Smart SMS Classification for Android Operating System Using Natural Language Processing

S.S. Sumanasekara*, M.W.S. Kaumada, N.E.C Jayasekara and S.T.C.I Wimaladharma

Department of Computer Science and Informatics, Uva Wellassa University, Sri Lanka

The use of Short Message Service (SMS) is increasing as more people exchange SMS messages very frequently due to the rapid increase of mobile phone usage and the simplicity in sending SMS messages. However, this has led to an increase in mobile device attacks using SMS Spam. The two main categories of SMS Messages are spam messages and ham (legitimate) messages. Up to now, several kinds of research were done on SMS classification but all of them are on spam filtering techniques by using various algorithms and machine learning techniques. In this paper, we present a novel approach that can detect and filter both spam and ham messages into a better organization under six different predefined categories named as Primary for legitimate messages, Bank and Finance, Social and Web, Promotions, Service Provider Messages, and Spam Messages by using Natural Language Processing for Android Operating System. A smart messaging application that can properly organize SMS into categories will help to identify the SMS easily as they are classified under different tabs. Even though SMS can be identified and categorized manually with little or no effort by people, it remains difficult for mobile phones. A dataset is created according to the Sri Lankan context and various experiments are performed to evaluate the performance of the SMS Classification. Initially, the features were selected based on the behavior of messages and extracted the features from the dataset to get the feature vectors. Naive Bayes and Support Vector Machines algorithms were used to select the best classification algorithm. With the highest accuracy rate, the Support Vector Machines algorithm is selected to train the model while k-Fold cross-validation is used to perform the validation. Our proposed approach achieved a 93% accuracy rate and the model is deployed in the Android environment and its performance is confirmed using a proof of concept.

*Keywords***:** SMS classification, Natural language processing, Support vector machines, Naive bayes algorithm, Android