Visual spoofing in content-based spam detection

Mark Sokolov, Kehinde Olufowobi, and Nic Herndon
East Carolina University

13^{th} International Conference on Security of Information and Networks
November 4 – 7, 2020
Istanbul, Turkey

Overview

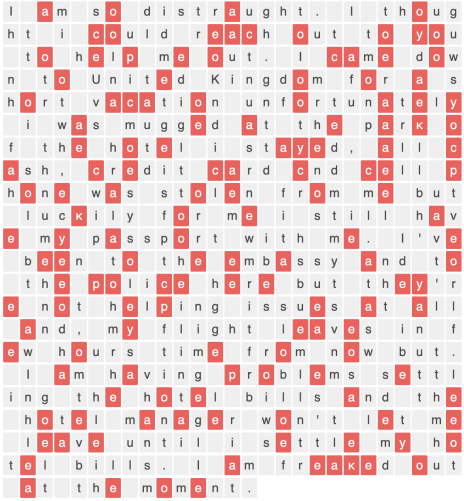
1. Introduction
   - Unchanged vs. changed email
2. Experimental design
   - Data source and format
   - Workflow
3. Results
   - Accuracy
   - Microsoft mail
   - Google mail
   - Other domains
4. Summary

Introduction

Subject: Please send money

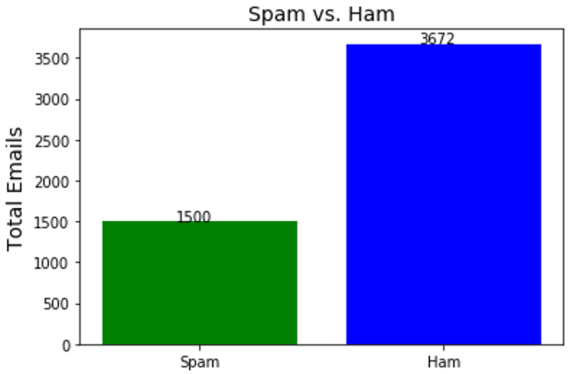
Body: I am so distraught. I thought i could reach out to you to help me out. I came down to United Kingdom for a short vacation unfortunately i was mugged at the part of the hotel i stayed, all cash, credit card and cell phone was stolen from me but luckily for me i still have my passport with me. I've been to the embassy and to the police here but they're not helping issues at all end, my flight leaves in few hours time from now but. I am having problems settling the hotel bills and the hotel manager won't let me leave until i settle my hotel bills. I am freaked out at the moment.

