Data Pre-Processing in Spam Detection

Anjali Sharma
Bansthali Vidyapith, Jaipur Campus

Manisha
Bansthali Vidyapith, Jaipur Campus

Dr. Manisha
Bansthali Vidyapith, Jaipur Campus

Dr. Rekha Jain
Bansthali Vidyapith, Jaipur Campus

Abstract

Nowadays, most of the people have access to the Internet, and digital world has become one of the most important parts of everybody's life. People not only use the Internet for fun and entertainment, but also for business, banking, stock marketing, searching and so on. Hence, the usage of the Internet is growing rapidly. One of the threats for such technology is spam. Spam is a junk mail/message or unsolicited mail/message. Spam is basically an online communication send to the user without permission. Spam has increased tremendously in the last few years. Today more than 85% of mail /messages received by users are spam. These days, spam is a very serious problem because spamming has become a very profitable business for spammers. Spam email takes on various forms like adult content, selling products or services, job offers etc. Spam costs the sender very little to send but most of the costs are paid by the recipient or the service providers rather than by the sender The cost of spam can also be measured in lost human time, lost server time and loss of valuable mail/messages. In filtering of spam, the data cleaning of the textual information is very critical and important. Main objective of data pre-processing in spam detection is to remove data which do not give useful information about the class of the document. In this paper, the focus is on various pre-processing steps of text data such as noise elimination, stop word removal, and stemming. For stemming, Porter's algorithm has been used. Further, some results, after applying all the data pre-processing steps have been displayed.

Keywords: Spam, Spam detection, Data pre-processing

I. INTRODUCTION

Spam refers to unsolicited commercial email. Also known as junk mail, spam floods Internet users electronic mailboxes. These junk mails can contain various types of messages such as pornography, commercial advertising, doubtful product, viruses or quasi legal services [1].

II. SPAM DETECTION STEPS

The basic steps of spam detection are classified as Data Pre-processing Steps, Representation of Data, and Classification.
In this paper we discuss data cleaning steps mainly.

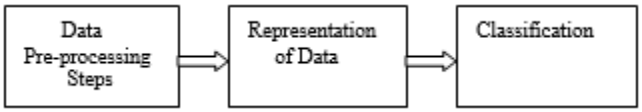

Fig. 1: Main Steps in the Spam Detection

A. Data Pre-Processing Steps:

In filtering of spam, the pre-processing of the textual information is very critical and important. Main objective of text data pre-processing is to remove data which do not give useful information regarding the class of the document. Furthermore we also want to remove data that is redundant. Most widely used data cleaning steps in the textual retrieval tasks are removing of stop words and performing stemming to reduce the vocabulary [2]. In addition to these two steps we also removed the words that have length lesser than or equal to two.

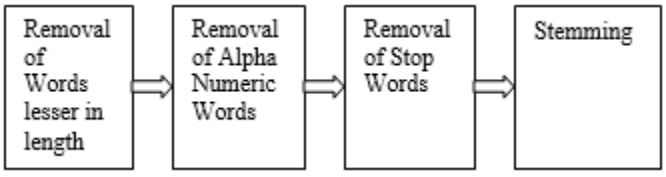

Fig. 2: Data Pre-processing Steps

B. Representation of Data:

The Next main task was the representation of data. The data representation step is needed because it's very hard to do computations with the textual data. The representation should be such that it should reveal the actual statistics of the textual data. Data representation should be in a manner so that the actual statistics of the textual data is converted to proper numbers. Furthermore it should facilitate the classification tasks and should be simple enough to implement.
There exist many term weighting methods which will calculate the weight for term differently such as Boolean Weighting, Term frequency, Term Document Frequency inverse document frequency (TF-IDF).

C. Classification:

In Simple terms classification is a task of learning data patterns that are present in the data from the previous known instances and associating those data patterns with the classes. Later on when given an unknown instance it will search for data patterns and thus will predict the class based on the absence or presence of data patterns.

III. METHODOLOGY OF DATA PRE-PROCESSING IN SPAM DETECTION

The basic data pre-processing steps of spam detection are:

1) The words having length <=2 are removed.
2) All the special characters are removed.
3) Stop words are removed.
4) Porter's Stemming Algorithm is applied to bring the word in their most basic form.
5) The word frequency of all the words.
6) The normalized term frequency of all the words.
7) The inverse document frequency of all the words.
8) Term Document Frequency inverse document frequency (TF-IDF)

A. Removal of Words Lesser in Length:

Investigation of English vocabulary shows that almost all such words whose length are lesser than or equal to two contains no useful information regarding class of the document. Examples includes a, is, an, of, to, as, on etc. though there are words which have length of three and are useless like the, for, was, etc but removing all such words will cost us losing some words that are very useful in our domain, like sex, see, sir, fre (often fre is used instead of free to deceive the automatic learning filter). All email of the data set were passed through a filter which removed the words that have length lesser than or equal to two. This removed bundle of words from the corpus that were useless and reduced the size of the corpus to great extend.

