Spam detection

Laruina Mirko
Artificial Intelligence and Data Engineering

Motivation

Being in a connected world means staying in touch with friends and colleagues with many and different forms of communication. For school and business related activities, the email is the preferred mean and we would want its experience to be as smooth as possible.

Mail providers have been offering spam filtering services for many years now. The reason is simple: we don't want to receive mails which are uncalled for and that may expose ourselves to all sort of scams and rip-offs.

The aim of this project is to build an IMAP mail client application supported by a spam classification system guided by Artificial Intelligence

Dataset

Raw email data, labeled as spam or "ham", are quite common and can be found over the Internet.

For our purposes, we will use mails coming from two different sources.

SpamAssassin:

- https://www.kaggle.com/veleon/ham-and-spam-dataset
- 2501 ham, 501 spam

Enron:

- http://www2.aueb.gr/users/ion/data/enron-spam/
- 19088 ham, 32988 spam

Pre-processing (I)


...
From: "Justin MacCarthy" <macarthy@iol.ie>
To: "Ilug@Linux.Ie" <ilug@linux.ie>
Date: Thu, 29 Aug 2002 16:37:26 +0100
...
Subject: [ILUG] Looking for a file / directory in zip file
Sender: ilug-admin@linux.ie
Errors-To: ilug-admin@linux.ie
X-Mailman-Version: 1.1
Precedence: bulk
List-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

Is there a way to look for a particular file or directory in 100's of zip files??
Something like zgrep but for the filename instead of a word
Thanks Justin

Since we are using raw email data, we have to clean them by removing unnecessary SMTP headers (which are also different between the two datasets) and produce a data format we can actually use.

We will preserve info on the sender, receiver, the subject of the mail and its body. A binary attribute "html" will mark the mails appropriately. The label will be expressed under the attribute "spam"

Sender	Receiver	Subject	Html	Body	Spam
macarthy@iol.ie	ilug@linux.ie	[ILUG] Looking for a file / directory in zip file	False	Is there a way to look for a particular file or directory in 100's of zip files?? …	False

Pre-processing (II)

In a second phase, all the mails containing HTML will be parsed and only the text content of the body will be used in the final dataset

```
 <td width=3D"242">=20
                <b><font size=3D"2" face=3D"Verdana,
Arial, =
Helvetica, sans-serif">
                <ul>
                   <li>No Hidden Fees</li>
                   <li>Over 16,000 Credits</li>
                   <li>Over 300 Courses</li>
                </ul>
                </font></b>
               </td>
```

No Hidden Fees
Over 16,000 Credits
Over 300 Courses

Text-mining process

A text-mining approach will be applied to obtain a vector representation of the string content, which will be needed to apply the classification algorithms.

Steps:

- Tokenization
- Stop-word filtering
- Stemming
- Feature representation

Classification (I)

