SJSU Student Research Competition Finalist

A SERP-Mining Approach for Classification of DNS Requests

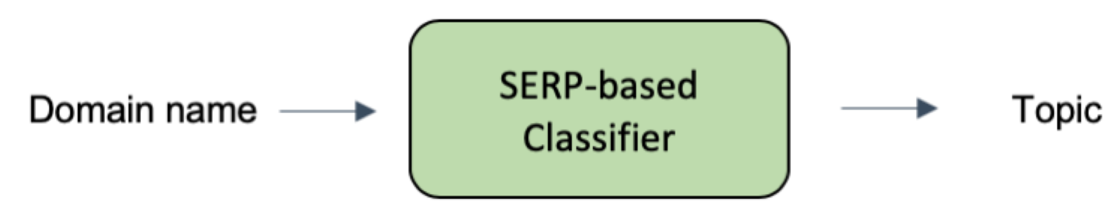
Junlan Lu, Nikhil Saunshi Takappa,
Aldrich Mangune and Dr. Magdalini Eirinaki
College of Engineering

Abstract

In this work, we present and evaluate a machine learning framework that takes as input a domain name (based on the respective DNS request) and outputs the content category it belongs to. We evaluate several options for feature engineering and classification to find the most optimal setup for the specific problem domain. We also address the problem of data collection and preprocessing. We propose a SERP (Search Engine Response Pages)-mining approach to collect and label an appropriate dataset. Our experimental evaluation uncovers several interesting insights and forms the basis for further work into this interesting domain. The problem we addressed is summarized in the High-level architecture diagram.

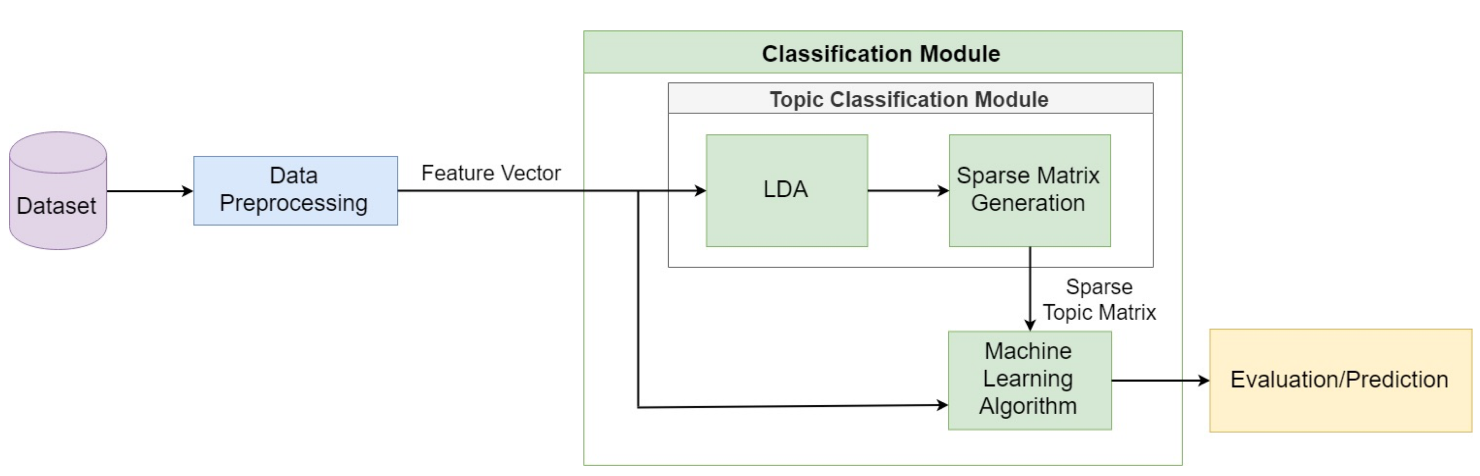
Motivation and Contribution

- There exists several categories of web pages that belong to "borderline" categories (e.g. websites selling illegal substances or weapons) and might be of interest for any public or private organization to monitor as outgoing traffic.
- We built a machine learning framework for classifying DNS requests into topic categories, including data collection, pre-processing, and classification through various configurations.



