Spam E-mail Analysis

Issue

Large quantities of junk email that contain everything from:

- Advertisements for products/web sites, make money fast schemes, chain letters.
- Invitations to join in a fantastic new venture to pornography, frequently with explicit sexual language and attached photographs.

Those spam emails are not just annoying and, in some cases, even offensives, they are dangerous, it consumes network traffic and mail servers' storage. Furthermore, spam has become a major component of several attack vectors including attacks such as phishing, cross-site scripting, cross-site request forgery and malware infection.

Spamming takes many forms, through SMS messages and also through emails. It is possible to classify messages, either emails or SMS as spam or non-spam (ham in this case) using machine learning.

Business question

How to identify which emails are spam or ham (non-spam)?

Goal

Determine whether a given email is spam or not based on words indicators by the creation of an algorithm

Hypotheses

- The most spam received the most difficult to identify desired email.
- The most effective countermeasures the less spam you get.
- Unsolicited e-mail consisted mostly of messages from prankers, chain letters, and inappropriate messages.
- With better filtering the spammers will be out of business.
- The most common spam words are for free – fantastic deal – money making – cash – incredible deal – save big money -stock alert – eliminate debt – get paid – Notspam – we have spam – join millions – you're a winner.
- The most straightforward filtering solutions involve filtering messages from known spam senders based on information in message headers.

Collect and manage data

The dataset contains a significant number of instances, 5,574 records that have previously been labeled as either spam or not-spam (ham).

The file will consist of two columns. One with the labels and another with the messages or data.

```
v1,v2,,,
ham,"Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amo
ham,Ok lar... Joking wif u oni...,,,
spam,Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry
ham,U dun say so early hor... U c already then say...,,,
ham,"Nah I don't think he goes to usf, he lives around here though",,,
spam,"FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it st
ham,Even my brother is not like to speak with me. They treat me like aids patent.,,,
ham,As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune
spam,WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim
spam,Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for
ham,"I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough
spam,"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days,
spam,"URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to
ham,I've been searching for the right words to thank you for this breather. I promise i wont take your help
ham,I HAVE A DATE ON SUNDAY WITH WILL!!,,,
spam,"XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> ht
ham,Oh k...i'm watching here:),,,
ham,Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.,,,
ham,Fine if thatåÕs the way u feel. thatåÕs the way its gota b,,,
spam,"England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 8707
ham,Is that seriously how you spell his name?,,,
ham,I‰Û÷m going to try for 2 months ha ha only joking,,,
```

Methodology

Build an efficient method to identify spam email based on the analysis of the content of email messages by developing classification models

Classification Models

Those algorithms are used to predict the class the data belongs: ham/spam.

Logistic Regression

One of the most common, successful and transparent ways to do the required binary classification to “ham” and “spam” is via logistic function which takes as input the e-mail content and outputs the probability of desired or unwanted e-mail received.

- What is the probability of an email to be spam or ham?

Support Vector Classifier

It is a supervised machine learning algorithm capable of performing classification, that uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. SVC represent a good alternative given that spam problems are medium complexity with a datasets of not having excessively large dimensionality and size.

Random Forest

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

Features

CountVectorizer

It is used to transform text string to a vector of term (token counts). Each column in represents a unique Word and the number of times that it appears. For example:

```
Unique words:
{'jurong': 4223, 'point': 5736, 'crazy': 2271, 'available': 1271, 'bugis': 1703, 'great': 3534, 'world': 8221, 'la': 4348,
'buffet': 1701, 'cine': 1994, 'got': 3494, 'amore': 1051, 'wat': 8020, 'lar': 4384, 'joking': 4191, 'wif': 8128, 'oni': 536
4, 'free': 3265, 'entry': 2875, 'wkly': 8179, 'comp': 2110, 'win': 8140, 'fa': 3005, 'cup': 2329, 'final': 3121, 'tkts': 751
4, '21st': 411, '2005': 402, 'text': 7383, '87121': 784, 'receive': 6110, 'question': 6005, 'std': 7023, 'txt': 7696, 'rat
e': 6057, 'apply': 1128, '08452810075over18': 77, 'dun': 2738, 'say': 6445, 'early': 2757, 'hor': 3814, 'nah': 5088, 'don':
2651, 'think': 7438, 'goes': 3458, 'usf': 7831, 'lives': 4534, 'freemsg': 3272, 'hey': 3731, 'darling': 2386, 'week': 8065,
```

