Spam email Classification using Support Vector Machine

Aldrin C. Racusa
Department of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Manila
Manila, Philippines
acracusa@up.edu.ph

Lorenz Timothy Barco Ranera
Department of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Manila
Manila, Philippines
lbranera@up.edu.ph

***Index Terms*—Support Vector Machine, Spam Classification, Email, Support Vector, Machine Learning**

I. Introduction

Spam, referring to the sending of unsolicited email, is a crucial and increasing problem for both home users and companies. [1]. Recipients of spam often have had their email addresses obtained by spambots, which are automated programs that crawl the internet looking for email addresses. [2]

While there are many ways to prevent spam, one of the common techniques used is the Spam Filtering. A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. [3]

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. [4]

This paper aims to present a model using Support Vector Machine algorithm that would classify whether a specific email is a spam email or not.

II. Dataset

The dataset the is used in this paper is *TREC 2007 Public Corpus*. There are 75,419 email messages with 50199 being a spam email and 25220 for not spam email (ham).

III. Preprocessing

In Natural Language Processing, there are various preprocessing techniques implemented as preliminary procedures in handling text data, that may be words or documents, before representing them into the vector representations, such as bag of words model, TF-IDF, and document embedding. The rationale for this is because some words that are insignificant may appear in our corpus that could negatively affect the training of our dataset and eventually the model and its accuracy.

A. Removal of Insignificant Words

There are various texts that are insignificant however appears frequently like the string *"x x x"*. The possessive moneme *" 's "* may frequently appear in emails but it does not give any significant unique meaning to the email document. Other examples would be strings like *(a)*, *(b)*, *(c)*, *(1)*, *(2)*.

B. Tokenization

Tokenization is the process where a text, which is a string, is broken down into words and punctuation. Each unit is called *tokens* [8] [9]. One may think of it as decomposing a large string into an array of string, which are words of the document. This step is necessary because texts are converted into an array of words in NLP as a data structure.

For example, let us consider a short document which contains the text *"I am a Computer Science student"*. Tokenizing the said document will result in a [*'I'*,*'am'*, *'a'*, *'Computer'*, *'Science'*, *'Student'*, *"."*].

There is a python implementation of this available from the Natural Language Toolkit (NLTK) package. There is also a R programming language implementation of this available from the Text Mining (TM) package.

C. Lowercasing

Lowercasing is a preprocessing technique in which all words which is capitalized or has a capitalized character is converted in its lowercase form. This technique avoids the the case where words with different forms of cases are treated differently [10].

For example, let us consider a short email which contains the text *"Know what? Thanks to Anatrim, my marriage was luckily saved! I fell down into this circle, depression more eating more depression. My wife was about to leave me as I was turning in overweight psycho."* Without the use of lower casing, the words *"Anatrim"* and *"anatrim"* are treated differently. Applying lowercasing, the document is now *"know what? thanks to anatrim, my marriage was luckily saved! i fell down into this circle, depression more eating more depression. my wife was about to leave me as i was turning in overweight psycho."*

D. Removal of Non-alphabetical symbols

HTML tags, Non-UTF characters, Foreign language characters such as Japanese and German, strings with numerical values embedded, and removal of non-alphabetical symbols, mostly numbers and punctuation marks, are done as a preprocessing technique to the corpus. Punctuation symbols removed because they do not infer any significant meaning or contribute to the document's content at all depending on the domain of the document. In this study, all punctuation symbols, HTML tags, Non-UTF characters, and Foreign language characters are removed. [10] [**?**]. The same case can be said to numbers.

E. Removal of Stop-words

There are some words that appear almost in every text documents but does not provide any significant meaning. These words may also affect the performance of the model when employing Machine Learning and other NLP techniques. These words are called *stop words*. In effect, the vocabulary size of unique words in a corpus will also decrease leaving out all significant words [8] [9].

Linking verbs, conjunctions, and pronouns are examples of stop words in the English language. These words such as *"is"*,*"he"*, and *"and"* frequently appear in almost all of documents. In computing similarity between two documents, the two documents may be considered *"similar"* but it just so happens that both contains a huge frequency of these kinds of stop words. The same can be argued in classification of documents [8]. In topic discovery, these words cannot provide any meaningful *"topics"* in analyzing the distribution of topics in a documents. Overall, stop words are not significant.

