Abstract

The mobile phone market has experienced a substantial growth over recent years. In second quarter of 2013, a total of 432.1 million mobile phones have shipped, which shows a 6.0% year over year increase. As the utilization of mobile phone devices has become commonplace, Short Message Service (SMS) has grown into a multi-billion dollars commercial industry. SMS is a text communication platform that allows mobile phone users to exchange short text messages (usually less than 160 seven-bit characters). It is the most widely used data application with an estimated 3.5 billion active users, or about 80% of all mobile phone subscribers at the end of 2010. As the popularity of the platform has increased, we have seen a surge in the number of unsolicited commercial advertisements sent to mobile phones using text messaging. Additionally, SMS Spam is particularly more irritating than email spams, since in some countries they contribute to a cost for the receiver as well. These factors along with limited availability of mobile phone spam-filtering software makes spam detection for text messages an interesting problem to look into.

Our work involves the extraction features for classifying the data set using Naïve Bayes, Random Forest, Support Vector Machine and the ensemble of these classifiers in WEKA through the knowledge we got by going through so many research papers related to SMS spam detection. We also used Rough Set Theory (RST) for generating the rule set for the further data analysis.