A Comparison of Machine Learning Algorithms for Multilingual Phishing Detection

Dakota Staples, Paul Cook, Saqib Hakak

Faculty of Computer Science, University of New Brunswick

dstaples@unb.ca, paul.cook@unb.ca, saqib.hakak@unb.ca

Introduction

- Phishing emails are increasing in volume
- Little data exists for phishing emails in languages besides English
- Most research only focuses and trains systems on English due to the lack of data
- We evaluate different systems to detect phishing emails in multiple languages. For our experiment we use English, French and Russian as this is the data that is available.
- The system is widely zero shot, with the model never seeing the testing language during training at all, except for the three monolingual tests

Data

- English Train - 3983 emails
- English Test – 996 emails
- French Train – 472 emails
- French Test – 119 emails
- Russian Train – 175 emails
- Russian Test – 44 emails
- EnglishFrench – 5570 emails
- EnglishRussian – 5198 emails
- FrenchRussian – 810 emails

Data Acquisition

- English data was taken from the Enron Spam dataset, specifically Enron Spam 1.
- French data was acquired from other researchers who collected the spam emails. Benign emails in this set were translated from the TREC07 dataset.
- Russian data was acquired from other researchers who collected the spam emails. Benign emails in this set were translated from the Enron dataset.

	En/En	Fr/En	Ru/En	Fr,Ru/En	En/Fr	Fr/Fr	Ru/Fr	En,Ru/Fr	En/Ru	Fr/Ru	Ru/Ru	En,Fr/Ru
GPT2	0.99	0.72	0.66	0.61	0.5	1	0.56	0.48	0.7	0.54	0.95	0.68
GPT3	0.99	0.67	0.28	0.76	0.78	1	0.63	0.68	0.81	0.77	1	0.77
XLMR	0.99	0.72	0.71	0.99	0.68	0.98	0.68	0.99	0.95	0.5	0.97	0.95
LR	0.93	0.62	0.4	0.74	0.79	0.96	0.45	0.59	0.5	0.36	0.5	0.55
RF	0.91	0.64	0.28	0.67	0.41	0.94	0.43	0.41	0.54	0.45	0.5	0.5
SVM	0.78	0.73	0.45	0.75	0.62	0.88	0.49	0.63	0.47	0.52	0.5	0.63
MFC Baseline	0.71	0.71	0.71	0.71	0.52	0.52	0.52	0.52	0.5	0.5	0.5	0.5

Future Direction

- Leverage new models such as GPT-4 to see how they perform.
- Look at obtaining more data in multiple languages

Conclusion

- Monolingual spam detection is an easy problem
- Multilingual spam detection is a hard task
- By training on multiple languages, we can improve the accuracy of our models on average