Content-based Classification of Fraudulent Webshops

UNIVERSITY OF AMSTERDAM

System and Network Engineering

SIDN

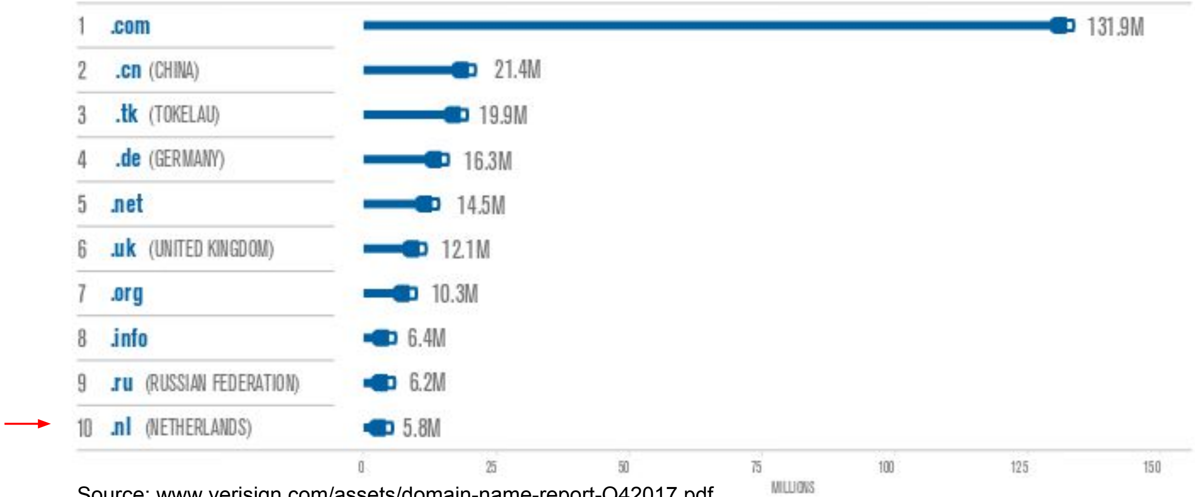
Mick Cox & Sjors Haanen

RP30

July 5th 2018

The .nl Top Level Domain (TLD)

- > 5.8 million registered domain names (Q1 2018)[1]
- 10th largest TLD (Q4 2017)[2]
- Good reputation for e-commerce[3]
- Maintained by SIDN



Source: www.verisign.com/assets/domain-name-report-Q42017.pdf

Problem statement

Fraudulent webshops:

- Luxury goods, high discounts
- Payment by credit card
- Risk: money scams, identity & credit card theft
- *Spoof* and *Concocted* sites

