INFORMATICS INSTITUTE OF TECHNOLOGY
in collaboration with
the University of Westminster, UK

BEng (Hons) DEGREE PROGRAMME in
Software Engineering

A DYNAMIC REAL-TIME SPAM DETECTION FRAMEWORK

A dissertation by
H.B.R.A.K.R.V.K Bandara

Supervised by
Mr. Champika Samarasinghe

Submitted in partial fulfilment of the requirements for the
BEng (Hons) Software Engineering
Department of Computing

W1582949 | 2015072

May 2019

Abstract

Online social networks such as Facebook, Twitter and LinkedIn have had a huge rise of users all over the world for the past few years. The constant growth of Twitter OSN due to many active users and modern features that revolutionise how humans communicate make Twitter an attractive platform for spammers. While most Twitter posts are relevant and helpful, some of these tweets contain spam which can be a waste of time, resources and potentially compromise the users' security or result in stolen data that might be sensitive and private.

According to existing work, when trying to identify spam, it has been found that the characteristics of the spam tweets and spammers keep changing over time and therefore most machine learning classifiers of spam detection are not updated with new or changed spam tweets. This issue can be referred to as the "Twitter Spam Drift" problem. For this research, a spam detection framework has been developed that can detect spam while avoiding the spam drift problem. The identification of spam drift will be done for the tweets that will be analysed in real-time using NLP.

The developed spam detection framework will consist of three modules: *spam detector* module operating in real-time mode, *drift detector* module which will identify drifted spam tweets using NLP, and *review and retrain* module where the drifted spam tweets are reviewed and consequently will be used to re-train the tweet classifier.

Experiments on a small-scale dataset show that the framework will be able to detect spam continuously while overcoming the spam drift issue for tweets in the Twitter OSN. The successful results of the research satisfied experts who evaluated the system while suggesting room for further investigation and enhancements

Key Words: Machine Learning, Text Classification, Twitter Spam Detection, Spam Drift Detection, Natural Language Processing, Fuzzy Logic