IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 116-119
www.iosrjournals.org

Spam Detection using Natural Language Processing

[1]Rohit Giyanani, [2]Mukti Desai
[1]*Thadomal Shahani Engineering College P. G. Kher Marg, (32nd Road), TPS-III off Linking Road. Bandra (West), Mumbai - 400050.*
[2]*D. J. Sanghvi College of Engineering Plot No.U-15, J.V.P.D. Scheme, Bhaktivedanta Swami Marg, Vile Parle (West), Mumbai-400 056.*

Abstract: *Spam mails can be referred as unsolicited bulk email. These messages are used to advertise products and services for phishing purposes or to lead recipients to malicious sites with unethical intentions. Although numerous techniques to block spam e-mails have been developed, we still receive them quite often. The reason behind this is mainly ability of the spammers to manipulate the filters. Therefore we present a method based on Natural Language Processing (NLP) for the filtration of spam emails in order to enhance online security. The technique presented in this paper is a stepwise approach which blocks spam emails based on the sender as well as the content of the mail.*

Keywords: *Natural Language Processing (NLP), spam detection, online security, spam filtering.*

I. Introduction

i) Spamming

With the popularity of the Internet, email is a part of our daily life. It is the most widely used medium for communication worldwide because of its cost effectiveness, reliability, quickness and easy accessibility. Email is prone to spam emails because of its wide usage and all of its benefits as a genuine medium of communication.

Internet Spam is one or more unsolicited messages sent or posted as a part of larger collection of messages, all having substantially identical content.

Most spam messages take the form of advertising or promotional materials like debt reduction plans, getting-rich quick schemes, gambling opportunities, pornography, online dating, health-related products etc.

The major technical disadvantages of spam messages are wastage of network resources (bandwidth), wastage of time, damage to the PC's & laptops due to viruses.

Spammers generally have a designed personalized template emails to deliver their messages using a bulk mailing software. It is widely assumed that most of the spam messages are sent directly from a collection of bots.

ii) NLP

NLP belongs to the CS taxonomy as the child of Artificial Intelligence (AI).

Natural Language Processing is a technique for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

'Naturally occurring texts' can be of any language, mode, genre, etc. The texts can be oral or written and must be in a language used by humans to communicate to one another. Significantly the text being analyzed should not be specifically constructed for the purpose of the analysis, but rather it should be collected from actual usage.

In simple terms, NLP is the use of computers to process written and spoken language for some useful purpose: to translate languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines.

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. In this paper, statistical approach is used for the proposed solution. Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena.

In this paper, section II explains about the different approaches that have been developed and are utilized for NLP; section III consists of the proposed model for spam detection using NLP engine; finally, section IV contains the conclusion for the same.

II. Related Work

i) N gram Modeling

An N-gram is an N-character slice of a longer string. Although ideally the term implies that notion of any co-occurring set of characters in a string can be included, but in this paper, we shall refer the term for contiguous but overlapping slices only.

N-grams of several different lengths simultaneously are used in the processing while detecting spam emails. Even blanks are appended to the beginning and ending of the string to support pattern matching efficiently.

Let us replace blank spaces by the underscore character ("_") for the n grams of the word "STAY".

Bi-grams: _S, ST, TA, SY, Y_

Tri-grams: _ _S, _ST, STA, TAY, AY_, Y_

Quad-grams: _ _ _ S, _ _ ST, _STA, STAY, TAY_, AY_ _, Y_ _ _

This approach is extremely beneficial as it allows the use of the same word or phrase in different context when their meanings vary substantially.

ii) Word Stemming

Content based spam filters are not beneficial if they are unable to understand the meaning of the words or phrases in an email. Nowadays, spammers change one or more characters of offensive words in their spam in order to penetrate content based filters. But the important thing to observe is that the spammers change the words in such a way that a human being can understand the meaning the words without any difficulty. Based on the above mentioned observations, a rule based word stemming technique that can match words those both look alike and sound alike is developed.

The following steps are commonly used as a word stemming algorithm:

1) Remove all non-alpha characters. (Allow som characters like '/' '\ ' '|' etc. which can be used together to look like some characters, such as \/ for 'V')
2) Remove all vowels from the word except the initial one.
3) Replace all digits and symbols by similar looking digits or characters.
4) Replace consecutive repeated characters by a single character and vice versa.
5) Use phonetic algorithms like sound ex on the resultant string.
6) Give it a numeric value depending on the operations performed over it.
7) Update the database for that particular keyword.

iii) Bayesian Classification

The purpose of spam filters is to decide whether an incoming message is legitimate (i.e., ham) or unsolicited (i.e., spam). There are many different types of filter systems, including:

Word lists: Simple and complex lists of words that are known to be associated with spam.

Black lists and white lists: These lists contain known IP addresses of spam and non-spam senders respectively.

The training phase of Bayesian spam filter maintains a database to keep a track of the total number of spam and ham messages to be used to train the Bayesian spam filter. The training phase of the filter consists of splitting the decoded message into single tokens, which are the words that make up the message. For each token, a record in the token database is updated that maintains two counts: the number of spam messages and the number of ham messages in which that token has been observed so far.

Once a Bayesian spam filter has created a token database, messages can be analyzed. Just like the training phase, the message is first decoded and split into single tokens. For each token, a spam probability is calculated based on the number of spam and ham messages that have contained this token out of the total number of spam and ham messages that have been used to train the Bayesian spam filter. The following formula is frequently used for this calculation:

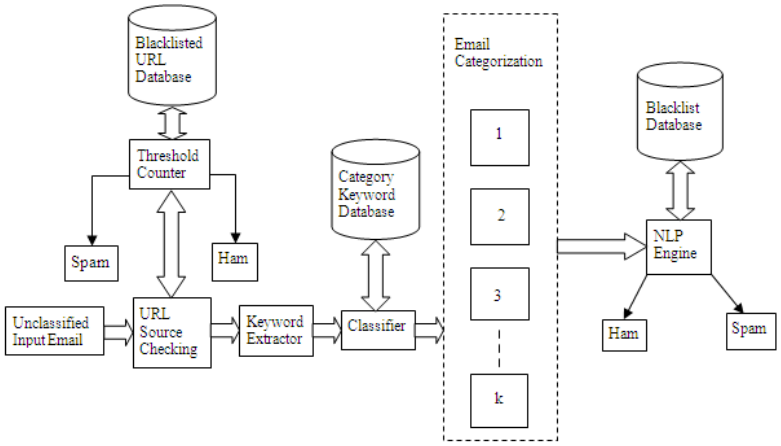
$$P_{spam}(token) = \frac{\frac{n_{spam}(token)}{n_{spam}}}{\frac{n_{spam}(token)}{n_{spam}} + \frac{n_{ham}(token)}{n_{ham}}}$$

III. Proposed Model

i) Components

Email Input: the unclassified input given to the spam detection model.
URL Source Check: checking the incoming Email's source URL.
URL Blacklist Database: the database containing all the URLs which have been detected during training phase to be spam.
Threshold Counter: This is a counter which keeps track of the number of emails sent over a period of time t.
Keyword Extractor: tokenizes the entire message into tokens and send these tokens as keywords to the classifier.
Classifier: works on Naive Bayes' Classification. It takes the keywords as input and classifies the email into a specific Category.
NLP Engine: It takes as input the unclassified Email and its category and then processes it further using the statistical NLP approach.

ii) Block Diagram



iii) Working

1. An unclassified email is taken as input for processing.
2. The URL Source checking block checks the URL of the incoming email and searches URL Blacklist Database to find a match. If the search is successful then, it categorizes the email as a spam or considers it to be ham and processes further.
3. At the same time there is a Threshold Counter which keeps track of the number of emails coming from this source over a period of time t.

When emails from a particular URL or IP address go beyond a specific large number, that is, this counter reaches its threshold value and will directly categorize the email as spam and along with that, it also updates the black list.

4. The Email is then passed to the Keyword Extractor which will tokenize the message into keywords.
5. These keywords are passed to the Classifier which will classify the mail into a specific category E.
6. The email is then passed to the NLP Engine along with its category. This is where the core process of Content Analysis takes place through various algorithms and techniques to give the classified output of the email being spam or legitimate.

IV. Conclusion & Future Scope

Spam mails are a serious concern to and a major annoyance for many Internet users. The mode proposed as a solution in this paper is highly beneficial because it introduces a threshold counter which helps overcome congestion on the web server and also maintain the spam filter efficiency but at the same time, it also requires overhead storage space for the databases.

Since NLP is a relatively underdeveloped area for research, further enhancements can be made in the field of spam detection for online security using Natural Language Processing in future.

References

[1]. Auto-Coding and Natural Language Processing by Richard Wolowitz - 3M Health Information System - White Paper 2011.
[2]. Security Focus Report – Spam in Today's Business World by TREND LABS – Global Technical Support and R & D Center of TREND MICRO - White Paper 2011.
[3]. Christoph Karlberger, Gunther Bayler, Christopher Kruegel, and Engin - "Exploiting Redundancy in Natural Language to Penetrate Bayesian Spam Filters" at Kirda Secure Systems Lab Technical University Vienna.
[4]. Shabbir Ahmed and Farzana Mithun – "Word Stemming to Enhance Spam Filtering" at Department of Computer Science & Engineering, University of Dhaka, Bangladesh.
[5]. "Blocking over 98% of Spam using Bayesian Filtering Technology", GFI Software, http://www.secinf.net/anti_spam/Blocking_Spam_Bayesian_Filtering.html, Oct. 2003.
[6]. R. Hall. "How to Avoid Unwanted E-Mail", Communications of the ACM, 41(3), 88-95 (1998).
[7]. William B. Cavnar and John M. Trenkle - "N-Gram-Based Text Categorization" at Environmental Research Institute of Michigan.