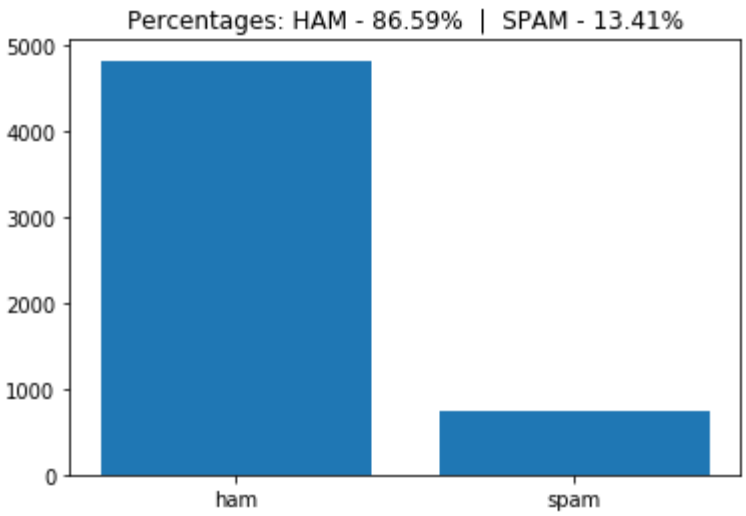
Stop words

Stop words are basically a set of commonly used words in any language. Removing the words that are very commonly used, the focus will be on the important words, keywords.

```
Stop words
frozenset({'then', 'may', 'our', 'all', 'done', 'thru', 'than', 'latter', 're', 'during', 'cannot', 'thus', 'became', 'nowhe
re', 'us', 'find', 'towards', 'while', 'mine', 'moreover', 'three', 'in', 'everywhere', 'such', 'at', 'elsewhere', 'someho
w', 'former', 'here', 'll', 'whoever', 'of', 'else', 'thereafter', 'bottom', 'hereby', 'though', 'her', 'sometimes', 'mostl
y', 'but', 'me', 'your', 'first', 'besides', 'without', 'another', 'must', 'wherever', 'ever', 'not', 'give', 'top', 'wher
e', 'there', 'he', 'third', 'ten', 'when', 'we', 'part', 'move', 'what', 'five', 'should', 'one', 'twelve', 'ltd', 'hereafte
r', 'forty', 'get', 'because', 'under', 'namely', 'how', 'myself', 'down', 'a', 'interest', 'eight', 'although', 'few', 'wa
```

Insights and Visualizations

Ratio between spam and non-spam



86.59 % are desired emails.

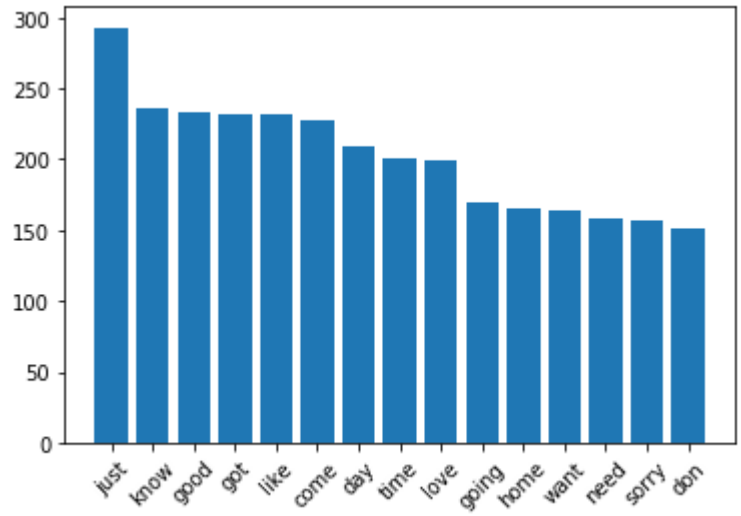
13.41 % are spam.