Unchanged vs. changed email

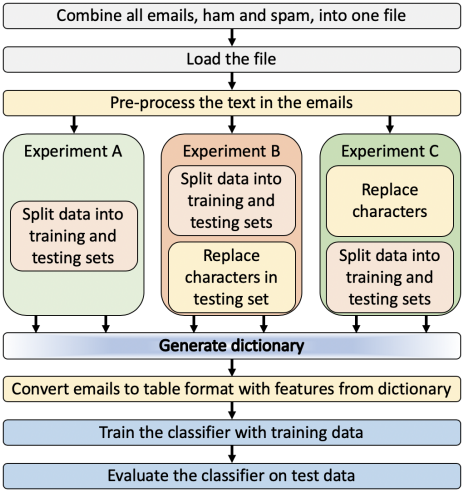


Data source

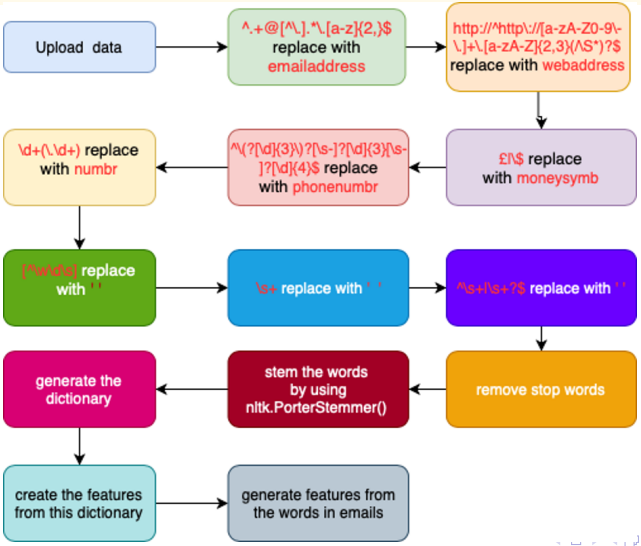
- Emails from the Enron1 dataset



Workflow


Combine all emails, ham and spam, into one file
Load the file
Pre-process the text in the emails
Experiment A
Split data into training and testing sets
Experiment B
Split data into training and testing sets
Replace characters in testing set
Experiment C
Replace characters
Split data into training and testing sets
Generate dictionary
Convert emails to table format with features from dictionary
Train the classifier with training data
Evaluate the classifier on test data

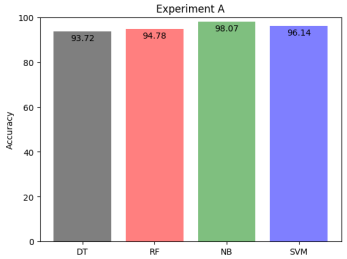
Pre-processing



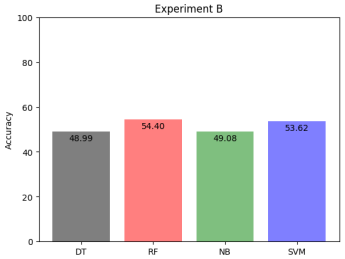
Analysis details

- Machine learning algorithms used
  - Naïve Bayes (NB)
  - Support vector machine (SVM)
  - Decision tree (DT)
  - Random forest (RF)
- Evaluation metrics
  - **Accuracy**
  - Precision
  - Recall
  - F-1 score
  - Confusion matrix
  - Production testing (Microsoft)

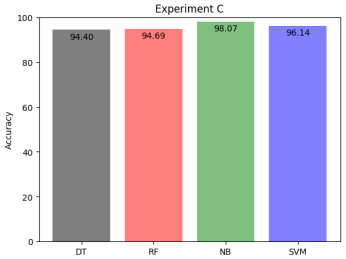
Results and discussion



Control experiment involving entirely unmodified datasets – the default encoding of characters in both the training and testing sets is preserved

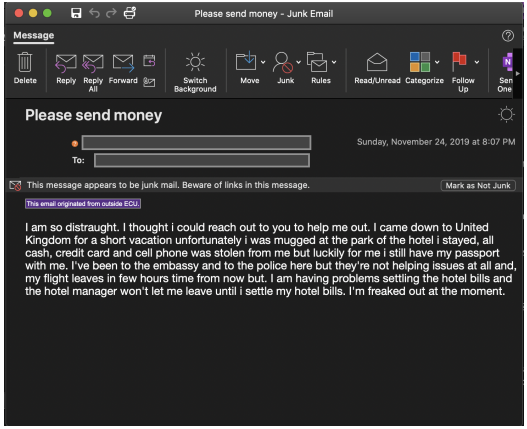


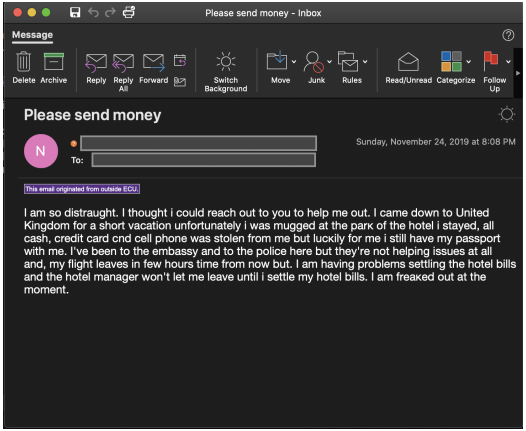
Encoding of the training set is preserved and the encoding of certain characters in the testing set uses corresponding confusables from the Cyrillic alphabet



