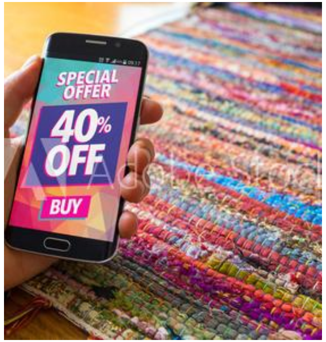
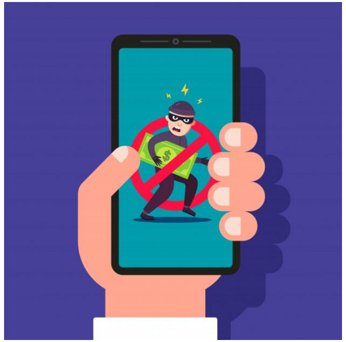
Predicting Spam Message from SMS Board

Group Member: Winnie Shao, Wenjia Dou, Yihui Zhang

Motivation and Backgrounds





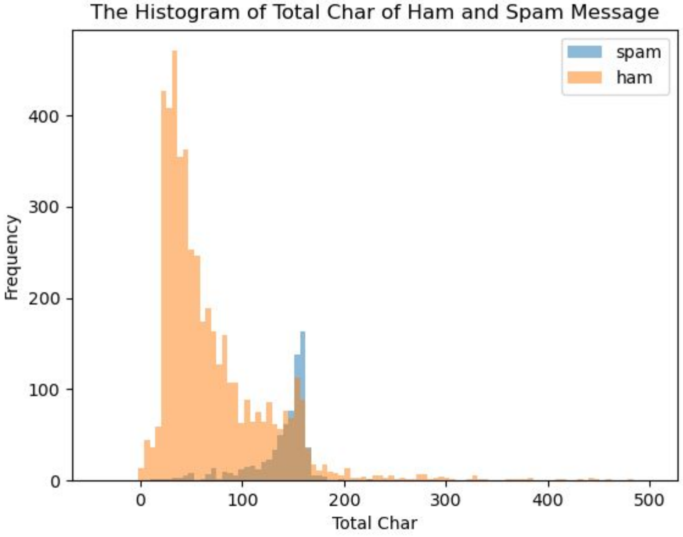
Research Question

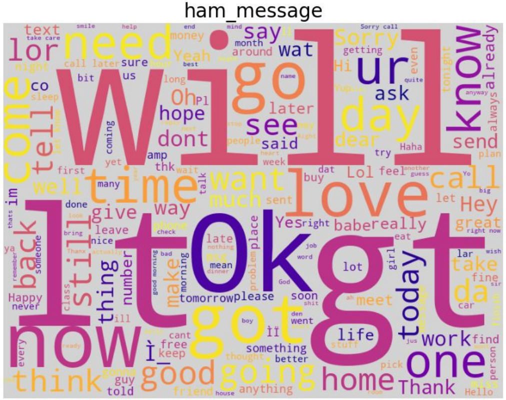
1. What information can we get to distinguish the difference between ham and spam messages if we only perform basic exploratory data analysis?
2. How can we process the text message to make them machine-readable?
3. What model is most effective in predicting spam messages?
4. What features are most important in predicting spam messages?

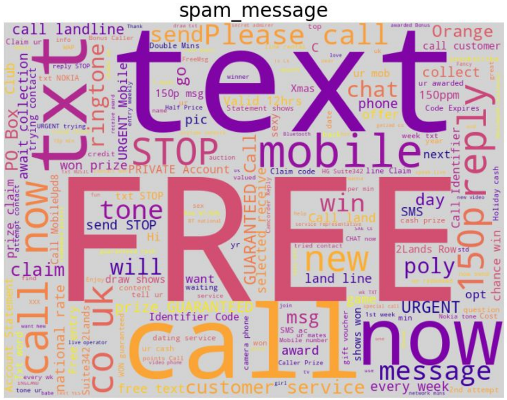
Methodology

1. Data cleaning and exploratory analysis.
2. Converting raw text messages into five numerical features.
3. Finding the best prediction model by training four machine learning models, including Decision Tree Classifier, Naive Bayes Classifier, SVM Classifier, and Random Forest Classifier.
4. Finding the most significant predictor for spam messages by using the 5 features one by one with the most effective model we obtained.

