ARTIFICIAL INTELLIGENCE PROJECT

Spam Detection using Natural Language Processing

Table of Contents

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.

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

1. Introduction

Nowadays, e-mail provides many ways to send millions of advertisement at no cost to sender. As a result, many unsolicited bulk e-mail, also known as spam e-mail spread widely and become serious threat to not only the Internet but also to society. For example, when user received large amount of e-mail spam, the chance of the user forgot to read a non-spam message increase. As a result, many e-mail readers have to spend their time removing unwanted messages. E-mail spam also may cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content. Over the past many years, many approaches have been provided to block e-mail spam .

For filtering, some email spam are not being labelled as spam because the e-mail filtering does not detect that email as spam. Some existing problems are regarding accuracy for email spam filtering that might introduce some error. Several machine learning algorithms have been used in spam e-mail filtering, but Na¨ıve Bayes algorithm is particularly popular in commercial and open-source spam filters [2]. This is because of its simplicity, which make them easy to implement and just need short training time or fast evaluation to filter email spam. The filter requires training that can be provided by a previous set of spam and non-spam messages. It keeps track of each word that occurs only in spam, in non-spam messages, and in both. Na¨ıve Bayes can be used in different datasets where each of them has different features and attribute3.

2. Methodology :

This section describes the methodology that is used for the research. The methodology that is used for the filtering method is machine learning techniques that divide by three phases.The methodology is used for the process of e-mail spam filtering based on Na¨ıve Bayes algorithm.

3.1. Na¨ıve Bayes classifier

The Na¨ıve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset [4]. In this research, Na¨ıve Bayes classifier use **bag of words** features to identify spam e-mail and a text is representing as the bag of its word. The bag of words is always used in methods of document classification, where the frequency of occurrence of each word is used as a feature for training classifier. This bag of words features are included in the chosen datasets. Na¨ıve Bayes technique used Bayes theorem to determine that probabilities spam e-mail. Some words have particular probabilities of occurring in spam e-mail or non-spam e-mail. Example, suppose that we know exactly, that the word **Free** could never occur in a non-spam e-mail. Then, when we saw a message containing this word, we could tell for sure that were spam email. Bayesian spam filters have learned a very high spam probability for the words such as Free , but a very low spam probability for words seen in non-spam e-mail, such as the names of friend and family member. So, to calculate the probability that e-mail is spam or non-spam Na¨ıve Bayes technique used Bayes theorem as shown in formula below.

$$P(spam \mid word) = \frac{\mathrm{P(spam).P(word|spam)}}{\mathrm{P(spam).P(word|spam) + P(non-spam).P(word|non-spam)}}$$

Where:

(i) P(spamword) is probability that an e-mail has particular word given the e-mail is spam.

(ii) P(spam) is probability that any given message is spam.

(iii) P(wordspam) is probability that the particular word appears in spam message.

(iv) P(non−spam) is the probability that any particular word is not spam.

(v) P(wordnon − spam) is the probability that the particular word appears in non-spam message.

To achieve the objective, the research and procedure is conducted in three phases. The phases involved are as follows:

(i) Phase 1: Pre-processing

(ii) Phase 2: Feature Selection

(iii) Phase 3: Na¨ıve Bayes Classifier

The following sections will explain the activities that involve in each phases in order to develop this project. Figure shows the process for e-mail spam filtering based on Na¨ıve Bayes algorithm.

3.2. Pre-processing

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values . Pre-processing of e-mails in next step of training filter, some words **like conjunction words, articles are removed from email body because those words are not useful in classification.**

As mentioned earlier, we are using **nltk tool** to facilitate the experiments.

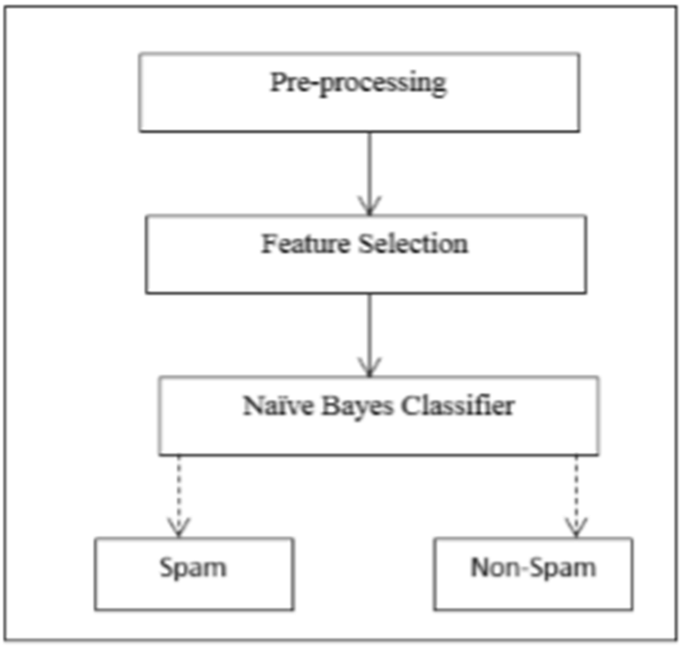


Figure 2. Process of E-mail spam filtering based on Naïve Bayes Algorithm

3.3 Feature Selection :

After the pre-processing step, we apply the feature selection algorithm, the algorithm which deploy here is Best First Feature Selection algorithm.

