SPAM FILTERING USING STATISTICAL NATURAL LANGUAGE PROCESSING

SHIVANI SHAH, NIMESH BUMB & KIRAN BHOWMICK

Department of Computer Engineering, D. J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

ABSTRACT

Spam is an important issue in the field of computer security because it is used to spread many other security hazards like worms, viruses or phishing. The increasing volume of spam has become the main problem concerning the email users and the classical way of filtering spam by blacklisting the spamming links is also not addressing this problem fully. This paper proposes the use of statistical Natural Language Processing (NLP) for purpose of building efficient Spam Filters. Statistical NLP uses Bayesian Classification, Maximum Entropy and Word Stemming algorithms to process the content and the sender link of the email, and based on the results classifies the email as legitimate or spam.

KEYWORDS: Spam, Spam Filters, Statistical Natural Language Processing, Bayesian Classification, Maximum Entropy and Word Stemming