Exploratory Analysis







Machine Readable Features

Converting the raw text messages into five numeric features.

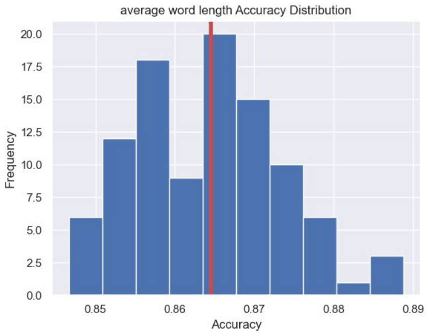
1. **word count**: total number of words of each message.
2. **average word length**: total length of words divided by the total number of words.
3. **url count**: total number of tokens that start with “http” (case non-sensitive).
4. **spc**: total numbers of special punctuations except for commas and periods.
5. **tf-idf**: a sparse matrix contains TF-IDF scores between each message and each term.

Machine Learning Models

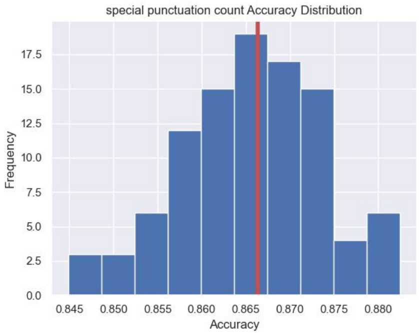
We split the language feature generated data into a 20% training set and an 80% testing set and then find the model with the most accurate score.

Decision Tree (with max depth 12)	Naive Bayes	SVM	Random Forest (with max depth 60)
0.9614349775785	0.8672645739910	0.9820627802691	0.9704035874439

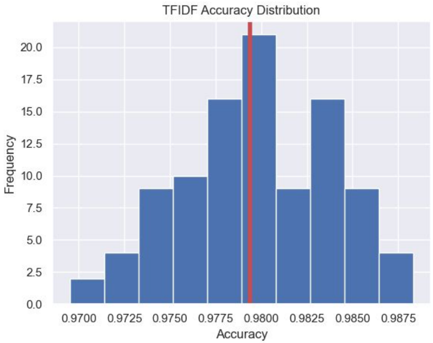
Most Important Feature(s)



average word length Accuracy Distribution

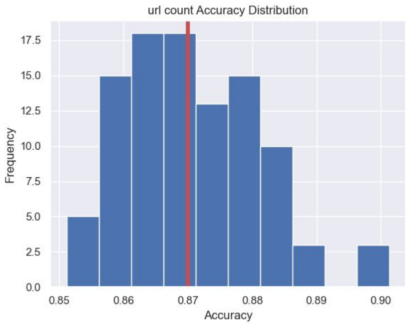


special punctuation count Accuracy Distribution

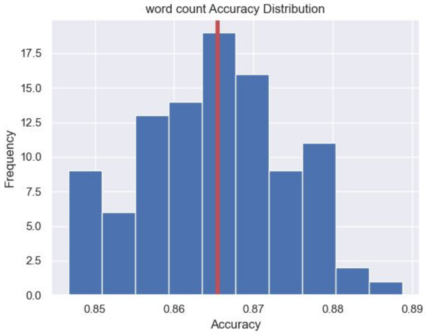


TFIDF Accuracy Distribution

Highest accuracy!



url count Accuracy Distribution



word count Accuracy Distribution

Future Work

1. Normalization/scalization on our features
2. Implementing KNN/XGBoost model
3. Finding more potential features

Thank you!