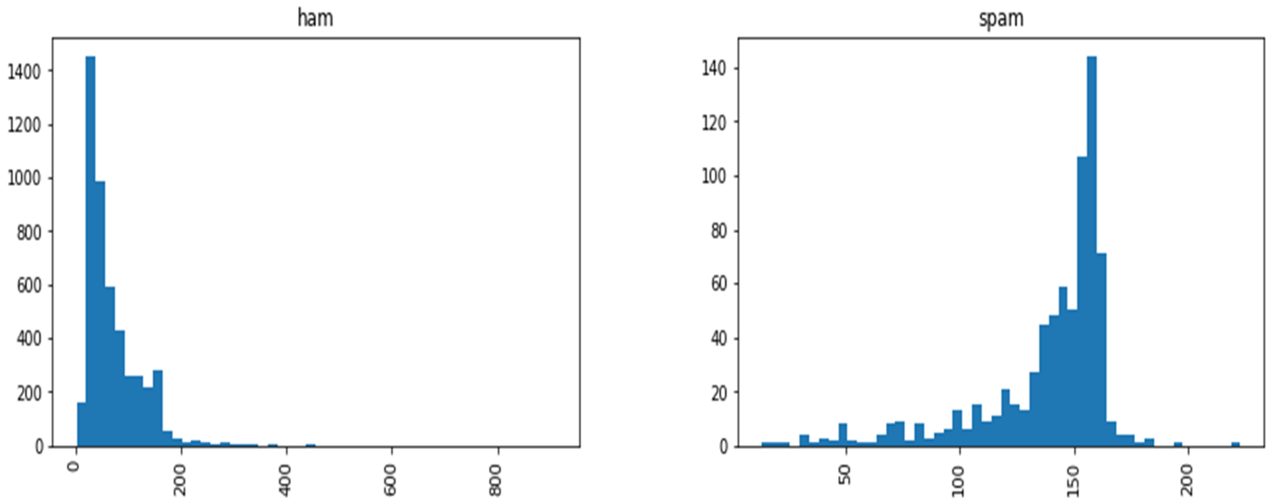
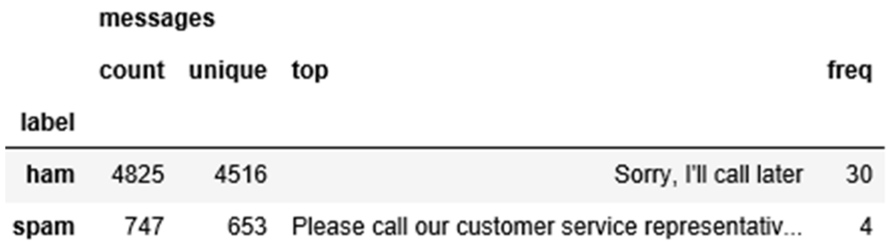


Figure:- figure showing the histogram plot **length vs frequency** of the both type of messages **'ham'** and **'spam'**.



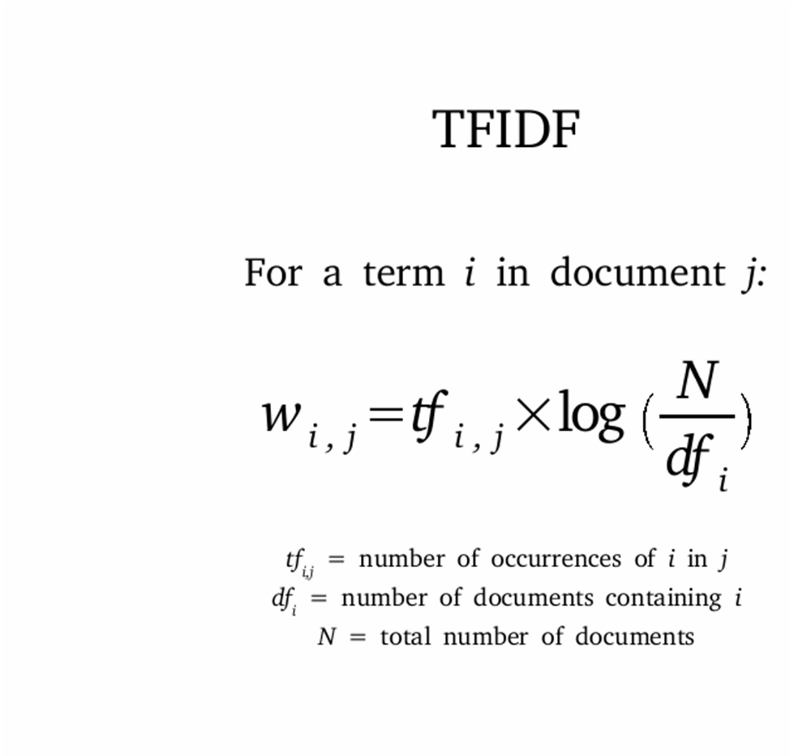
	messages			
	count	unique	top	freq
label				
ham	4825	4516	Sorry, I'll call later	30
spam	747	653	Please call our customer service representativ...	4

Figure:- this figure showing total number of ham and spam messages.

