The Human Language Technologies Lab
at Penn State
http://hltlab.org

PennState College of the Liberal Arts
PennState College of Information Sciences and Technology

A Multilingual Comparison of Email Scams

Duo Pan[1], Ellen Poplavska[1], Yichen Yu[2], Susan Strauss[2] and Shomir Wilson[1]
[1]College of Information Sciences and Technology, Pennsylvania State University [2]College 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.

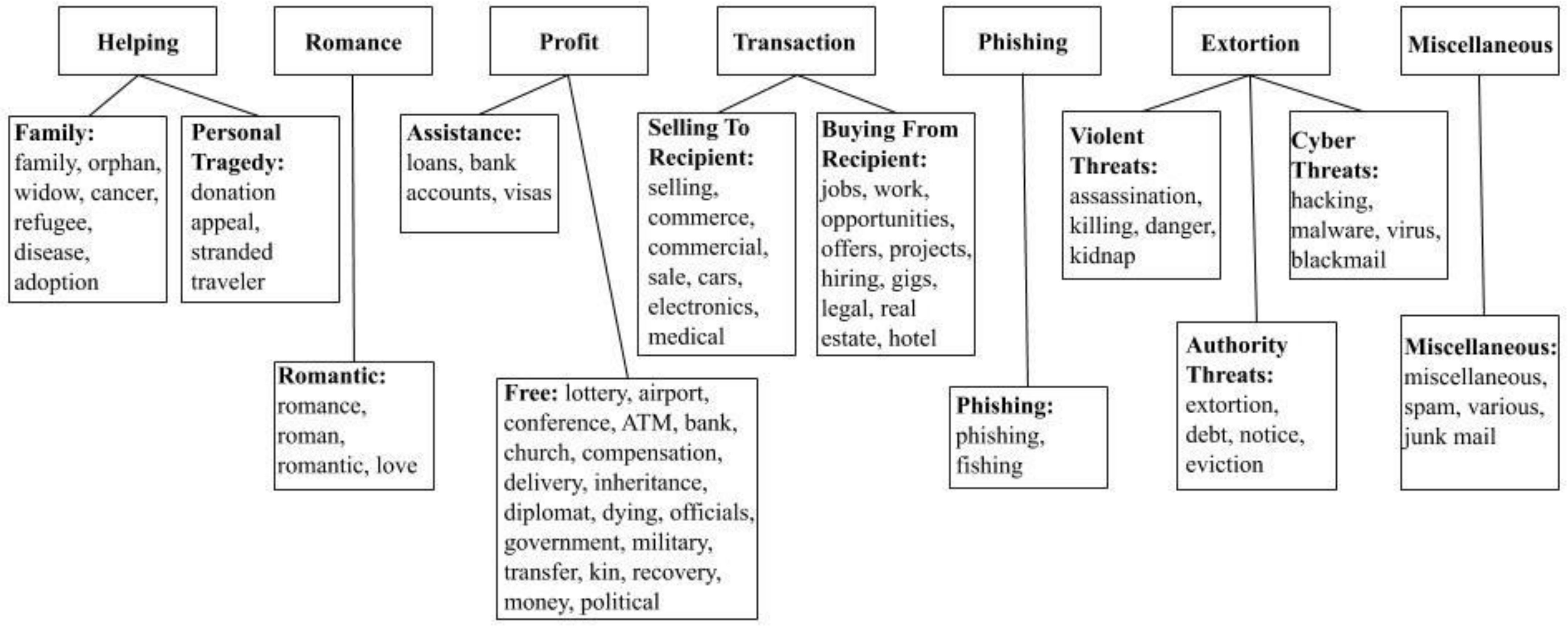
Contributions

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.



Helping
Family: family, orphan, widow, cancer, refugee, disease, adoption
Personal Tragedy: donation appeal, stranded traveler

Romance
Romantic: romance, roman, romantic, love

Profit
Assistance: loans, bank accounts, visas
Free: lottery, airport, conference, ATM, bank, church, compensation, delivery, inheritance, diplomat, dying, officials, government, military, transfer, kin, recovery, money, political

Transaction
Selling To Recipient: selling, commerce, commercial, sale, cars, electronics, medical
Buying From Recipient: jobs, work, opportunities, offers, projects, hiring, gigs, legal, real estate, hotel

Phishing
Phishing: phishing, fishing

Extortion
Violent Threats: assassination, killing, danger, kidnap
Cyber Threats: hacking, malware, virus, blackmail
Authority Threats: extortion, debt, notice, eviction

