incidents from the past to the present have demonstrated that Twitter spam does, in fact, affect the user experience and poses significant threats beyond social media platforms. For example, in April 2021, India was hit by a disinformation warfare campaign in which spammers used Twitter by changing their username and profile picture to a well-known figure in order to spread controversial fake news that could lead to riots and large-scale disruption. In another case, spammers abused Twitter by distributing fake or misleading posts and videos using a hashtag or keyword (Mishra, 2021).

Suspicious accounts that send duplicate or the same content to multiple users or post tweets that only include URL contents can be marked and reported as spammers in the current Twitter feature for further action by users. Twitter has used blacklisting services such as Trend Micro's Web Reputation Energy for spam filtering purposes. However, spammers' attack strategies are constantly changing, and blacklist services have their own limitations, making them unable to

detect spams earlier. As a result, researchers supported Machine Learning (ML) methods for identifying the underlying patterns of spammers' activities to detect spam more efficiently. (Sepideh Bazzaz Abkenar, 2020).

Social media network is an online platform that permit many people to interact remotely. There are various types of social medias accessible today, each of them comes with its own set of features and functionalities based on the intention for which it is expected. The simplicity of these networks, combined with the proliferation of personal devices like smartphones that enable continuous network access, encourages users to overcome some of the communicative difficulty that exist. Accordingly, individuals are emboldened to share private information with unknown being such as human or system (F. Concone, 2019). Machine learning algorithms, on the other hand, utilize data to discover unique patterns in data such as graphic images, words, and phrase, and even figures. A machine learning algorithm can only be fed digitally stored data. Many of today's recommendation systems, such as those algorithms on Netflix, Spotify, and Youtube; search engines used on Google and Baidu; and interactive social-media feeds such as Facebook and Twitter, are all operated by machine learning. Each of the platforms starts by gathering as much information about user as possible. The collected information comprises what user like to watch, how user react to status updates, as well as the links user click. Then, machine learning is used by the platforms to create highly educated speculations about what user might want next based on their records (B. Mukunthan, 2020).

1.2 Problem Statement

Among the various Twitter assessments, spam account identification is quite possibly the most explored and important. Spammers are entities, regardless of whether genuine individuals or robotized bots, whose objective is to over and again share messages containing undesirable substance for business or hostile purposes, like connects to pernicious sites, to spread malware, phishing assaults, and other hurtful action. Twitter spam detection is a continuous fight among cops and looters. To debilitate vindictive practices, interpersonal organizations are continually developing, and therefore, spammers have advanced too, embracing more complex methods that make it simple to avoid security systems. Following various studies, the researchers concluded that many works on social spam detection have been conducted; nonetheless, most prior work on

social media such as Twitter spam has focused on the strategies and procedures for spam detection and evasion on a solitary social network. These works have been discovered as being done for Facebook, MySpace, or Twitter. Various classifier models presented by a variety of researchers have previously been tested in spam detection, and it has been discovered that selecting the right one for the same purpose is a significant challenge. Spam is an evolving issue on the Internet by and large, and Twitter is no special case. Furthermore, Twitter spam is far more effective than email spam. Other researchers have proposed a variety of methods for dealing with Twitter spam, including detecting spammers dependent on tweeting history or social characteristics, recognizing sporadic activities, and grouping tweet-inserted URLs.

As the improvement of new spam recognition strategies requires the utilization of stable and commented on datasets to assess their presentation, such dynamism delivers the datasets in the writing rapidly old and almost pointless. Besides, giving the ground truth to a huge measure of information is a tedious errand that is as yet done physically in most of cases.

1.3 1.3 Research Questions

Based on the above problem statement, there are a few research questions formed as displayed in Table 1.1.

Table 1.1: Research Question

No.	Research Questions
1	Which analysis model is used to detect spam in Twitter?
2	Which model used is the most accurate to detect spam in Twitter?
3	What is the most accurate analysis model to detect spam in Twitter?

1.4 Research Objectives

Based on the research questions listed in Table 1.1, a couple of research objectives are revealed in Table 1.2 to solve the research questions from Table 1.1.

Table 1.2: Research Objectives

No.	Research Objectives
1	To analyze machine learning classification model such as Naïve Bayes, Support Vector Machine (SVM) and Random Forest.
2	To develop a comparison to measure the accuracy test for spam and non-spam tweets between machine learning models.
3	To propose the highest accuracy machine learning model in detecting Twitter spam.

1.5 Research Summary Matrix

The summary for research question and research objectives of the research is exhibited in Table 1.3.

Table 1.3 Summary of Research Question and Research Objectives

Research Questions	Research Objectives
Which analysis model is used?	To analyze machine learning classification model such as Naïve Bayes, Support Vector Machine (SVM) and Random Forest.
Which model that were applied is most accurate to detect spam in Twitter?	To develop a comparison to measure the accuracy test for spam and non-spam tweets between machine learning models.
What is the most accurate analysis model to detect spam in Twitter?	To propose the highest accuracy machine learning model in detecting Twitter spam.