New York Yankees New Era Fijne Zijde 950 Snapback Hat Navy Mannen Hoeden

~~€35.22~~ €19.60

Zonnebril Heren Ray-Ban Aviator zonnebril Gold Goud stocker en ligne

~~153.98€~~ 26.58€

Korting: 83%

Jeffrey Campbell Dames Blauw Laarzen Ronde Neus Denim

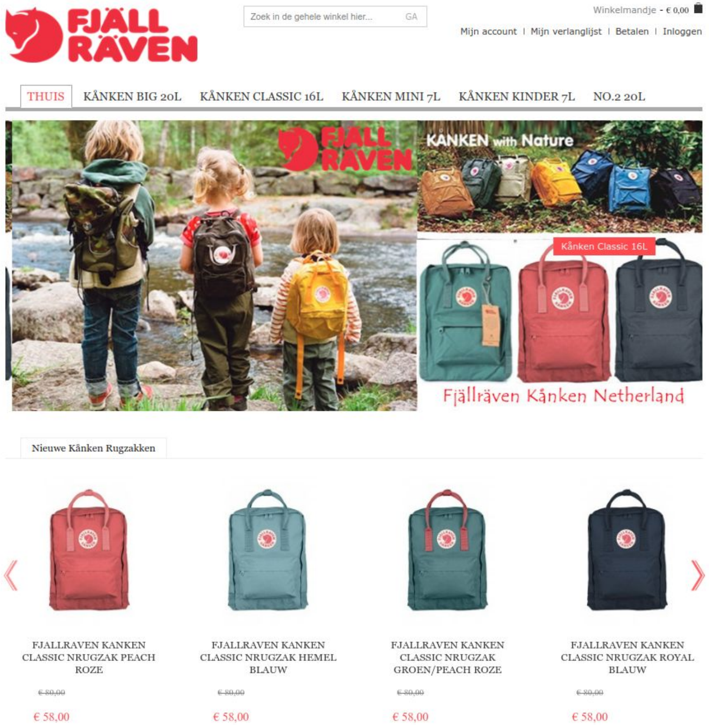
~~€196.11~~ €78.44

FJALLRAVEN KANKEN CLASSIC NRUGZAK GROEN/PEACH ROZE

~~€ 80,00~~

€ 58,00

Examples: Spoof & Concocted



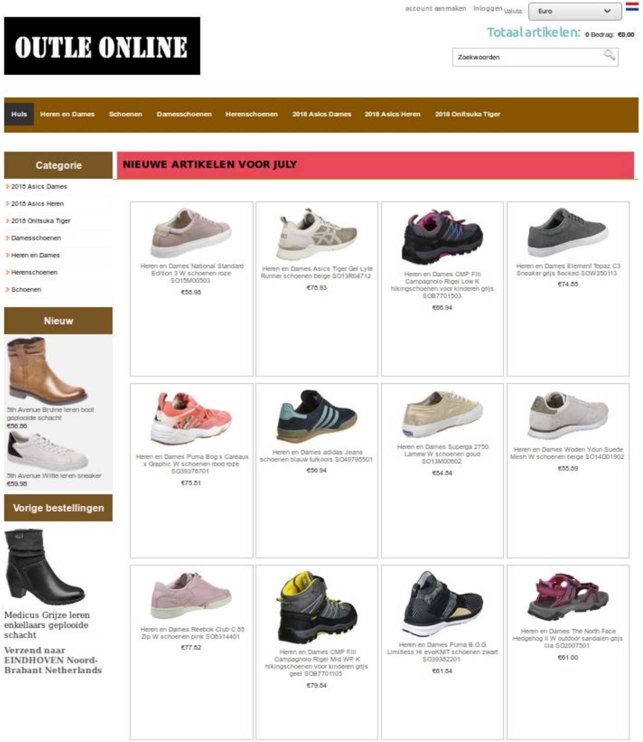
FJALLRAVEN KANKEN CLASSIC NRUGZAK PEACH ROZE
~~€ 80,00~~
€ 58,00

FJALLRAVEN KANKEN CLASSIC NRUGZAK HEMEL BLAUW
~~€ 80,00~~
€ 58,00

FJALLRAVEN KANKEN CLASSIC NRUGZAK GROEN/PEACH ROZE
~~€ 80,00~~
€ 58,00

FJALLRAVEN KANKEN CLASSIC NRUGZAK ROYAL BLAUW
~~€ 80,00~~
€ 58,00

Source: www.fjallraven-kanken.nl



NIEUWE ARTIKELEN VOOR JULY

Heren en Dames National Standard Edition 3 W schoenen roze SO15M00503
€58.98

Heren en Dames Asics Tiger Gel Lyte Runner schoenen beige SO13R04712
€78.93

Heren en Dames CMP Fiji Campagnolo Rigel Low K hikingschoenen voor kinderen grijs SOB7701503
€66.94

Heren en Dames Element Topaz C3 Sneaker grijs flecked SOW350113
€74.85

Heren en Dames Puma Bog x Careaux x Graphic W schoenen rood roze SO39376701
€75.81

Heren en Dames adidas Jeans schoenen blauw turkoois SO49795501
€56.94

Heren en Dames Superga 2750 Lamew W schoenen goud SO13M00602
€64.84

Heren en Dames Woden Ydun Suede Mesh W schoenen beige SO14G01902
€55.59

Heren en Dames Reebok Club C 85 Zip W schoenen pink SO5314401
€77.82

Heren en Dames CMP Fiji Campagnolo Rigel Mid WP K hikingschoenen voor kinderen grijs geel SOB7701105
€79.84

Heren en Dames Puma B.O.G. Limitless Hi evoKNIT schoenen zwart SO39352201
€61.84

Heren en Dames The North Face Hedgehog II W outdoor sandalen grijs lila SO2007501
€61.00

5th Avenue Bruine leren boot geplooide schacht
€56.56

5th Avenue Witte leren sneaker
€59.95

Medicus Grijze leren enkellaars geplooide schacht

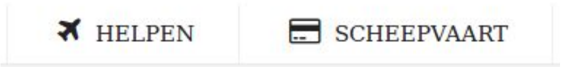
Verzend naar EINDHOVEN Noord-Brabant Netherlands

Source: www.autorijschoolmathieu.nl

Operators

- Many websites, same operator:
  - Same technology
  - Similar translation mistakes
  - Possibly 'fraudulent webshop as a service'

zencart
The art of e-commerce



- Likely foreign actors:
  - Hosting: often geolocated in Russia[4]
  - WHOIS
  - Code comments

Prior work

- 2016: SIDN Labs: nDEWS[4]
- 2017: Sahoo et al.: survey on malicious URL detection[5]
- 2018: Consumentenbond: identified 2000 fraudulent webshops[6]
- 2018: CrimeBusterBot classifier uses different sources[7]
- 2018 (ongoing): Classification on DNS and network data (Thijs Brands, TUDelft)

Motivation

Keeping .nl clean is in the interest of:

- SIDN
- the registrants
- the end user

SIDN dataset: a crawl of the .nl TLD (June 2018) is used to perform a classification.

Research Question

Is it possible to reliably classify fraudulent webshops in the .nl TLD, based on web content?

Datasets

Fraudulent webshops ('*nep*', 3634 observations):

- Consumentenbond
- CrimeBusterBot
- SIDN dataset

General websites ('*web*', 3650 observations):

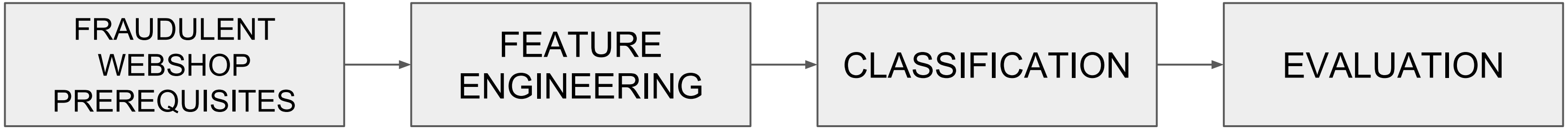
- Random sample SIDN dataset

Both manually sanitized

Approach

- Possible biased dataset. Is it representative?
- Matching technical implementations is circumventable

