1.2 Project statement and contributions

Active learning is a machine learning setting that allows the learning algorithm to submit queries to obtain the label of unlabeled instances. This paper designs an active learning architecture to help with fake web-shop detection.

First, we evaluate how active learning performs using the existing labeled data made available by DNS Belgium. In this experiment, we put the active learner in a situation where it does not have access to the labels of most instances until it issues a query to get them. This experiment is the opportunity to compare the performance of multiple query selection algorithms.

Three different learners are trained, each with a different query selection algorithm, to allow us to compare their performances.

Second, we apply the active learning architecture to the entire .be zone to attempt to uncover previously undetected fake web-shops. For this experiment, we build a dataset using multiple sources for labeled instances, including search engine results.
Both of those experiments are implemented, taking inspiration from the existing solution used by DNS Belgium and using Scikit-Learn. The results of the experiments are presented and interpreted. Next, we put our results into perspective and highlight some considerations to take into account when interpreting them. Finally, we suggest directions for future research related to the detection of fake web-shops using machine learning.

Chapter 2 discusses the previous research in the context of fake web-shop detection and introduces the key concepts of active learning;

Chapter 3 presents the existing features used by DNS Belgium as well as their existing classifier and its performance;

Chapter 4 describes how we propose to improve the existing classifier. It presents the additional features that we suggest, the architecture for the active learner and a description of the two experiments we conduct;

Chapter 5 presents the results of the two experiments conducted;

Chapter 6 discusses the results obtained and the limitations of validity of the experiments and their results;

Chapter 7 suggests directions for future research in the context of fake web-shop detection;

Chapter 8 summarizes the most important results and concludes this thesis.

Domains are cheap. Operators of malicious websites may prefer to register many domains as they are relatively inexpensive. If some are taken down, the remaining ones are enough to remain profitable.

Registrar concentration. More than 87% of the suspicious domains are registered by ten different registrars. The most used registrar is among the cheapest registrars and offers an API for bulk registration, which probably helps fake web-shop operators to automate the process.

Few content-management systems. The home pages of the web-shops are similar yet different. The websites seem to use few content-management systems and share some common design features.

Most domains are drop-catch. Most domains used for fake web-shops were domains that became available and were immediately registered: this practice is called “drop catching”. This allows the domain’s new owner to take advantage of the previously built trust in the domain name.

2.1.2 Analysis of the features used

Registration features

Feature	Type	Reference
Private or China-registered WHOIS [12]	Boolean	[45]
Domain age	Boolean	[9, 10, 45]
Re-registered domain	Boolean	[4, 44]
Re-registration delay (less than x days)	Boolean	[4]
Domain was transferred to another registrar	Boolean	[4]
Registration hour	Numerical	[4, 44]
Registrar	Categorical	[44]
Email provider of registrant	Categorical	[44]
Reported domain score	Numerical	[44]
Ratio of lowercase characters in registrant’s name	Numerical	[44]
Registrant country	Categorical	[10]

URL features

Feature	Type	Reference
Keywords in domain name	Boolean	[10, 45]
Length of domain name	Numerical	[45]
Explicit IP or port in the URL	Boolean	[3]
Presence of spelling mistakes	Boolean	[24]
Suspicious characters in the URL	Boolean	[24]

Product features

Feature	Type	Reference
Number of currencies	Numerical	[4, 10, 11, 45]
Percent savings average	Numerical	[10, 45]
Number of duplicated prices	Numerical	[45]
Unique brand mention count	Numerical	[45]
Number of products	Numerical	[33]
Percentage of discounted products	Numerical	[10]
Products present on the home page	Boolean	[10]
Number of numerical strings (similar to prices)	Numerical	[4]

Merchant features

Feature	Type	Reference
Presence of a (free) **email address**	Boolean	[4, 10, 33, 45]
Number of links to social media	Numerical	[4, 10, 11, 33]
Presence of deep links to social media	Boolean	[4, 11]
Business registration number / VAT found	Boolean	[10, 11, 33]
Presence of a phone number	Boolean	[4, 10, 11, 33]
Presence of an address	Boolean	[10, 11, 33]
Presence of bank account number	Boolean	[11]
Link to physical stores	Boolean	[10]
Jobs offerings	Boolean	[10]
Presence of a link to a mobile app	Boolean	[10]
Trust-mark logo misuses	Boolean	[43]

Payment features

According to Carpineto and Romano [10], fraudulent web-shops tend to use payment methods such as Western Union as transfers cannot be canceled or reversed. Mostard et al. [33] counted the number of payments methods mentioned in the HTML page as they expect fraudulent web-shop to accept less payments methods than legitimate ones.

Page-level features

Feature	Type	Reference
Large iframes	Boolean	[45]
HTTP headers	Categorical	[3]
Count of the HTML tags	Numerical	[3, 4, 11]
Number of internal / **external** links	Numerical	[3, 4, 11, 33]
Number of unique hostnames in the page links	Numerical	[3]
HTML size / Total word count	Numerical	[3, 33]
Presence of a shopping cart system	Boolean	[33]
Presence of copyright	Boolean	[33]
Links to known malicious web pages	Boolean	[24]
Number of images	Numerical	[4, 11]
Presence of meta Open Graph tags	Boolean	[11]
Distance between HTML title and the URL	Numerical	[4, 11]
Lexical diversity	Numerical	[11]
Number of words in the meta description	Numerical	[4, 11]
Number of meta keywords	Numerical	[4, 11]
Bag-of-words of HTML body	Vector	[3]
TF-IDF of HTML body	Vector	[3, 24]

Website-level features

Features	Type	Reference
Website in Alexa top sites	Boolean	[10, 45]
Existence of MX record	Boolean	[4, 44]
Presence / Issuer of TLS certificate	Boolean / Categorical	[3, 4, 44]
Autonomous system / Location of the website	Categorical	[4, 10, 44]
Number of pages found	Numerical	[33]
Number of open ports found	Numerical	[33]
Presence of analytic trackers or tracking cookies	Boolean	[10, 11, 19, 33]
Notice and consent banner for cookie law compliance	Boolean	[10]
Website uses cloaking techniques	Boolean	[10]

Visual features

As already mentioned, Mostard et al. [33] use the discrepancy between visual and contextual information about social media and payment method as a feature.
Bannur et al. [3] used the scale-invariant feature transform (SIFT) [31] on screenshots of web pages. The resulting descriptors can be used to compare different web pages together. Those descriptors can also be compared with a database of known logos to detect presence on the web page.
Finally, Kazemian and Ahmed [24] used the speeded-up robust features (SURF) [6], a feature descriptor like SIFT [31] but faster. They grouped pages into clusters of similar looking pages based on the screenshots. The clusters were then fed into the machine learning model to improve the classification.

2.2 Active learning

In order to train a binary classifier, a set of labeled instances is needed to train and test the model. However, in many cases, obtaining labeled data is expensive or time-consuming, but unlabeled data is abundant or can be obtained for cheap. In our current setting, we are facing a similar situation: DNS Belgium has a large amount of data about the websites in the .be zone but labeling instances for classification requires a human annotator and is time-consuming. Active learning relies on the idea that if a machine learning algorithm can choose data it learns from, greater accuracy is achievable with fewer labeled instances needed. Leveraging active learning in our context may be interesting as the machine the learning algorithm will be presented with many unlabeled instances to learn from.

Settles [37] presents the three main scenarios in which an active learning algorithm may submit queries to an oracle. The three scenarios are (i) membership query synthesis, (ii) stream-based selective sampling and (iii) pool-based active learning.

(i) membership query synthesis is simply asking for label of existing unlabeled data that was collected
(ii) stream-based selective sampling is asking for a label generated from a natural distribution, supposing obtaining an unlabel data is easy
(iii) pool-based sampling divides data into label and unlabelled. Then based on labelled data, select most promising unlabelled data to query

Dataset

Label	Count
Fake web-shop	1843
Legitimate web-shop	1085
Not a web-shop	77

Please note that when we refer to legitimate websites, we also include the legitimate web-shops in the set. In other words, the set of legitimate web-shops is a subset of the legitimate websites

Overview of features

Page 19-21

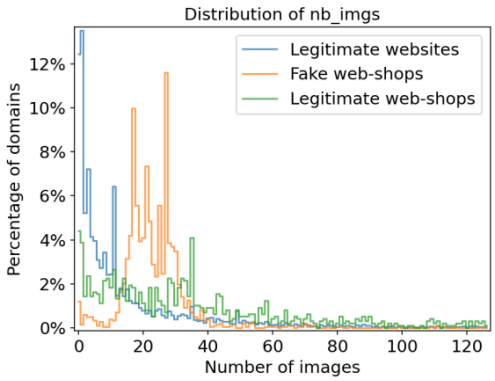


Figure 3.1: The distribution of the number of images on the fake web-shops is more compact than on the legitimate websites. Outliers on the x-axis have been removed for readability.

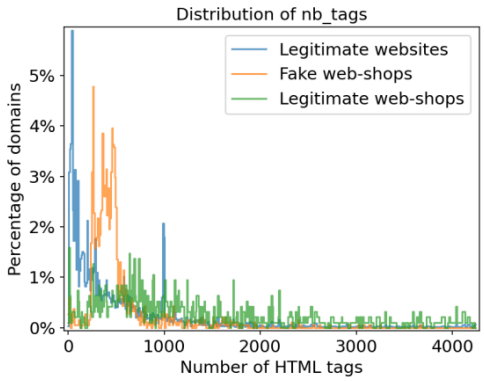
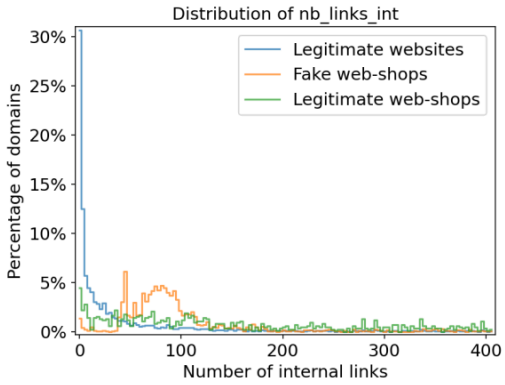
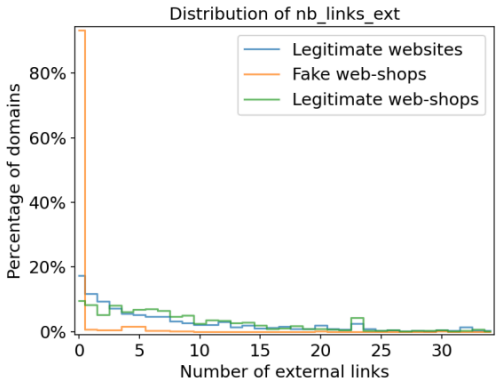


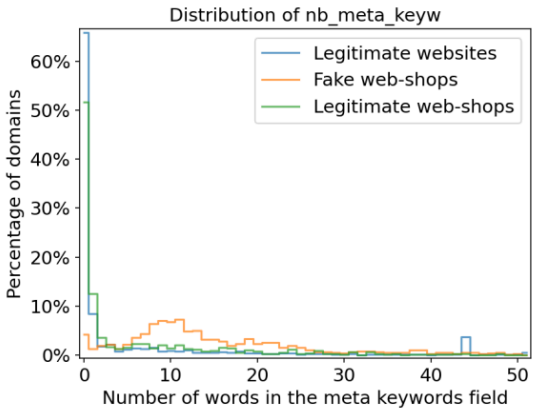
Figure 3.2: The distribution of the number of HTML tags on the fake web-shops is more compact than on the legitimate websites. Outliers on the x-axis have been removed for readability.



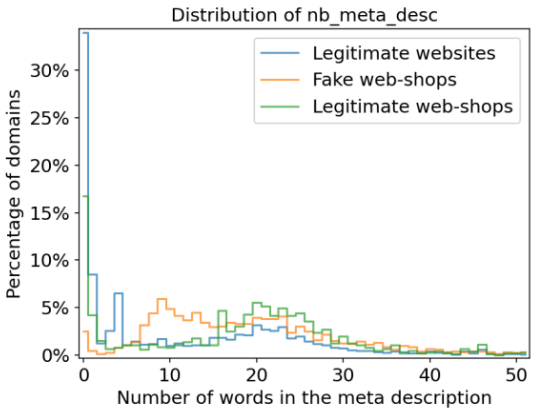
(a) Fake web-shops tend to have more internal links than legitimate websites.



(b) Very few fake web-shops have external links on their front page, which is more common on legitimate websites.



(a) Meta keywords are extensively used by fake web-shops but not much on legitimate websites. Search engines often use meta keywords to index the pages.



(b) Around a third of legitimate websites do not use the meta description field, while nearly all fake web-shops do.

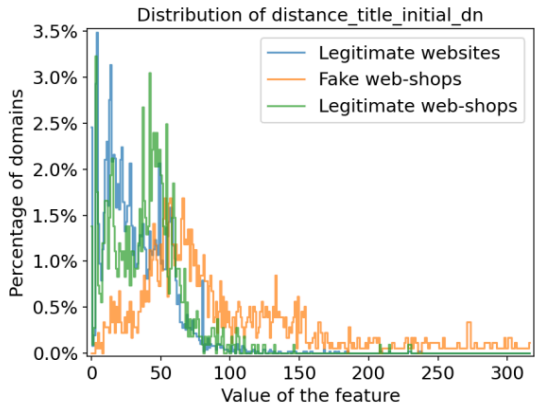


Figure 3.5: The edit distance between the title and the domain name (before following redirections) is higher for fake web-shops than for legitimate websites. This could indicate that the title and the domain name do not carry the same information for fake web-shops. Outliers on the x-axis have been removed for readability.

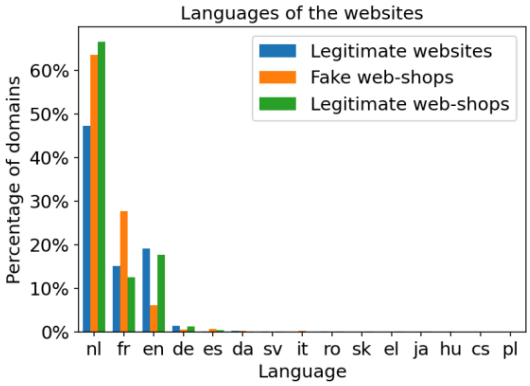
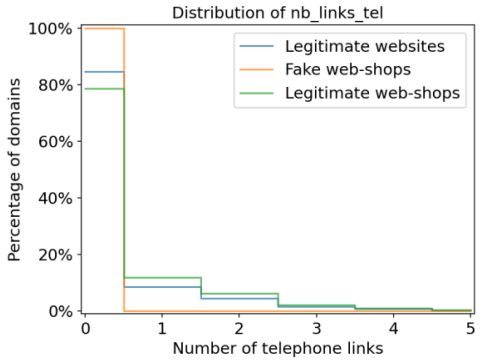
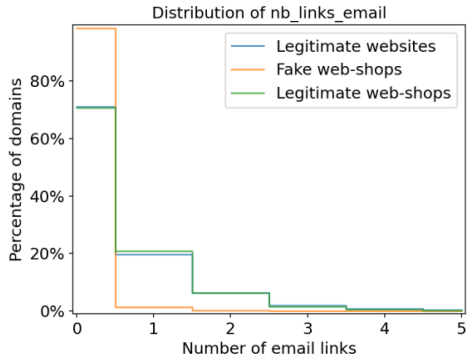


Figure 3.6: Fake web-shops and legitimate websites usually use the same languages. Only the 15 most used languages are depicted for readability.

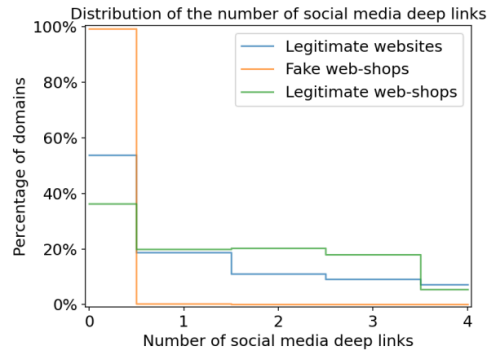


(a) Number of `tel:` links on the home page of the website.

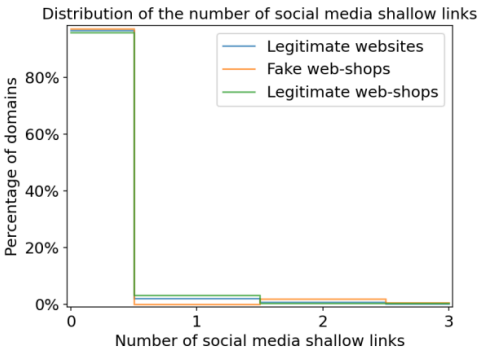


(b) Number of `mailto:` links on the home page of the website.

Figure 3.7: `tel:` and `mailto:` links are not very popular on website home pages. Fake web-shops tend not to use any of them while legitimate websites use them more, especially `mailto:` links. Outliers on the x-axis have been removed for readability.

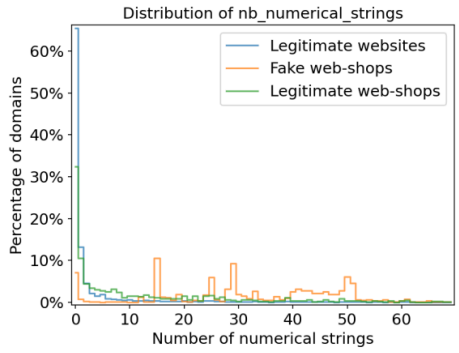


(a) Number of deep social media links on the home page of the websites.

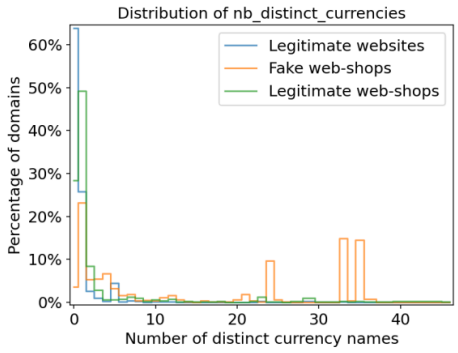


(b) Number of shallow social media links on the home page of the websites.

Figure 3.8: Websites, in general, do not often use shallows links to social media. Fake web-shops hardly ever use deep links to social media, while more than half of the legitimate web-shops do. Outliers on the x-axis have been removed for readability.

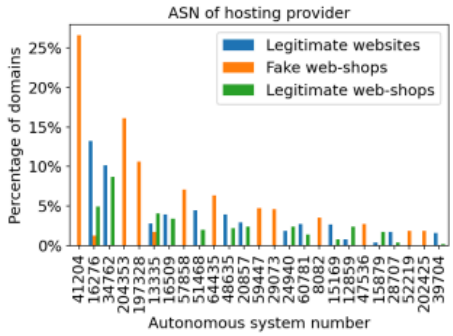


(a) Distribution of the number of numerical strings on the home page of the websites.

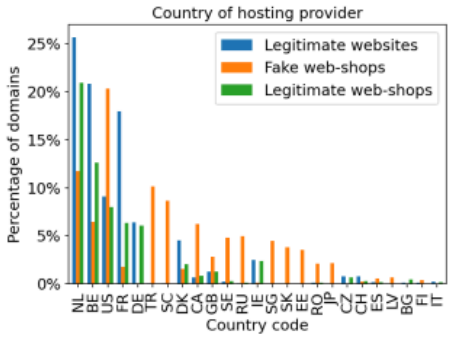


(b) Distribution of the number of distinct currencies on the home page of the website.

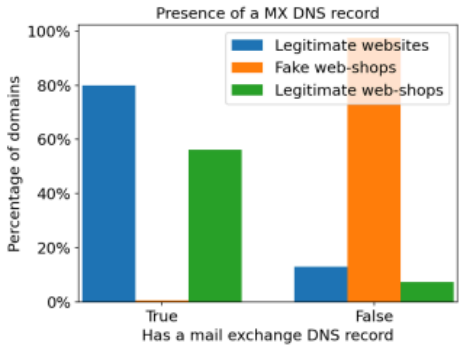
Figure 3.9: Fake web-shops seem to usually display more prices on their home page than legitimate websites. They also display many different currencies on their home page. Outliers on the x-axis have been removed for readability.



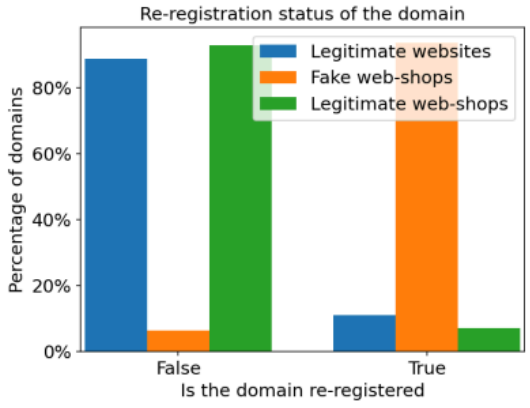
(a) Multiple autonomous systems only host fake web-shops. Fake web-shops are concentrated in a few autonomous systems, with legitimate websites spread across more autonomous systems.



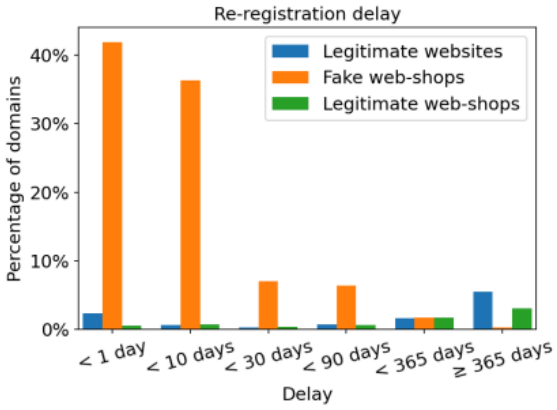
(b) Some countries only host fake web-shops while legitimate domains are concentrated in a few countries.



(c) Fake web-shops usually do not set up MX records. The sum for the categories does not always equal 100% because the data is missing for some entries in the dataset.

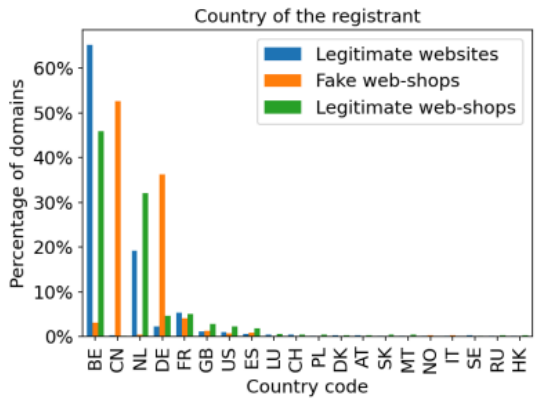


(a) Nearly all the fake web-shops are using re-registered domains.

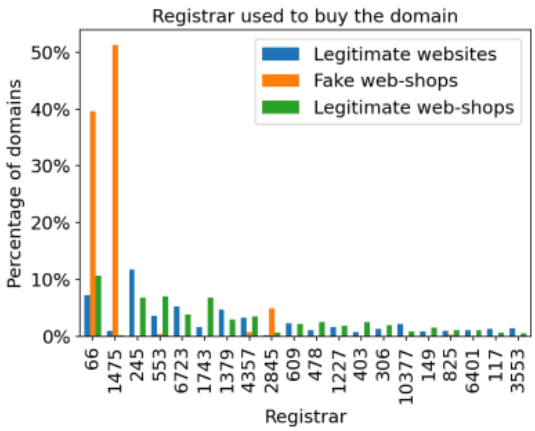


(b) Fake web-shops tend to re-register domains quickly after they become available. This is known as *drop-catching*. Re-registered legitimate web-shops do not exhibit such behavior.

Figure 3.11: The data in our baseline dataset contains cases of *drop-catching*, almost exclusively for fake web-shops.



(a) Registrant report that they live mainly in China and Germany. This information is provided by the registrants when buying a domain name.



(b) Two registrars account for around 90% of the fake web-shops.

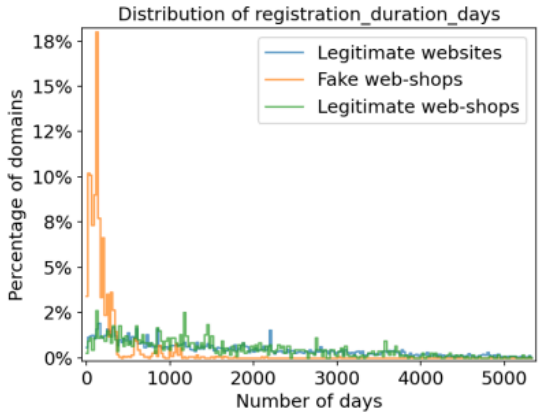


Figure 3.13: Fake web-shops are generally not very old when detected. The age of legitimate websites is more diverse. Outliers on the x-axis have been removed for readability.

3.3.1 General architecture

The current solution used by DNS Belgium to build a classifier uses 80% of the labeled data as a training set and 20% as a test set. The features are then pre-processed using the following steps.

- The missing boolean and numeric features are imputed with the median value.
- The numeric features are standardized by removing the mean and scaling to unit variance.
- The textual fields (`meta_text`, `body_text` and `title`) are independently turned into TF-IDF features. Only words of at least three characters and appearing in at least 0.1% and at most 90% of the documents are used.
- The list of hosts in `external_hosts` is converted to a string (all the elements are grouped in a single string, separated by a space) and turned into TF-IDF features.
- Other features, such as categorical features, are dropped by the current implementation of the classifier.

Since registration data is stored in a different location than the crawler data, registration features are not always used when performing fake web-shop predictions with the classifier. The website-level features are also often left out because they increase the query
time due to the structure of the stored data.
Once the features are pre-processed, a Random Forest Classifier is trained. A grid search with a set of hyper-parameters is used to find the best combination of hyperparameters with a 5-fold cross-validation

3.3.2 Performance evaluation

We trained a Random Forest Classifier as described above, using 51 features (see subsection 3.1.2). Table 3.3 summarizes the metrics of the trained classifier on the baseline
dataset. The most significant features (excluding TF-IDF features), based on the mean decrease in impurity, are depicted in figure 3.14. Those features are consistent with the distributions observed in section 3.2.

The trained Random Forest Classifier achieves a high recall and a perfect precision so, at first glance, this should lead DNS Belgium to find most of the fake web-shops in the .be zone. However, in practice, the classifier mainly reports false positive instances. (the classifier frequently flags website as fake when it is safe)

The poor performance of the classifier when applied to actual data has already been analyzed by Batsleer [4]. He identified three possible causes for the problem, which we think are still relevant.
1. Fake web-shops tend to use other TLDs than .be as the operators know that DNS

Belgium actively takes countermeasures against them.
2. Some parts of the .be zone are not well represented in our baseline dataset, causing the classifier to produce unexpected results when predicting the class for instances lying in those parts of the zone.
3. The strategy of fake web-shop operators evolves over time to evade detection. As most of the labeled fake web-shops were crawled in 2019, they may not represent the newest types of fake web-shops operating in the .be zone.

4.1.1 Open Graph tags

The Open Graph protocol “enables any web page to become a rich object in a social graph” [17]. Websites use this protocol to enable social media to display their content as rich objects instead of plain links once shared on a platform. As mentioned by Cox and Haanen [11], legitimate websites may have an incentive to adopt such protocols to increase their visibility on social media and attract more viewers. On the other hand, fake web-shops operators may also be interested in being represented on social media, but this requires more effort and will probably not be implemented by many.

The Open Graph tags are included in the HTML source code as meta tags in the head of the web page. Each tag is represented as a meta HTML tag with a property name and the value for that property. Listing 1 shows an example of how Open Graph tags can be used to represent a movie on social media. The property attribute always starts with og: (to indicate that this tag is an Open Graph tag) followed by the property name. The property name can be composed of multiple levels: a website can include the tag og:image to specify the image to use when sharing the content to a social media and include the tag og:image:alt to specify an alternative description of the image for accessibility purposes.

```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
<meta property="og:url" content="https://www.imdb.com/title/tt0117500/" />
<meta property="og:image"
↪  content="https://ia.media-imdb.com/images/rock.jpg" />
...
</head>
...
</html>
```

Listing 1: Example of Open Graph tags for the movie The Rock [17]

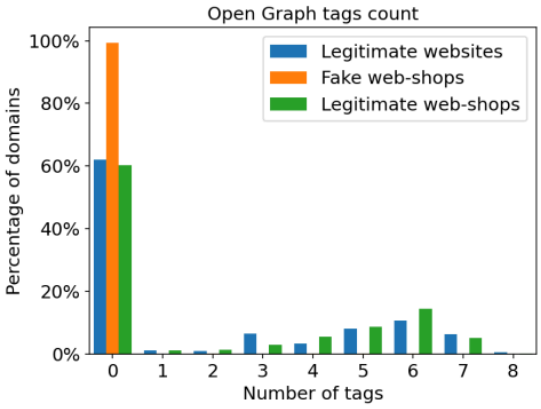


Figure 4.1: Almost none of the fake web-shop in our baseline dataset use Open Graph tags on the front page, while around 40% of legitimate websites do.

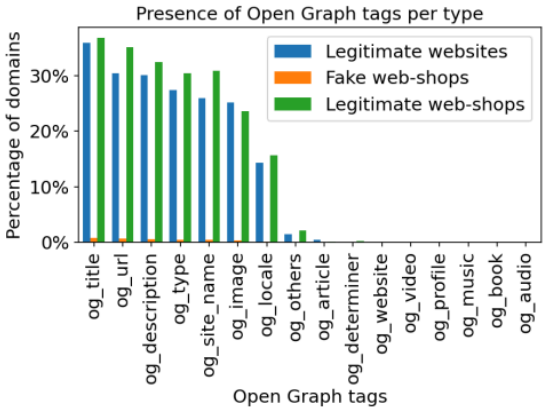


Figure 4.2: Six Open Graph tags are used by more than 20% of the legitimate websites. Only a small set of Open Graph tags are used in practice.

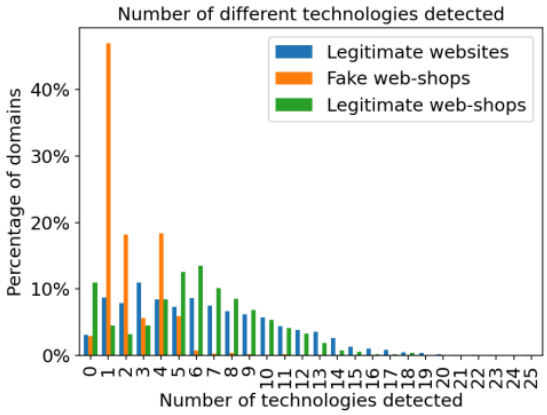


Figure 4.3: Fake web-shops usually use one to five different technologies on their website. The number of technologies on legitimate websites varies more and is generally comprised between 0 and 14.

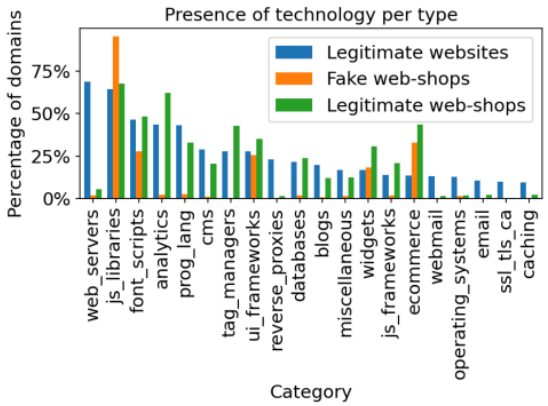


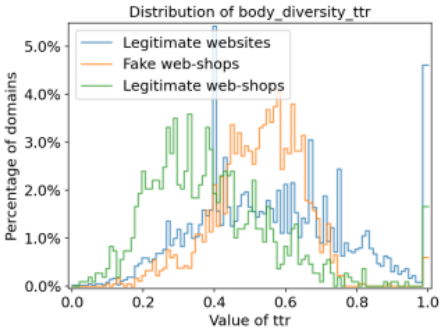
Figure 4.4: Five technologies are mainly used on fake web-shops, while legitimate websites use a more diverse set of technologies. Please note that the way this information was computed for websites crawled in 2019 (i.e., most of the fake web-shops in the dataset and some legitimate websites) may bias the result, see sub-section 4.1.2.

4.1.3 Lexical diversity of the body text

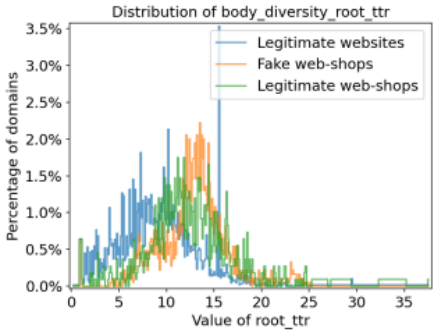
Cox and Haanen [11] suggest using lexical diversity as a feature. Their hypothesis is that fake web-shops will have lower lexical diversity than legitimate websites because they
offer little other content than the products on sale. To evaluate the lexical diversity of a document, multiple metrics exist. We focus on metrics using the Type-Token Ratio (TTR),
which is defined as the ratio between the number of types (i.e., the number of unique words
in the text) and the number of tokens (i.e., the total number of words in the text).

The following measures for lexical diversity were compared:

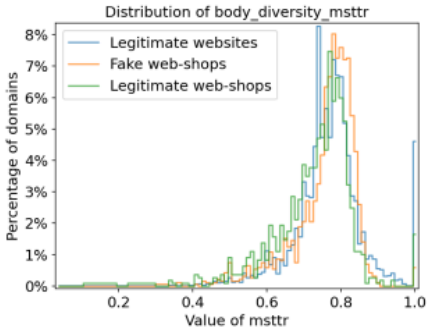
- `ttr` defined as $\frac{nb\_types}{nb\_tokens}$.
- `root_ttr` defined as $\frac{nb\_types}{\sqrt{nb\_tokens}}$.
- `msttr` defined as the average of the `ttr` computed for all the non-overlapping segments of 50 words in the text.
- `mattr` defined as the average of the `ttr` computed for all the (possibly overlapping) windows of 50 words in the text.



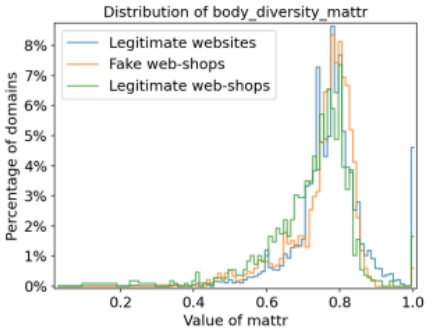
(a) Distribution of the `ttr` computed on the body text of the HTML pages.



(b) Distribution of the `root_ttr` computed on the body text of the HTML pages.



(c) Distribution of the `msttr` computed on the body text of the HTML pages.



(d) Distribution of the `mattr` computed on the body text of the HTML pages.

Figure 4.5: Distributions of the different metrics used to reflect the lexical diversity of the body text.

4.1.4 Semantic similarity between the domain name and the title

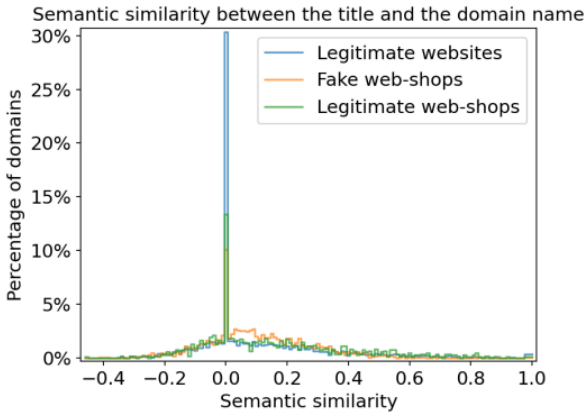
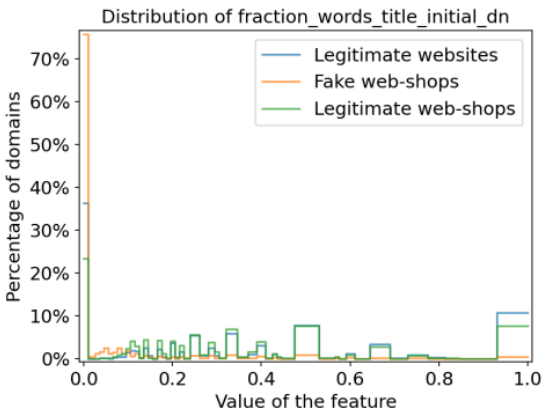


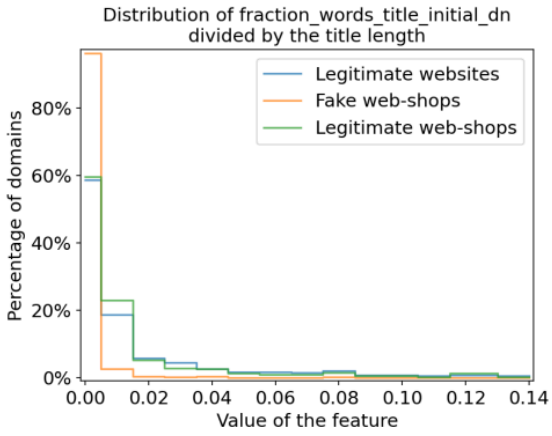
Figure 4.6: Legitimate websites generally have a somewhat higher semantic similarity between the title and the domain name. When the similarity cannot be computed, a score of 0 is attributed for the instance, which explains the peak.

Feature	Information gain	
	Absolute feature	Relative feature
`fraction_words_title_initial_dn`	0.225	0.311
`distance_title_initial_dn`	0.335	0.400
`longest_subsequence_title_initial_dn`	0.079	0.424

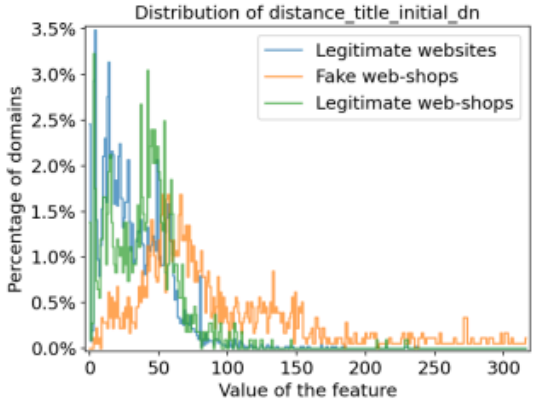
Table 4.2: Those features bring more valuable information when divided by the HTML title length. The values are computed when separating the legitimate websites from the fake web-shops.



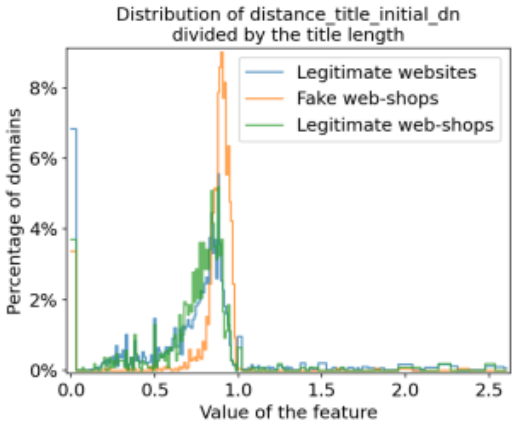
(a) Distribution of the `fraction_words_title_initial_dn` computed on the baseline dataset.



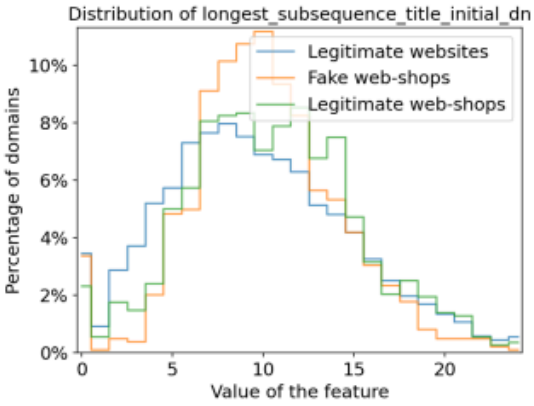
(b) Distribution of the `fraction_words_title_initial_dn` divided by the length of the title of the HTML computed on the baseline dataset.



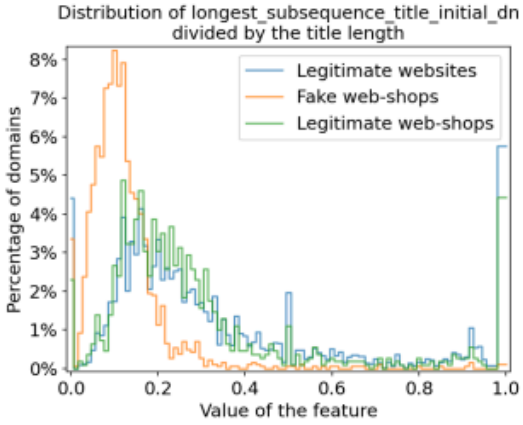
(c) Distribution of the `distance_title_initial_dn` computed on the baseline dataset.



(d) Distribution of the `distance_title_initial_dn` divided by the length of the title of the HTML computed on the baseline dataset.



(e) Distribution of the `longest_subsequence_title_initial_dn` computed on the baseline dataset.



(f) Distribution of the `longest_subsequence_title_initial_dn` divided by the length of the title of the HTML computed on the baseline dataset.