There is a collection of English stop words that can be imported from NLTK in Python. There is also a collection of English stop words that can be imported from the TM package in R.

F. Stemming

Stemming is a preprocessing technique which converts or reduces words into its simplest form. This technique helps reducing the number of the features in a corpus [8] [9] [11]. This is not problematic since the various forms of words have similar meanings even if they are in different forms [12]. The different forms of a particular word will be treated as different words without stemming.

For example, let us consider the words *"runs"*, *"ran"*, and *"running"*. The three mentioned words will be converted into the simplified word *"run"*.

There is a python and R implementation of this available from the NLTK package/TM package respectively with different algorithms like the Porter Stemmer algorithm.

G. Removal of Less Occurring Words

Words that occurs in less than a number of a particular number of documents are removed in the document. This will remove the cases where words are misspelled which has a significant small number of occurrence. This technique will also remove rare words that may not significant context enough [**?**].

IV. Support Vector Machine

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. [5]

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N being the number of features) that distinctly classifies the data points. [6]

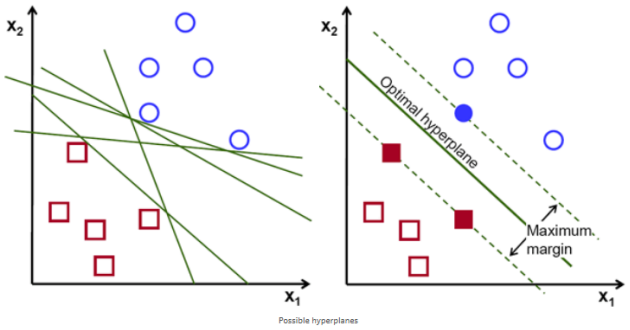

Fig. 1. Hyperplane

Some datasets are difficult to be seperated by a plane hence sometimes it is necessary that they are translated in another dimension. This technique is called *kernel trick*. There are four common kernel types used in SVM: *Linear*, *Radial Basis Function (RBF)*, *Polynomial*, and *Sigmoid* [7].

The most common parameters of SVM are *gamma* and *cost*. *Gamma* is the coefficient used in kernel types: *RBF, Polynomial* and *Sigmoid*. The gamma value influences correctness and fitness of the hyperplane in its dataset.

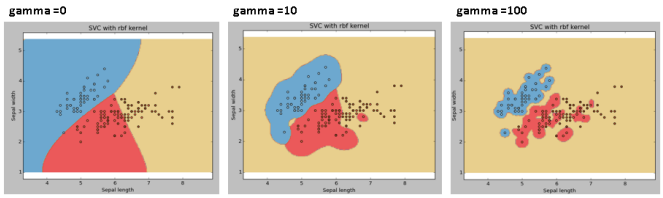

Fig. 2. Gamma parameter

Cost is the parameter that is used for the error term. It both influences the smoothness of the hyperplane and its correctness in classifying the labels.

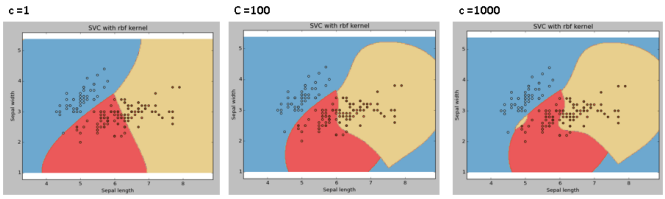

Fig. 3. Cost parameter

V. Methodology

The dataset that was preprocessed has been turned into a TF-IDF format using the *TfidfTransformer* in *sklearn* package.

The python library *pandas* was used in order to read the csv file and the library *sci-kit learn* was used to implement SVM, and a Pipeline to group the following: *tfidfTransformer*, *svm.SVC*, and *CountVectorizer*.

We then used a hyper-parameter tuning technique called *Grid Search* in order to get the best parameter that can produce the highest accuracy.

In order to implement the Grid Search, the dataset has been reduced to 1000 and 2000 rows respectively. Because the drawback of Grid Search is it is computationally expensive. Because it exhaustively get all possible combinations of the given parameter list in order to find the "*best*" parameters.

