Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230843

Cross-platform spam messages classification based on the multiple machine learning algorithms

Mengliang Tan

The Department of Applied Mathematic, University of California, San Diego, America

metan@ucsd.edu

Abstract. The proliferation of spam messages has had a detrimental impact on users' experience of emails and social media. Consequently, it is imperative to implement effective spam filtering mechanisms to enhance online experiences. Internet companies have leveraged machine learning algorithms to detect and thwart spam messages. Given the multitude of popular social media platforms, it is critical to evaluate the efficacy of prevalent machine learning algorithms across diverse online platforms. This study seeks to assess the performance of Support Vector Machine, Linear Regression, and Random Forest on social media. To this end, datasets containing spam and non-spam messages sourced from YouTube comment sections and Twitter will be procured. The text data will be transformed using a vectorizer to enable interpretation by machine learning models. Three models employing SVM, Linear Regression, and Random Forest will be trained and deployed to test their effectiveness. The models will be applied to detect spam messages in the test dataset, YouTube comment set, and Tweet set. The performance of the models will be evaluated based on accuracy, F1 score, and precision score. The findings indicate that the models' performance on various social media datasets is not satisfactory, as there is a significant reduction in accuracy.

Keywords: brain tumour, MRI, CNN, machine learning, deep learning.

1. Introduction

Spam emails, which are unwanted emails sent in bulk to users' mailboxes, pose significant challenges to users' email experience. These emails typically arrive in large quantities and occupy considerable space in the users' inboxes, often disrupting their ability to access important messages. It has been easier for spammers to get their hands on personal information such as email addresses. It can be leaked to spammers through computer viruses, information leaks from big internet companies, some websites that force you to submit your personal information to them, etc. That is why, in recent years, the volume of spam emails has been continuously increasing. Spam messages are responsible for 45.1% of the world's email traffic by March 2021 [1]. Hardware efficiency can also be prevented from being fully utilized by having a large inflow of spam emails taking up storage space and CPU. Some spam messages may contain Trojan downloaders in the mail, which will be disguised as the file of bills, then it will implement viruses in the users' computers after being downloaded [2]. These spam messages can also be used as a way for scammers to conduct fraudulent practices such as cheating individuals into sharing sensitive personal information like passwords, Bank Verification Numbers, and credit card numbers, which can cause actual financial loss [3].

In contemporary times, machine learning has emerged as a dynamic and diverse field. Numerous techniques have been put into use by major inter companies such as Google and Yahoo which are two of the biggest email services providers. The role played by Machine Learning algorithms in this context is to generate rules for spam filters to recognize spam and non-spam messages. The way to achieve this is that algorithms will be "fed" with training samples, and from these samples, they will learn the pattern of text data and recognize the rules of classification. After years of development and progress in the field, the power of spam filtering keeps increasing. For example, the machine learning model deployed by Google can reach 99% accuracy in filtering out spam emails [3].

In 1995, Support Vector Machine (SVM) created by Vapnik has later been widely applied and performed greatly in many areas such as function approximation, modeling, optimization control, and binary classification which is essential for spam filtering [4]. The formulation of SVM utilizes the Structural Risk Minimization principle, which minimizes an upper bound on the expected risk and has been proven to be better than the traditional Empirical Risk Minimization principle which works by minimizing the error of the training data [5]. Logistic Regression is a very commonly used machine learning algorithm with very low time complexity [6]. For its capability in data analysis, a lot of software companies have a Logistic regression model as their products' innate feature [7]. Random forest is also a very popular machine-learning algorithm. The Random Forest algorithm was created by Breiman and Cutler, and it was extremely well at handling massive data sets even under the circumstance when sets got missing data [8].

Despite the fact model trained by these machine learning algorithms has shown excellent performance, researchers have paid less attention to the accuracy and efficiency of machine learning's prediction of spam comments across different groups of data sets. With the proliferation of social platforms and the increasing amount of text data being transmitted online, the need for machine learning models that can recognize unwanted messages across different platforms has grown. This research will mainly focus on using Support Vector Machine, Logistic regression, and Random forest as the algorithm to train machine learning models with various data sets of text messages from different social platforms such as Twitter and YouTube comment sections and evaluate their performance.

