Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 9, 2022, page 139-147

Analysis on Spam Email by Statistical Learning

Loh Wei Kit, Adina Najwa Kamarudin*
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
*Corresponding author: adina.najwa@utm.my

Abstract
Email is a convenient communication way for most people. However, many people have took advantage to gain illegitimate benefits, prank, and spread fake news and inaccurate information. These are called spam emails that containing useless information. Many methods have emerged in detecting spam emails including blacklist, while list, filter suspicious email address and others. One of the most common technology is by using the statistical learning method as the classifier. There are few techniques have been used in this study which are Naïve Bayes model, Support Vector Machines, Random Forest and Long Short-term memory model. Specifically, the purpose of this study is to identify the key feature of spam email using these methods. The performance of all methods is accessed and compared to select the best model. The key feature is identified by obtaining the weight of the model and rearrangement is performed to find the important list of words as key feature for classification. The comparison among the models are assessed using confusion matrix, accuracy, recall, precision and F1 score. Based on the results, Naive Bayes model is outperform compared with the other models with highest accuracy of 98.23%, highest precision of 99.32%, high recall with 94.55% and highest F1 score with 0.9688.

Keywords: Naïve Bayes; Support Vector Machine; Random forest; Long Short-term memory

1. Introduction

With the advancement of the internet, the way people communicate is becoming more and more convenient. The email is one of the formal communication mediums all over the world. The use of email for illegitimate gains or black trick become common with the popularity of email. Solic et al. [1] highlight that spam is a business which can earn millions of Euros. Spam email is an unsolicited email that is useless for email users. The spam email is usually sent out in a large amount, and this could cause inconvenience for the email users. Internet Security Threat Report [2][3][4] shows that the overall email spam rate keeps decreasing from 2012 to 2015, but will increase from 2015 to 2019. This data set implies that we still need to improve the technology to cope with spam email issues. There are several general ways to filter spam and normal email, while machine learning classifier is one of the most famous and effective way. The objective of this study is identifying the main feature (keywords) of spam email by Naive Bayes, Support Vector Machine, Random Forest. Besides that, find the best method to identify spam email among these model and Long Short-Term Memory artificial neural network also of one the objective in this study. The spam email data from https://www.kaggle.com/nitishabharathi/email-spam-dataset is used

2. Literature Review

2.1 Spam email identification

2.1.1 Non-machine based

List-based filter method is a non-machine learning method. It creates three lists which are black list, white list and grey list. Email and IP addresses that listed in black list are categorised as spam while white list contains all email and IP addresses that are not categorised as spam. Those email and IP addresses that are unfamiliar and from unknown source will be categorised as spam emails by grey list. Iyengar et al. [5] had improved the email detection system by integrated spam filter using is the combination of the list based method and Bayesian classifier algorithm. This result the accuracy of spam email identification is increased from 96.46 percent to 97.3 percent. There are two common methods found in the content-based filter which are word-based filter and heuristic rule-based filter. The main difference between list-based filter and content-based filter is that list-based filter categorises the email by their email addresses or IP addresses. Meanwhile, the content-based filter categorises the email by its content. The list- based method has limitation with respect to its word dictionary, this heuristic method has limitation with the design of the heuristic rule. Razak and Mohamad [6] suggested the information in the header which is Received field, From field and Receiver address (To, CC, and BCC) is useful to identify the spam email. The HELLO verification technique is applied. Other than that, the number of receivers also efficient used to investigate spamming behaviour.

