**Spam Classification : Detecting Malicious Web Sites from Suspicious URLs**

By - Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker

Department of Computer Science and Engineering

University of California, San Diego

Team Members -

Swarna B – 01FB16EEC312

Safal Deepak Pansare – 01FB16EEC242

Swapneel – 01FB16EEC310

**Objective of the paper (from the original paper) -**

Malicious Web sites are a cornerstone of Internet criminal activities. As a result, there has been broad interest in developing systems to prevent the end user from visiting such sites. In this paper,

we describe an approach to this problem based on automated URL classification, using statistical methods to discover the tell-tale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly predictive models by extracting and automatically analyzing tens of thousands of features potentially indicative of suspicious URLs. The resulting classifiers ob-

tain 95–99% accuracy, detecting large numbers of malicious Websites from their URLs, with only modest false positives.

**Properties and methods being used in the original paper -**

Lexical features, Hos-Based Features, Computer Network features like – IP addresses, WHOIS, DNS, as well as location based properties.

Classification models used – Naive Bayes, Support Vector Machines, Logistic Regression

Dataset Source – random.yahoo, DMOZ open source directory, PhishTank as well as SpamScatter

**Our Approach for extension of the paper -**

We plan to use other methods like Gradient Boosting, Random Forest and Decision Trees for further improvement of the model. By doing so, we’ll be able to compare the performance of all the models and the best model with the most accurate results could be selected.

We also plan on making sure that after the model has done the classification, we are able to show the output to the user in the form of (say) 0 as a benign website and 1 as a malicious website.

Given that the internet space is ever expanding, we’ll try to link the newly updated databases in real-time to our model. So, that it can learn new websites as soon as they get added to the database.