Most common e-mail word counts

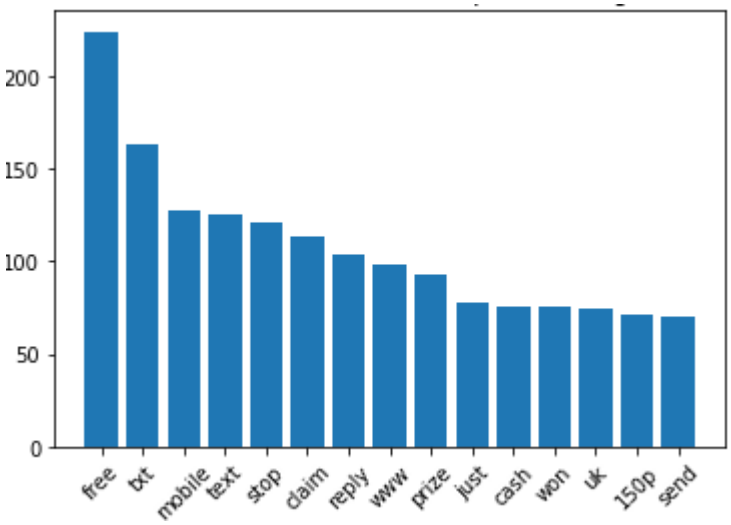
This analysis was done with all lowercase words, excluding all character symbols and applying some stop words.

The countermeasures were executed by counting words with same meaning and weight.

15 most common words on ham (non-spam) e-mail



15 most common words on spam e-mail



"free", "mobile", "txt", "prize", "won" are words that can indicate an award or where to call a contact to claim the reward or service offered at the e-mail content, which are typical standards for spam messages.

Results

Features

- CountVectorizer() function was used with fit_transform to convert the features to binary.
- Stop words were removed from the prediction.
- The label was set up using a vector: 1=spam, 0=ham.

Models

Taking a sample of 30% of the data for each model, meaning 1,672 messages:

- For SVC model the gridSearch() function was created to train the model with the most optimal parameters:

```
                precision    recall  f1-score   support

             0       0.98      1.00      0.99      1453
             1       1.00      0.88      0.94       219

    accuracy                           0.98      1672
   macro avg       0.99      0.94      0.96      1672
weighted avg       0.98      0.98      0.98      1672

Accuracy:     0.9844497607655502
Confusion Matrix:
[[1453    0]
 [  26  193]]
```

From 1453 messages that are non-spam, 100% were predicted as non-spam.
From 219 messages that are spam:

i. *Right predictions: 193 (88%) e-mails were predicted as spam, but, e-mails were*
ii. *False positives (marking good mail as spam): 26 (12%) e-mails were predicted wrong.*

- Default parameters for Random Forest:

```
                   precision    recall  f1-score   support

                0       0.96      1.00      0.98      1453
                1       1.00      0.71      0.83       219

         accuracy                           0.96      1672
        macro avg       0.98      0.85      0.90      1672
     weighted avg       0.96      0.96      0.96      1672

     Accuracy:     0.9617224880382775
     Confusion Matrix:
     [[1453    0]
      [  64  155]]
```

From 1453 messages that are non-spam, 100% were predicted as non-spam.
From 219 messages that are spam:

i. Right predictions: 155 (71%) e-mails were predicted as spam, but, e-mails were
ii. False positives (marking good mail as spam): 64 (29%) e-mails were predicted wrong.

