# CSC 722 Project Proposal
# Title - An evaluation of SMS Spam Classification Techniques

**Team Members:**

| | |
|---|---|
| **Name - Amritanshu Agrawal** | **Unity ID - aagrawa8** |
| **Name - Guilherme Ferreira** | **Unity ID - gferrei** |

## INTRODUCTION

The growth of mobile phone users in the last two decades has lead to a dramatic increase in SMS text messaging, and consequently SMS spam messages. Research has shown that 82.1% of text messaging users tend to open every single SMS message they receive [1], exposing them to spam messages that can possibly contain scam links, pornographic material, malware or ransomware. In practice, fighting mobile phone spam is difficult by several factors, including the lower rate of SMS's that has allowed many users and service providers to ignore the issue, and the limited availability of mobile phone spam-filtering software.

## RELATED WORK

There have been studies done in different classification task like identifying the phishing emails [2] which have compared multiple learners to see which one performs the best which is similar task to classify spam SMS. Nagwani et al.[3] compared the SMS spam classification using Bi-level text classification and clustering techniques. Mahmoud et al.[4] and Cormack et al.[5] played with different feature engineering to classify spam SMS. We wish to combine Nagwani et al., and Cormack et al. work and extend their analysis using various other techniques.

## DATASET

The dataset used in this evaluation is the SMS Spam Collection Dataset [6], from Kaggle, a popular platform for open source datasets. It contains 5574 SMS messages in the form of raw text classified as either spam or ham (legitimate) messages, where 4,827 are legitimate messages and 747 are spam. It was collected by Almeida et al.[7]

## APPROACH

We plan on evaluating several text preprocessing (such as stemming, stopwords removal, etc) and feature extraction techniques (term frequency, term frequency-inverse document frequency, row normalization, feature reductions, etc.) in order to create useful features to serve as input for the different predictors. As the dataset does contain class imbalance, we will also be using class imbalancing techniques such as SMOTE [8].

After preprocessing, we will run an extensive comparison of classification algorithms (some are svm, decision tree, k-nearest neighbors, etc.) in order to determine which ones outperform the others for this particular task. Finally we will analyze the results and see if any insights can be deduced from it.

## **REFERENCES**

[1] http://www.shiftcomm.com/blog/what-is-the-open-rate-of-sms-text-messaging/

[2] Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "A comparison of machine learning techniques for phishing detection." In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pp. 60-69. ACM, 2007.

[3] Nagwani, Naresh Kumar, and Aakanksha Sharaff. "SMS spam filtering and thread identification using bi-level text classification and clustering techniques." Journal of Information Science 43.1 (2017): 75-87.

[4] Mahmoud, Tarek M., and Ahmed M. Mahfouz. "SMS spam filtering technique based on artificial immune system." IJCSI International Journal of Computer Science Issues 9.1 (2012): 589-597.

[5] Cormack, Gordon V., José María Gómez Hidalgo, and Enrique Puertas Sánz. "Feature engineering for mobile (SMS) spam filtering." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[6] Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

[7] https://www.kaggle.com/uciml/sms-spam-collection-dataset

[8] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.