2.1.2 Machine learning based

Ratnesh Kumar Dubey1 [7] applied Naïve Bayes and support vector machines to recognise the spam email and SMS. They use an SMS data set containing 5574 perceptions as their training and test data. Overall, both methods show high accuracy in the classification of spam email. The Naïve Bayes model shows 99.49 percent accuracy which is slightly higher than the support vector machine accuracy. Hussain and Qamar [8] have tested ten text classification methods in spam email. These 10 methods are Latent Dirichlet Allocation (LDA), Chi-square Automatic Interaction Detector (CHAID), Iterative Dichotomiser (ID3), random forest, decision stump, decision tree, k-nearest neighbor, support vector machine, Naïve Bayes method and Random Tree method. The accuracy of the ten methods shown above highlight that the Naïve Bayes method has the highest accuracy (95.71%) and it is followed by the support vector machine (85.24%). Random Forest method with 82.37 percent. The latent Dirichlet allocation shows the lowest accuracy, which is 55.5 percent. From the result, it can be concluded that Naïve Bayes, support vector machine and random forest method with high accuracy and these methods should be conducted in this research study. Liu and Guo [9] had improved the Bidirectional Long short-term memory artificial neural network by combine then convolutional layer and attention mechanism to become attention-based bidirectional long short-term memory with convolution layer.

3. Methodology

3.1 Naïve Bayes Model

The Naïve Bayes method is a method based on Bayes' theorem which is defined as the following

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

P(A) is the probability of event A occurs

P(B) is the probability of event B occurs

$P(A|B)$ is the probability of event A occurring under condition event B occurs

$P(B|A)$ is the probability of event B occurring under condition event A occurs

The Bayes theorem shows us that the posterior probability of class can be calculated through the product of likelihood $P(B|A)$ and class prior probability of class $P(A)$ divided by the prior probability of the predictor $P(B)$. This implies that the posterior probability of class is directly proportional to the product of likelihood and class prior probability of class, so we can calculate the score of the object classified to the relevant class.

3.2 Support Vector Machine

The equation of the hyperplane is shown as below.

$$w \cdot x + b = 0 \tag{2}$$

Where w is the normal vector to the hyperplane and b is the constant. This hyperplane is used to classify the data by separating it into different groups by their class. The optimal hyperplane of support vector machine can reach by solve the soft margin optimization problem.

The soft margin optimization problem is built as below.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{Max}} M$$

Subject to

$$\sum_{j=1}^{p}\beta_j^2 = 1, y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1-\epsilon_i), \epsilon_i \geq 0, \quad \sum_{i=1}^{n}\epsilon_i \leq C \tag{3}$$

The C in the model represents the budget for the amount of the individual data that can violate the margin and the hyperplane.

3.3 Random Forest

The first step to build a regression tree is to decide which feature becomes the root. The factors that will be considered in arranging the order of nodes in the decision tree is the node purity. Gini Impurity is a way to measure the node purity. Gini impurity method is calculated as below.

$$G = \sum_{k=1}^{K} \hat{P}_{mk}\left(1 - \hat{P}_{mk}\right) \tag{4}$$

Where $\hat{P}_{mk}$ = probability of a classification m and k

Then Gini Impurity is calculated for all feature with different category, the lowest value of gini impurity of the feature of selected category will be the root of the classification tree. For each internal node, same method which is the Gini impurity to decide the feature under the condition of the last level of node or root. Random forest is an improvement from the bagged tree method. The bagged tree method generated the bootstrapped training data set to reduce the variance. The bootstrapped training dataset is generated by selecting the sample data in the training set randomly with replacement. In fundamentals of bagging method, the subset of the features may be randomly selected at each candidate split and the. Classification tree algorithm is applied to generate individual decision trees for each bootstrapped training data with random subset of the feature. The majority vote method will be applied to the result of each individual decision tree to get the final result.

3.4 Long Short-Term Memory

LSTM model has the main phase as shown in the below.

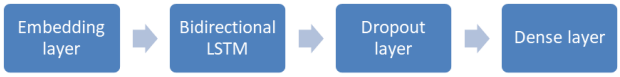

Figure 1: Design of LSTM model

Embedding layer is a layer that able to gather the word with similar meaning close together in high dimension space. Besides that, it also able to reduce the input dimension used to train the model. The embedding matrix with n*m dimension is created in this layer. The n is the input word size. While, the m is the number of dimensions of embedding space we design.

