SMS SPAM FILTER MODEL

BY PHONG PHAM

THE PROBLEM

- Singtel is concerned with an increase in customer complaints about the number of spam SMSs they receive
- The company wants to find data-driven solutions to reduce spam on the network and encourage customers to continue their mobile plans.

OBJECTIVES

- To create a spam filter system that can reduce the number of spam SMSes by 75% by the end of 2021.

STRATEGIES

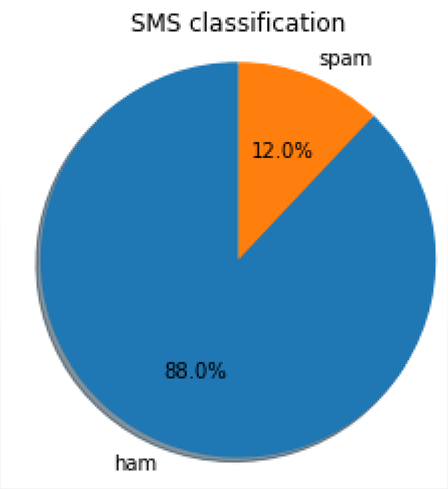
- Analyze 5,574 SMSs from free for research sources on the Internet, with the messages having been correctly categorized as either spam or ham.
- Come up with a prediction model that can be used in an in-house spam filter system

DATA INFORMATION

- Dataset downloaded from: SMS Spam Collection
- 5,547 records and 2 columns (message content and spam flag)
- Target variable is spam flag, which is a categorical variable

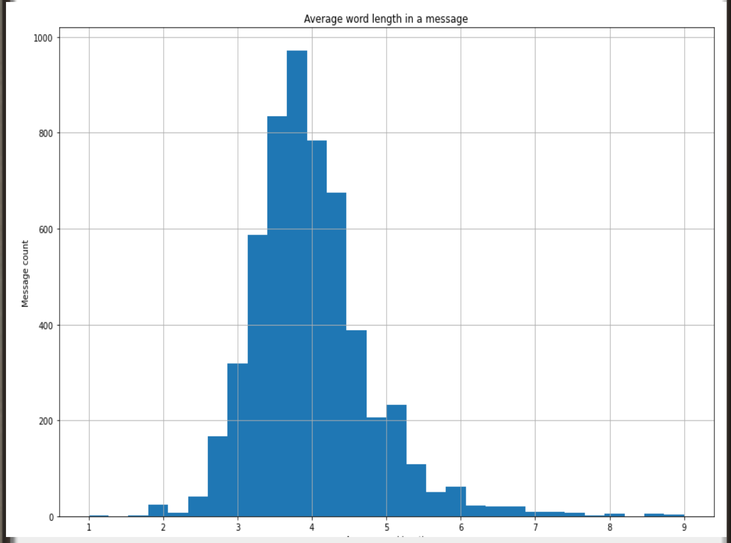
DATA EXPLORATION: CLASS IMBALANCE

- Only 12% of all SMSs are classified as spam.
- Might lead to prediction inaccuracies with too many false positive (i.e. actual non-spam but labeled as "spam")



