CSCU9YS: Computer Security and Forensics

#### Spam email classification using Machine Learning

#### Anna Rasburn (2411187)

# **Milestone 1**

## **Specifications**

The problem to solve is classification of emails into spam or ham. This process takes several stages, firstly beginning with pre-processing the body of email. This pre-processing consists of the stages: removing stop words, numbers and punctuation and lemmatization. Once you have completed the pre-processing stage, for each email you create a word dictionary which has a record of the frequency of certain words in the email. For each email, the stage of feature Extraction consists of the process of comparing the email’s dictionary to three thousand of the most common English words. This stage also includes representing an email using a number so spam is represented by -1 and ham is represented by 1. Then there is the classification stage which consists of three machine learning techniques: SVM, Naïve Bayes and Random Forest which classifies the email is ham or not. The measurement metrics of each classifier is calculated which consist of F1, precision, recall and accuracy.

## **Flowchart**

A close up of text on a white background

Description automatically generated

## **Pseudo Code**

Get Data{

Iterate through files

Make a list of words from each email

}

Prepare email for Data pre-processing{

Remove Stop words

Remove Lemmas

Remove punctuation

Remove numbers

}

Create Word Dictionary {

Gather words from all emails

Find frequency of each word in the larger set

}

Feature Extraction {

Create a file of 30,000 common words

Compare the body of email to the words in the file

}

Machine Learning Classifiers{

SVM{

Use Library  
}

Naïve Bayes{

Use Library

}

Random Forest{

Use Library

}

Accuracy{

SVM{

Precision

Recall

F1

}

Naïve Bayes{

Precision

Recall

F1

}

Random Forest{

Precision

Recall

F1

}

}

# **Milestone 2**

## **Text Pre-Processing**

Before you can use the data for classification you must first pre-process the data. This stage consists of three main step.

The first step is to remove stop words from the body of the email. Stop words are words like ‘and’, ‘of’, ‘in’ which have a high frequency in the body of the email. The Stop words have no strong indication of whether an email is spam or not and due to the high frequency of these words, this would skew the classification therefore they should be removed.

The second step is to remove punctuation. Similarly to the stage above, punctuation offers no clear indication of whether an email is spam or not. The high frequency of the punctuation would also skew the clusters in the classification.

The third step is lemmatization which is the process of the getting the a word down to it’s original mean so multiple words can be grouped together into one ‘lemma’ [1]. For example, ‘includes’, ‘included’ and ‘including’ are grouped together into one lemma ‘include’. This step reduces the number of words in our dataset and increases their frequency making them easier to cluster and classify.

I also decided within this step was to remove numbers. This is because I thought this because the numbers did not seem to give any clear pattern to whether they belonged to a spam or not spam email and they increased the frequency and size of the word dictionary.

## **Machine Learning methods**

### SVM

Support Vector Machines is a supervised machine learning technique used the most commonly for classifiers and regression. SVM is based on hyperplanes which is splits a dataset into two classes so in this case it would be classes of spam and ham. Support vectors in and SVM are calculated by which points from the class which is the closest to the hyperplane.

The advantages of SVM are they are very accurate for smaller datasets especially if there are a small number of classifiers.

The disadvantage of SVM is that it is not very accurate for large datasets and less effective for ‘noisy’ dataset.

### Naïve Bayes

Similarly to Support Vector Machine, Naïve Bayes is a supervised machine learning technique. It uses the Bayes Theorem.

Multinomial Naïve Bayes has a multinomial distribution which is useful for data that needs to be counts such as word counts in a text and creating a word dictionary.

### Random Forest

Random Forest is

## **Dataset**

STATE WHETHER YOU CREATE A DICTIONARY FOR ALL EMAILS OR ONE EACH

FEATURE EXTRACTION -> THE 30,000 WORDS, PUT A NUMBER REPRESENTING WORD FREQUENCY. ARRAY OF [-1, 1] TO BE WHETHER SPAM OR NOT

STEP 4: TRAINING THE CLASSIFIER – SVM AND NB. IS THE MODEL WORKING OR NOT - CORE OF THE ALGORITHM

HOW MANY TEST EMAISL DO YOU HAVE?

WRITE WHAT EACH OF THE STEPS MEAN

## **Measurement Metric**

### Accuracy

### Precision

### Recall

### F1 Score

# **Milestone 4**

Results

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Finding

Discussions

# References:

[1]"Stemming and lemmatization", *Nlp.stanford.edu*, 2018. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. [Accessed: 07- Nov- 2018].

[2]Y. Yang and B. Klimt, "The Enron Corpus: A New Dataset for Email Classification Research", 2018.

[3]"sklearn.svm.SVC — scikit-learn 0.20.0 documentation", *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Accessed: 13- Nov- 2018].

[4]"1.11. Ensemble methods — scikit-learn 0.20.0 documentation", *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html#forest. [Accessed: 14- Nov- 2018].

[5]"1.9. Naive Bayes — scikit-learn 0.20.0 documentation", *Scikit-learn.org*, 2018. [Online]. Available: https://scikit-learn.org/stable/modules/naive\_bayes.html#multinomial-naive-bayes. [Accessed: 15- Nov- 2018].