Mark Sokolov, MS in Software Engineering
Advisor: Dr. Nic Herndon

Visual Spoofing in Content-Based Spam Detection

Visual spoofing in content-based spam

- Visual spoofing is a type of threat where the similarities of characters and letters from different languages are used to confuse or trick users.
- The content-based filtering is also known as cognitive filtering that recommends items based on a comparison between the content of the items and a user profile items.

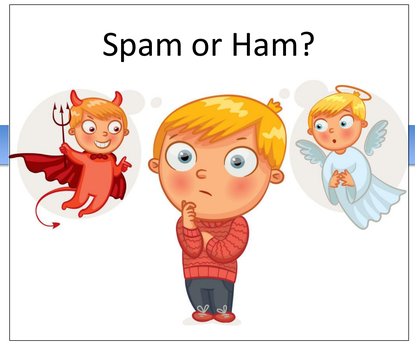
Background

- The Unicode standard provides a unique number for every character, enabling disparate computing systems to exchange and interpret text in the same way.
- Owing to ease of access, the number of email users today continues to grow rapidly. The use of email extends beyond just messages in text form; it is also used for sharing other types of data – images, videos, archive files, etc.
- Scammers constantly refine their approach of using fraudulent emails to induce individuals to reveal sensitive or private information – a practice known as phishing.
- 22% of organizations see phishing as their greatest security threat
- 94% of malware was delivered via email
- 65% of attacker groups used spear phishing as the primary infection vector

Research goals

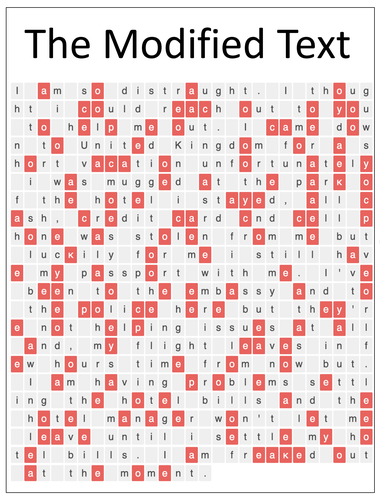
- Identify new problems in the emails spam filtering.
- Classifying spam emails as ham emails and propose methods to address this threat.
- Protect people from cybercriminals and losing personal information such as bank account and card data, logins and passwords of Internet services, and so on.

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 park of the hotel i stayed, all cash, credit card cnd 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 and, 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.

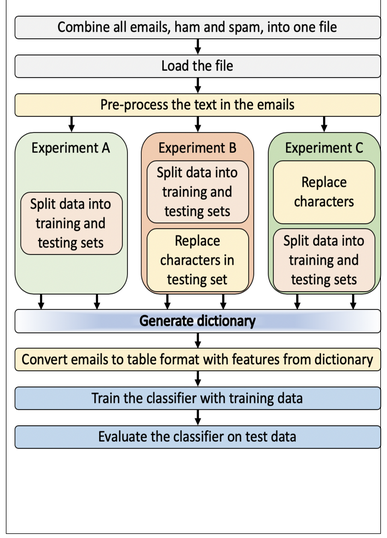


I am so distraught. I thought i could reach out to you to help me out. I came down to United Kingdom for a short vacation unfortunately i was mugged at the park of the hotel i stayed, all cash, credit card and cell phone was stolen from me but luckily for me i still have my passport with me. I've been to the embassy and to the police here but they're not helping issues at all and, 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.

The Modified Text



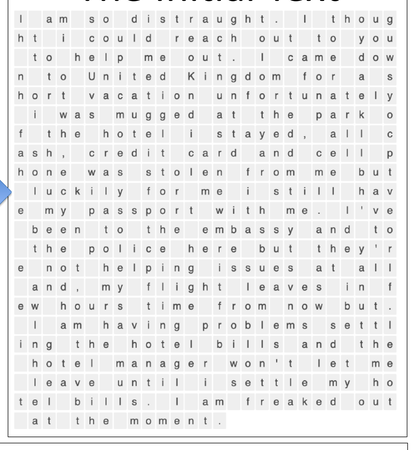
The Experimental Design

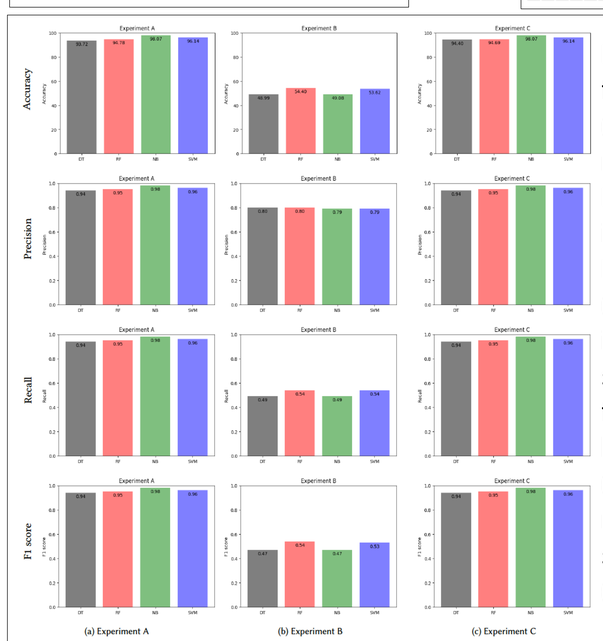


The Characters Replaced, and their Replacements

Character	Replacement
a	U+0430
e	U+0435
k	U+043A
o	U+043E
p	U+0440
c	U+0441
y	U+0443

The Initial Text





Evaluation metrics for Experiment A (no characters replaced), Experiment B (characters replaced in the test set), and Experiment C (characters replaced in train and test sets) from the following machine algorithms used: decision tree (DT), random forest (RF), naïve Bayes (NB), and support vector machine (SVM).

Experiment A

Each email text was preserved in its original form – i.e., no characters were replaced. We trained and evaluated our set of classifiers using data from this distribution only.

Experiment B

We modified our testing data by introducing corresponding Cyrillic letters in place of the Latin letters 'a', 'e', 'k', 'o', 'p', 'c', and 'y'. As desired. This resulted in single and mixed-script confusables in our testing dataset. The intent here was to simulate a visual spoofing effect in the email messages that our testing set consists of. With the original character encoding in our training dataset still preserved, we trained all classifiers using data from the same distribution as in experiment A. However, model evaluation is done using data effectively from a different distribution than the one used in experiment A.

Experiment C

We modified both our training and testing datasets to replicate visual spoofing in the entire distribution used to develop our model, using the set of confusables introduced in experiment B. Training and evaluation was performed using data from this modified distribution only. As a result, unlike experiments A or B, each of our models would simulate spam filters designed to classify emails that contain visual spoofing.

Conclusion and future work

- These experiments indicate that using a classifier trained on data using Latin alphabet, to classify a message with a combination of Latin and Cyrillic letters leads to much lower classification accuracy compared to the same classifier used with a message with Latin characters only.
- In future work we plan to evaluate this approach with characters from multiple alphabets. In addition, we would like to investigate the impact of this method with other applications used for text communication.