Cross-validation was performed where 25% of the dataset are used for testing and 75% are used for training in pipeline.

The dataset for the Pipeline without Grid Search has a 53991 rows for training set, and 17997 for the test set. While the Pipeline with Grid Search has a 750 rows for training and 250 rows for testing for a total of 1000 rows, and lastly, 1500 rows for training set and 500 rows for testing set for a total of 2000 rows.

There is a total of 3 SVM models generated and analyzed in this study.

VI. Results and Discussion

The following are the results of the three models. Cross-validation was performed where 25% (17997 emails) of the dataset are used for testing and 75% (53991 emails) are used for training.

The first model is Support Vector Machine with a parameter of *gamma* = auto, *kernel* = linear, and *Cost* = 1. The model got an accuracy of 0.996, precision of 1.0 and a recall of 1.0. The below in Table I and the confusion matrix is shown below in Table II.

TABLE I
SVM with kernel=linear, gamma=auto, cost=1

Accuracy	0.996
Precision	1.0
Recall	1.0

TABLE II
Confusion Matrix

	Actual: Ham (H)	Actual: Spam (S)
Pred: H	6131	56
Pred: S	22	11788

The second model is Support Vector Machine with Grid Search performed on 1000 observations. The output "best" parameter of the Grid Search are *gamma* = 0.01, *kernel* = linear, and *Cost* = 2. The model also has an accuracy of 0.98, precision of 0.98 and a recall of 0.98. The results are shown below in Table III and the confusion matrix is shown below in Table IV.

TABLE III
SVM with kernel=linear, gamma=0.01, cost=2, 1000 observations

Accuracy	0.98
Precision	0.98
Recall	0.98

TABLE IV
Confusion Matrix

	Actual: H	Actual: S
Pred: H	59	4
Pred: S	1	186

The third model is a Support Vector Machine with Grid Search performed on 2000 observations. The output "best" parameter of the Grid Search are *gamma* = 0.01, *kernel* = linear, and *Cost* = 2. The model also has an accuracy of 0.994, a precision of 0.99 and a recall of 0.99. The results are shown below in Table V and the confusion matrix is shown below in Table VI.

TABLE V
SVM with kernel=linear, gamma=0.01, cost=2, 2000 observations

Accuracy	0.994
Precision	0.99
Recall	0.99

TABLE VI
Confusion Matrix

	Actual: H	Actual: S
Pred: H	133	2
Pred: S	1	364

The first model is the best model among the three because it has the highest accuracy (0.996).

VII. Conclusion

It is possible to classify *Ham* and *Spam* emails using Machine Learning specifically Support Vector Machine technique.

In this study, we are able to generate a model with accuracy of 0.996% at best. Various and intense preprocessing techniques, and a "good" parameter should be considered first in the corpus in order to generate a model with that accuracy.

REFERENCES

[1] M. Siponen ; C. Stucke *Effective Anti-Spam Strategies in Companies: An International Study*. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)

[2] "Email spam ". Retrieved in https://searchsecurity.techtarget.com/definition/spam

[3] "Spam Filter". Retrieved in https://searchmidmarketsecurity.techtarget.com/definition/spam-filter

[4] "Support Vector Machine". Retrieved in http://www.statsoft.com/textbook/support-vector-machines

[5] "SVM wiki". Retrieved in https://en.wikipedia.org/wiki/Support-vector$_machine$

[6] "Support Vector Machine Introduction". Retrieved in https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[7] S. Ray. "Understanding Support Vector Machine algorithm from examples (along with code)," 2017. Retrieved in https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[8] V. Gurusamy and S. Kannan, "Preprocessing techniques for text mining ," October 2014

[9] A. I. Kadhim1, Y.-N. Cheah, and N. H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," in 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014.

[10] M. J. Denny and A. Spirling, "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it," 2017.

[11] S. Vijayarani1, M. J. Ilamathi, and Nithya, "Preprocessing techniques for text mining - an overview," International Journal of Computer Science and Communication Networks, vol. 5, no. 1, pp. 7-16.

[12] A. G. Jivani, "A comparative study of stemming algorithms," International Journal of Computer Technology and Applications, vol. 2, no. 6, pp. 1930-1938, 2011.

[13] P. Naval, "CS 280: Learning as Inference"