1.6 Scope of the Research

In this study, three different machine learning classifier models are used to detect spam in Twitter. The main scope for using three different machine learning classifiers such as Naïve Bayes, Support Vector Machine (SVM) and Random Forest is due to lack of researchers comparing these three models in spam detection study despite their popularity and high prediction accuracy.

In machine learning, there are countless diverse forms of classification tasks that could be encountered, such as Naïve Bayes, Support Vector Machine and Random Forests. Naïve Bayes is a probabilistic algorithm centered on the Bayes Theorem that is applied to execute a wide scope of classification activities. The Bayes Theorem is a facile mathematical equation for computing restrictive probabilities. Support Vector Machine (SVM), alternatively, is a vigorous machine learning model that shows high precision with different classification issues and is commonly used in an assortment of entrenched applications. In a variety of cases and applications, SVMs have shown high classification precision ratios, outflanking other famous classification algorithms. Random forests (RF) is a versatile, user-friendly algorithm that consistently creates excellent results even without the presence of hyper-parameter tuning. Random Forests creates a "forest" out of an ensemble of decision trees, which are typically trained using the "bagging" method hence also called as ensemble or bagging method (Donges, 2021).

Python is another important tool in this study as it is used to implement the classifier models in code form. Python is a sophisticated beneficial programming language where its language builds and object-oriented approach are aimed to help developers recorded as a hard copy clear, consistent code for both little and enormous scope projects. Python is ideal for a wide array of machine learning (ML) and artificial intelligence (AI) projects because it is a stable, flexible, and simple programming language. In fact, there are numerous Python machine learning and AI libraries and packages available that will be used in this study to analyze machine learning classifiers such as Naïve Bayes, SVM, and Random Forest. The Python programming language is also used in this study for data visualization, with modules such as Pandas and Plotly generating a graph plot to measure the accuracy difference between the machine learning classifiers analyzed.

The parameters from the classification report result of machine learning models will be used to determine which model has the highest accuracy value. In this study, we will be focusing on precision and accuracy parameters in order to distinguish which machine learning models is the most accurate. Precision parameter will determine how many positive classes are predicted to be true while accuracy value will be generated from the total of spam and non-spam tweets predicted correctly.

1.6.1 Research Contribution

This research will attempt to aid a social media network or platform in detecting spam threats. This research will also benefit users by allowing them to detect and respond to potential spam threats progressively, resulting in a safer social networking exposure to users. Additionally, cyberlaw and forensic enforcement agencies can use the implemented method to identify behaviors and patterns in social media networks. Furthermore, manually categorizing the tweets extracted from the Twitter platform into Ham and Spam categories will produce more accurate data as an output. The system is trained and tested using the labelled datasets.

1.6.2 Keywords

This section emphasizes keywords that are related to or relevant to the research. Keywords are important phases, words, or concepts in research.

1. Twitter: Twitter, a social media networking web site established in 2006, can be utilized to get news, follow celebrities and organizations, or stay in contact with old acquaintances (Forsey, 2019). Twitter is also a combination of social media network such as Instagram and Facebook, as well as technologies such as instant messaging, to form chains of users who can interact anytime and anywhere with short messages known as tweets.

2. Machine Learning: Machine Learning is an artificial intelligence (AI) application that permits systems to consequently learn, create, and improve from training without having to do anything explicitly programmed (Varone M., 2020). Machine learning algorithms allow computers to practice or train on data inputs and afterward apply arithmetic analysis to output values that fall within a specific scale.

3. Twitter Spam: Twitter spam has become a major problem in recent years as spammers on Twitter tweet for a variety of reasons, including spreading advertisements, disseminating pornography, spreading viruses, phishing, or essentially subverting a system's status. If a tweet is not entirely composed of text, it is considered spam. Instead, it could incorporate a mention, an URL, a hashtag, or a graphic image. (Niddal Imam, 2019).

1.7 Report Organization

Each chapter of the report is summarized in this section. This report is divided into seven chapters, which are explained below.

1.7.1 Chapter 1: Introduction

Chapter 1 portrays how to detect spam on Twitter using a machine learning approach. This section likewise incorporates the project's problem statement, project question, project objective, project scope, and project contribution.

1.7.2 Chapter 2: Literature Review

Chapter 2 discusses some peer-reviewed papers on spam detection in social media using a machine learning approach. It includes existing machine learning classifying techniques, their strengths, weaknesses, and limitations, as well as proposed techniques and improvements to existing techniques.

1.7.3 Chapter 3: Project Methodology

Chapter 3 justifies the methodology and actions undertaken to carry out this research. This chapter also includes a milestone and a Gantt chart for the research to ensure that the tasks assigned are completed on time and smoothly.

1.7.4 Chapter 4: Implementation

This chapter focuses on the code and development of spam detection on Twitter using machine learning approaches. This chapter will also look at the system's expected outcome.

1.7.5 Chapter 5: Testing and Analysis

Chapter 5 sets the machine learning approach to spam detection to the test and analyses the results. A few tests and analyses will be performed, and the resulting results will be explained in this chapter.