SMS Spam Classification

# **Kyle Morris Winter 2020 https://github.com/jkylemorris/MDSC-Portfolio-Kyle-Morris/**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

I’ll be doing some classification on spam SMS messages. <https://towardsdatascience.com/create-a-sms-spam-classifier-in-python-b4b015f7404b> is an example done with a different dataset.

The FTC has addressed spam text messages and given some tips to spot them in the wild: <https://www.consumer.ftc.gov/articles/how-recognize-and-report-spam-text-messages>

Spam text messages were on the rise as early as 2019: <https://www.usatoday.com/story/tech/columnist/2019/08/29/spam-text-messages-6-clever-ways-stop-them/2141700001/>

Naïve Bayes is a popular choice for classifying text: <https://towardsdatascience.com/implementing-a-naive-bayes-classifier-for-text-categorization-in-five-steps-f9192cdd54c3>

Some more info on logistic regression, the workhorse of classification systems: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

Our dataset we will be using along with some info about it: <https://www.kaggle.com/uciml/sms-spam-collection-dataset>

Some business uses of machine learning classifications in the real world: <https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications>

Machine learning versus rules-based systems: <https://monkeylearn.com/text-classification/>

Some info on the Natural Language ToolKit - <https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a>

12 best text classification tools - <https://lionbridge.ai/articles/12-best-text-classification-tools-and-services/>

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

<https://www.kaggle.com/uciml/sms-spam-collection-dataset>

It’s a collection of SMS texts, including spam ones. We’re going to be attempting to classify text messages as spam!

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

Can we train a machine learning model on actual spam messages, so that we can proactively identify spam messages before they reach our inbox? Email does a similar job with white/blacklists, text analysis, etc, but I haven’t seen it used as much on SMS.

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

It’s a classification problem, so logistic regression, naïve bayes, are some of the techniques I’ll look at. I would probably do a couple of models so as to better test which method would work best with our data. By classifying based on the text, we may be able to identify Spam messages with some degree of accuracy.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

Interestingly enough, our dataset originates from Singaporean English, so I wonder if a model trained on that text would perform as well on English language spam? I might have to run some tests on that after the bulk of work is done. I suspect that rabbit hole would suck up some time if I got too involved in it!

# Concluding Remarks

Tie it all together. Think of this section as your final report’s abstract.

Linear regression is a useful tool to solve classification problems using a supervised machine learning process. By taking the time to classify the initial dataset, we can save time by classifying future data with, depending on our model, some degree of accuracy. Some business uses for this would be useful for routing support emails to the correct team i.e. certain priorities assigned to tickets based on the text used in the support email.