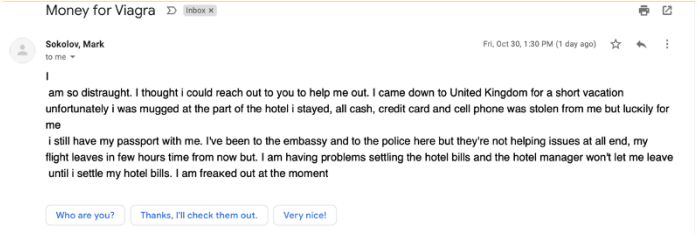
Confusables introduced in training and testing sets so that each model is trained and evaluated with data from a single, mixed-script that contains confusables

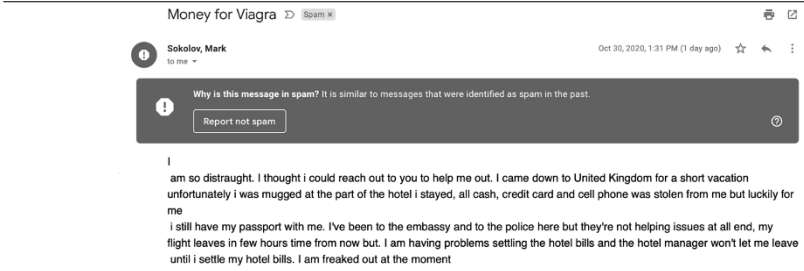
Junk mail vs. inbox – Microsoft mail



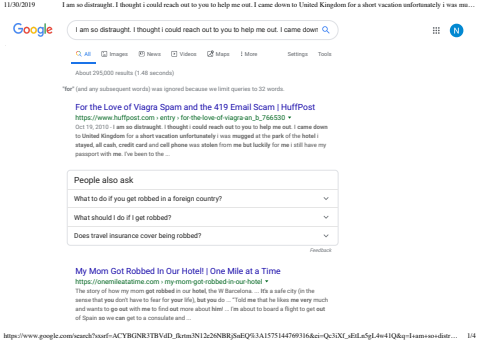


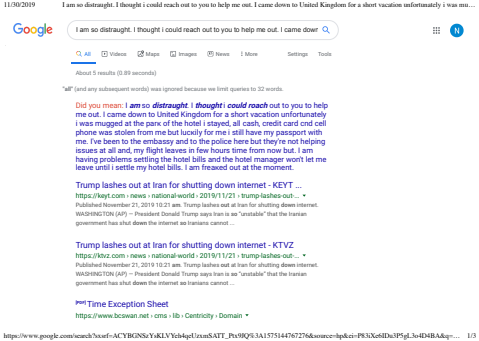
Junk mail vs. inbox – Gmail



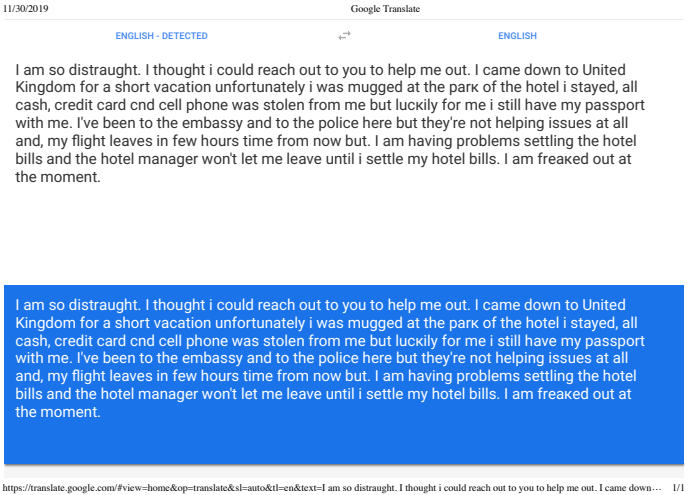


Google search





Google translate



Summary

Questions

Mark Sokolov
sokolovm19@students.ecu.edu

https://github.com/sokolovm19/Visual_Spoofing