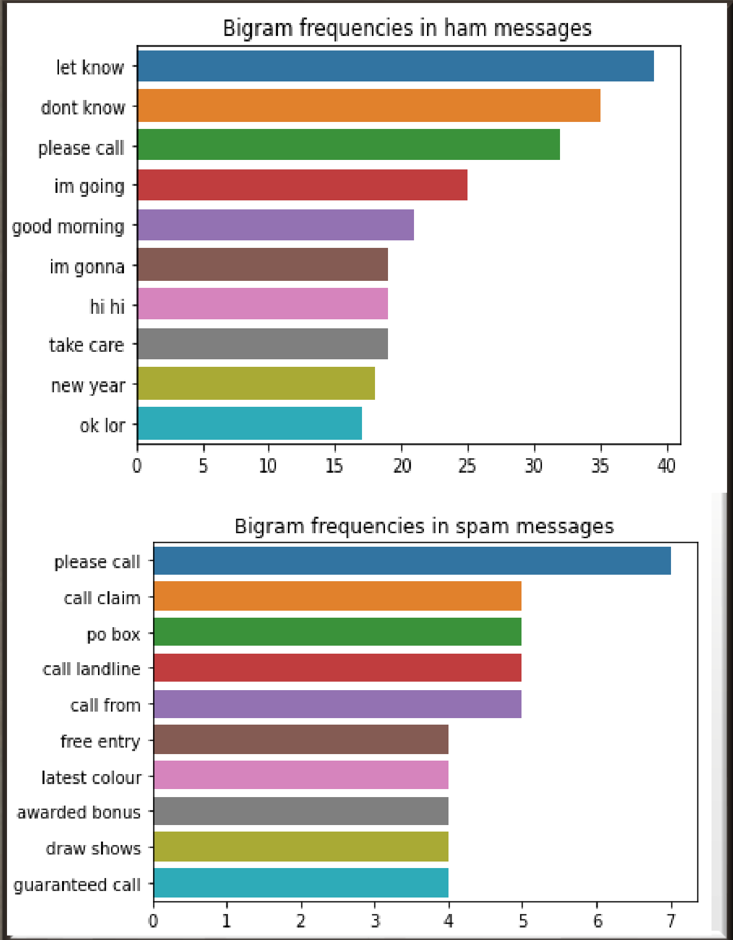
DATA EXPLORATION: AVERAGE WORD LENGTH

- After removing outliers, the average word length of messages ranges from 1-9 letters
- A considerable number of messages have an average word length of between 3-5 letters



DATA EXPLORATION: BIGRAMS

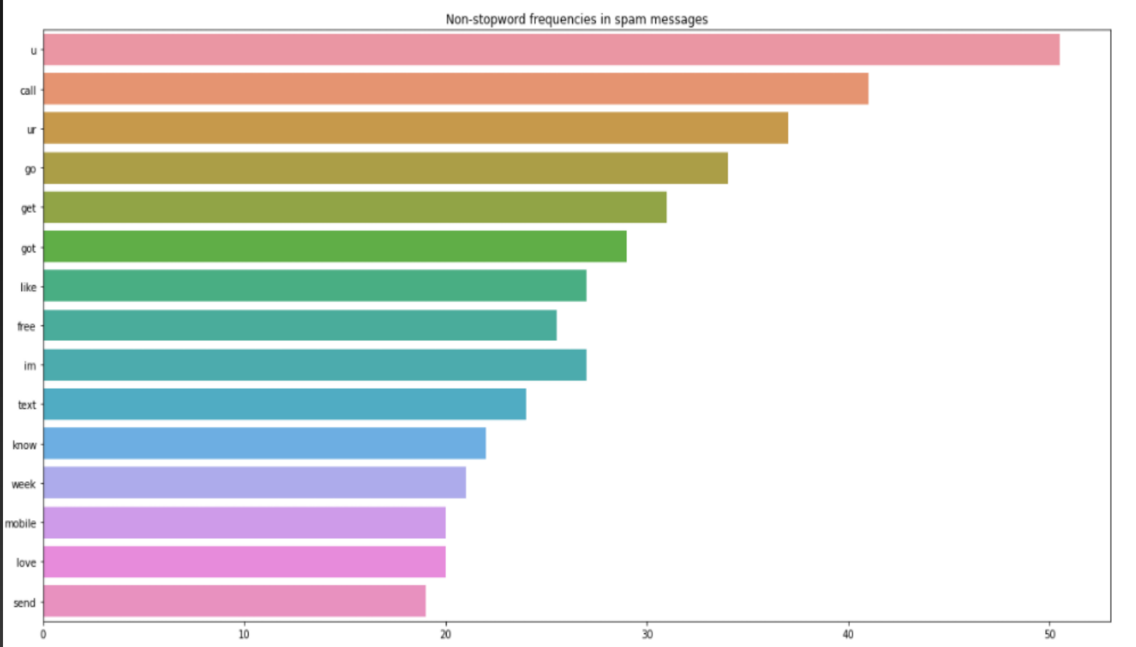
- Top spam bigrams are word combinations acting as calls-to-action
  - Likely to involve some type of scams
- For non-spam, more conversational bigrams appear the most



