INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 4)
Available online at: www.ijariit.com

Advanced detection of spam and email filtering using natural language processing algorithms

Sujesh Shankar
sujesh14mandal@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu

ABSTRACT

Unsolicited bulk emails from random email addresses sent to a user's inbox are generally called junk or spam emails. 45% of all emails sent are spam and 14.5 billion spam emails are sent every single day. Around 36% of spam emails is content related to sales, advertising, and promotions that the recipient explicitly did not opt to receive. However, not all spam emails are used for this purpose. Spam emails are also sent for phishing purposes that deceive users and lead the recipients to malicious websites with unethical intentions. Numerous techniques have been developed to block such spam emails but a majority of users still receive them. This is because of the ability of the spammers to manipulate the filters. Spam costs businesses a whopping $20.5 billion every year. Even worse is that the cost of spam is likely to continue rising. Data indicates that losses to business will grow to $257 billion annually within a few years if the current rate of spam email is not decreased. To curb this problem, we present a method based on Natural Language Processing (NLP) for the filtration of spam emails in order to enhance online security. The technique proposed in this research paper is an approach which stepwise blocks spam mail based on the sender's email address along with the content of the email. This paper presents a proposed NLP system using N-gram model, Word Stemming algorithm and Bayesian Classification algorithm for detection of spam content and effectively filtering it.

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

1. INTRODUCTION

1.1 SPAM

Email is an essential part of our daily life as the internet becomes widely available and popular. Because of its cost-effectiveness, reliability, speed and easy accessibility, the internet has become the most widely used medium for communication worldwide. As the internet grows, our email and all of its benefits as a genuine medium of communication become equally prone to spam content. Internet spam is a phenomenon where one or more unsolicited emails are sent as a part of a larger collection of messages having substantially similar or identical content. Advertising or promotional material like getting rich schemes, debt reduction plans, pornography, online dating, gambling as well as health-related products is what spam usually comprises of. Wastage of network resources (bandwidth), waste of time, and damage to machines due to several viruses are the major technical disadvantages of spam mail. Generally, a personalized template email using bulk mailing software is what spammers use to deliver such emails. It's assumed that a collection of bots are developed for the purpose of sending such spam messages.

1.2 Natural Language Processing [NLP]

Natural Language Processing (NLP) belongs to the computer science taxonomy, also referred to as the child of Artificial Intelligence (AI). NLP is a technique to analyze and represent naturally occurring texts at multiple levels of linguistic analysis to achieve human-like language processing for the sole purpose of understanding and executing a wide range of tasks. Texts that occur naturally can be of any language, mode, and genre respectively. They also can be oral or written but must be in a language used by humans to communicate with other. Consequently, the text used for analysis should not be constructed specifically for the purpose of analysis but obtained and collected from actual usage. In simpler terms, using computers to process is written and orally spoken languages, to translate languages, obtaining information from the internet on text data banks to answer particular questions and to converse with machines is what Natural Language Processing is used for.

There are four main categories where Natural Language Processing falls into, Symbolic, Statistical, Connectionist, and Hybrid. We are proposing a Statistical approach in this research paper. A statistical approach leverages multiple mathematical techniques and uses a large corpus for developing generalized models of linguistic phenomena. Section 2.1 elaborates on the various relevant NLP models developed and used today whereas section 3 elaborates on the proposed model for spam filtering using the NLP engines. Section 4 explains the relevant conclusions of the proposed model.

As displayed in Figure 1, a typical Spam Filtering system consists of a natural language processing and artificial intelligence module. The diagram shows how queries are processed inside an NLP application. It is required to summarize input queries less than ten words before entering

them into the system. After that input queries are passed into the natural language processing module and all the queries are categorized as nouns/verbs and plural/singular terms. The gathered data is then passed into an artificial intelligence module to identify what commands must be executed. Post that, all the required commands are identified, combined and sent to the execution module. Ultimately the execution module executes all the requested commands.

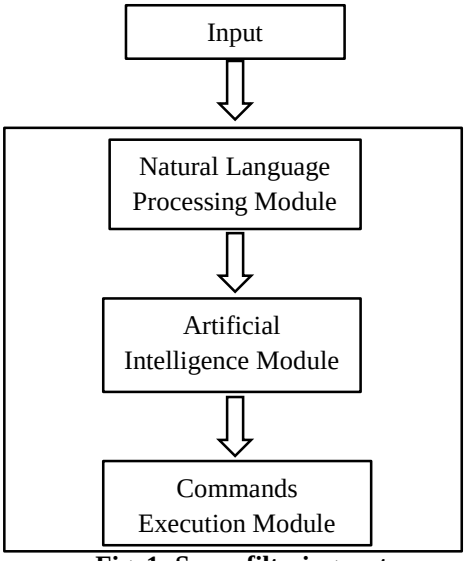

Fig. 1: Spam filtering system

2. LITERATURE REVIEW

Multiple research papers on Natural Language Processing were perused to gather information needed to develop a spam filtering application. In recent years, a considerable amount of resources have been deployed to develop intelligent systems. Majority of the applications are focussed on assisting users with artificial intelligence. The most popular personal assistant applications like Apple's Siri, Google Assistant, Cortana, Amazon Echo and ReQall work on NLP algorithms. Siri, the first modern virtual assistant for a mobile device [1] was developed by Apple. Commonly, these virtual assistant applications work via text or voice commands by executing tasks such as clicking and uploading photos. Natural Language Processing algorithms are used to match the text or voice commands. Manual tasks like taking notes, information organization, calendar scheduling, reminders and settings alarms are usually performed by such assistant applications. Breakthrough developments in Machine Learning, Artificial Intelligence, and Natural Language Processing truly contribute to building such personal assistant applications. Businesses can minimize the effort for customer servicing by using such applications as such systems are fast and efficient compared to humans. In the subsequent papers, a new language called CyberMate Scripting Language (CSL) is being developed, which can be used to model business information with a system [11].

To develop the natural language processing functionality for the spam filtering system, Part-of-Speech (POS) tagging module of NLP library is used. The POS tagger is an application that reads the text and assigns parts of speech to each word, nouns, verbs and adjectives [12] inclusive. A POS tag value is assigned to each of these words and returned by the natural language framework. The spam filter identifies these tags and generates a relevant PowerShell command for user input. Eventually, these generated commands are sent to the command execution module.

Table 1: NLP framework tags supported and meaning of corresponding POS tag term.

Tag	Description
IN	Proposition, Subordinating conjunction
JJ	Adjective
NN	Noun, Singular or Mass
NNS	Noun, Plural
NNP	Proper noun, Singular
POS	Possessive ending
PRP	Personal pronoun
RB	Adverb
RBR	Adverb, Comparative
RBS	Adverb, Superlative
VB	Verb, Base form
VBD	Verb, Past tense
VBG	Verb, Gerund or present participle
VBN	Verb, Past participle
VBP	Verb, Non-third person singular present
VBZ	Verb, Third person singular present

Natural Language Processing systems have Text Processing that processes text documents (typically unstructured texts) and involves a number of stages of processing.

1. Cleaning: Unwanted control characters are removed.
2. Tokenization: Breaking the stream of text, which are the minimal units of features, into tokens.
3. End-of-Sentence Detection: Sentence boundaries are identified and marked.
4. Part-of-Speech Tagging: Adding a tag that indicates a token for each part of speech.
5. Phrase Detection: Units that consist of multiple words are identified and marked – typically they consist of noun phrases, but need not always be nouns.
6. Entity Detection: Entities that consist of people names, places, organizations other proper nouns are identified and marked.
7. Categorization: Typically marks what category something belongs to; commonly categorization is used primarily for named entities, example: proper nouns.
8. Event Detection: Events are identified and marked. They generally correspond to verbs.
9. Relation Detection: Relations, which are connections between two or more entities or between entities and events are identified and marked.
10. Extraction: The identified entities are extracted from the document and stored externally. Entities like events, relations, and any other identified concepts (like dates) are included.

2.1. Existing NLP models

2.1.1. N-Gram modelling

N-gram is an N-character slice of a much longer string. The term implicates the notion of a co-occurring set of characters in a string that can be included. However, we will refer this term for contiguous overlapping slices only in this paper. While detecting spam emails, N-grams of different lengths are simultaneously used in the process. To support pattern matching efficiently, even blanks are appended to the beginning and end of the string.

Let's the underscore character ("_") for the n-grams of the word "LOVE" be replaced by the blank spaces:

Bi-grams: _L, LO, OV, LE, E_

Tri-grams: _ _L, _LO, LOV, OVE, VE_, E_

Quad-grams: _ _ _ L, _ _ LO, _LOV, LOVE, OVE_, VE_ _, E_ _ _

We can make use of the same word or phrase in a different context even when the meanings are different or vary substantially. Therefore this approach is beneficial.

2.1.2. Word stemming

Spammers have been routinely changing one or more characters of words in their spam content in order to penetrate content-based filters. This is why content-based filters are not efficient if they are not able to differentiate and understand the meaning of words or phrases in such spam emails. The most important thing to notice is that spammers alter words ways that are easily understandable to humans. A rule-based stemming technique is devised so that it can match words that look alike, sound alike and mean alike. The following steps are used in a word stemming algorithm:

1. All Non-Alpha Characters are deleted. Some characters like '|', '\', '/', etc can be used together to look like characters, for example, /\/ for 'N'
2. Vowels are all deleted except the initial word.
3. Similar looking digits and characters are all replaced by other similar digits and characters.
4. Characters that are repeated consecutively are replaced by a single character and vice versa.
5. Phonetic syntax or models, for example, sound ex, are used on the resultant string.
6. Depending on the operations performed over it, a numeric value is assigned.
7. The database is updated for that specific keyword.

2.1.3. Bayesian classification

Spam filters are set for the sole purpose of analyzing an incoming message and classifying if its legitimate (ham) or unsolicited (spam). There are many different types of filter systems, including:

1. Word lists: Simple and complex lists of words that are known to be associated with spam.
2. Blacklists and whitelists: These lists contain known IP addresses of spam and non-spam senders respectively.

A database is maintained by training the Bayesian spam filter to store and track the total number of spam and ham messages. Splitting the decoded message into single tokens, words which make up the message is performed during the training phase of the filter. Every token has a record in the database of tokens and is subsequently updated for two counts, during the number of spam messages and during the number of ham messages in which that token has been observed, respectively. Once the token database has been created by the Bayesian spam filter, messages can then be analyzed. The message is first decoded and split into the token. A spam probability is calculated for each token 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 formula used frequently for this calculation is as follows:

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

$N_{spam}(token)$ and $N_{ham}(token)$: The total number of times a token appeared in a spam or ham mail.
N_{spam} and N_{ham} : The total numbers of spam and ham tokens in this formula.

3. PROPOSED SYSTEM

3.1 Components

1. Email Input: An unclassified input given to the spam detection model.
2. URL Source Check: Checking the incoming email source URL.
3. URL Blacklist Database: Database containing all the URLs which have been detected during the training phase to be spam.
4. Threshold Counter: A counter which keeps track of the number of emails sent over a period of time T.
5. Keyword Extractor: The entire message is tokenized into tokens and sends these tokens as keywords to the classifier.
6. Classifier: This works on the Naive Bayes Classification. It takes as input the keywords and classifies the email into a specific category.
7. NLP Engine: The unclassified email is taken as input and its category is then processed further using the statistical NLP approach.

3.2. Block diagram

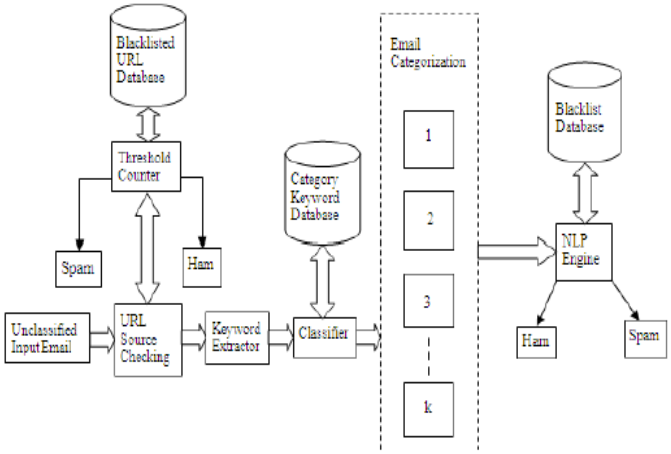

Fig. 2: Block diagram

3.3. WORKING

1. The initial input taken for processing is an unclassified email.
2. The URL source checking block searches the incoming email URL in the URL blacklist database to find a match. If the match is successful, the email is categorized as spam. If not, it is considered to be ham and processed further.
3. During the time step 1 and 2 are performed, a Threshold Counter keeps track of all emails coming from various sources over a period of time T. If emails from a particular source URL or IP address exceed the threshold value, which in itself is a large value, the email is categorized as spam and added to list of spam URLs in the blacklist database.
4. The email is fed to a Keyword Extractor which tokenizes the messages into specific keywords.
5. The keywords are then passed on to a Classifier which classifies these emails into a specific category, for example, Category Z.

6. The category Z emails are then passed to the NLP engine where the core process of content analysis takes place. Various NLP algorithms and techniques give the classified output of the email to be spam or legitimate.

4. CONCLUSION

Spam emails are a serious concern as they can hurt productivity and could also result in security breaches. Apart from this, spam emails are a major source of annoyance for many users on the internet. This proposed solution introduces a threshold counter which overcomes the congestion problem caused on the web server and helps maintain the spam filter efficiency. Thus, it is highly beneficial. However, it also requires overhead storage space for the databases. Since Natural Language Processing is a relatively underdeveloped area for research, further enhancements can be made to the proposed system for spam detection and email filtering in the field of online security.

5. REFERENCES

[1] 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.

[2] "Blocking over 98% of Spam using Bayesian Filtering Technology", GFI Software, http://www.secinf.net/anti_spam/Blocking_Spam_Bayesian_Filtering.html, Oct. 2003.

[3] R. Hall. "How to Avoid Unwanted E-Mail", Communications of the ACM, 41(3), 88-95 (1998).

[4] William B. Cavnar and John M. Trenkle - "N-Gram-Based Text Categorization" at Environmental Research Institute of Michigan.

[5] Jeff Dunn. (6. 2, 2011). iPhone 4S hands-on. [Online]. Available: https://www.engadget.com/2011/10/04/iphone-4s-handson/

[6] 25 Tasks You Can Outsource to a General Virtual Assistant. [Online]. Available: http://www.chrisducker.com/25tasks/

[7] N. T. Weerawarna; H. M. H. R. B. Haththella; A. R. G. K. B. R. Ambadeniya; L. H. S. S. Chandrasiri, CyberMate ~ Artificial Intelligent business help desk assistant with instant messaging services, 16-19 Aug.2011.

[8] Stanford NLP Group, Stanford Log-linear Part-Of-Speech Tagger. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml

