SPAMBRELLA : MACHINE LEARNING BASED SINHALA SPAM COMMENTS DETECTION SYSTEM FOR YOUTUBE

Rishoni Nathaliya De Silva

A dissertation submitted in partial fulfilment of the requirement for Bachelor of Engineering (Honours) degree in Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2020

ABSTRACT

Identifying and moderate Sinhala spam content on social media is challenging for users. YouTube has shown up as a main rival in the video sharing space.One of the most usable features of YouTube is that users can comment on others' videos. This feature permits users to collaborate with others and share their sentiments, opinions , and so on.This has become an open door for malicious users to share promotional , harmful , mis-driving substances known as spam content. Spam can be considered as harmful,misusing,cyber threat because spam has the potential of cyber security threat for end users. Detecting these spam contents is difficult, due to language dependent limitations. Therefore, the requirement for automatic identification of spam comments on social media has become of utmost importance. Simple keyword spotting procedures cannot be used to identify the exact intention of a comment. Proposed system addresses the mentioned issue by building an ensemble spam classification model with machine learning that can be used to classify spam comments in Sinhala language. This study has been able to develop different pre-processing techniques for the Sinhala sentence normalization. The unique features of the Sinhala spam comments were investigated for the feature extraction phase. With the use of different natural language techniques, the content was classified for the domains spam and non-spam. Different feature extraction techniques used and ensemble/single classifiers used for classifying the text and enhancing the performance of the system. The trained model was then able to classify racist comments with a 88.0% accuracy in experimental results.The Project evaluation was conducted along with self and expert evaluation.

Therefore, the requirement for automatic identification of racist comments on social media has become of utmost importance.However, simple keyword spotting techniques cannot be used to accurately identify the exact intent of a comment. In this paper, we address this issue by building a text analytics model with machine learning that can be used to filter racist comments in Sinhala language.

A Two -Class Support Vector Machine was trained with a set of carefully chosen comments from Facebook that were labelled as racist and non-racist based on