DATA EXPLORATION: NON-STOPWORD FREQUENCIES

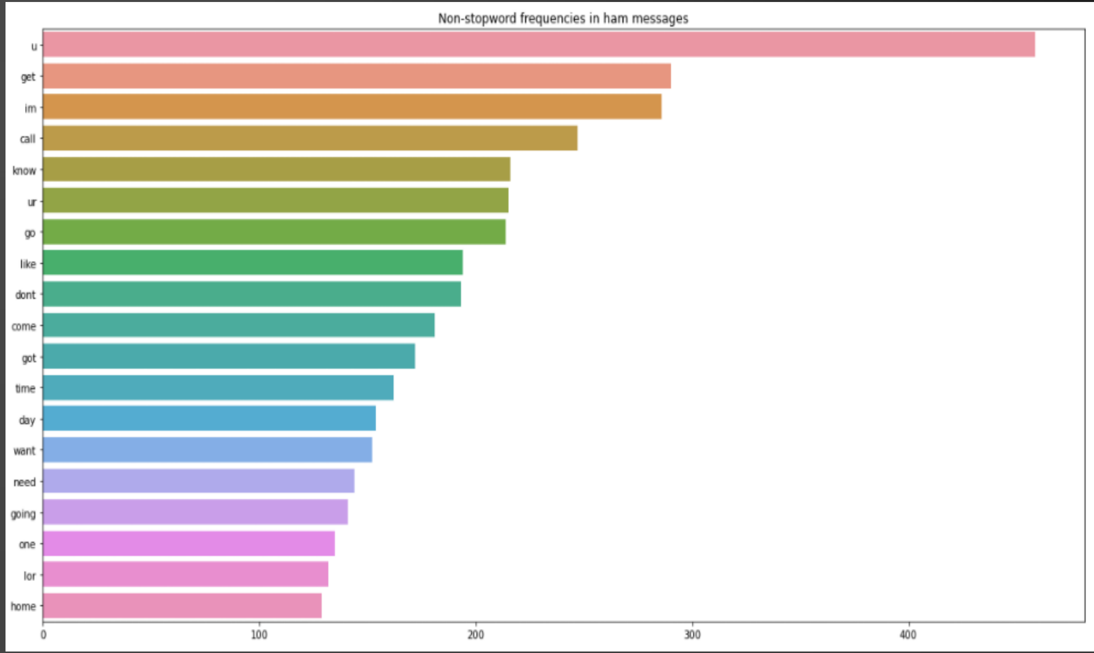
Spam:

- **Mainly calls-to-action to encourage recipients (“u” and “ur”) to contact spammers (“call”, “text”, “send”)**
- **“Free”: entice recipients to take action.**



Non-spam:

- **Mainly conversational language**
- **“Call”: when something needs to be discussed at length**



MACHINE LEARNING: PREPROCESSING STEPS

1. Remove stopwords, numbers, and emoticons
2. Lemmatization: find the root (lemma) of each word
3. One-hot encoding for "flag" column
4. Use CountVectorizer and TF-IDF: how relevant a word is to each message
5. Oversampling the minority class (Spam) using SMOTE
6. Split data into training and testing sets (with a 65:35 ratio)

MACHINE LEARNING: MODELING

- The following classification algorithms were used
  - Logistic Regression
  - K-Nearest Neighbor (KNN)
  - Random Forest
  - Gradient Boosting
- Accuracy, precision, cross validation scores were used to evaluate each model

MACHINE LEARNING: MODELING

Algorithm	Accuracy	Precision	CV test score
Logistic Regression	0.8633	0.7949	0.8494
K-Nearest Neighbors	0.53	0.5181	0.7723
Random Forest	0.9237	0.936	0.926
Gradient Boosting	0.8290	0.86	0.8415

- Random Forest and Logistic Regression yielded the best accuracy and CV test scores
- Hyperparameter tuning done on these 2 methods using randomized cross validation

MACHINE LEARNING: MODELING

- Results from using optimal parameters for Random Forest and Logistic Regression

Algorithm	Accuracy	Precision
Random Forest	0.9335	0.9418
Logistic Regression	0.8783	0.8218

- The fewer false positives (higher precision) the better the prediction model

  ➔ Random Forest should be used for future predictions

CONCLUSIONS

- More than 5,000 messages were engineered into word vectors, which were used in predicting the spam flag
- Random Forest provided the best results out of 4 supervised classification models

SOLUTIONS

- Build an in-house spam filter using the Random Forest prediction model
- Survey customers every quarter to measure the efficacy of the SMSs spam filter

LIMITATIONS AND FUTURE ACTIONS

- Data mainly consists of SMSs sent in Singapore
  - → Useful to incorporate data from other sources in case the spam filter usage is expanded to other countries
- Prediction model can be recalibrated with email data and used for email spam filter in the future