High-level architecture diagram

System Architecture



The system architecture of the overall framework, containing the DNS Classification module, data pre-processing, feature engineering and classification steps

SERP Dataset

- A total of 112 categories to be classified, with 11,278 instances
- Of those categories, 82 fall under "general" content and 30 fall under "borderline" categories to be monitored.

SAMPLE INSTANCES FROM OUR DATASET

TITLE	*DESCRIPTION*	*CATEGORY*
Amazon Advertising	Start advertising with our self-service solutions ... Combine sight, sound, and motion in ads on Amazon sites, devices like Fire Tablet, and across the web.	Advertising Site
Roku Advertising	If you decline, your information won't be tracked when you visit this website. ... Roku Advertising delivers relevant audiences and measurable results. ... our robust advertising platform offers brands the ability to reach the growing audience that...	Advertising Site

Experimental Results

Table 1: Accuracy of LDA-enhanced ML classification. Results report metric scores for different passes p, different number of topics n and top words t for newsgroup20 and Yelp datasets.

Dataset			*Newsgroup20*					*Yelp*			
Model			rf	knn	svc	nb	lr	rf	knn	nb	lr
p10	n10	t10	0.085	0.046	0.049	0.072	0.066	0.437	0.231	0.159	0.436
		t50	0.258	0.242	0.05	0.241	0.164	0.435	0.443	0.327	0.447
	n20	t10	0.087	0.054	-	0.088	0.068	0.441	0.418	0.333	0.437
		t50	0.328	0.315	0.048	0.307	0.145	0.375	0.38	0.32	0.443
p20	n10	t10	0.078	0.052	-	0.066	0.049	0.453	0.212	0.392	0.452
		t50	0.303	0.287	-	0.274	0.146	0.451	0.44	0.31	0.447
	n20	t10	0.103	0.056	-	0.1	0.073	0.446	0.361	0.36	0.446
		t50	0.329	0.325	-	0.305	0.183	0.373	0.379	0.271	0.45
p50	n10	t10	0.083	0.038	-	0.067	0.048	**0.457**	0.234	0.451	0.456
		t50	0.306	0.298	0.051	0.296	0.185	0.439	0.442	0.305	0.443
	n20	t10	0.089	0.069	-	0.089	0.074	0.438	0.306	0.348	0.438
		t50	**0.352**	0.329	0.048	0.316	0.098	0.402	0.403	0.303	0.443
p100	n10	t10	0.079	0.058	-	0.066	0.054	0.451	0.128	0.44	0.45
		t50	0.308	0.305	0.049	0.293	0.118	0.432	0.432	0.313	0.449
	n20	t10	0.103	0.055	-	0.094	0.059	0.448	0.377	0.339	0.448
		t50	0.34	0.321	0.05	0.3	0.21	0.444	0.425	0.343	0.443

Table 2: Accuracy of LDA-enhanced ML classification. Results report metric scores for different passes p, different number of topics n and top words t for url-title and url-description datasets.

Dataset			*Title*					*Description*				
Model			rf	knn	svc	nb	lr	rf	knn	svc	nb	lr
p10	n10	t10	0.037	0.017	0.017	0.02	0.022	0.026	0.013	0.021	0.019	0.021
		t50	0.3	0.25	0.018	0.085	0.018	0.088	0.066	0.019	0.08	0.019
	n20	t10	0.025	0.01	0.017	0.013	0.022	0.025	0.008	0.018	0.009	0.018
		t50	0.275	0.229	0.022	0.095	0.052	0.134	0.101	0.018	0.092	0.018
p20	n10	t10	0.048	0.018	0.016	0.026	0.027	0.024	0.013	0.014	0.016	0.014
		t50	**0.285**	0.239	0.017	0.099	0.017	0.095	0.081	0.017	0.092	0.017
	n20	t10	0.024	0.007	0.017	0.012	0.024	0.028	0.011	0.022	0.01	0.022
		t50	0.274	0.235	0.023	0.108	0.057	0.145	0.121	0.017	0.129	0.017
p50	n10	t10	0.045	0.019	0.02	0.026	0.029	0.019	0.012	0.017	0.019	0.017
		t50	0.287	0.246	0.021	0.087	0.023	0.102	0.083	0.016	0.085	0.016
	n20	t10	0.021	0.016	0.015	0.015	0.021	0.03	0.012	0.021	0.009	0.021
		t50	**0.285**	0.231	0.018	0.104	0.045	**0.156**	0.131	0.019	0.137	0.019
p100	n10	t10	0.035	0.02	0.016	0.019	0.026	0.03	0.01	0.021	0.017	0.021
		t50	0.255	0.204	0.021	0.098	0.024	0.117	0.087	0.019	0.111	0.019
	n20	t10	0.027	0.011	0.019	0.015	0.022	0.025	0.01	0.019	0.008	0.02
		t50	0.268	0.206	0.015	0.103	0.036	0.145	0.129	0.016	0.127	0.016