- Default parameters for Logistic Regression:

```
                   precision    recall  f1-score   support

                0       0.98      1.00      0.99      1453
                1       1.00      0.84      0.91       219

         accuracy                           0.98      1672
        macro avg       0.99      0.92      0.95      1672
     weighted avg       0.98      0.98      0.98      1672

     Accuracy:     0.979066985645933
     Confusion Matrix:
     [[1453    0]
      [  35  184]]
```

From 1453 messages that are non-spam, 100% were predicted as non-spam.
From 219 messages that are spam:

i. Right predictions: 184 (84%) e-mails were predicted as spam, but, e-mails were
ii. False positives (marking good mail as spam): 35 (16%) e-mails were predicted wrong.

Best Model

The Grid Search technique generates the best combination of parameters to get the best model outcome. In order to get the highest score for both categories: ham and spam, a set of kernel, gamma and C types with all possible combinations of each metric (precision and recall) were tested:

Precision

Precision average for both categories by each parameter combination:

```
# Tuning hyper-parameters for precision

The best parameter set is:

{'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}

Grid scores on development set:

0.833 (+/-0.401) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.432 (+/-0.001) for {'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
0.977 (+/-0.011) for {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
0.833 (+/-0.401) for {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
0.977 (+/-0.013) for {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.979 (+/-0.006) for {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'}
0.978 (+/-0.016) for {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
0.980 (+/-0.009) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
0.979 (+/-0.013) for {'C': 1, 'kernel': 'linear'}
0.979 (+/-0.014) for {'C': 10, 'kernel': 'linear'}
0.979 (+/-0.014) for {'C': 100, 'kernel': 'linear'}
0.979 (+/-0.014) for {'C': 1000, 'kernel': 'linear'}

Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1453
           1       1.00      0.88      0.93       219

    accuracy                           0.98      1672
   macro avg       0.99      0.94      0.96      1672
weighted avg       0.98      0.98      0.98      1672

Accuracy:     0.9838516746411483
```

Recall

Recall average for both categories by each parameter combination:

```
# Tuning hyper-parameters for recall

The best parameter set is:

{'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}

Grid scores on development set:

0.508 (+/-0.010) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.500 (+/-0.000) for {'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
0.882 (+/-0.026) for {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
0.508 (+/-0.010) for {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
0.934 (+/-0.026) for {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.876 (+/-0.037) for {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'}
0.942 (+/-0.026) for {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
0.925 (+/-0.018) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
0.933 (+/-0.021) for {'C': 1, 'kernel': 'linear'}
0.930 (+/-0.027) for {'C': 10, 'kernel': 'linear'}
0.930 (+/-0.027) for {'C': 100, 'kernel': 'linear'}
0.930 (+/-0.027) for {'C': 1000, 'kernel': 'linear'}

Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1453
           1       0.99      0.87      0.92       219

    accuracy                           0.98      1672
   macro avg       0.98      0.93      0.96      1672
weighted avg       0.98      0.98      0.98      1672

Accuracy:     0.9814593301435407
```

Once the model was trained, the most optimal parameters for *precision* and *recall* parameters are: ***C = 100 , gamma = 0.0001 y kernel = rbf.***

Proven Hypotheses

- Once an unsolicited message is detected, setting the words to distinguish spam email from non-spam email will improve the model prediction.
- The most common spam words identified through countermeasures, the most spam emails are filtered.
- The most common spam word is “free”.

Disproven Hypotheses

- Unsolicited emails can be consisted or any topic.
- *With better filtering the spammers will be out of business*: Research into the economics of the spam industry suggests as filters improve, the information assets of spammers become more valuable and lead to more, not less, overall spamming activity.

Recommendations

- First, use the spam filtering and constantly train your model by adding new data to improve the spam-ham prediction:
    i. When you find spam in your inbox, don't just delete it. Select it and tell your mail client that this particular message is spam. You also need to train the client about your false positives. Once a day, go through your spam folder looking for messages that don't belong there. When you find one, select it and tell the client that it made a mistake.
- Hide your email address: The more people have your email address, the more spam you're going to get.