**Spam SMS filtering Using Machine Learning**

**Problem Statement**

Spam is unsolicited and unwanted messages sent electronically. Email spam is sent/received over the Internet while SMS spam is typically transmitted over a mobile network. Traditional email spammers are moving to the mobile networks as the return from the email channel is diminishing due to effective filtering, industry collaboration and user awareness. The Short Messaging Service (SMS) mobile communication system is attractive for criminal gangs for a number of reasons. It is becoming cost effective to target SMS because of the availability of unlimited pre-pay SMS packages in countries such as India, Pakistan, China, and increasingly the US. In addition, SMS can result in higher response rates than email spam as SMS is a trusted service with subscribers comfortable with using it for confidential information exchange. According to the GSMA it is inevitable that mobile network operators across the globe will see a rise in the volume and sophistication of SMS attacks in 2011.

**Background**

Early work proposing the application of automatic text classifi- cation techniques to SMS spam filtering includes work by Xiang, Chowdhury, and Ali (2004) who suggested that Support Vector Machines (SVMs) would be appropriate for the problem but did not evaluate their use, and work by Healy, Delany, and Zamolotskikh (2005) that considered using k-NN classifiers. Gómez Hidalgo, Bringas, Sánz, and Garcı´a (2006) evaluated a number of classification algorithms on two SMS spam datasets and concluded that these techniques can be effectively transferred from email to SMS spam filtering, with SVMs being the most suitable. Work by Cai, Tang, and Hu (2008) on a Chinese spam dataset used the simpler and lesser used Winnow algorithm (Littlestone, 1988), a linear classifier that has shown good performance in high dimensional feature domains with irrelevant features.

Wu, Wu, and Chen (2008) used a Bayes learner to extract keywords for monitoring traffic centrally, allowing a spamminess score to be assigned, however this work was not evaluated. Jie, Bei, and Wenjing (2010) added a cost function to a Naive Bayes filter which assigned a high cost to false positives. This translates into a high spam classification threshold, and a higher threshold results in higher spam precision. Longzhen, An, and Longjun (2009) proposed using a k-nearest neighbour algorithm (k-NN) as part of a multi-filtering approach. After black- and white-listing, a message is first classified by a filter using rough sets, which provide approximate descriptions of concepts. If this filter classifies the message as spam, it is then passed to the k-NN classifier for fi- nal classification. An evaluation on a data set of 550 spam SMS and 200 non-spam SMS with k = 12 showed that this dual filtering method is faster and more accurate than using k-NN alone.

**Methodology**

Although supervised learning techniques feature in the majority of recent work in SMS spam filtering there have been other machine learning approaches investigated. Rather than using the more standard text classification approaches, the SMS messages may be represented as a character-based vector which was projected into a smaller normalised feature space and clustered to identify clusters of spam and non-spam messages. New messages are classified based on their distance from the known spam and non-spam clusters. This approach is motivated by the lack of keywords available for the normal classification algorithms due to the short length of the SMS messages, but the efficacy of this approach at classifying spam was not evaluated. The behaviour of spam senders over time can be indicative of whether a given message is spam or not. Hu and Yan (2010) add a frequency analysis of SMS traffic to an existing spam filter with the goal of improving the central system’s real-time processing speed. By considering the frequency of spam messages received during different time periods and at different locations, they focus filtering on specific time periods and locations. Their approach improves the throughput of the system greatly, but at the cost of a large decrease in spam detection and a significant rise to 2.5% in the false positive rate.

Non content-based technologies such as social network analysis has become popular in the email filtering area. Network analysis approaches are address-based filtering approaches which aim to predict whether a sender is a potential spammer or not. This is different from the objective of the content and collaborative filtering techniques, which is to predict whether the message itself is spam or not. There is some evidence of the start of the use of these techniques for SMS filtering.

**Hybrid solutions:** With spam filtering there is no single solution that works. It is likely that some types of SMS spam can be better filtered by certain methods, so similar to the email domain, we see hybrid solutions as a promising avenue. Given that SMS filtering must happen under very strict processing time restrictions, content-based and collaborative filters could be usefully augmented with simple, less resource-intensive filtering methods such as blacklisting or traffic profiling.

**Advanced address-based filtering:** The move in recent times in email spam filtering has been towards advanced address-based filtering approaches including social network analysis and reputation-based filtering. These techniques should be considered in the mobile domain also but the lack of adequate data will hamper such efforts.

**Scalability and real-world deployment:** The work reviewed here represents research prototypes and solutions prepared under controlled laboratory conditions. Our benchmark implementation, which classifies messages in under 2 ms on average, is within the requirements for a system handling real-life SMS traffic volumes in terms of messages per second per node. This indicates that a filter based on support vector machines may be a feasible solution. For any real-world deployment however, the issues of scale and robustness become crucial, and high-speed databases, clustering, as well as efficient data structures and implementations will be required.

**Industry collaboration:** Progress in this field will have to be validated by real-world trial deployments, which only the network operators can facilitate. As volumes of spam increase, the promise of content-based filtering should make such collaboration attractive to industry.

**Experimental Design**

Base filters We selected for evaluation five spam filtering approaches that compare favourably with others in the literature:

Bogofilter, a widely deployed open source spam filter. Although dubbed ‘Bayesian’ Bogofilter employs a novel method for combining features that has come to be known as χ2. Bogofilter’s features consist of words consisting of alphanumeric sequences of characters in the text. Words appearing in specific header fields are distinguished from those appearing elsewhere. For example, the word “ platypus” would be treated as “ Subject:platypus” were it to occur in the subject field of the message, “From:platypus” were it to occur in the from field, and so on. Feature selection is effected separately on each message; the 300 features most strongly indicating each class are chosen.