2. Method

2.1. Dataset Preparation

In this study, six data sets provided by Kaggle were utilized [9-13]. All data sets were in the form of comma-separated values files. The content of these files included messages that were labeled as spam and non-spam from Twitter and YouTube comment sections. There were two Twitter sets with one having 11787 messages and another one having 5169 messages. For YouTube comments, there were three sets. Two of them were comments from videos related to Kate Perry and Eminem. Kate Perry set had 348 messages. Eminem set had 446 messages. The third one had 4673 messages. The remaining data set would be the spam training set which would be used to train the machine learning model. It had 5169 messages. It is noteworthy that all of the data sets were binary.

In terms of the processing part of this study, all six csv. Files would be read into data frames by using pd.readcsv(). 4470 messages of the spam training set would be used as the training set, and the rest was for testing the model. Then, it was needed to merge two Tweets sets and three YouTube comments sets. In YouTube comments message sets, columns of comment IDs, dates, and authors were not needed. Hence, they would be first removed by using pandas.drop() and had columns of class and content swapped for further integration. After this, pandas.merge() method was used here to merge all three datasets. A similar process would be applied to two other two data sets of Tweets. Columns not related to the work would be removed and then two sets would be merged. After the integration, drop_duplicates() was utilized to remove repeating messages if they existed. Now there were three sets: the spam training set, the integrated YouTube comment set, and Tweet set.

The machine clearly could not directly interpret the text data of the files. To solve this, CountVectorizer() would be applied here to create a vectorizer to transform text data into numerical data, thereby the machine could count the number of presences of different words in the file. For the step of vectorization, the fit_transform() method was utilized to fit the vectorizer to the training data and also transform text data. This vectorizer would be saved.

2.2. Machine learning model

Support Vector Machine is a machine learning method that is good at dealing with classification and regression problems, and it uses hypothesis space of a linear function in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [5]. SVM functions in the classification process by seeking a hyperplane to distinguish between points belonging to the first and second classes. Furthermore, it also maximizes the distance between points of different classes and the hyperplane. Random forest performs binary classifications by constructing decision trees for class prediction and then the output would be determined by choosing the class that gotten chosen by the most number of trees [14]. Logistic regression analyzes the relationship between several existing independent variables and produces a binary variable based on the interactions between the different variables [15].

Following the preprocessing part, the spam training set was divided into a training part and a testing part. For further implementations, scikit-learn, a machine learning library for python, would be imported.

After this, a model would be created utilizing the Support Vector Machine algorithm. To improve the accuracy of the algorithm, the hyperparameters for the SVM model were tuned to optimize the machine learning process this study just created. RandomizedSearchCV would be implemented here. There were many hyperparameters, and not all of them would be adjusted. RandomizedSearhCV was a fast way to perform hyperparameters tuning. C-parameters and gamma parameters would be tuned to find the most optimized decision boundary which could separate data into two classes as accurately as it could while preventing the model from being overfitting. Upon completing hyperparameter tuning, the training data set was used to train the model using fit(). After training was completed, the accuracy score, precision score, and F1 score, of the model would be first evaluated on the testing part of the spam training set, then on YouTube’s comment set and Tweet set.

RandomForestClassifier() would be used to create a Random forest model. N_estimators would be set to 100 and random_state set to 42. Since this process was done in a different file, the previously saved vectorizer would be loaded, and then directly transform the text data of the training part. LogisticRegression() was used to create the Logistic regression model. These two models would then be tested like how it was done on the SVM model previously.

3. Result and discussion

After applying Support Vector Machine, Random Forest, and Logistic regression algorithms on three data sets, the experiment have shown metrics that reflect the performance of different machine learning algorithms on data from different online social platforms. In Table 2, The Random Forest model managed to reach 100% precision. As shown in Tables 1, Table 2, and Table 3, SVM had the highest accuracy and F-1 score on all three sets.

Table 1. Performance of the SVM model on different data sets.

Metrics of Performance	Spam-Training set	YouTube's comment set	Tweet set
Accuracy	0.985	0.577	0.650
Precision	0.978	0.895	0.684
F1-score	0.937	0.219	0.246

Table 2. Performance of the Random Forest model on different data sets.

Metrics of Performance	Spam-Training set	YouTube's comment set	Tweet set
Accuracy	0.974	0.541	0.646
Precision	1.000	0.807	0.705
F1-score	0.881	0.083	0.241

Table 3. Performance of the Logistic regression model on different data sets.