TFIDF(term frequency–inverse document frequency)

In information retrieval, **tf–idf** or **TFIDF**, short for **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. Tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.[2]

Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. tf–idf can be successfully used for **stop-words** filtering in various subject fields, including text summarization and classification.



TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

4. Experimental Setup

The experimental setting of the research is like follows:

4.1. The Evaluation Metric:

The Evaluation Metric Evaluation metrics are used to evaluate the performance of the spamCollection dataset.The most simple measure is filtering accuracy namely percentage of messages classified correctly.

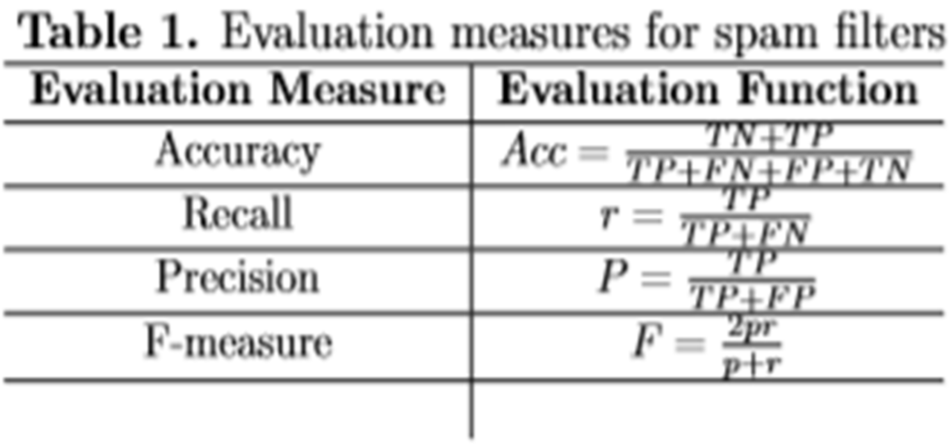


Table 1. Evaluation measures for spam filters

Evaluation Measure	Evaluation Function
Accuracy	$Acc = \frac{TN+TP}{TP+FN+FP+TN}$
Recall	$r = \frac{TP}{TP+FN}$
Precision	$P = \frac{TP}{TP+FP}$
F-measure	$F = \frac{2pr}{p+r}$

Where accuracy, recall, precision, F-measure, FP, FN, TP and TN are defined as follows:

(i) **Accuracy:** Accuracy is calculated as the total number of correct prediction divided by the total number of dataset.

(ii) **Recall:** Percentage spam message manage to block

(iii) **Precision:** Percentage of correct message for spam e-mail

(iv) **F-measure/F-score:** Weighted average of precision and recall

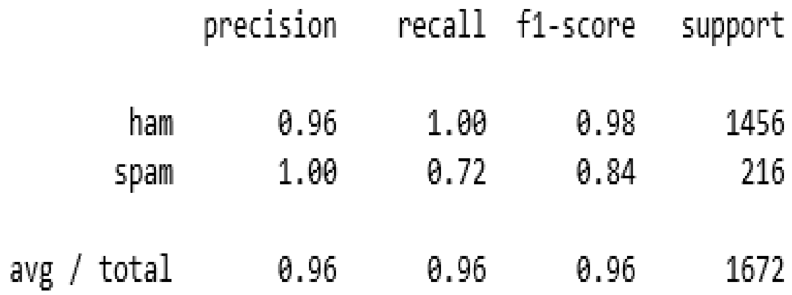
(v) **False Positive Rate (FP):** The number of misclassified non spam emails

(vi) **False Negative Rate (FN):** The number of misclassified spam emails

(vii) **True Positive (TP):** The number of spam messages are correctly classified as spam

(viii) **True Negative (TN):** The number of non-spam e-mail that is correctly classified as non-spam

RESULT:



```
             precision    recall  f1-score   support

        ham       0.96      1.00      0.98      1456
       spam       1.00      0.72      0.84       216

avg / total       0.96      0.96      0.96      1672
```

FIGURE:- Figure shows the evaluation measures of for spam filters using Naiive Bayes algorithm.

6. Conclusion

E-mail spam filtering is an important issue in the network security and machine learning techniques; Na¨ıve Bayes classifier that used has a very important role in this process of filtering e-mail spam. The quality of performance Na¨ıve Bayes classifier is also based on datasets that used. As can see, dataset that have fewer instances of e-mails and attributes can give good performance for Na¨ıve Bayes classifier.

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.

PROJECT REPORT

Prepared By - MAHESH SINGH DASILA

CSE-A

160102022