**GR5067 Final Research Report: Spam Classification**

Cynthia Wang

Columbia University

**Abstract**

Mobile devices make communication easier, but at the same time they bring risks to their users. Spam messaging has been a concern for all the mobile users because it can cause negative consequences such as leakage of personal information. This research project aims to build spam classifiers using machine learning models. Four models are used in this research: random forest, logistic regression, multinomial Naïve Bayes, and support vector machine. Among all the models, logistic regression has the highest accuracy and support vector machine is the best at detecting spam messages. This project shows that building spam classifiers using machine learning techniques is feasible. The spam classifiers can significantly reduce the risks associated with spam messaging.

**Introduction**

We are now living in a century in which mobile devices are our primary way for communication. SMS text messages, phone calls and emails are all popular and convenient ways for daily communication. According to a report by Cloudmark, people send over 200,000 SMS text messages every second (“SMS Spam Overview”, n.d.). The response rate of SMS text messages is also higher than the response rates of other means of communication: 90% of the text messages are opened within 15 minutes of receipt, but only 20-25% of the emails are opened within 24 hours of receipt. However, even though the convenience of text messages significantly improves the efficiency of our communication, it also carries more risks for us when the text messages are misused. It is mentioned in the report that the number of SMS spam campaigns grew by 300% from 2011 to 2012. Among the spam messages people received, 92% of them are fraud/scam. In the United States, the top three categories of spam are “receive a gift card”, “iPhone/iPad test and keep”, and “need cash now”. Spam messages can cause negative consequences for the text message receivers. Some spam messages collect people’s personal information such as SSN and banking information. Some of them offer “free” products or services in the beginning but then ask people to pay for other products later. The messages may even contain malicious contents that might expose receivers’ phones to malware.

Due to the security issues that SMS texts messages can cause, mobile carriers want to be able to detect the spam messages and reduce the risks associated with them to protect their cell phone users. This project aims to build classifiers that can classify SMS text messages into legitimate messages and spam messages and compare the performance of these classifiers. With a spam classifier, the mobile carriers can reduce the spam messages that people receive and improve the security of text messages.

**Method**

**Data**

The dataset that this project uses is SMS Spam Collection from UCI Machine Learning Repository (“SMS Spam Collection v.1”, n.d.). The dataset contains 747 spam SMS messages and 4827 legitimate messages. It is downloaded as a text file. Each line in the file is a text message with its label in the beginning of the line. Legitimate messages are marked as “ham” and the spam messages are marked as “spam.”

**Methodology**

The dataset was first cleaned and then converted to a data frame with one column containing the text messages and one column containing the corresponding labels. The punctuation marks were removed in the first step. Numbers were kept in the messages because they might be important in detecting spam messages. Then the text messages were tokenized and stop words were removed in the messages using the NLTK package. Spacy was used for word lemmatization. After that, the dataset was split into training and testing data: 80% of the data were used for training and 20% of the data were used for testing. The training data was transformed into a vector using TF-IDF. Principle component analysis was then performed to reduce the number of features in the vector with the target variance set as 95%. This process reduced number of features from 6789 to 2196.

Four models were used to build the spam classifiers: random forest, logistic regression, multinomial Naïve Bayes, and support vector machine. Grid search was performed on random forest, logistic regression and support vector machine for hyperparameter tuning. The 5-fold validation scores were generated in the grid search processes for these three models, and the validation score for multinomial Naïve Bayes was calculated separately later. Then the models were trained using the best sets of parameters returned by grid search. All the models were trained using the vector that was transformed by PCA except multinomial Naïve Bayes. Since PCA would result in features with negative values which are not accepted in multinomial Naïve Bayes, this model was trained using the vector returned by TF-IDF transformation. Lastly, the models are tested on testing data to check their performance.

**Results**

The performance measures of the models are presented in Table 1. Random forest has 93.5% accuracy, 100% precision, 54.7% recall, and 70.7% F-score. The model performs well in terms of accuracy, but the false negative rate is high. Among the messages that are actually spam, only 54.7% of the messages are marked as “spam” by the model. Therefore, random forest model is not really good at detecting the spam messages. However, the 100% recall rate shows that this model doesn’t return any false positive result. The model doesn’t mark legitimate messages as spam messages. This is important for a spam classification model because people don’t want to discard the messages that are actually useful. Logistic regression has 98.5% accuracy, 98.0% precision, 91.3% recall, and 94.5% F-score. The accuracy is highest among all the models, and the 91.3% recall rate indicates the model is relatively good at capturing the spam messages. The precision of this model is not perfect but also pretty high. Multinomial Naïve Bayes has 97.0% accuracy, 100.0% precision, 78.9% recall, and 88.2% F-score. The precision is also 100% like the random forest model, but the recall is relatively low among the models. Similar to the random forest model, multinomial Naïve Bayes is good at preventing legitimate messages from being marked as spam, but relatively weak at detecting messages that are actually spam. One thing to notice here is that the multinomial Naïve Bayes model was trained by a vector with more features than the one used by other models, so its accuracy would actually be lower it was trained by the vector used in other models. Support vector machine has 93.5% accuracy, 98.7% precision, 94.4% recall, and 96.5% F-score. This model has the highest recall rate and the highest f-score. The recall and precision are balanced well for this model. As for the validation of these models, the validation scores from 5-fold cross validations are not significantly different from the models’ accuracy scores. Therefore, it is safe to say that the models are not overfitted.

A screenshot of a cell phone

Description automatically generated

Table 1. Model Performance.

**Conclusion**

This project’s goal is to build classifiers for spam messages. Four machine learning models are trained to build the classifiers. Among all the trained models, logistic regression has the highest accuracy. Random forest and multinomial Naïve Bayes have the highest precision, which means they are best at preventing false positive results. Support vector machine has the highest recall, which means it’s best at detecting spam messages among all the models. Considering the fact that we may want to keep false positive rate low in the case of spam classification, multinomial Naïve Bayes and random forest are good options for building the spam classifier. Support vector machine and logistic regression don’t have perfect precision, but they are good at detecting spam messages. Thus, they are also good candidates for spam classifiers. The results of this project show that building a well-performed spam classifier with machine learning models is feasible. Implementation of a spam classifier can greatly reduce the inconvenience and risks caused by spam messages.

**Reference**

SMS Spam Overview. (n.d.). Retrieved from

<https://www.cloudmark.com/en/s/resources/whitepapers/sms-spam-overview>.

SMS Spam Collection v.1. (n.d.). Retrieved from

http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/.