For example,

- Input string x= " Do humans only use ten percent of their brains ? "
- Output string x= " humans only use ten percent their brains ? "

B. Removal of Alpha Numeric Words:

There were many words found in the corpus that were alpha numeric. Removal of those terms was important as they do not keep on repeating in the corpus and they are just added in the emails to deceive the filter so that our classifier fails to find patterns in the given email.

They do not keep on repeating in the email instances. In this sense they can be considered as unique terms. They are present in large numbers in the corpus and adding them to our features set will have drastic increase in the features set size with little of information.

Counting the number of alpha numeric words in subject line or in the entire email might be helpful as spam's are reported to contain large number of alpha numeric words. So a single feature containing the number of alpha numeric words in an email might be helpful.

For example,

- Input string x= " humans only use ten percent their brains ? "
- Output string x= humans only use ten percent their brains

C. Removal of Stop Words:

In information textual retrieval there are words that do not carry any useful information and hence are ignored during spam detection. In general and for document classification tasks we consider them as words intended to provide structure of the language rather than the content and mostly include pronouns, prepositions and conjunctions.

For example,

- Input string x= humans only use ten percent their brains
- Output string x= humans ten percent brains

Table – 1
Stop Word List for Experiment Set

then, there, that, which, the, those, now, when, which, was, were, been, had, have, has, will, subject, here, they, them, may, can, for, such, and, are, but, not, with, your, alone, anyways, along, anywhere, able, already, apart, about, also, appear, above, although, appreciate, according, always, appropriate, between, be, beyond, became, both, because brief, become, but, becomes, by, becoming, before

D. Stemming:

The main pre-processing tasks applied in textual information retrieval tasks is the stemming. Stemming is a process of reducing words to its basic form by stripping the plural from nouns (e.g. "apples" to "apple"), the suffixes from verbs (e.g. "measuring" to "measure") or other affixes [3]. For example, applies, applying & applied matches apply. Originally proposed by Porter on 1980, it defines stemming as a process for removing the commoner morphological and in-flexional endings from words in English [5]. In the context of document classification we can define it to be a process of representing words and its variants with its root. We used the porter stemming algorithms.

For example,

- Input string x= humans ten percent brains
- Output string x= human ten percent brain

Table2 shows some examples of the words after being stemmed with porter's algorithm.

Table – 2:
Few Examples of Words with their Stems

Words	*Stem*
ponies	*poni*
caress	*caress*
cats	*cat*
feed	*fe*
agreed	*agre*
plastered	*plaster*
motoring	*motor*
sing	*sing*
conflated	*conflat*
troubling	*troubl*
sized	*size*
hopping	*hop*
tanned	*tan*
falling	*fall*

E. Term Frequency:

Term frequency counts the number of occurrences of term in a text document [6].

Mathematically it can be represented as:

Term _ Frequency _ Wij = tfij ….Eq(1)

where, tfij as the frequency of term i in document j

F. Term Document Frequency Inverse Document Frequency (TF-IDF):

Tf-Idf weighting represent that those terms whose presence is in lesser number of text documents(e-mails) can discriminate well between the classes[4].

In Tf-Idf, we have found normalized term frequency, inverse document frequency and Tf-Idf of each word in document.

$$TFIDF\_W_{ij} = tf_{ij} \times idf$$

$$TFIDF\_W_{ij} = tf_{ij} \times log\ (N/n_i) \qquad \ldots.Eq(2)$$

where, tf_{ij} is normalized term frequency

tf_{ij} = *Number of times term t appears in a document) / (Total number of terms in the document)*

N is the total number of documents or emails in the corpus

n_i is the number of documents in the corpus where term i appears.

IV. EXPERIMENTS AND RESULTS

In this section, the data sets of selected emails that used to conduct experiments are presented. Next, a pre-processing of our data by the system are given. Finally, a set of experiments are presented followed by the results and their discussion.

A. The Data Sets:

Data set of five different spam e-mails is used to conduct experiments. Table 3 shows data sets of five spam emails.