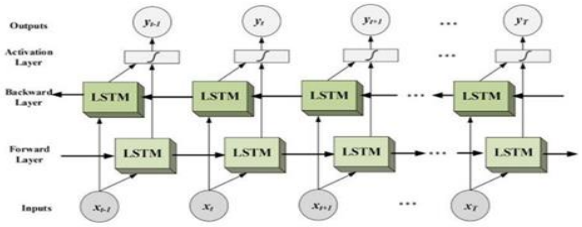

Figure 2: Design of LSTM model Bidirectional LSTM (source: Verma, Y. (2021, November 20))

The main different between bidirectional LSTM and standard LSTM is standard LSTM is involved input flow only in one direction. The standard LSTM input only flow in one direction while Bidirectional LSTM flow in both direction as figure 6 shows below. Bidirectional LSTM considers the information come from both directions into output.

The dropout layer is used to prevent the overfitting problem in model. The dropout effect is achieved through the following mathematics equation. Bernoulli distribution is used in dropout layer. The dropout concept is shows as below

neuron after dropout $= ka(n(x)) = \begin{cases} a(n(x)) & for\ k = 1 \\ 0 & for\ k = 0 \end{cases}$, while the probability to exist k=1 is p, where p is the dropout rate from Bernoulli distribution.

The last layer in the LSTM model is dense layer. It mainly functions as changing the dimensionality of the output from previous layer into suitable dimension so that our model extracts the correlations between those features. The sigmoid function is used in LTSM because our research is binary classification.

3.5 Model performance

The model will be evaluated in four measurement which are accuracy, precision, recall and F1 score. The confusion matrix is used to calculated those four measurements.

Table 1: Confusion Matrix

		Predict result	
		Positive	Negative
Actual Values	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Table 2: Formula of accuracy, precision, recall and F1 score.

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ (5)	Precision = $\frac{TP}{TP + TN}$ (6)
$Recall = \frac{TP}{TP + FN}$ (7)	$F1\ score = \frac{2 * precision * recall}{precision + recall}$ (8)

4. Results and discussion

4.1 Text pre-processing

The first step in text pre-processing is expand the contraction in text. The second step is converting the upper case to lower case. Third step applied to text pre-processing is removed the useless and not significant symbol. The final step for text pre-processing is lemmatization. The example is shows as below.

'on sun jul 28 2002 at 07 09 29pm 0100 kevin lyda mentioned there is a tutorial on linux ie for it thanks donnacha but since i am not an apache person this only got me part of the way step by step this is what i did cd root wget http www remotecommunications com apache mod_gzip src 1 3 19 1a mod_gzip c gz gunzip mod_gzip c gz apxs ic mod_gzip c thanks kevin mod_gzip for the terminally lazy kate irish linux user group ilug linux ie http www linux ie mailman listinfo ilug for un subscription information list maintainer listmaster linux ie'

Figure 3: Clean raw data

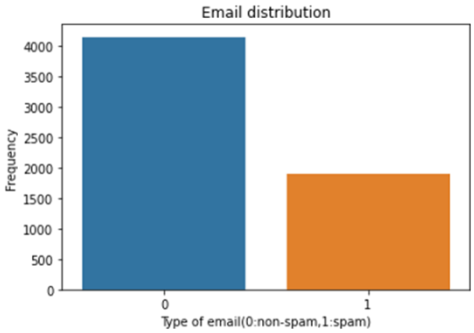

Figure 4: The frequency of spam and non-spam email after remove the missing value

After removing the missing value, about 4150 non-spam email and 1896 spam email are left.

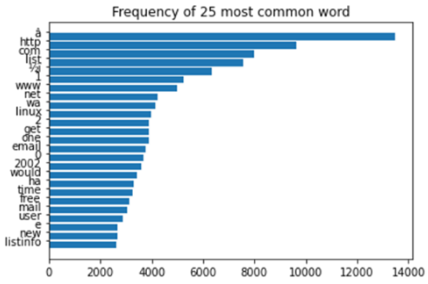

Figure 5: The frequency of top 25 most common word in email dataset

The bar chart above shows the top 25 frequency common word exist in the dataset. The most common word is ‘â’ and its number are far ahead compared to another word. After the word ‘â’, it follows by the frequency word that exist in website address which is ‘http’ and ‘com’.

4.2 Naïve Bayes

4.2.1 Important feature

Some of the words for example ‘the’, ‘to’, ‘and’, ‘in’, ‘is’ have high and almost similar value of likelihood in both spam and non-spam emails. This is reasonable because those words are frequency used words in English. However, the large difference of likelihood for word like ‘you’ with 0.007568 in non-spam and 0.020299 in spam emails, may implies that the word ‘you’ is relative more in spam email. Besides that, the word ‘½ï ‘are highly exist in spam email with 0.007371 likelihood but cannot found in top 25 non-spam key feature. After inspection, it is found that the email with the word ‘½ï ‘is usually a bunch of garbled characters. Obviously, the email with word ‘½ï ‘is useless and can be categorise into spam email group. Besides that, the words with ‘http’ with 0.01034 likelihood and ‘www’ with 0.00508 likelihood are importance in categorising the email as non-spam email. These two words may imply that the email with uniform resource locator address has high probability that it is not a spam email. The word ‘500’, ‘price’, ‘offer’, 'advertising' and 'membership' are the frequent word exist in advertisement which are categorised as spam email. Besides the advertisement, the word ‘investment’ and ‘interest’ are usually existing in the very suspicious email. It worth to mention that the word ‘receive’ is exist in both advertisement and suspicious email in a high frequency. For, non-spam email key feature, the words ‘org’, ‘subscription’ and 'newsisfree' are the keyword of the website to which the data owner is subscribed.

4.3 Support Vector Machines

4.3.1 Important Feature

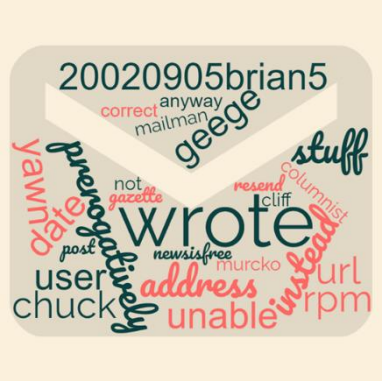

Figure 6: Wordcloud of spam email key features

The most important key feature for spam email is ‘remove’ with highest coefficient which is 0.3552. This word is found mostly in advertisement emails in which it frequently exists in the end of the email together with the link for unsubscribe option. Other important features for spam email with high coefficients are the word ‘click’ and ‘offer’ which also appear mostly in advertisement emails with 0.2582 and 0.1771. The word ‘click’ accompanied mostly with the link for purchasing the product. The interesting fact found in the non-spam email is the word ‘wrote’ only exist in the non-spam email. After investigate the content of the email, the word ’wrote’ is the frequently exist in notification of the massages or user reply from different website. Besides that, the words ‘greege’ is found to be the name of user’s friend, who is email data provider. Another word found as important feature in non-spam email is ‘newisfree’. This word describes the subscription of the email users.

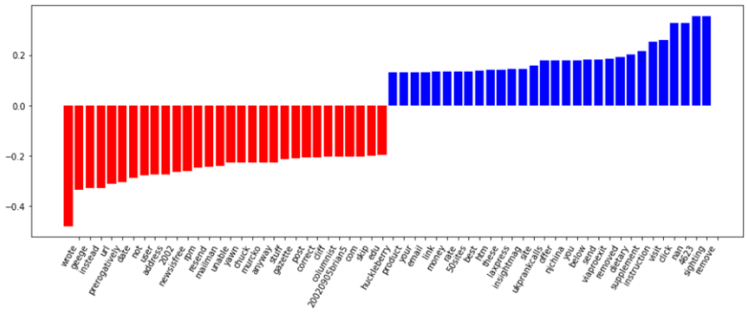

Figure 7: Bar chart for weight coefficient of the key feature

Red colour bar represents the non-spam email’s key feature coefficient and blue colour bar represent the spam email’s key feature coefficient. The bar chart above shows the distribution of the weight coefficient for each feature found from support vector machine model. The positive coefficient weight represents the key feature of the spam email while the negative coefficient weight represents the key feature of the non-spam email. Only top 30 weight coefficient from each group displayed to visualize. According to the bar chart, the highest positive weight coefficient is the keyword ‘remove’ with 0.3552 while the highest negative weight coefficient is the keyword ‘wrote’ with -0.48095. Among the 30-positive coefficient, the keyword ‘sighting’ with 0.3529 close to keyword ‘remove’ with 0.3552. Then the coefficient decreases to 0.328518 for keyword ‘4623’ and 0.328518 for keyword ‘nan’. Next, the remaining keyword from ‘supplement’ to ‘offer’ show similar weight coefficient within range of 0.2029 to 0.1771. For non-spam emails, there is the big difference between the keyword ‘wrote’ and other remaining word. The second highest is ‘geege’ with -0.33691 and it is slowly decrease until the last keyword ‘huckleberry’. In conclusion, the keyword ‘remove’ and

'sighting' are the most important features to categorised the email to spam email while 'wrote' are the most important feature to categorised the email to non-spam email.

4.4 Random Forest

The highest value of the feature importance is 0.028235 which is for the keyword 'the'. However, 'the' keyword is high frequently exist in both spam and non-spam email therefore it is difficult to interpret this result without further investigation. The other significant keywords is '½ï' that can found in the random forest feature importance. This keyword '½ï' is the important keyword to categorise emails as spam email.

4.5 Long Short-Term Memory

4.5.1 Model Identification

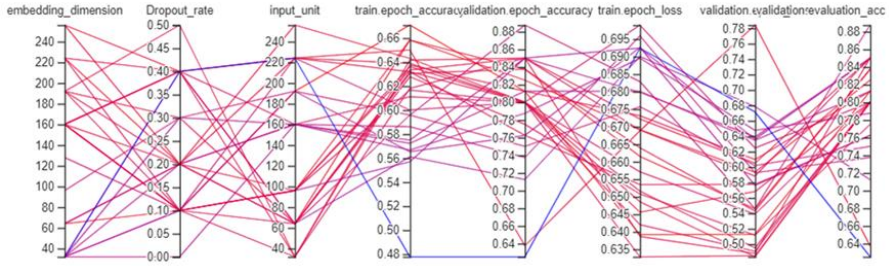

Figure 8: The parallel coordinate graph above highlights the relation of every layer in LSTM and accuracy.

The combination of Bidirectional LSTM layer with 96 number of embedding dimensions, 192 number of units in Bidirectional LSTM and 0.3 dropout rate is selected as our Bidirectional LSTM model.

4.6 Comparison of the models

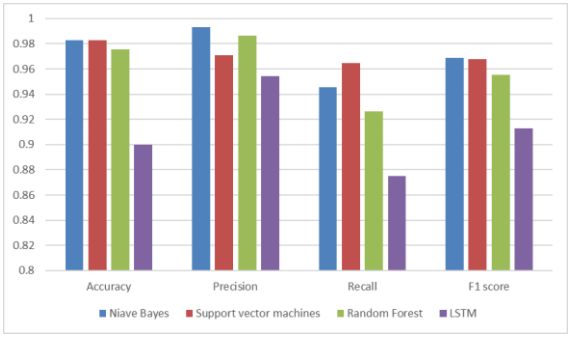

Figure 9: Performance of Naïve Bayes, Support vector machines, Random Forest and LSTM

Figure 24 shows the performance of Naïve Bayes, support vector machines, random forest and LSTM. The performance of the model is measured by accuracy, precision, recall and F1 score. Overall, the performance of all models is at least 86%. Based on the accuracy of the model, Naïve Bayes and support vector machine models are outperform with 98.28% and 98.19% respectively. Meanwhile LSTM model has low accuracy percentage of 90%. With respect to the precision accuracy measures, Naïve Bayes model again has the highest percentage of 99.32% followed by random forest model with 98.63% and the lowest is LSTM with 95.25%. Slightly different in recall measure, in which support vector machine shows the best performance with 96.47% with the least performance is 87.5% for LSTM. Finally, F1 score for Naive Bayes and support vector machine model are outperformed with 96.88%. In conclusion, the Naive Bayes and support vector machine model are the best model in classify the emails into spam and non-spam email.

Conclusion

The one of the purposes of this research is to identify the feature keywords of spam and non-spam email by statistical method. The Naïve Bayes model has identified few words such as '½ï 'that usually exist in spam email with garbled word, '500', 'price', 'offer', 'advertising' and 'membership' that frequently exist in spam email advertisement. The non-spam email feature keywords found by Naïve Bayes are the frequency words found in website address which are 'www' and 'http'. By observations, those words are frequently found in the notification email from user's subscribed website. Support vector machine also also found the important features of spam and non-spam email. The key feature of spam email identified are 'remove', 'click' and 'offer'. While, the key feature 'wrote', greege' and 'newsisfree' are found for non-spam email. Last, the keyword '½ï 'is also have a high weight value in random forest model which this key feature also identified by Naïve Bayes model. Besides that, the key feature word 'free' and 'com' also are the important features for random forest model.The second objective to achieve in this research is find the best method to identify spam email among Naive Bayes, Support Vector Machine, Random Forest and Long short-term memory artificial neural network. This research found that the Naïve Bayes model consist of highest accuracy. From the comparison of performance measurement, the Naïve Bayes model accuracy has accuracy higher than support vector machine, random forest and LSTM model by 0.009,0.0073 and 0.0828 respectively. Besides that, the precision of the Naïve Bayes model also highest among the four model which is 0.9932. Even the Naïve Bayes model has the highest in accuracy and precision but recall performance of the Naive Bayes model slightly lower than support vector machine by 1.92 %. Last factor to measure their performance is F1 score. The Naïve Bayes model score the highest value in this part with 0.9688. Overall, the Naïve Bayes model is found to be selected as the best model among those four models.

Acknowledgement

The researcher very grateful to those who helped this research.

References

[1]. K. Solic, D. Sebo, F. Jovic and V. Ilakovac, "Possible Decrease of Spam in the Email Communication", Proceedings IEEE MIPRO, (Opatia), pp. 170-173, May 2011.
[2]. Internet Security Threat Report 2013 :: Volume 18. (2013, March). Symantec. https://www.insight.com/content/dam/insight/en_US/pdfs/symantec/symantec-corp-internet-security-threat-report-volume-18.pdf
[3]. Internet Security Threat Report 2016 :: Volume 21. (2016, April). Symantec. https://docs.broadcom.com/doc/istr-16-april-volume-21-en
[4]. Internet Security Threat Report Volume 24. (2019, February). Symantec. https://docs.broadcom.com/doc/istr-24-2019-en
[5]. Iyengar, A., Kalpana, G., Kalyankumar, S., & GunaNandhini, S. (2017). Integrated SPAM detection for multilingual emails. 2017 International Conference on Information Communication and Embedded Systems (ICICES). https://doi.org/10.1109/icices.2017.8070784
[6]. Razak, S., & Mohamad, A. F. (2013). Identification of spam email based on information from email header. 2013 13th International Conference on Intellient Systems Design and Applications. https://doi.org/10.1109/isda.2013.6920762
[7]. Ratnesh Kumar Dubey, Shubha Mishra, Dilip Kumar Choubey,"Recognizing Spam Emails/SMS Using Naive Bayes and Support Vector Machine", Complex Systems and Complexity Science Journal, Volume-8, Issue-4, October 2021
[8]. Roohi Hussain, Usman Qamar, "An Approach to Detect Spam Emails by Using Majority Votin", The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014), November 2014
[9]. Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing, 337, 325–333 https://doi.org/10.1016/j.neucom.2019.01.078