Abstract

A large percentage of fraudulent spam mails are believed to originate from Nigeria or from Nigerians in remote locations. These mails (popularly referred to as 419 spa m) come in broad categories but all with the intent of defrauding the e-mail user. Testing the validity of senders and receivers address has proven to be effective in tracking spam mails. This approach will not filter out ordinary e-mails since typical e-mail users will always include their true e-mail addresses to facilitate replies. Checking the IP- addresses from where 419 mails originate is one way of ascertaining their actual origin in order to build a database of mails abuse, blacklisting keeping in mind that blacklisted IP addresses could be used to stop the delivery of for their mails from such addresses in foture. This research examines features selected specifically from the content analysis of Nigeria spam e-mail. A domain specific statistical content analysis tool (e-STAT) was developed and implemented using Bayesian statistical technique. e-STAT was tested and trained with a sizeable balanced corpus of Nigerian 419 spam e-mails and ham e-mails. Analysis of mail classification using e-STAT revealed current concept drift patterns among Nigerian 419 spammers and provided a blacklist of about 2,173 e-mail sender 's addresses, 563 URI,s within spam mails and a total of 13,491 bag-of-words common to Nigerian spam e-mails. The research obtained results that will guide foture research in the domain of 419 mails in designing effective spam filter s and electronic mail classifiers.

Keywords: Address, Bayesian, E-mai/s, Fraudulent, Nigeria, Statistical, 419 spam, Technique.