Table – 3:
The Data Set of Spam emails (corpora)

Email 1	*Call FREEPHONE 0800 542 0578 now !*
Email 2	*85233 FREE > Ringtone! Reply REAL*
Email 3	*FROM 88066 LOST £12 HELP*
Email 4	*2/2 146tf150p*
Email 5	*ringtoneking 84484*

B. Workflow of Data Pre-processing of Emails:

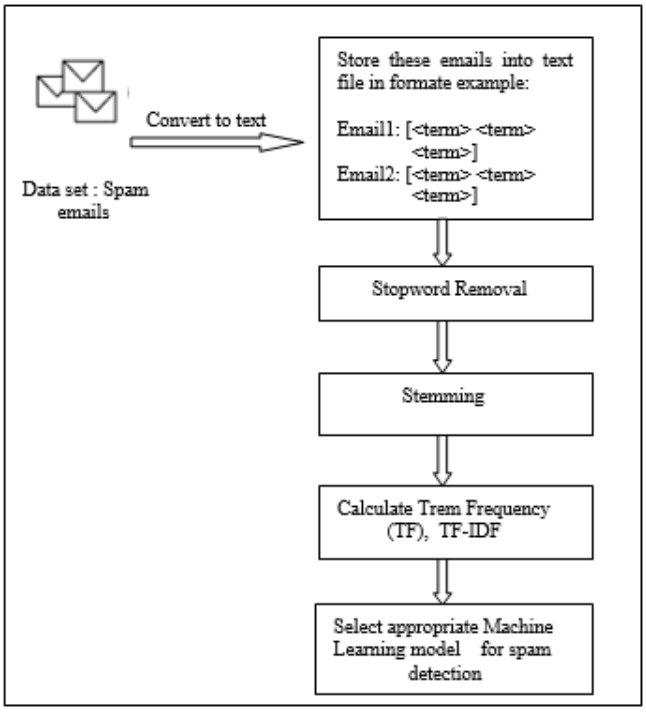

Fig. 3: Workflow of Data Pre-processing in Spam Detection

C. Results:

1) *Email 1:*

Call FREEPHONE 0800 542 0578 now !

Email Terms	*0578*	*0800*	*542*	*call*	*freephon*	*now*
Word Frequency	*1*	*1*	*1*	*1*	*1*	*1*
Normalized Term Frequency	*0.166*	*0.166*	*0.166*	*0.166*	*0.166*	*0.166*
TF_!DF	*0.116*	*0.116*	*0.116*	*0.116*	*0.116*	*0.116*

2) *Email 2:*

85233 FREE > Ringtone ! Reply REAL

Email Terms	*85233*	*free*	*rington*	*repli*	*real*
Word Frequency	*1*	*1*	*1*	*1*	*1*
Normalized Term Frequency	*0.2*	*0.2*	*0.2*	*0.2*	*0.2*
TF_IDF	*0.047*	*0.047*	*0.047*	*0.047*	*0.047*

V. CONCLUSION

In this paper, we have seen that pre-processing steps play a crucial role in classifying emails as spam or non-spam emails given. Words in the message are pre-processed before using classifier. The word goes through stop words removal steps and then stemming process step to extract the word root or stem.

REFERENCES

[1] Masurah Mohamad, Khairulliza Ahmad Salleh, "Independent Feature Selection as Spam-Filtering Technique: An Evaluation of Neural Network", Malaysia.

[2] Nouman Azam, "Comparative Study of Features Space Reduction Techniques for Spam Detection", Department of Computer Engineering College of Electrical and Mechanical Engineering National University of Sciences and Technology.

[3] Thamarai Subramaniam, Hamid A. Jalab and Alaa Y. Taqa , "Overview of textual anti-spam filtering techniques", International Journal of the Physical Sciences Vol. 5(12), pp. 1869-1882, 4 October, 2010.

[4] M. Basavaraju, Dr. R. Prabhakar, "A Novel Method of Spam Mail Detection using Text BasedClustering Approach ", International Journal of Computer Applications (0975 – 8887) Volume 5– No.4, August 2010.

[5] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin, **"**Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks" , IJCSI International Journal of Computer Science Issues, Egypt**,** Vol. 10, Issue 2, No 1, March 2013.

[6] El-Sayed M. El-Alfy, "Learning Methods For Spam Filtering", College of Computer Sciences and Engineering King Fahd University of Petroleum and Minerals, Saudi Arabia, ISBN: 978-1-61122-759-8 2011 Nova Science Publishers, Inc.