



Using Lexical Features for URL Classification- A Machine Learning Approach

Apoorva Joshi
FireEye

Levi Lloyd
Lawrence Livermore National Lab

Paul Westin
FireEye

Srini Seethapathy
FireEye

OUTLINE

- Motivation
- Previous Work
- Deployment Specification
- Requirements and Goals
- Tasks and Tools
- Observations
- Conclusions
- Future Work

MOTIVATION

From: "Help Desk" <[@marshall.edu](mailto:helpdesk@marshall.edu)>
Date: November 20, 2017 at 11:12:45 AM EST
To: [@live.marshall.edu](mailto:live.marshall.edu)
Subject: Termination Notice
Reply-To: masonjohn459@gmail.com



A notice from my IT Department

Our record indicates that you recently made a request to terminate your Office365 email. And this process has begun by our administrator.

If this request was made accidentally and you have no knowledge of it, you are advised to cancel the request now

Please give us 24 hours to terminate your account OR.

[Sign in here to cancel termination](#)

Failure to cancel termination will result in the termination of your account <http://bit.do/dUcQw>

Thanks

↑
Hover mouse over URL;
not a marshall.edu site

Want to go to UWF for free Inbox x

5:35 PM (0 minutes ago) ☆

to me ▾

Just click the link below and provide your social security number and banking info:

<http://uwf.edu/media/university-of-west-florida/offices/controller/forms/TITLE-IV-AUTHORIZATION-FORM.pdf>

real_form.com/please-trust-me

Hovering the mouse over a link will display the real URL below.

AT&T Automatic Billing Message - Message (HTML)

Message

From: AT&T Reach Out And Touch Someone [REDACTED] Sent: Tue 7/25/2017 5:14 PM
To: [REDACTED]
Cc:
Subject: AT&T Automatic Billing Message

Message | ATT00001.txt | ATT00002.htm

AT&T Official - Entertainment, TV, Wireless & Internet

Dear Customer,
Your monthly wireless bill for your account is now available.
Total Balance Due: \$491.57
[View Billing Statement](#)
Electronic bill message.

Thank you,
AT&T Online Services

Hover your mouse over links to see the true URL

<http://koyalantie.fi/XOON622261/>

PREVIOUS WORK

Blacklists



Host information, network traffic etc.



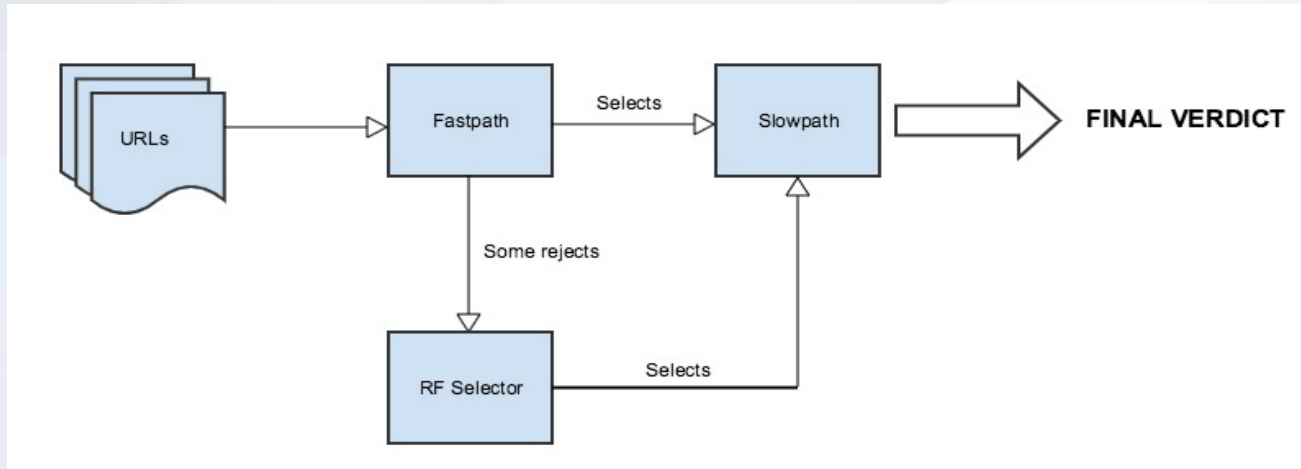
HTML and Javascript content



LEXICAL FEATURES ONLY???

DEPLOYMENT SPECIFICATION

- Model should run as plugin for FAUDE (FireEye Advanced URL Detection Engine)
- Should correct FNs from fastpath analysis
- URLs to be sent for slowpath analysis based on the model verdict



REQUIREMENTS AND GOALS

- Model should act as a means of down selection and/or detection
- False Negative Rate should be very low
- False Positive Rate such that the model results in at most 20% increase in current load
- Model latency should be in the order of 10^{-1} ms

THE DATASET

- ~5.5 million labelled URLs
- 60% benign, 40% malicious URLs
- Collected from different sources – Openphish, Alexa whitelists, FireEye products and honeypots

TASKS AND TOOLS

Feature vectors



TASK	TOOLS
N-grams of URLs	NLTK, mmh3
Extract lexical features	urllib
Modelling	Random Forest

FEATURE VECTORS

'www.google.com'

[('w','w','w'), ('w','w','.')]...

