YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD

SYAZA LIYANA BINTI MUHAMAD SHAPEE

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: [***YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD***]

SESI PENGAJIAN: [**2020 / 2021]**

Saya: SYAZA LIYANA BINTI MUHAMAD SHAPEE
(HURUF BESAR)
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	

(TANDATANGAN PELAJAR)

Alamat tetap: NO.3, SEKUNTUM 21,
TAMAN BUKIT DAHLIA 81700
PASIR GUDANG, JOHOR

Tarikh: 7 SEPTEMBER 2021

(TANDATANGAN PENYELIA)

MR. NOR AZMAN BIN MAT ARIFF

Nama Penyelia

Tarikh: 8/9/2021

YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD

SYAZA LIYANA BINTI MUHAMAD SHAPEE

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

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2021

DECLARATION

I hereby declare that this project report entitled

YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT:  Date : 7 SEPTEMBER 2021

SYAZA LIYANA BINTI MUHAMAD SHAPEE

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 Network) with Honours.

SUPERVISOR :  Date : 8/9/2021

MR. NOR AZMAN BIN MAT ARIFF

DEDICATION

To my beloved parents, Zaleha binti Ali and Muhamad Shapee bin Ghazali who inspired me to be strong despite of many obstacles in life, for their prayers and their overwhelming support morally and financially. My sisters and brother, Syaza Hazirah, Syaza Najihah and Muhammad Adam have never left my side and are very special.

To my fellow friends, for being there for me throughout the entire bachelor program and their cooperation while conducting the research.

To my supervisor, for encouraging and believing in me to complete this research.

ACKNOWLEDGEMENTS

All praises be to Almighty Allah S.W.T who has blessed me with the belief, strength and capabilities to understand, learn and complete this research. Peace and prayers be upon our most beloved Prophet Muhammad S.A.W, the most beautiful soul, whose sayings, actions and stories have deeply inspired me enough to believe that there are no limitations to what I can achieve when we are fully committed to accomplish something, knowing that Allah is on my side.

I also admire the help and the guidance of my supervisor, Mr. Nor Azman bin Mat Ariff for his guidance, encouragement and patience from all aspects during the preparation of this research are highly appreciated.

I am blessed to have had such wonderful, loving and supporting parents, Zaleha binti Ali and Muhammad Shapee bin Ghazali for the education they gave me at home as I was growing up and the education, they paid for till I graduated. They have been my pillar of strength and till this very day, every small achievement I make, they always want to be the first to know and to congratulate.

To my fellow friends who have helped in the strenuous process in collecting information and preparing this research. May Allah bless you all for your patience and selfless commitment.

ABSTRACT

The number of YouTube users is constantly rising. However, such success is not without its drawbacks. Spam has become a common form of attack and threat, and most YouTube users are unaware of it. Receiving and being overwhelmed with unnecessary spam regularly has become one of the most internet-disruptive topics in today's world. The Support Vector Machine (SVM) is used in this study to develop a YouTube detection framework. The YouTube spam datasets were obtained from the UCI Machine Learning Repository. This project aims to show that an SVM model can accurately predict YouTube spam in a comment. Based on the SVM model, this research could produce a system that can detect spam and legitimate comments on YouTube.

ABSTRAK

Bilangan pengguna YouTube terus meningkat. Namun, kejayaan itu bukan tanpa kekurangannya. Spam telah menjadi bentuk serangan dan ancaman yang biasa, dan kebanyakan pengguna YouTube tidak menyedarinya. Menerima dan dibanjiri dengan spam yang tidak perlu secara berkala telah menjadi salah satu topik yang mengganggu internet di dunia sekarang. Mesin Vektor Sokongan (SVM) digunakan dalam kajian ini untuk mengembangkan kerangka pengesanan YouTube. Set data spam YouTube diperoleh dari UCI Machine Learning Repository. Projek ini bertujuan untuk menunjukkan bahawa model SVM dapat meramalkan spam YouTube dengan tepat dalam komen. Berdasarkan model SVM, penyelidikan ini dapat menghasilkan sistem yang dapat mengesan spam dan komen yang sah di YouTube.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASCII	-	**American Standard Code For Information Interchange**
BOW	-	**Bag-Of-Word**
CS	-	**Chi-Square**
DNS	-	**Domain Name System**
DOS	-	**Denial Of Service Attack**
Email	-	**Electronic Mail**
FN	-	**False Negatives**
FP	-	**False Positives**
FS	-	**Feature Selection**
FYP	-	**Final Year Project**
GIF	-	**Graphic Interchange Format**
HTML	-	**Hypertext Markup Language**
IG	-	**Information Gain**
IDE	-	**Integrated Drive Electronics**
Weka	-	**Waikato Environment for Knowledge Analysis**
SVM	-	**Support Vector Machine**