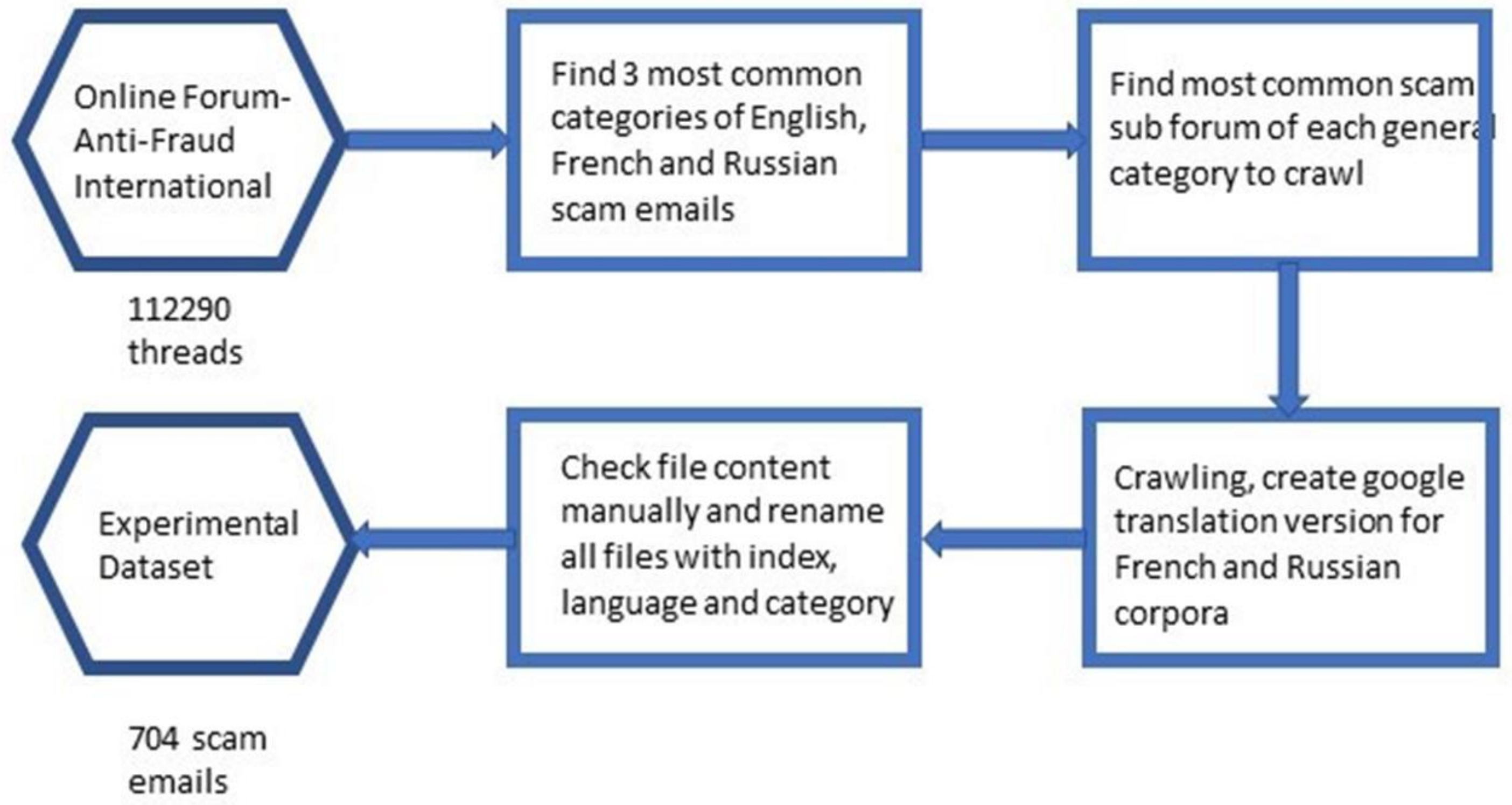
Miscellaneous
Miscellaneous: miscellaneous, spam, various, junk mail

An illustration of forum title keywords and how they fall into the category system

Datasets

Number of Crawled Scam Emails in Each Language				
Language	Top Categories[1]	Top Forums[2]	Number	Total
English	Profit	Government	98	294
	Transaction	Business	99	
	Helping	Orphans	97	
French	Profit	Dying	100	299
	Romance	Romantic	99	
	Helping	Orphans	100	
Russian	Profit	Lottery	70	111
	Romance	Romantic	27	
	Helping	Cancer Victims	14	

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



Online Forum-Anti-Fraud International
112290 threads
Find 3 most common categories of English, French and Russian scam emails
Find most common scam sub forum of each general category to crawl
Crawling, create google translation version for French and Russian corpora
Check file content manually and rename all files with index, language and category
Experimental Dataset
704 scam emails

The pipeline from online forum to dataset

Discussion

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.

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