1000-long mmh3 hash-based one-hot representation

N-grams

mmh3 based
encoding

23 URL Lexical
features

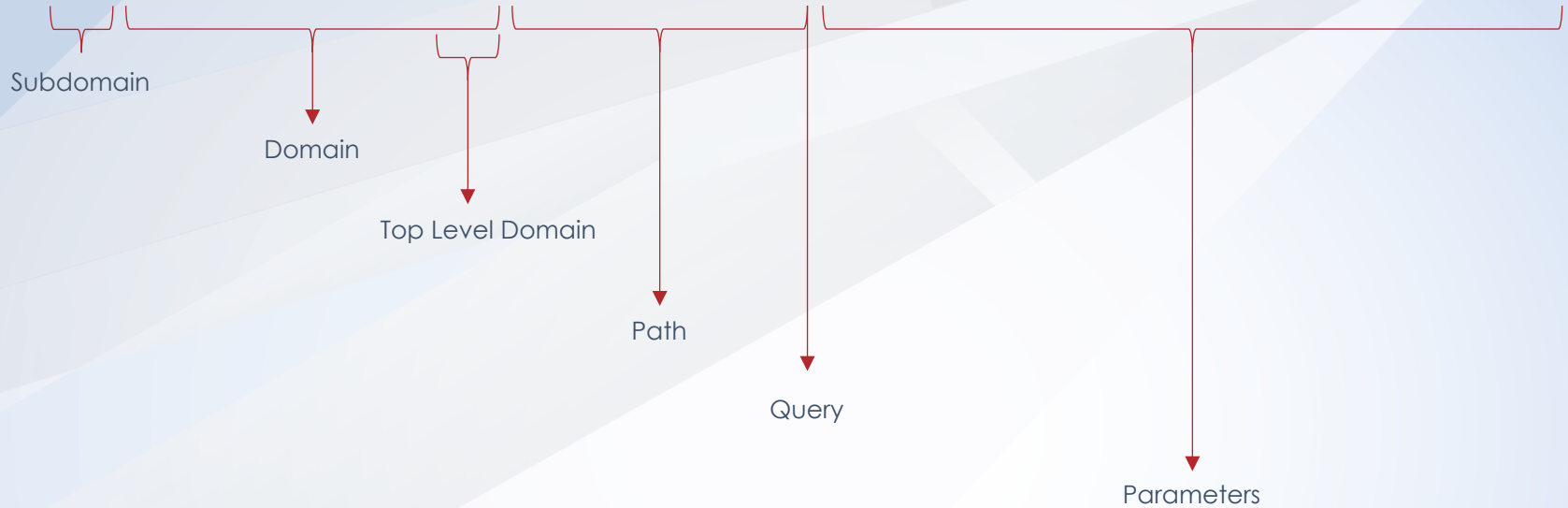
FEATURE VECTOR

Length of domain, number of sub-domains, special characters in URL path etc.

1023-long vector representation of the URL

FEATURE VECTORS

<http://www.video.platinumindustrialcoatings.com/wp-content/plugins.php?to=calgaryps3&message=28dd33dc15e8c68934883418341967>



Complete list of features: <https://arxiv.org/abs/1910.06277>

MODELLING

- Simple Classifiers- Logistic Regression, Naïve Bayes
- Bagging and Boosting Classifiers- Random Forest, Gradient Boost, Adaboost
- Metrics- Accuracy, AUC, FNR

OBSERVATIONS

ALGORITHM	ACCURACY (%)	AUC	FNR (%)
Logistic Regression	87	0.96	4.75
Naïve Bayes	70	0.74	10.38
Random Forest	92	0.99	0.38
Gradient Boost	90	0.92	9
Adaboost	90	0.9	10

OBSERVATIONS

Suspicious URL patterns:

- TLDs in shady list- .biz, .info, .ru, .cn
- Keywords, special characters in URL path
- IP address in primary domain
- High entropy hostnames
- Uppercase or single character directory



OBSERVATIONS

Number of trigrams	Number of lexical features	Accuracy (%)	FPR (%)	FNR (%)
1000	0	85	29.8	0.4
1000	23	92	16.8	0.38
300	23	93	12.5	0.93
100	23	94	11.5	1.09
0	23	95	8.15	1.11

The Random Forest feature importances also showed that it was focusing on both ngram and lexical features

OBSERVATIONS

Max depth	Accuracy (%)	FNR (%)
5	72	1.13
15	88	0.48
20	92	0.38
27	94	0.73
30	95	0.75

Tuning other parameters had no real effect on the evaluation metrics

CONCLUSIONS

- ~22% increase in detections for < 20% increase in load
- Reduction in FNs
- Purely lexical models can be used for fast verdicts on URLs
- Alternative to heuristic-based downselection which needs manual updates

FUTURE WORK

- Deep Learning approach
- Augment the model with new features as necessary
- Cache model verdicts

ACKNOWLEDGMENTS

Levi Lloyd

Paul Westin

Srini Seethapathy

FireEye





Thank You!

Email: apoorva.joshi@fireeye.com