Table 3: Precision score, F1 score and cross validation accuracy for *title* and *description* input datasets over all DNS categories

Dataset	*Title*			*Description*		
Model	Precision	F1	Accuracy	Precision	F1-score	Accuracy
rf.w2v	0.84	0.83	0.94	0.76	0.73	0.92
rf.tfidf	0.89	**0.88**	**0.95**	0.85	0.84	**0.94**
knn.w2v	0.81	0.79	0.93	0.74	0.71	0.91
knn.tfidf	**0.91**	0.82	0.78	0.84	0.81	**0.94**
svc.w2v	0.86	0.85	0.93	0.82	0.8	0.89
svc.tfidf	**0.91**	**0.88**	0.93	**0.86**	0.84	**0.94**
nb.w2v	0.84	0.80	0.82	0.76	0.70	0.73
nb.tfidf	0.88	0.85	0.90	0.85	0.76	0.89
lr.w2v	0.86	0.85	0.89	0.81	0.79	0.89
lr.tfidf	0.86	0.85	0.94	**0.86**	**0.85**	**0.94**

Table 4: Precision score, F1 score and cross validation accuracy for "Borderline" subsets.

Dataset	*Title*			*Description*		
Model	Precision	F1	Accuracy	Precision	F1	Accuracy
rf.w2v	0.80	0.78	**0.93**	0.79	0.78	0.94
rf.tfidf	0.81	0.78	0.90	**0.85**	0.82	0.94
knn.w2v	0.70	0.67	0.78	0.77	0.76	0.94
knn.tfidf	0.73	0.72	0.92	0.79	0.77	0.94
svc.w2v	0.79	0.77	0.91	0.83	0.81	0.90
svc.tfidf	0.82	0.79	0.88	**0.85**	**0.83**	0.94
nb.w2v	0.77	0.73	0.77	0.77	0.73	0.80
nb.tfidf	0.80	0.75	0.85	0.84	0.76	0.90
lr.w2v	0.80	0.79	**0.93**	0.84	**0.83**	0.94
lr.tfidf	**0.85**	**0.82**	0.86	**0.85**	**0.83**	**0.95**

Table 5: Precision score, F1 score and cross validation accuracy for "General" subset

Dataset	*Title*			*Description*		
Model	Precision	F1	Accuracy	Precision	F1	Accuracy
rf.w2v	0.82	0.81	0.94	0.75	0.71	0.91
rf.tfidf	0.90	0.88	0.95	0.84	0.83	0.94
knn.w2v	0.78	0.76	0.92	0.74	0.69	0.90
knn.tfidf	0.90	0.84	0.94	0.83	0.81	0.94
svc.w2v	0.86	0.85	0.93	0.80	0.78	0.87
svc.tfidf	0.90	0.87	0.93	**0.88**	**0.85**	**0.95**
nb.w2v	0.83	0.80	0.81	0.74	0.69	0.73
nb.tfidf	0.89	0.85	0.85	0.85	0.77	0.77
lr.w2v	0.85	0.84	0.89	0.81	0.80	0.90
lr.tfidf	**0.91**	**0.89**	**0.95**	0.87	**0.85**	0.94

Conclusions

- Considering the multiple configurations used, Random forest, logistic regression, and SVM were the best performing classifiers and LDA performed less than expected, reinforcing the saying that "simpler is better" in machine learning applications.
- We also observed that the borderline instance classification does not follow the same patterns as the regular ones, with the title of a URL being a more weak indicator of the class label than its description.

IEEE Big Data 2019 Conference Paper: Lu, Junlan & Saunshi, Nikhil & Mangune, Aldrich & Eirinaki, Magdalini & Yu, Bin & Liu, Cricket. (2019). A SERP-Mining Approach for Classification of DNS Requests.