A Multilingual Comparison of Email Scams

Duo Pan¹, Ellen Poplavská¹, Yichen Yu², Susan Strauss² and Shomir Wilson¹

¹College of Information Sciences and Technology, Pennsylvania State University  ²College of The Liberal Arts, Pennsylvania State University

Motivation

The existing research on scam emails focuses on scams written in English.

Can we build a multilingual scam email dataset to find a more culturally-aware approach to email safety education?

Results

1. English vs. French orphan scams
   English scammers preferred Gmail (66%); French scammers preferred Yahoo Mail (61%). English scammers created more varied scam scenarios.

2. French vs. Russian romantic scams
   Frequent locations in French romance scams: Canada, London, and Cotonou. Russian romance scams are similar to advance-fee scams because 48% of senders' IP addresses are from Senegal.

Contributions

1. Built first multilingual scam email dataset
2. Created categories for sorting scam emails consistently across all languages
3. Provided insights on developing culturally-aware scam education

Annotation Schemes

The first annotation scheme is a list of seven mutually exclusive categories for sorting the sub-forums on the Anti-Fraud International website into different scam types.

The second annotation scheme is a list of words frequently found in scam emails about different topics across different languages.

1. Cultural differences between different languages' scam emails
   The English corpus has more transaction scams, while the French and Russian corpora have more romance scams.

2. Developing insights for education
   When writing educational materials in different languages, researchers should focus on different types of scams, providing specific guidance for users who speak different languages.

Datasets

1. From the first Annotation Scheme   2. From the Anti-Fraud International online forum

Number of Crawled Scam Emails in Each Language

<table>
<thead>
<tr>
<th>Language</th>
<th>Top Categories¹</th>
<th>Top Forums²</th>
<th>Number</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Profit, Romance, Helping</td>
<td>Dying, Romantic, Orphans</td>
<td>100, 99, 100</td>
<td>299</td>
</tr>
<tr>
<td>Russian</td>
<td>Profit, Romance, Helping</td>
<td>Lottery, Romantic, Cancer Victims</td>
<td>70, 27, 14</td>
<td>111</td>
</tr>
</tbody>
</table>

Future Work

1. Continuing to analyze categories and languages within our dataset
2. Crawling more emails in other languages, such as Chinese and Korean
3. Creating protocols for scam educators to provide more diverse examples and warnings tailored to users’ languages