Our approach: Target the prerequisites to build fraudulent webshops

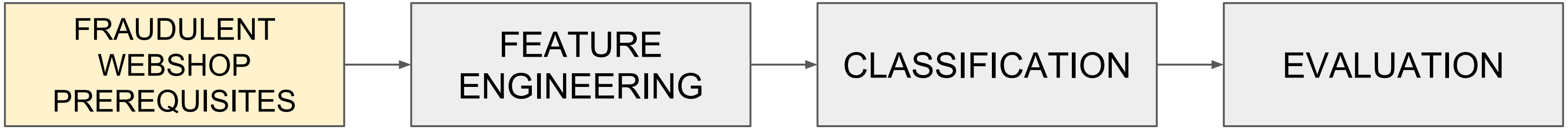


Method

Prerequisites

Fraudulent webshop prerequisites

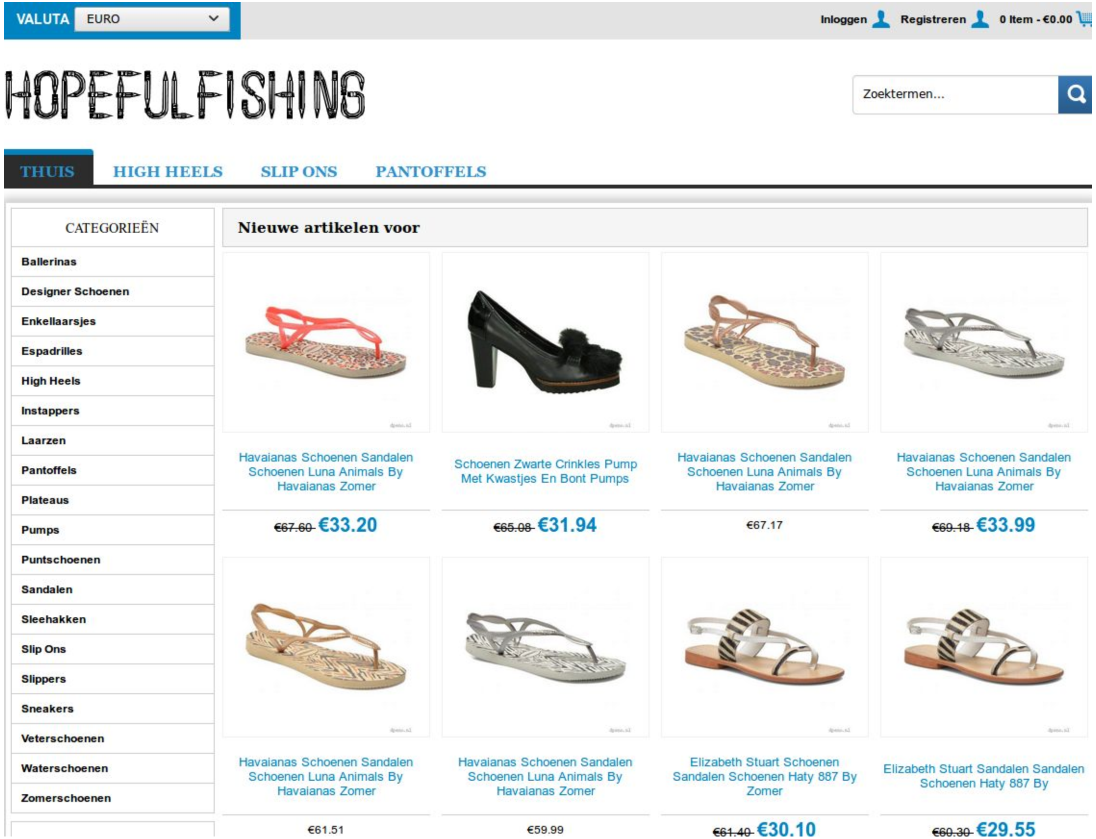
- Customer attraction
- SEO score
- Scalability



Method

Customer attraction

- Popular brands
- Attractive discounts
- High stock, many sizes
- Social media buttons
- Webshop logo



CATEGORIEËN

- Ballerinas
- Designer Schoenen
- Enkellaarsjes
- Espadrilles
- High Heels
- Instappers
- Laarzen
- Pantoffels
- Plateaus
- Pumps
- Puntschoenen
- Sandalen
- Sleehakken
- Slip Ons
- Slippers
- Sneakers
- Veterschoenen
- Waterschoenen
- Zomerschoenen

Nieuwe artikelen voor

Havaianas Schoenen Sandalen Schoenen Luna Animals By Havaianas Zomer
~~€67.60~~ €33.20

Schoenen Zwarte Crinkles Pump Met Kwastjes En Bont Pumps
~~€65.08~~ €31.94

Havaianas Schoenen Sandalen Schoenen Luna Animals By Havaianas Zomer
€67.17

Havaianas Schoenen Sandalen Schoenen Luna Animals By Havaianas Zomer
~~€69.18~~ €33.99

Havaianas Schoenen Sandalen Schoenen Luna Animals By Havaianas Zomer
€61.51

Havaianas Schoenen Sandalen Schoenen Luna Animals By Havaianas Zomer
€59.99

Elizabeth Stuart Schoenen Sandalen Schoenen Haty 887 By Zomer
~~€61.40~~ €30.10

Elizabeth Stuart Sandalen Sandalen Schoenen Haty 887 By
~~€60.30~~ €29.55

Source: www.hopefulfishing.nl

SEO score

- Dependant on visibility in search engines
- Using recently expired Dutch domain names
- Registration likely by drop catchers
- Intel on SEO by third parties (majestic.com)

Example domain names:

- autorijschoolmathieu.nl
- bestratingengroendienstverlening.nl
- stichtingmali.nl
- blaasorkestdacapo.nl
- psycholoog-ermelo.nl

Scalability

Scalability strategy: replication

- Simple, generic webshops
- No time to tweak each webshop
- High risk of takedowns / short lived
- Evade manual work, automate everything
- Operators may control many webshops (also in other TLDs)

Trompetforum.nl

kraamcentrumdebakermat.nl

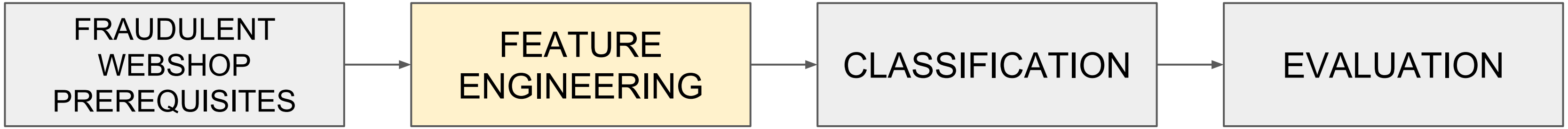
condoomshopthofje.nl

seks-therapeut.nl

trompetforu
kraamcentru
condoomsh
seks-therap

Feature Engineering

Model characteristics in measurable features



FRAUDULENT WEBSHOP PREREQUISITES

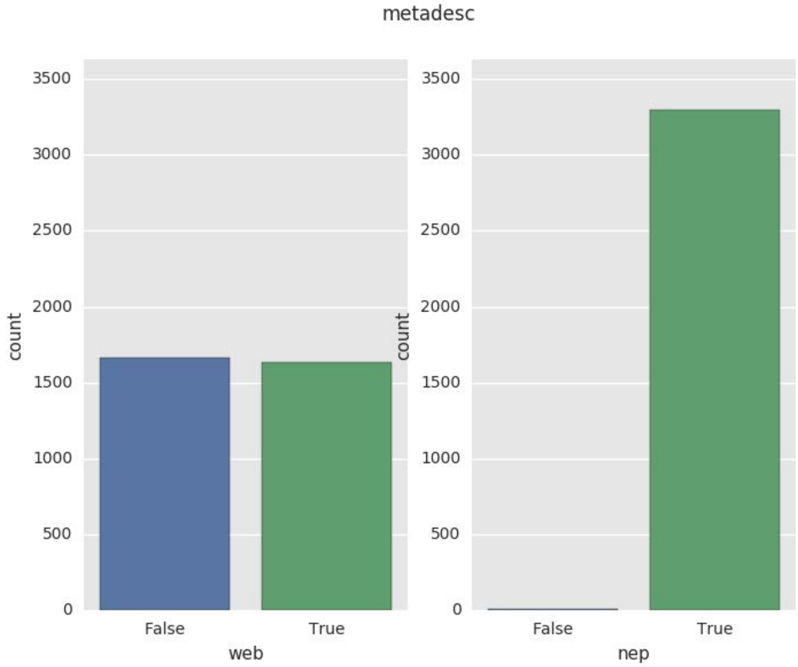
FEATURE ENGINEERING

CLASSIFICATION

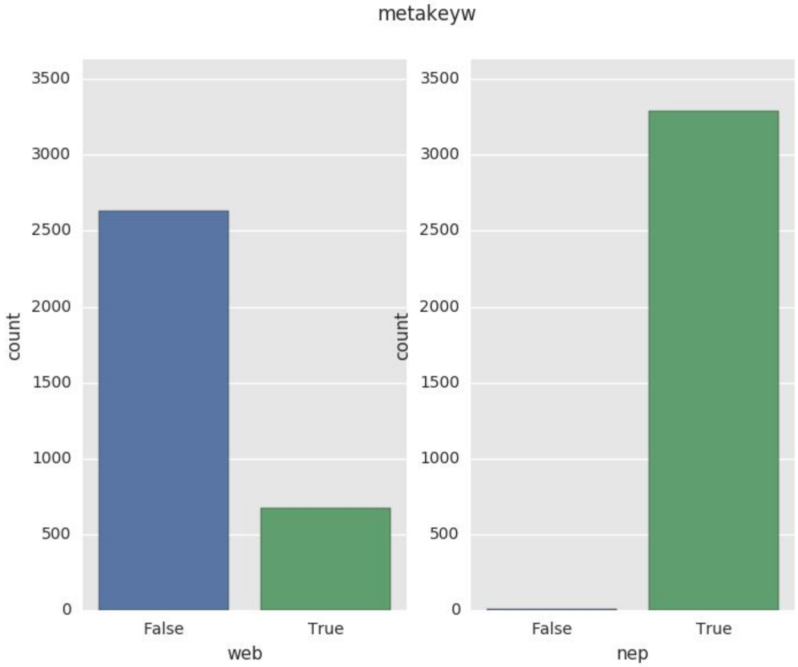
EVALUATION

Method

Meta tags



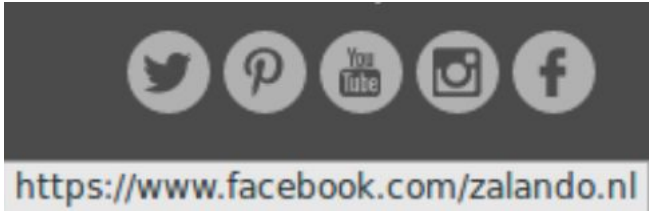
Meta Description



Meta Keywords

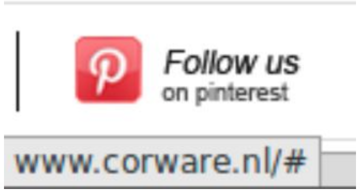
Social media linking

Genuine:

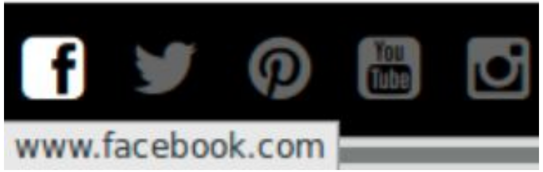


https://www.facebook.com/zalando.nl

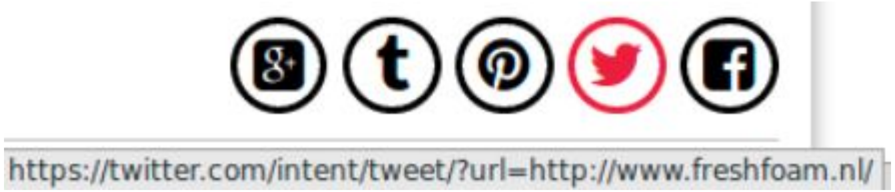
Possibly fraudulent



www.corware.nl/#

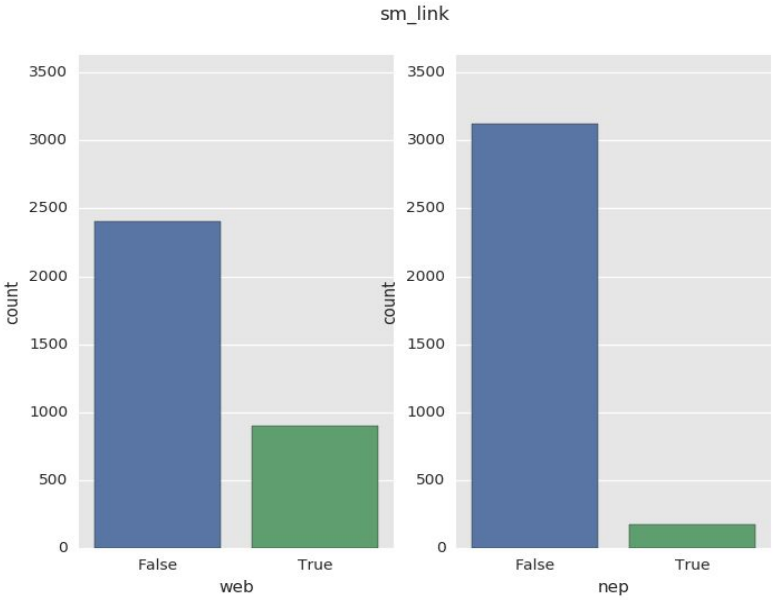


www.facebook.com

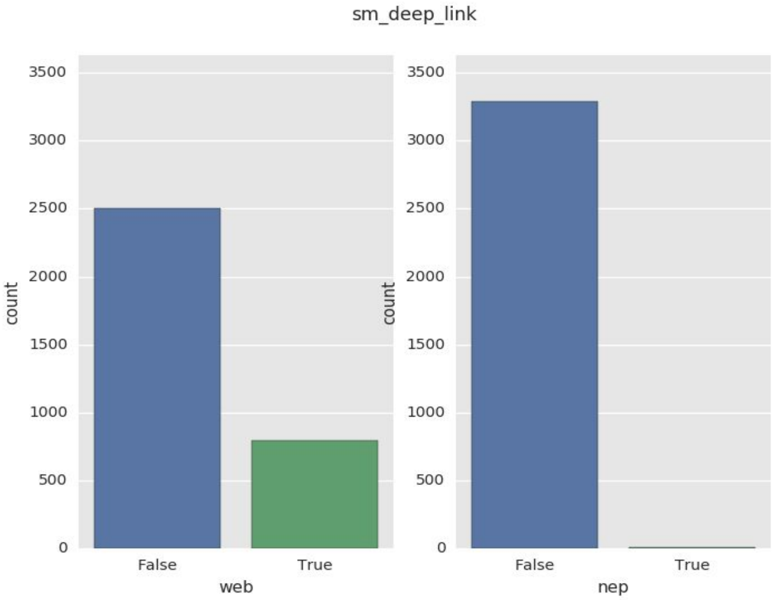


https://twitter.com/intent/tweet/?url=http://www.freshfoam.nl/

Social media linking

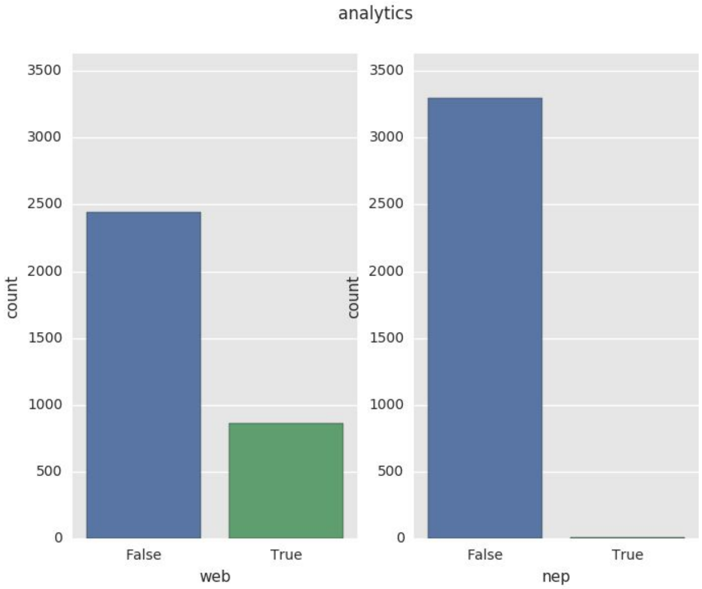


Social Media Links

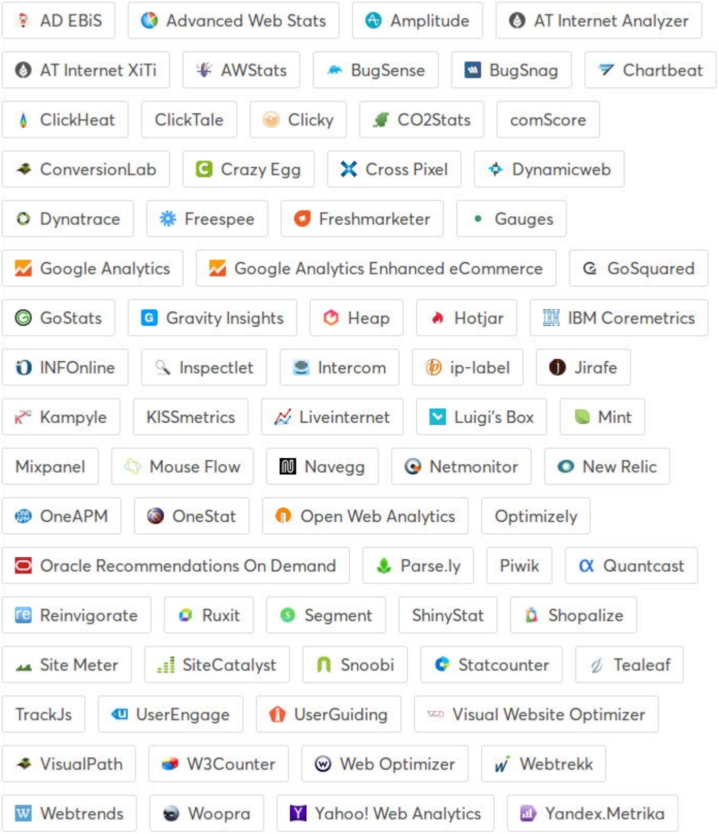


Social Media Deep Links

Web analytics



Analytics Integration



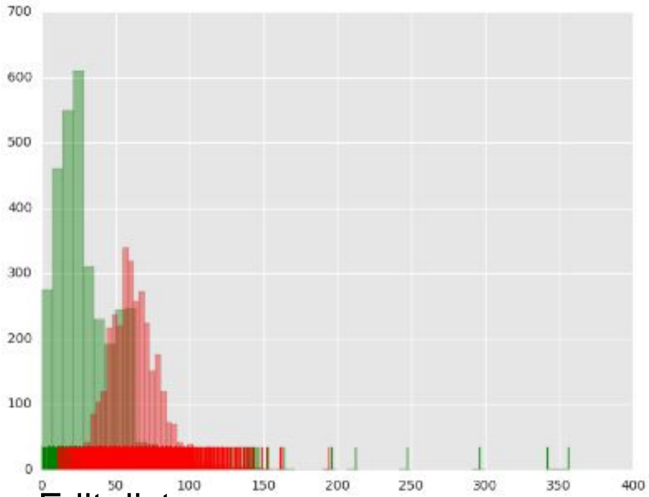
- AD EBiS
- Advanced Web Stats
- Amplitude
- AT Internet Analyzer
- AT Internet XiTi
- AWStats
- BugSense
- BugSnag
- Chartbeat
- ClickHeat
- ClickTale
- Clicky
- CO2Stats
- comScore
- ConversionLab
- Crazy Egg
- Cross Pixel
- Dynamicweb
- Dynatrace
- Freespee
- Freshmarketer
- Gauges
- Google Analytics
- Google Analytics Enhanced eCommerce
- GoSquared
- GoStats
- Gravity Insights
- Heap
- Hotjar
- IBM Coremetrics
- INFOnline
- Inspectlet
- Intercom
- ip-label
- Jirafe
- Kampyle
- KISSmetrics
- Liveinternet
- Luigi's Box
- Mint
- Mixpanel
- Mouse Flow
- Navegg
- Netmonitor
- New Relic
- OneAPM
- OneStat
- Open Web Analytics
- Optimizely
- Oracle Recommendations On Demand
- Parse.ly
- Piwik
- Quantcast
- Reinvigorate
- Ruxit
- Segment
- ShinyStat
- Shopalize
- Site Meter
- SiteCatalyst
- Snoobi
- Statcounter
- Tealeaf
- TrackJs
- UserEngage
- UserGuiding
- Visual Website Optimizer
- VisualPath
- W3Counter
- Web Optimizer
- Webtrekk
- Webtrends
- Woopra
- Yahoo! Web Analytics
- Yandex.Metrika

Source: Wappalyzer.com

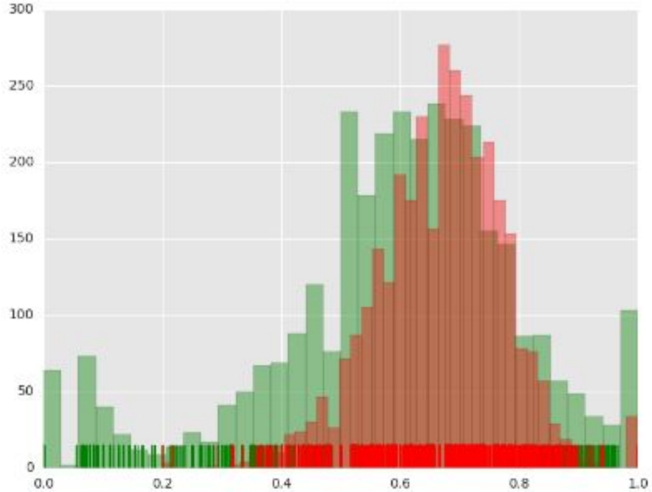
Domain/title string distance

Syntactical difference

```
www.autorijschoolmathieu.nl

Damesschoenen van aQa COGNAC (A3433-Z23A25)
/ Van Mierlo Schoenen

Levenshtein distance: 60

Jaccard distance: 20


www.rabobank.nl

Rabobank - Particulieren

Levenshtein distance: 20

Jaccard distance: 20
```



Edit distance



Jaccard distance

Domain/title similarity

Semantic difference

- Calculate similarity score
- Using word2vec word embeddings
- Model pretrained on SoNaR 500 and Wikipedia (NL) corpus [8]

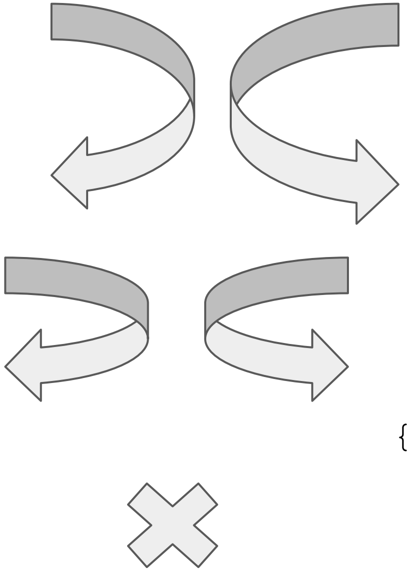
Segmentation Algorithm

1. Split domain in all possible substrings
2. Filter stop words
3. Filter to dictionary
4. Take longest subword
    - Filter all subwords not element of longest subword
    - Recursive step

Domain/title similarity contd.

```
autorijschoo           torijscho
autorijschool          torijschoo
autorijschoolm         torijschool
autorijschoolma        orijschoolm
autorijschoolmat       rijschoolma
autorijschoolmath      ijschoolmat
autorijschoolmathi     schoolmath
autorijschoolmathie    choolmath
autorijschoolmathieu   hoolmath
                       oolmath
```

Damesschoenen van aQa COGNAC (A3433-Z23A25) / Van Mierlo Schoenen



{'Mierlo', 'Van', 'COGNAC', 'Schoenen', 'Damesschoenen', 'van', 'aQa', 'A3433', 'Z23A25'}

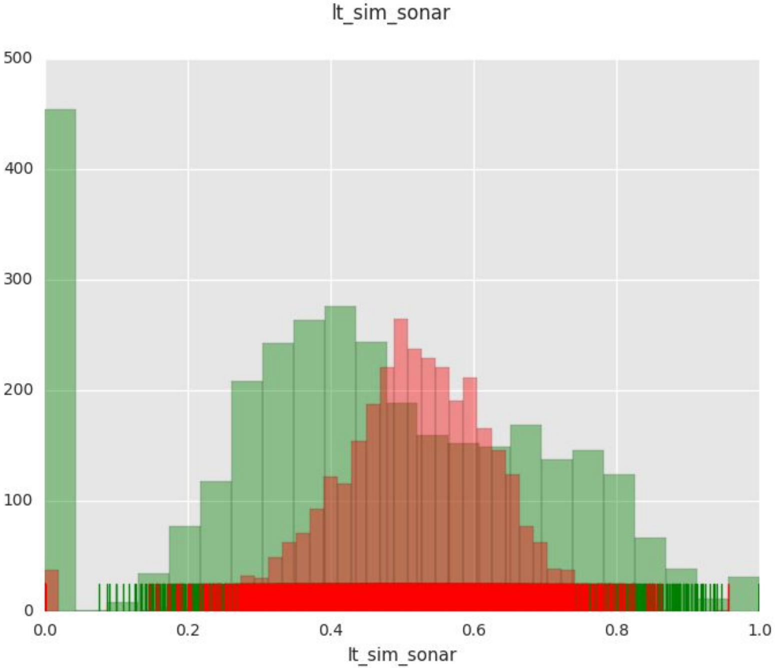
{'rijs', 'auto', 'mathieu', 'u', 'eu', 'mat', 'ijs', 'autorijschool', 'rij', 'ij', 'school, 'rijschool'}

{'Mierlo', 'Damesschoenen', 'Schoenen', 'COGNAC'}

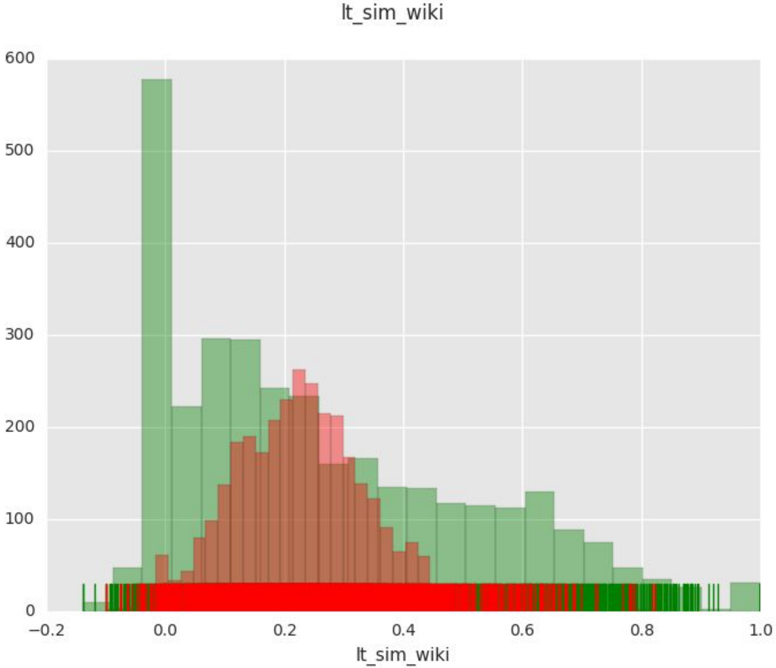
{'mathieu','autorijschool'}

sonar: 0.30163282278907727 wiki: 0.21168016747044305

Domain/title similarity contd.



Similarity on Sonar Corpus



Similarity on Wikipedia Corpus

Feature overview

Table 1: Overview of used features

Fraudulent webshop prerequisites	**Feature**
Customer attraction	Currency symbol count
	Image Count
SEO	Meta Description / Keyword: token count
	Domain label / title edit distance
	Domain label / title similarity
	CSS & Javascript includes: count
Scalability	Meta Open Graph
	Web analytics
	Anchor tags (internal/external)
	Pattern match: Phone / Address / Postcode / Place / IBAN
	Lexical Diversity (Total/Unique)
	Social Media links & deeplinks

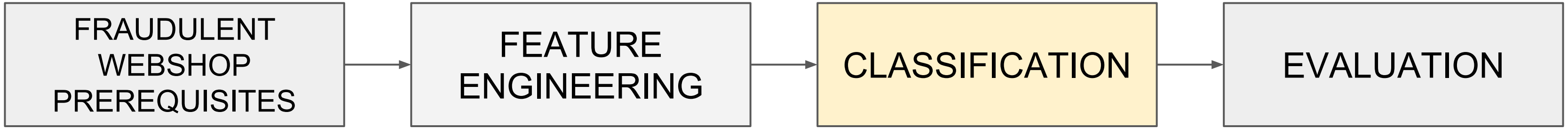
Classification

Experiment 1: Labeled dataset

- 10-fold cross validation
- 3000 train/ 300 test
- AdaBoost Algorithm

Experiment 2: .nl zone

- 4.9 million valid page sources
- Seven different algorithms
- Confidence score



FRAUDULENT WEBSHOP PREREQUISITES → FEATURE ENGINEERING → CLASSIFICATION → EVALUATION

Method

Results (experiment 1)

Table 2: Averages on AdaBoost 10 fold cross validation using 6600 observations: even class, default parameters

Average	**AdaBoost**
Accuracy	0.9934
Recall	0.9909
Precision	0.9941
F_1 Score	0.9915

Table 3: Most informative features

Rank	**Feature**	**Weight**
1	analytics	0.1207
2	currencycnt	0.1048
3	distance_edit	0.0986
4	sm_deep_link	0.0779
5	links_external	0.0615
6	scriptscnt	0.0600
7	links_hash	0.0538
8	lexunq	0.0421
9	lt_sim_wiki	0.0420
10	distance_jaccard	0.0419

Results (experiment 2)

Table 4: Classification SIDN dataset*

	Classified fraudulent (positive)	**Classified normal (negative)**	**Pct positive**
Majority vote (4/7)	73,519	4,839,753	~1.496%
Unanimous vote (7/7)	1522	4,911,750	~0.03%

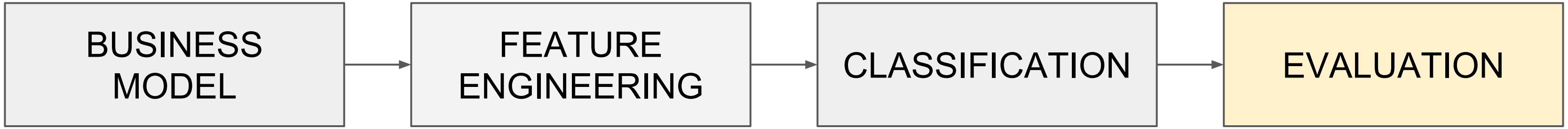
Table 5: Precision

	True positive	**False Positive**	**Precision**
Majority vote (sampled 5000)	4	60	~6.667%
Unanimous vote (7/7)	1303	219	85.61%

***** Domains with included page source in the SIDN dataset

Evaluation

Discussion, Future Work & Conclusion



BUSINESS MODEL

FEATURE ENGINEERING

CLASSIFICATION

EVALUATION

Method

Discussion & Future Work

Yes, content-based classification of fraudulent webshops is possible.

- Results labeled set are high. Unlabeled still shows false positives
- Did we correct our initial dataset?

- Only classified index pages
- Algorithm selection, tuning and data preprocessing
- Combine results with other perspectives
- Other applications of semantic similarity?
- Many features still left undiscovered
    - Payment processing
    - Translated text recognition
    - NLP on Dutch grammar

Conclusion

Our contributions

- Shown that content-based classification can be done
- Introduced semantic similarity to represent website content
- Resulting classification as a basis for future work

References I

1 - SIDN Labs (2018). ".nl stats and data". https://stats.sidnlabs.nl/en/registration.html

2 - Verisign Inc. (2018). "The Domain Name Industry Brief". https://www.verisign.com/assets/domain-name-report-Q42017.pdf

3 - United Nations Conference on Trade and Development (2017). "UNCTAD B2C E-COMMERCE INDEX". http://unctad.org/en/PublicationsLibrary/tn_unctad_ict4d09_en.pdf

4 - Moura, G.C. M., Müller, M., Wullink, M, Hesselman, C. (2016). "nDEWS: a new domains early warning system for TLDs" In: IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2016), co-located with IEEE/IFIP Network Operations and Management Symposium (NOMS 2016). Istanbul, Turkey, May 2016.

References II

5 - Sahoo, Doyen and Liu, Chenghao and Hoi, Steven CH (2017). "Malicious URL detection using machine learning: A survey" arXiv preprint arXiv:1701.07179.

6 - Consumentenbond (2018). "Consumentenbond laat 850 foute webwinkels offline halen" https://www.consumentenbond.nl/nieuws/2018/consumentenbond-laat-850-foute-webwinkels-offline-halen

7 - Richard Garsthagen (2018). "CrimeBusterBot". https://github.com/AnykeyNL/CrimeBusterBot

8 - Stephan Tulkens and Chris Emmery and Walter Daelemans (2016). "Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource". https://github.com/clips/dutchembeddings