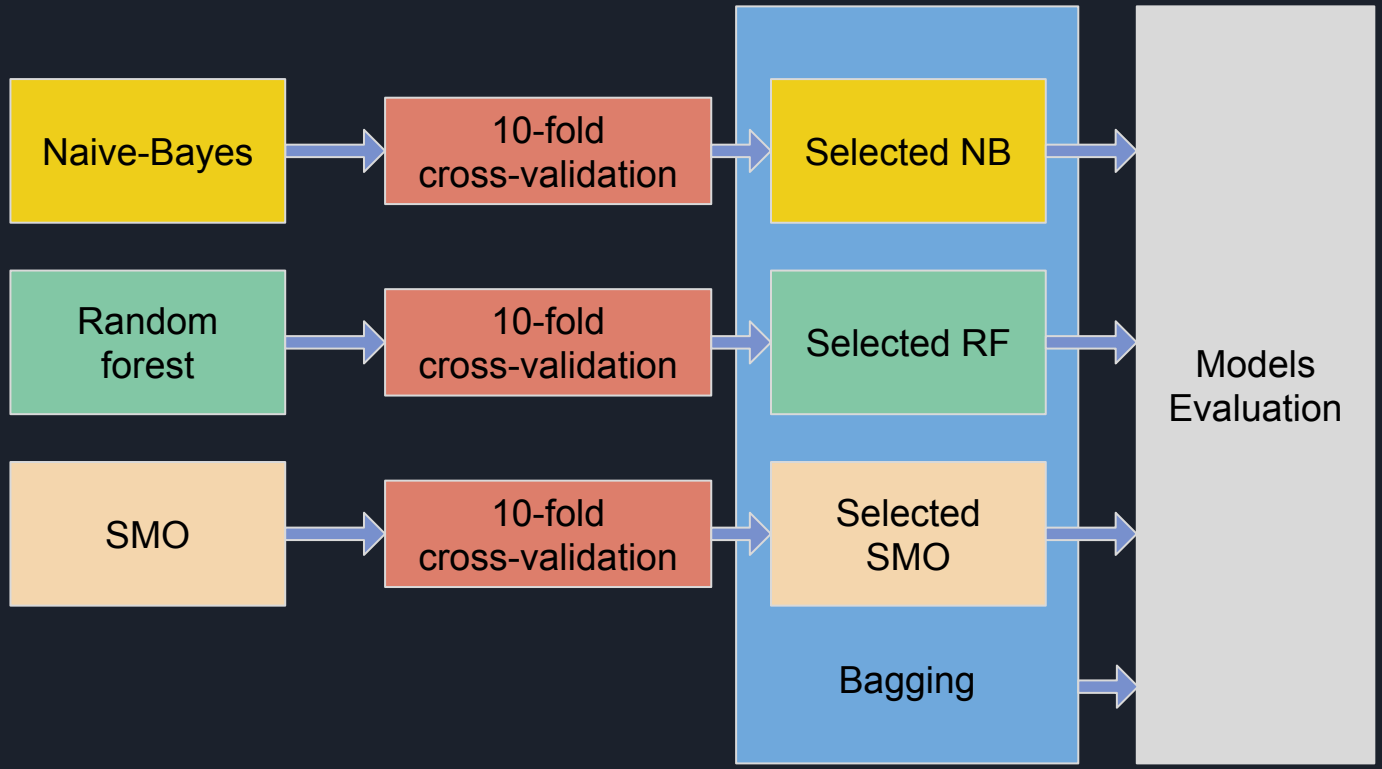
Multiple classifiers (Naive Bayes, Random Forest and SMO) will be trained with the previously processed data.

For each type of classifier, we will apply 10-fold cross validation, selecting the classifier with the least variance and the more generalization capability.

An ensemble bagging classifier (with majority vote) will be built with the best classifier (per type).

All the classifiers will be evaluated and compared to select the one more suited for our needs.

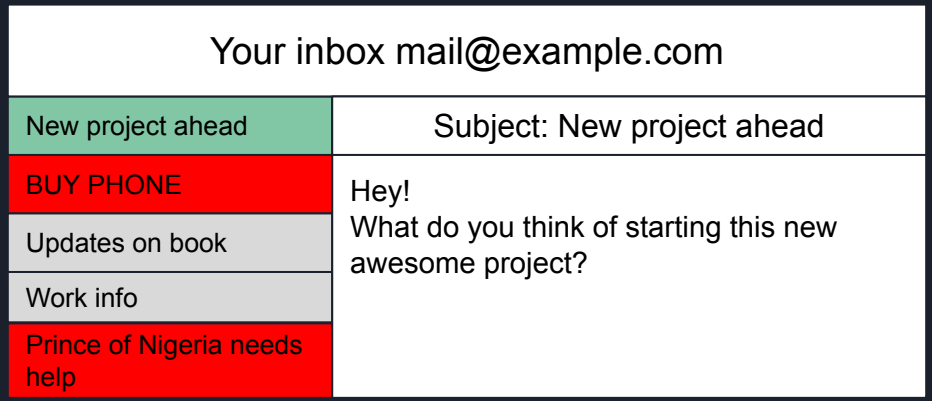
Classification (II)



Application

The application will consist of a simple mail client (using IMAP protocol) with a list of mails and the possibility to read their content.

Through our classifier, the mail detected as spam will be marked in red in the list, while a warning will be displayed when it is opened.



Thank you!