Metrics of Performance	Spam-Training set	YouTube's comment set	Tweet set
Accuracy	0.983	0.563	0.646
Precision	0.978	0.875	0.759
F1-score	0.930	0.170	0.178

The study's results, as presented in Tables 1, Table 2, and Table 3, indicated that the three models performed exceptionally well in terms of their accuracy, precision, and F1 score, with minimal disparities between their metrics. However, when handling datasets from Twitter and YouTube comment sections, all three models experienced a significant decline in their performance metrics, particularly in F1 score and accuracy. The F1 score, which combines the precision and recall scores to evaluate the models' ability to identify true positives while minimizing false positives and false negatives, revealed a marked discrepancy between the precision score and F1 score for all three models, with the former being significantly higher. The Random Forest model exhibited the largest drop in F1 score when applied to the YouTube comments dataset, indicating potential instability when encountering various datasets. While the other two models also displayed low F1 scores, they remained relatively consistent. Interestingly, precision exhibited the smallest decline compared to the other metrics for all three models, implying that the models' ability to avoid false positives was less affected during cross-platform examinations.

Comparing all three models, it is indicated that cross-platform examinations can have a huge influence on their performance. None of them produced an optimistic performance, and none of them has a huge advantage over the other two models. The SVM model is slightly better than the other two since its performance is relatively consistent on different datasets and has a higher F1 score which means more balanced in avoiding false negatives and false positives.

4. Conclusion

This study investigated the performance of three machine learning models, namely SVM, Logistic regression, and Random forest, on various social media platforms. The models were tested on different datasets obtained from social media platforms, and various metrics of binary classification were evaluated to assess their accuracy and efficacy. The proposed method was validated through a series of experiments, and the findings revealed that the performance of machine learning models on social media platforms significantly dropped when the models were not previously trained using datasets from the corresponding social platform. The results indicated no significant differences between the performance of the three models, with SVM exhibiting a slightly superior performance. However, the study could have benefited from a more diverse range of social media datasets beyond Twitter and YouTube. Future research will focus on improving the accuracy of machine learning algorithms when handling data from other platforms.

References

[1] Kudupudi N NAIR S 2021 Spam message detection using logistic regression International Journal of Advanced Computer Science and Applications 9(9): 815-818

[2] Spam Report 2016 https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2016/08/07185333/Spam-report_Q2-2016_final_ENG.pdf

[3] Dada E G Bassi J S Chiroma H et al. 2019 Machine learning for email spam filtering: review, approaches and open research problems Heliyon 5(6): e01802

[4] Hsu W C Yu T Y 2010 E-mail Spam Filtering Based on Support Vector Machines with Taguchi Method for Parameter Selection J. Convergence Inf. Technol 5(8): 78-88.

[5] Jakkula V 2006 Tutorial on support vector machine (svm) School of EECS, Washington State University 37(2.5): 3

[6] Mrisho Z K Ndibwile J D Sam A E 2021 Low Time Complexity Model for Email Spam Detection using Logistic Regression International Journal of Advanced Computer Science and Applications 12(12)

[7] Berrou B K Al Kalbani K Antonijevic M et al. 2023 Training a Logistic Regression Machine Learning Model for Spam Email Detection Using the Teaching-Learning-Based-Optimization Algorithm Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022). Springer Nature 104: 306

[8] Reddy K N Kakulapati V 2021 Classification of Spam Messages using Random Forest Algorithm Resesearchgate

[9] Kaggle 2018 https://www.kaggle.com/datasets/goneee/youtube-spam-classifiedcomments?select=Youtube02-KatyPerry.csv

[10] Kaggle 2017 https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

[11] Kaggle 2023 https://www.kaggle.com/datasets/greyhatboy/twitter-spam-dataset

[12] Kaggle 2021 https://www.kaggle.com/datasets/fahmisulthoni/tweet-spam

[13] Kaggle 2022 https://www.kaggle.com/datasets/madhuragl/5000-youtube-spamnot-spam-dataset

[14] Kontsewaya Y Antonov E Artamonov A 2021 Evaluating the effectiveness of machine learning methods for spam detection Procedia Computer Science 190: 479-486

[15] Lawton G et al. 2022 What Is Logistic Regression? - Definition from Searchbusinessanalytics Business Analytics, TechTarget, 20 Jan. https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression.