KU LEUVEN

The detection of fake webshops in the .be zone

Senne Batsleer

Thesis submitted for the degree of
Master of Science in Artificial Intelligence ,
option Big Data Analytics

Thesis supervisors:
Prof. dr. Jesse Davis
Maarten Bosteels (DNS Belgium)

Assessor:
Dr. Ir. Lieven Desmet

Mentor:
Ir. Pieter Robberechts

Academic year 2020 – 2021

Preface

I would like to thank everybody who helped me complete this thesis. First of all, I am grateful to my promotor, Prof. dr. Jesse Davis, and DNS Belgium for providing such an interesting thesis topic. It intrigued me from the start and I never regretted choosing it. Second, I would like to thank my assessor, dr. Ir. Lieven Desmet, for taking his time to read this thesis and evaluating my contributions to the subject. My sincerest gratitude goes to my mentors, Maarten Bosteels and Ir. Pieter Robberechts, who guided me throughout the year in our weekly meetings and always provided me with useful feedback. Additionally, I thank Maarten for navigating me through DNS Belgium's databases and helping me out whenever I had questions. I would also like to express thanks to Quentin from DNS Belgium, who resolved issues with DNS Belgium's IT infrastructure on many occasions.

I owe many thanks to my parents for allowing me to pursue my Advanced Master's Degree and providing me with everything I need. Furthermore, I am very grateful to my family, friends, and girlfriend Jill for their love and support. Finally, I thank my friends Jan and Jan for the pleasant working environment and my friend Jef for the relaxing walks and jog sessions.

Senne Batsleer

Contents

Abstract

Belgium's e-commerce market is growing, with consumers spending more money online almost every year. Unfortunately though, some webshops are fraudulent and sell counterfeit goods or do not deliver goods at all. DNS Belgium aims to detect and suspend this kind of fake webshops before they can harm unsuspecting internet users. Since there are more than 1,000,000 active domains in the .be zone, manual verification of webshops is infeasible. Instead, DNS Belgium collected 3900 examples of both fake webshops and benign domains, and trained a supervised classifier on features extracted from their registration records and HTML content.

This thesis first presents an extensive analysis of the features and performance of DNS Belgium's existing classifier. Inspired by previous research, we propose and implement several new features. We show that the feature distributions suggest the presence of clusters of similar fake webshops in the labeled dataset. Furthermore, we demonstrate that this may negatively impact the classifier's performance when fake webshop operators try new tactics and new clusters emerge.

Another drawback of DNS Belgium's approach is the need for manual labeling, which is a time-intensive and error-prone process. Previous research efforts in the context of fake webshop detection suffered from the same inconvenience. In this thesis, we attempt to alleviate this issue by learning from Positive and Unlabeled (PU) data, which is a machine learning technique that does not require negative labels. The ability to include domains with unknown labels enables us to significantly increase the number of domains we can incorporate in the learning process. More specifically, we include all domains in the .be zone that use some kind of e-commerce technology.

Learning from PU data requires assumptions on which fake webshops are selected to be labeled and which are not. We try out two methods that assume every webshop is equally likely to be labeled and one method that assumes the probability to be labeled depends on the webshop's features. For each of the three methods, we manually verified the 500 highest-scoring domains. This resulted in 98 domains that were marked suspicious, with 58 of them originating from the best performing method. We show that our approach yields similar precision as DNS Belgium's classifier and generalizes well, but suffers from missing fake webshops in the unlabeled data.

Chapter 1

Introduction

1.1 Problem statement

Online shopping is becoming increasingly popular in Belgium. A survey conducted in 2019 by Comeos VZW [13] found that 7 out of 10 Belgians purchased online goods in the preceding year, an increase of 17% over five years. Almost half of these online buyers spend more than €150 per month, compared to one out of three buyers five years earlier. Furthermore, the study shows that the lack of trust in e-commerce dropped over the years, as well as the reluctance to share personal information. Prominent drivers for e-commerce are convenience (i.e., time-saving, home delivery, 24/7 access to shops ...), reduced costs (i.e., lower prices and special discounts) and ease of purchase management (i.e., more choice, products/service comparison, more product information ...) [13]. Comeos's recent follow-up study [14] concluded that the current Covid-19 crisis boosts e-commerce even more.

Unfortunately, as the popularity of e-commerce is rising, so is the presence of fake webshops. In 2017, Dataprovider.com estimated that up to 10% of the 85,000 Dutch webshops were fraudulent [6]. A year later, Test Aankoop claimed that out of 700,000 fraudulent webshops worldwide, an estimated 11,000 would be operating in the .be domain [52]. These fake webshops aim to scam consumers in one of two ways: either they deliver counterfeit goods, or they do not deliver the sold goods at all. Furthermore, customers are at risk of identity theft and credit card fraud. In 2016, 3% of online buyers in the Netherlands reported to have experienced financial damage due to these practices [11]. Finally, counterfeit products may pose health and safety risks.

Fake webshops do not only pose problems to customers but also brand owners suffer considerable damage [36]. First of all, they lose revenue from missed sales. Second, an abundance of cheap counterfeit products may degrade the perceived value of a certain brand and exert downward pressure on the price of legitimate products. Finally, to compete with the advertisement campaigns of counterfeiters, marketing costs are likely to increase.

To the untrained eye, fake webshops do not differ significantly from legitimate ones. Europol provides a list of possible red flags, such as unreasonably high discounts, lack of contact information, and spelling mistakes [22]. However, as it is difficult to prove with certainty that a webshop is fake, taking them down can take some time. Nevertheless, a joint effort by Europol, the US National Intellectual Property Rights Coordination Centre and 27 EU member states led to the seizure of over 20,000 domain names in 2017 [23]. Two years later, DNS Belgium took down 2090 fake webshops in the .be domain [17]. Although these are promising results, the battle against counterfeiters and fake webshops is not over yet.

1.2 Research context

The research in this master's thesis is conducted on behalf of DNS Belgium, which is the Top Level Domain (TLD) registry of the .be, .vlaanderen, and .brussels TLDs. This means DNS Belgium operates and maintains the databases of all domain names registered with one of these TLDs. Part of their mission is to provide safe internet to all its users and to ensure consumer trust in the .be domain [18]. In this respect, they aim to detect and block fake webshops as soon as possible, ideally before any internet user is scammed.

Since it is impossible to manually inspect each newly registered domain, a machine learning-based detection tool to track down fraudulent webshops is desirable. To this end, DNS Belgium can extract features from two main data sources. First, they have access to historical registration data of all .be, .vlaanderen, and .brussels second-level domains (such as example.be). This involves information regarding both the registrant (i.e., the user) and the registrar (i.e., the company registering the domain name at the registry on behalf of the user). Second, they developed a crawler to collect HTML content and take screenshots of websites. In 2019, they trained a Random Forest classifier [8] on features extracted from these data sources, which led to the discovery of around 3700 fake webshops in the .be domain so far. The current training database contains 3900 examples, with nearly equal amounts of positives and negatives. The positives were collected after customer complaints to the FPS Economy and by the early versions of their classifier. The negatives contain both legitimate webshops, partly provided by employees of DNS Belgium, and random non-malicious websites in the .be domain.

Quite some research efforts have been put into the topic of detecting and thwarting fake webshops. Wang et al. [56] identified SEO campaigns adopted by known fake webshops and subjected these campaigns to legal and technical interventions. Tian et al. [53] tried to oppose fake webshops by intervening at the payment processing level. However, both research teams experienced that the fraudsters quickly came up with new techniques to evade detection. Wadleigh et al. [55] and Carpineto and

Romano [10] examined how different search engines and query types influence the presence of fake webshops in search results. Haanen and Cox [15] and Wabeke et al. [54] classified domains in the .nl domain on behalf of SIDN labs, the Dutch equivalent of DNS Belgium. While the goals and data sources of these two research teams were similar to ours, our research differs in the learning approach we take. As explained in the next section, we aim to address several issues in fake webshop detection that are ignored by traditional learning methods.

1.3 Goals and methodology

In this thesis, we take a two-step approach to detecting fake webshops. In the first stage, we separate webshops from other types of sites; in the second stage, legitimate webshops are distinguished from fake webshops. We first review existing literature to complement DNS Belgium's existing classifier with additional features that may be suited to discriminate fake webshops from legitimate ones. Next, we modify the existing classifier to better cope with several issues ignored in previous research.

First, gathering examples is time-consuming and not error-proof. Positive examples result from reports of scammed customers and detection by previously built classifiers, but manual verification is needed to ensure the positive label is correct. Gathering negative examples is even more time-consuming, as researchers have to find these themselves. Furthermore, websites can get hacked and turn malicious, invalidating the previously assigned negative label. We address this issue by switching from supervised machine learning techniques to Learning from Positive and Unlabeled Data, or PU learning in short [4]. In this setting, we only have access to examples with positive labels (i.e., fake webshops) and unlabeled examples, which can have positive and negative labels.

Second, our training set may be biased, as fake shops are probably not reported at random. For example, a scammed consumer may be more likely to report the webshop if he/she suffered a substantial financial loss. We examine this by comparing PU learning based on the Selected Completely At Random (SCAR) assumption with PU learning based on the Selected At Random (SAR) assumption [5].

Finally, fraudsters likely adapt their strategy over time to avoid detection mechanisms. Therefore, the classifier should generalize well if we want it to pick up on new tactics. We assess the generalization capabilities of the classifier by clustering the fake webshops and using different clusters for training and testing.

The structure of this thesis is as follows:

- Chapter 2 discusses previous research efforts in the context of fake webshop detection and provides an introduction to PU Learning;

- Chapter 3 analyzes the features and performance of DNS Belgium's existing classifier;
- Chapter 4 describes how we attempted to improve the existing classifier. It covers the additional features we implemented, the approach to gather unlabeled data and the clustering of the fake webshops;
- Chapter 5 provides an empirical evaluation of the implemented PU learning methods;
- Chapter 6 summarizes the most important results of this thesis and suggests future work.

Chapter 2

Related work

This chapter starts with a description of previous efforts to detect fake webshops. We discuss the features and techniques used for classification, as well as the obtained results. The second section of this chapter formalizes the concept of PU learning and introduces the assumptions that are typically made to facilitate learning.

2.1 Detection of fake webshops

This section first gives a broad overview of past research aimed at detecting and thwarting fake webshops. We discuss methods, observations, and important conclusions. In the second subsection, we discuss the features used for classification more in-depth.

2.1.1 Overview

Research on the detection of fake webshops started only several years ago. One of the first contributions was made by Wang et al. [56], who investigated black hat Search Engine Optimization (SEO) campaigns adopted by webshops selling counterfeit luxury goods. These campaigns aim at achieving a higher ranking for certain types of search queries, with the ultimate goal of attracting more users to the fake webshops. Based on textual features of HTML content, the authors were able to distinguish 52 SEO campaigns among known fake webshops. These campaigns were subsequently subjected to technical interventions by search engines and legal interventions by brand holders. While these interventions certainly disrupt counterfeit websites, the counterfeiters quickly adapted to them. The authors concluded that both technical and legal interventions should be employed more responsively and broadly to be truly effective.

Their work was complemented by Wadleigh et al. [55], who developed a binary classifier to predict for any website in the search results whether it sells counterfeit goods or not. It was trained on features based on WHOIS information, price information, and website content. Three types of classifiers were examined: Support

Vector Machines (SVM) [51], Generalized Linear Models (GLM) [44] and Adaptive Boosting (AdaBoost) [46]. The first two performed significantly better than the latter. By analyzing the search results and the detected counterfeit shops, the authors reached several interesting conclusions. First, they discovered that the type of search query influences the risk of coming across a counterfeit webshop in the search results. Second, they found that the occurrence rate of counterfeit webshops was lower for brands taking active countermeasures. Third, counterfeit websites were more likely to be registered and hosted in certain countries than in others. Finally, they found that legitimate webshops tended to be older than counterfeit webshops, which were often registered less than a year ago.

Carpineto and Romano [10] extended the work of Wadleigh et al. and also started from queries in search engines. They showed that not only the type of search query but also the choice of web search engine influences the risk of stumbling upon counterfeit websites. Search results were processed in a two-stage fashion: first, e-commerce websites were discriminated from non-e-commerce websites; second, a distinction was made between legitimate and fake webshops. Separate SVM-classifiers were trained for both of these tasks and both achieved accuracies around 90%.

An entirely different approach to combat the issue of counterfeit webshops was taken by Tian et al [53]. Previous work [37] had shown that credit card payments made up 95% of the revenue for illegal online pharmaceuticals. Payment card systems like Visa interconnect a range of banks and impose fines to banks that process transactions for merchants selling counterfeit goods. These merchants usually don't establish accounts at the bank directly, but through third-party payment processors. It is therefore important for banks to screen these payment processors thoroughly. In an attempt to disrupt the counterfeit industry, the researchers partnered with Visa and made a series of test purchases over a two-year period. Whenever they received counterfeit products, they reported them to Visa. The researchers observed that most banks soon improved their vetting of third-party payment processors. However, they also noticed the rise of "bullet-proof" payment processors, which employed techniques to detect and filter out test purchases. An arms race emerged, in which the researchers trained their volunteers to evade filtering, and the payment processors came up with better filtering methods. Their research shows that intervening at the payment level can be effective in the battle against counterfeit shops, although it is a time-consuming process to stay ahead of the filtering techniques.

Haanen and Cox [15] trained an AdaBoost classifier [46] on content-based features to identify fraudulent webshops. They identified three prerequisites to successfully run such a fraudulent business: customer attraction, high search engine scores, and the ability to scale to a large number of webshops. These prerequisites served as the basis for the selection of their features, which will be discussed later.

From our perspective, the most relevant research was conducted by Wabeke et al. [54] for SIDN, which is the Dutch equivalent of DNS Belgium. They focused on

detecting fraudulent webshops selling counterfeit luxury goods and developed two detection systems. The first one was based on their observation that many known suspicious webshops shared one specific feature: long HTML `<title>` elements listing a range of luxury brands and discount-related words. As explained by Wang et al. [56], this improves their ranking on search engines. However, it also provided the researchers with a simple heuristic. They constructed a list of suspicious words and flagged any webpage with more than 5 matching words in the HTML page title as suspicious. Using this detection system, named BrandCounter, more than 18,000 suspicious webshops were detected. During the analysis of these domains, the researchers made the following observations:

- As domains are cheap, counterfeiters may choose to register a large number of domains. Even if some are taken down, their practices will still be profitable. Furthermore, a lot of domains are not renewed after one year, indicating that domains are disposable. This is consistent with the observations made by Wadleigh et al. [55].
- Most domains were registered by a small concentration of registrars. The most popular registrar was very cheap and provided an API for bulk registration. Furthermore, the majority of webshops were hosted in a small number of Autonomous Systems.
- The homepages of the webshops were different, yet similar, i.e., it seemed as if only a few content management systems (CMS) were used.
- Most domains were *drop-catch* [25], meaning they were almost immediately re-registered after they expired and became available. This way, counterfeiters try to exploit the reputation built up by the previous owners of that domain name [32].
- Registration timing patterns and e-mail providers of registrants could be related to (east) China.

The researchers then reported the suspicious domains to the registrars where they were registered. For the registrar with the highest concentration of suspicious webshops, SIDN performed a follow-up on 4107 domains. At least 3700 of these were effectively taken down by the registrar.

After the first round of takedowns, the number of detected fake shops decreased quickly. In response, the researchers developed a new detection system based on a Support Vector Machine [51]. The selected features were based on the observations above and will be discussed later on in this chapter. This approach led to the discovery of another 894 fake webshops, each of which could have never been detected by their first classifier. Furthermore, registrars, hosting providers, and email providers were already more diversified. This clearly shows that counterfeiters had already adapted to the deployed detection methods.

Mostard et al. [40] were, to the best of our knowledge, the first researchers to include visual features in their classifier. They hypothesized that fraudulent online stores add logos of social media platforms and frequently used payment methods, in an attempt to better resemble legitimate webshops. However, a legitimate website is likely to add only social media on which they have an account and payment methods they support. Therefore the HTML content will likely contain links to the corresponding websites. On the other hand, fake webshops will likely not own social media accounts and will pretend to support more payment methods than they do. As such, the HTML content will not contain references to those websites. The researchers used a Convolutional Neural Network (CNN) to detect the presence of these logos and used the discrepancy between visual and contextual information as additional features. This enabled them to increase the F1-score of their Random Forest classifier [8] from 93% to 98%.

2.1.2 Features used for classification

We divide the encountered features for classification into the following categories: registration features, URL features, product features, merchant features, payment features, page-level features, website-level features, and visual features. We discuss for each of these categories why they can be useful and indicate in which of the research papers we encountered them. Apart from the research papers mentioned in the previous subsection, we also include features used by Kazemian and Ahmed [29] and by Bannur et al. [1]. Their research did not focus on detecting fake webshops specifically but on the broader class of malicious webpages. We only include features that are potentially relevant to our application.

Very similar features are grouped. Furthermore, features appearing in bold in the tables below signal that they were considered statistically significant in at least one research paper. Unfortunately, Carpineto and Romano [10], Bannur et al. [1] and Kazemian and Ahmed [29] did not provide a feature importance analysis. Wabeke et al. [54] and Haanen and Cox [15] only specified the relative importance of features, so we depicted their five most informative ones in bold.

Some features we encountered are not included in this overview, as they were not applicable in our setting. Examples include features related to website behavior in different search queries, as used by Carpineto and Romano [10].

Registration features

This category of features comprises information about registrants and registrars. While certain information is publicly available through the WHOIS protocol [16], operators of TLD registries like DNS Belgium and SIDN have more information at their disposal. Table 2.1 gives an overview of the registration features we encountered. As we discussed earlier, re-registration is an indicator of the *drop-catch* mechanism, and some registrars are more popular than others. Suspicious registration hours and

Table 2.1: Overview of registration features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Re-registration of domain name	Boolean	[54]
Registration hour	Categorical	[54]
Registrar	Categorical	[54]
E-mail provider of registrant	Categorical	[54]
Reported domains score	Numerical	[54]
Ratio of lowercase characters in registrant's name	Numerical	[54]
Registrant country	Categorical	[10]
Registered in China	Boolean	[55]
Domain age (less than ... years)	Numerical (/Boolean)	[10], [55]

email providers may be clues of registration from a specific country. The so-called "reported domains score" measures the ratio of reported malicious domains to the total number of domains registered by a certain registrar. The absence of capital letters in the registrant's name may hint at a fake and carelessly constructed name. Finally, as fake webshops are a relatively recent phenomenon, older domains may be less likely to be fake.

URL-level features

Table 2.2 summarizes the features that can be extracted from the URL or domain name of a website. The presence of suspicious words ("official", "replica", "cheap" or the targeted brand name) in the URL may indicate that a website tries to achieve high rankings in search engines. According to Wadleigh et al. [55], fraudulent webshops often use subdomains concatenating multiple words, so also the length of the domain name may reflect its trustworthiness. Another red flag is the presence of spelling mistakes in the URL.

Furthermore, for reregistered domains, the domain name of the fake webshop often does not reflect the content on the website. The syntactic difference between the HTML title and the domain name is typically larger for fake webshops than for legitimate ones, as demonstrated by Haanen and Cox [15]. Since similarly written words are not necessarily similar in meaning, also the semantic similarity between a domain name and the HTML title can be used as a feature. This can be measured by using a word embedding such as word2vec [38], which represents words in a high dimensional vector space. Models are trained in a way that words that often co-occur in a sentence lie close to each other in the vector space. The semantic similarity can

Table 2.2: Overview of URL-level features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Presence of suspicious words in URL	Boolean	[10], [55]
Length of domain name	Numerical	[55]
Presence of suspicious characters in URL	Boolean	[29]
Presence of spelling mistakes in URL	Boolean	[29]
Edit/Jaccard distance from domain name to HTML title	Numerical	[15]
Semantic similarity between domain and HTML title	Numerical	[15]

then be measured by calculating the cosine similarity or the Word Mover's Distance [30] between the embeddings of the HTML title and the domain name.

Product-related features

Table 2.3 outlines the features that can be extracted from the offered (counterfeit) products. Fraudulent webshops typically display unrealistically high discounts, both in absolute value and in percentages. The presence of multiple currencies can indicate that a single site is used to serve multiple countries, as is often the case for fake shops. A large number of currency symbols indicates the presence of many product offers on the homepage. According to Carpineto and Romano [10], homepages of legitimate webshops often act as a gate to their offers, while homepages of fake webshops often already contain product offers. A large number of duplicate prices could be a sign of lazy counterfeiters who copy and paste products without changing the prices. Finally, fake webshops are likely to mention a lot of unique brand names in an attempt to achieve high rankings in search results.

Merchant-related features

Legitimate webshops typically display a lot of information about the merchant's identity and business to reassure their customers. Table 2.4 presents the features that can be extracted from this information. Deep links to social media, as explained by Cox and Haanen [15], are hyperlinks to a specific webpage of a website, instead of to the index page of that website. For example, "www.facebook.com/nike/" is an example of a deep link, whereas "www.facebook.com" is not. As fake webshops will rarely own social media accounts, we can expect to find few deep links in their HTML.

Table 2.3: Overview of product-related features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Average price reduction	Numerical	[10]
Average percentage of price reduction	Numerical	[10], [55]
Percentage of discounted products	Numerical	[10]
Presence/**number of different currencies**	Boolean / Numerical	[10], [55], [15]
Total number of currency symbols	Numerical	[15]
Number of products	Numerical	[40]
Product offers on homepage	Boolean	[10]
Number of duplicate prices	Numerical	[55]
Number of unique brands mentioned	Numerical	[55]

Table 2.4: Overview of merchant-related features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Presence of a company name	Boolean	[40]
Presence of an address	Boolean	[10], [40], ([15])
Presence of a phone number	Boolean	[10], [15], [40]
Presence of a VAT number	Boolean	[10], [15], [40]
Presence/**number of links to social media**	Boolean / Numerical	[10], [15], [40]
Presence of deep links to social media	Boolean	[15]
Presence of link to mobile app	Boolean	[10]
Presence of link to a physical store	Boolean	[10]
Presence of a 'work with us' link	Boolean	[10]
Presence of an e-mailadres (of a free webmail provider)	Boolean	[10], [55], [40]
Presence of a banking number	Boolean	[15]

Payment features

According to Carpineto and Romano [10], Western Union is a preferred payment mechanism of scammers, as transfers can not be canceled or reversed. Additionally, the receiving merchant can essentially remain anonymous. Mostard et al. [40] used the number of payment methods found in the HTML as a feature, as they expected that fake webshops would support fewer payment methods than legitimate ones.

Page-level features

All features related to the general HTML structure of the website fall under the category of page-level features and are listed in Table 2.5. According to Carpineto and Romano [10], fake webshops are unlikely to display a notice & consent banner to comply with the cookie law. As shown in [9] and [41], large IFrames can be used to obfuscate malicious scripts in criminal websites. HTML meta descriptions and meta keywords are used by search engines to better index websites. High rankings in search engines are desirable for fake webshops, so this could be a popular tool to boost them. Haanen and Cox [15] found that fraudulent webshops used only a minimum of CSS and JavaScript, as these typically require some effort and imply customization. Meta Open Graph tags indicate optimizations by websites expecting to share content on Facebook, so these are also unlikely to be found on fake webshops. The number of images on a webpage may roughly approximate the number of offered products. The presence of a shopping cart system helps to distinguish webshops in general from regular websites. The number of links (internal, external, and others) is an indication of the size of the website, as well as of the effort that has been put into its creation. Furthermore, external links can be checked against blacklists of malicious URLs. The lexical diversity, i.e., the total number of words divided by the number of unique words, can be expected to be low for fraudulent webshops, as they mostly offer a range of similar products.

The body of the HTML can also be processed and turned into features. One possibility is to use a bag-of-words model, where the number of times every word appears is counted. An alternative is to use Term Frequency - Inverse Document Frequency (TFIDF) weighting, where each word is assigned a weight proportional to the number of times it appears on the page (i.e., the term frequency) and inversely proportional to the number of webpages in the training set it appears in. A large weight implies that a word occurs many times on the given page, while it's rarely mentioned on other pages.

Visual features

As explained earlier, Mostard et al. [40] used the discrepancy between visual and contextual information about available payment methods and social media accounts as features.

Table 2.5: Overview of page-level features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Website displays notice & consent banner	Boolean	[10]
Presence of large IFrames	Boolean	[55]
Presence/number of HTML tags (meta descriptions, meta keywords, forms, scripts, CSS ...)	Boolean / Numerical	[15], [1]
Presence of meta Open Graph tags	Boolean	[15]
Number of images	Numerical	[15]
Presence of shopping cart system	Boolean	[40]
Number of internal/**external**/ *mailto:*/*intent:*/map **links**	Numerical	[15], [40], [1]
External links to known malicious websites	Boolean	[29]
HTML size/word count	Numerical	[40], [1]
Lexical diversity	Numerical	[15]
Bag-of-words of HTML body		[1]
TF-IDF of HTML body		[1], [29]

Another approach, taken by Bannur et al. [1], is to use the Scale Invariant Feature Transform (SIFT) [34] on a screenshot of the webpage. This algorithm computes descriptors at regions of interest that are invariant to scaling, rotation, and affine distortion. The collection of these descriptors then allows detecting similar webpages. Alternatively, SIFT features can be computed on the included images and compared against a database of known descriptors of social media logos or payment method logos. The presence of these logos can then be used as a feature.

Finally, Kazemian and Ahmed [29] used Speeded Up Robust Features (SURF) [2], another feature descriptor with performance similar to SIFT, but faster. The authors clustered webpage screenshots based on their SURF descriptors and used the cluster-ID as a feature.

Website-level features

We categorize all remaining features as website-level features and summarize them in Table 2.6. As many fraudulent webshops do not configure mail servers, the presence of an MX record is a useful feature. Wabeke et al. [54] found that all TLS certificates of fake webshops were issued by a small number of TLS certificate issuers. As touched upon in the first part of this chapter, the AS or location of the hosting

Table 2.6: Overview of website-level features and the research paper(s) in which they were used.

Feature	Type	Research paper(s)
Existence of MX-record	Boolean	[54]
Presence/issuer of TLS-certificate (if any)	Boolean / Categorical	[54], [1]
Start- and enddate of SSL-certificate	Categorical	[54]
Autonomous System / location of hosting provider	Categorical	[54], [10]
Website sets tracking cookies	Boolean	[10]
Website found in Alexa top . . . sites	Boolean	[10], [55]
Presence of website analytics software	Boolean	[15], [40]
Number of open ports	Numerical	[40]

provider may also provide relevant information. According to Carpineto and Romano [10], fraudulent webshops usually do not set tracking cookies as they do not host third parties. Frequently visited websites, occurring in Alexa's list of most popular websites, are less likely to be fake webshops than websites that do not occur in that list. Legitimate websites may benefit from web analytics software to gain insights into customer behavior. While counterfeiters may also benefit from such software, setting up the software can be a time-consuming process. Legitimate webshops are likely more interested in the protection of their services and the sensitive data of their customers than fake webshops. Therefore it can be expected that legitimate webshops will have fewer open ports than fraudulent ones.

2.2 Learning from positive and unlabeled data

Traditional binary classification algorithms are trained on sets of fully labeled data with both positive and negative training examples, i.e., in a supervised way. However, as we explained earlier, it may be interesting to treat our training set as an instance of positive and unlabeled data (PU data). This section formalizes the problem of learning from PU data and explains the assumptions that are typically made. Additionally, we discuss PU learning techniques on a high level. We based this section on the survey conducted by Bekker and Davis [4], to which we refer for further information about PU learning.

2.2.1 Formalization

The goal of PU learning is to build a binary classifier, which distinguishes positive training examples from negatives ones based on their attributes. In general, we denote a training example as a tuple (x, y), where x is the attribute vector and y the class label. From now on, we let $y = 1$ correspond to a positive example and $y = 0$ to a negative example. Furthermore, let α denote the positive class prior, i.e., the intrinsic probability that a training example belongs to the positive class.

The supervised setting with fully labeled data assumes that the training set represents an independent and identically distributed (i.i.d.) sample of the true underlying distribution:

$$\begin{aligned} \mathbf{x} &\sim f(x) && (2.1)\\ &\sim \alpha f_+(x) + (1-\alpha) f_-(x), && (2.2) \end{aligned}$$

where $f(x), f_+(x)$ and $f_-(x)$ respectively represent the probability density functions of the true population, the true positives and the true negatives.

In the setting of PU learning, no negative examples and only some of the positive examples are labeled. We introduce a binary variable s indicating whether a training example was selected to be labeled, so that training examples are represented by a triplet (x, y, s). Since only positive examples can be labeled, we know that $Pr(y = 1|s = 1) = 1$. Additionally, we define the propensity score $e(x)$ [5] as the probability that a positive example with attribute vector x was selected to be labeled, i.e., $e(x) = Pr(s = 1|y = 1, x)$. Finally, let $c = Pr(s = 1|y = 1)$ express the label frequency, i.e., the ratio of positive examples that are labeled. Using this notation, we can derive the probability density function of the labeled distribution $f_l(x)$ as follows:

$$\begin{aligned} f_l(x) &= Pr(x|s = 1) && (2.3)\\ &= Pr(x|s = 1, y = 1) && (2.4)\\ &= \frac{Pr(s = 1|x, y = 1)}{Pr(s = 1|y = 1)} Pr(x|y = 1) && (2.5)\\ &= \frac{e(x)}{c} f_+(x). && (2.6) \end{aligned}$$

We distinguish two training set scenarios. In the so-called single-training-set scenario, we assume that one possesses a single training set that represents an i.i.d. sample from the true distribution. This means that a fraction αc of the training examples will be labeled. The distribution of attribute vectors in the training set can therefore be expressed as follows:

$$\begin{aligned} \mathbf{x} &\sim f(x) && (2.7)\\ &\sim \alpha f_+(x) + (1-\alpha) f_-(x) && (2.8)\\ &\sim \alpha c f_l(x) + (1-\alpha c) f_u(x), && (2.9) \end{aligned}$$

where $f_u(x)$ represents the probability density function of the unlabeled data. A second scenario in which PU data arises is the case-control scenario. Here we assume that one possesses two independently drawn datasets. One of these contains only positive examples, while the other contains only unlabeled examples and is an i.i.d. sample of the true distribution. Consequently, we can write:

$$\mathbf{x}|\mathbf{s} = \mathbf{0} \sim f_u(x) \tag{2.10}$$
$$\sim f(x) \tag{2.11}$$
$$\sim \alpha f_+(x) + (1-\alpha) f_-(x). \tag{2.12}$$

Depending on the scenario, the derivation of PU learning methods may differ. Nevertheless, most methods are applicable in both scenarios.

Note that in PU learning, the class prior α and the label frequency c are related. Given a PU dataset, one can calculate the expected value of one quantity if the value of the other is known. This can be seen as follows:

$$c = Pr(s = 1|y = 1) \tag{2.13}$$
$$= \frac{Pr(s = 1, y = 1)}{Pr(y = 1)} \tag{2.14}$$
$$= \frac{Pr(s = 1)}{Pr(y = 1)}. \tag{2.15}$$

The numerator in this equation can be calculated from the dataset. In the single-training-set scenario, the denominator is equal to the class prior α. In the case-control scenario, we can compute the denominator as follows:

$$Pr(y = 1) = Pr(y = 1|s = 0)Pr(s = 0) + Pr(y = 1|s = 1)P(s = 1) \tag{2.16}$$
$$= \alpha Pr(s = 0) + Pr(s = 1) \tag{2.17}$$

2.2.2 Assumptions in PU learning

In PU learning, we distinguish two reasons why an example is unlabeled: either the example is negative, or it is positive but not selected by the labeling mechanism. To enable learning in this setting, one needs to make assumptions about the labeling mechanism or the class distribution (or both). Furthermore, additional assumptions need to be made to estimate the class prior, which is an important input for many PU learning methods. For more information about the latter, we refer to [4].

Labeling mechanism assumptions

Most PU learning methods are based on the Selected Completely At Random or SCAR assumption [21]. This assumption states that the labeled examples are selected completely at random from the real distribution of positives, independent from their attributes. This is equivalent to stating that the propensity score $e(x)$ is constant:

$$e(x) = Pr(s = 1|x, y = 1) = Pr(s = 1|y = 1) = c. \tag{2.18}$$

Equation 2.6 then simplifies to $f_l(x) = f_+(x)$, meaning the probability density function of the labeled dataset accurately reflects the probability density function of the true positives.

A more general assumption about the labeling mechanism is the Selected At Random or SAR assumption [4], in which the propensity score depends on the attribute vector x. As a result, the labeled dataset is a biased sample from the real distribution of positives. A specific instance of SAR proposed by He et al. [26] assumes that the propensity score depends on the *probabilistic gap* $\Delta Pr(x) = Pr(y = 1|x) - Pr(y = 0|x)$. In particular, the propensity score is a non-negative, monotonically decreasing function of $\Delta Pr(x)$: the smaller the probabilistic gap, the smaller the probability of observing a label for a positive example.

Data assumptions

A common assumption made about the classes is that they are separable. This means there exists a theoretical classifier that can achieve 100% accuracy on the classification task. Another common assumption is to presume smoothness, i.e., if two attribute vectors are similar, then their probabilities of being positive are also similar.

2.2.3 Evaluation method

Common evaluation metrics for supervised classification tasks are the precision and the recall. The former measures how many of the positively classified examples truly belong to the positive class, while the latter quantifies how many examples of the positive class are classified correctly. More formally, we can define the precision as $p = Pr(y = 1 \mid \hat{y} = 1)$ and the recall as $r = Pr(\hat{y} = 1 \mid y = 1)$, where y and $\hat{y}$ correspond to the true and the predicted labels respectively. If one aims to optimize both the precision and the recall, the F_1 score is an appropriate measure. It is equal to the harmonic mean of p and r, i.e., $F_1 = \frac{2pr}{p+r}$, and is largest when both p and r are large.

It is not possible to estimate the precision directly from PU data, since we do not know the true labels for the unlabeled data. Consequently, it is also impossible to estimate the F1-score. Under the SCAR assumption we can estimate the recall though, i.e., $r = Pr(\hat{y} = 1 \mid s = 1)$. Furthermore, Lee and Liu [31] came up with an alternative to the F_1 score, which can be estimated from PU data and is high when both the precision and recall are high. We call this score F_1' and calculate it as

follows:

$$F_1' = \frac{pr}{Pr(y=1)} \tag{2.19}$$

$$= \frac{pr^2}{rPr(y=1)} \tag{2.20}$$

$$= \frac{Pr(y=1 \mid \hat{y}=1)r^2}{Pr(\hat{y}=1 \mid y=1)Pr(y=1)} \tag{2.21}$$

$$= \frac{Pr(y=1 \mid \hat{y}=1)r^2}{Pr(\hat{y}=1, y=1)} \tag{2.22}$$

$$= \frac{r^2}{Pr(\hat{y}=1)} \tag{2.23}$$

2.2.4 PU learning methods

Following Bekker and Davis [4], we can divide PU learning methods into three categories. The first category consists of two-step techniques, which start by identifying reliable negative examples and then train a supervised classifier on the reliable negative and positive examples. Two-step methods assume both separability and smoothness, such that all negatives in the unlabeled dataset are assumed to differ significantly from the labeled examples. The second category is made up of biased learners, which consider all unlabeled examples as negatives, albeit with class label noise. These learners make the SCAR assumption, as they consider a constant noise level for negative examples. The final category consists of class prior incorporation methods. This category relies on the estimation of the class prior, which is then used to directly apply the mathematics of the SCAR assumption. In the next paragraphs, we delve a little deeper into each of the three categories and present applications related to our setting, i.e., classification of webpages in an adversarial setting. For more elaborate overviews of the PU learning techniques and applications, we refer to Bekker and Davis [4] and Jaskie and Spanias [28].

Two-step techniques

Two-step techniques differ mainly in the first step, i.e., the identification of reliable negatives. These are typically defined as unlabeled data that are very different from the positive examples, but there is a lot of freedom in the choice of distance measure. The second step consists of training on the positives and the reliable negatives, which can be done using any supervised classification method. Popular choices are Support Vector Machines and Naive Bayes classifiers, but other methods are possible as well. Sometimes semi-supervised methods are used to also incorporate the remaining unlabeled examples.

We encountered several applications of two-step techniques that are related to our topic of interest. Yu et al. [60] developed the *Positive Example-Based Learning* (PEBL) framework for classifying webpages into multiple predefined classes. In the

first stage, they search for features that occur more often in the positive data than in the unlabeled data. Examples that do not possess such strongly positive features are then categorized as reliable negatives. In the second stage, they iteratively train an SVM classifier on the positives and the reliable negatives. After each iteration, the set of reliable negatives is extended with the unlabeled examples that were predicted negative by the SVM. In their experiments, PEBL achieved performances similar to a traditional SVM classifier in a supervised setting. Zhang et al. [61] used another two-step technique (and a class prior incorporation method) to detect potential malicious URLs. Finally, we found applications in fake review detection [43][27].

Biased learning

Recall that biased learning methods consider the unlabeled examples as negatives with class label noise. Therefore, we refer to the unlabeled examples as negatives in this context. The original biased learning technique for PU learning was introduced by Liu et al. [33] and consists of a biased SVM that places different penalties on misclassified positives and negatives. This boils down to placing more importance on the correct classification of positives, of which we know the class label is correct, than on the correct classification of negatives, which have class label noise. Lee and Liu [31] proposed a similar approach but in a logistic regression framework, which they called weighted logistic regression,

Even still, too much weight may be given to unlabeled negative examples that are actually positive. Mordelet and Vert [39] addressed this by proposing bagging SVM, which was inspired by Breiman's well-known bagging technique [7]. Bagging SVM first trains multiple biased SVMs that discriminate the positive examples from different subsets of the negatives, and then uses a voting mechanism to aggregate predictions of the individual classifiers. The authors refer to the fraction of positives hidden among the unlabeled data as *contamination* and reason that the amount of contamination in the different training sets is varied by subsampling the negatives. This induces variability in the classifiers, which can then be exploited by the aggregation procedure.

Claesen et al. [12] went one step further and proposed Robust Ensemble SVMs, which subsample also the positive examples. Their method has the same advantages as bagging SVM and can additionally cope with outliers or contamination in the set of positives.

Class prior incorporation methods

There are different types of class prior incorporation methods, but all of them require knowledge of the class prior, or equivalently, the label frequency c. The first type of method is called *postprocessing* and was introduced by Elkan and Noto [21]. They came up with the concept of a non-traditional classifier, which is trained to predict $Pr(s = 1 \mid x)$, i.e., the probability of being labeled. Under the SCAR assumption, we can then directly obtain the probability of an example being positive through

$Pr(y = 1 \mid x) = \frac{Pr(s=1 \mid x)}{c}$.

The second type consists of *preprocessing* methods, which transform the PU data to a new dataset that can be fed to a supervised classifier. This transformation can be a weighting of the data based on the label frequency [20], a duplication of the unlabeled data to let them count partially as positives and partially as negatives [21], or a reweighting that leads to the same expected empirical risk as the fully labeled dataset [19]. This last method of empirical risk minimization is particularly interesting since it has been adapted to work under the SAR assumption [5].

The final type of class prior incorporation methods adapts existing algorithms, such as decision tree learning algorithms, to allow learning from PU data. For more details, we refer to Bekker and Davis [4].

Chapter 3

Analysis of existing solution

This chapter analyses the existing solution implemented by DNS Belgium. First, we take a closer look at the ground truth labeling. Second, we discuss the features that were used for classification and inspect their distribution. Third, we review the supervised classifier and analyze its performance.

3.1 Ground truth

DNS Belgium performed the ground truth labeling for its supervised classifier during the last months of 2019. Table 3.1 provides an overview of the approximate labeling timeline. The first collection of fake webshops originated from the FPS Economy, which had received customer complaints about certain domains. Soon, DNS Belgium discovered that many of those domains were hosted in the same Autonomous Systems and countries. Building on this knowledge, they were able to detect more fake webshops residing in those ASes and countries. Later, they gradually extended the labeled dataset with fraudulent shops detected by early versions of their classifier. In total, they gathered 1981 labels of fake webshops.

The benign dataset is a mixture of both webshops and non-webshops. The 760 initial labels for benign webshops were provided by employees of DNS Belgium and it is therefore not surprising that the labeled dataset contains many well-known Belgian webshops, such as coolblue.be, decathlon.be and delhaize.be. The benign labels were extended with random domains from the .be domain, which were manually verified to be non-malicious. Currently, the number of benign labels amounts to 1919.

For each of the 3900 labeled domains, DNS Belgium fetched the HTML of the homepage and extracted some features. These were subsequently merged with features from their registration database. A description of the extracted features, categorized according to our analysis in the previous chapter, is given by Table 3.2.

Table 3.1: Timeline of the labeling process by DNS Belgium.

Approximate crawl date	Number of fake webshops	Number of benign domains
19/09/2019	1697	760
25/09/2019	163	403
09/10/2019	47	756
09/12/2019	74	0
Total	1981	1919

Table 3.2: Overview of features available for DNS Belgium's supervised classifier

Feature category	Feature name	Description	Type
Registration	`IS_TRANSFERED`	Whether domain was transfered to another registrar	Boolean
	`IS_REREGISTERED`	Whether domain was re-registered after it expired	Boolean
	`rereg_1d`, `rereg_10d`, `rereg_30d`, `rereg_90d`, `rereg_365d`, `rereg_older`	Indicators of period between domain expiration and re-registration	Boolean
Product-related	`nb_numerical_strings`	Number of numerical strings that resemble prices	Numerical
Merchant-related	`nb_links_tel`	Number of telephone links	Numerical
	`nb_links_email`	Number of email links	Numerical
Page-level	`nb_imgs`	Number of images	Numerical
	`nb_links_int`	Number of internal links	Numerical
	`nb_links_ext`	Number of external links	Numerical
	`nb_input_txt`	Number of text input fields	Numerical
	`nb_button`	Number of clickable buttons	Numerical
	`nb_meta_desc`	Number of words in content of meta description tag	Numerical
	`nb_meta_keyw`	Number of words in content of meta keywords tag	Numerical
	`nb_tags`	Total number of HTML elements	Numerical
	`body_text`	Text of HTML body	String
	`meta_text`	Content of meta description and meta keyword tags	String
	`title`	HTML title	String
Website-level	`has_mx`	Presence of mail exchanger (MX) record	Boolean
	`ssl_flag`	Presence of TLS-certificate	Boolean
	`as`	AS of hosting provider	Categorical
	`country_code`	Country code of hosting provider	Categorical

3.2 Feature distribution over labeled data

This section inspects the differences in the feature distributions for fake webshops and benign domains. It respectively discusses the registration, merchant-related, page-level, and website-level features.

3.2.1 Registration features

The majority of fake webshops are re-registered (Figure 3.1), so the fake webshop owners were probably not the original owners of the domain name. However, only a minority is transferred to another registrar (Figure 3.2). On the other hand, benign domains are less likely to change from owner and more likely to be transferred. This could be explained by the fact that different registrars offer different features, different support, and different pricing options. While a benign domain owner benefits from choosing a registrar that suits his or her needs, this may be of minor importance to a fake webshop owner, as long as the registrar is not too expensive. Furthermore, a fake webshop operator may not really care about the domain name itself and prefer to register another one instead of putting effort into the transfer. We also observe that the re-registration behavior is very different for both types of domains (Figure 3.3). The vast majority of fake webshops are re-registered within a day of expiration, which confirms the *drop-catch* behavior mentioned in the previous chapter. Benign domains do not exhibit such a pattern and are even more likely to be re-registered after a long time.

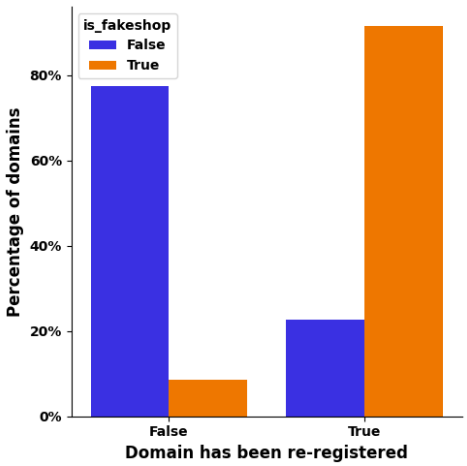

Figure 3.1: Fake webshops are far more likely to be re-registered than benign domains.

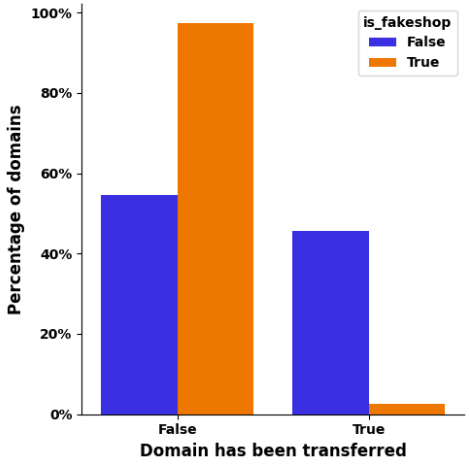

Figure 3.2: Fake webshops are seldomly transferred to another registrar, while this is quite common practice for benign domains.

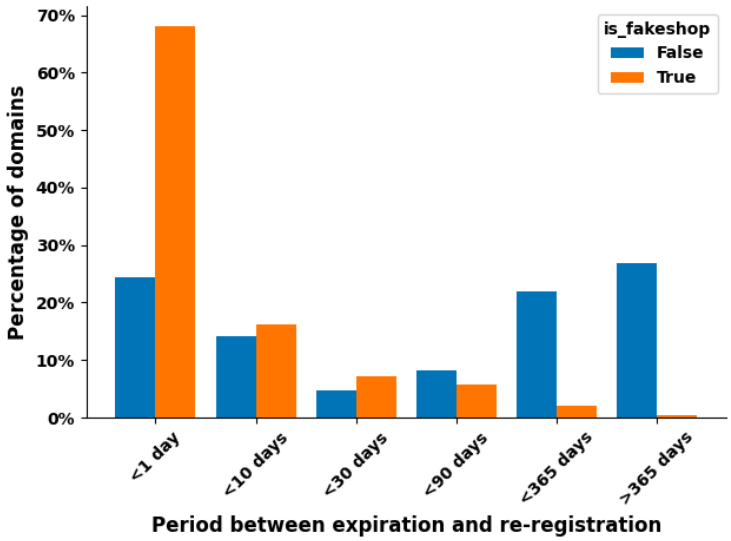

Figure 3.3: Fake webshops are mostly re-registered according to a drop-catch mechanism, while benign domains are more likely to be re-registered after a long time. Percentages are expressed with respect to the number of re-registered domains per category (and not the total number of domains per category).

3.2.2 Merchant-related features

The merchant-related features consist of the number of telephone and email links on the homepage. It is very unlikely to encounter either of those on a fake webshop, although their presence is also very limited in benign domains (Figures 3.4 and 3.5). We should note that these features truly represent links, i.e., HTML `<a>` tags where the `href` attribute starts with `tel:` or `mailto:`. Therefore, telephone numbers and email addresses that occur in plain text in the body are not picked up by these features.

3.2.3 Page-level features

The distributions of the number of images, internal links and tags are very compact for fake webshops, while they are spread out for benign domains (Figures 3.6 through 3.8). This raises the suspicion that the labeled dataset contains multiple clusters of similar fake webshops, while the benign domains are truly a diverse sample of the .be zone. Furthermore, it seems that many fake webshops are self-contained, in the sense that they rarely display links to other web pages (Figure 3.9). The use of text input fields and buttons is quite uncommon for both fake webshops and benign domains (Figures 3.10 and 3.11). Finally, the distributions of the meta tags usage hint again at clusters of fake webshops and more diverse benign domains (Figures 3.12 and 3.13).

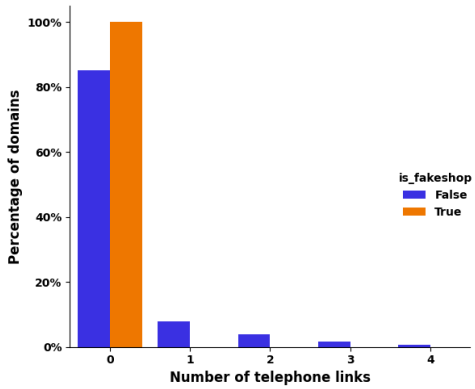

Figure 3.4: Telephone links occur only rarely on the homepage of a website, irrespective of whether it is fraudulent or benign.

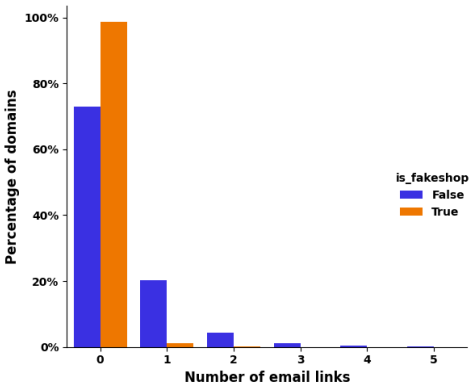

Figure 3.5: Email links occur only rarely on the homepage of a website, irrespective of whether it is fraudulent or benign.

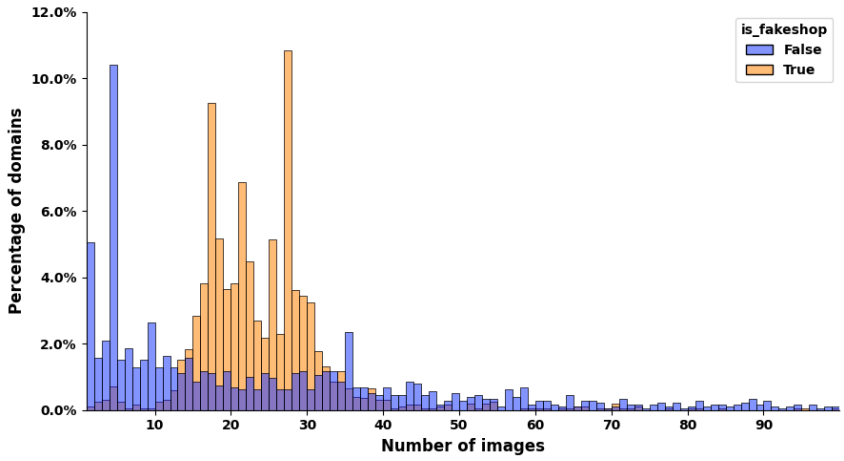

Figure 3.6: The compact distribution of the number of images for fake webshops hints at the presence of clusters of similar fake webshops. Outliers on the x-axis are excluded to enhance the clarity of the figure.

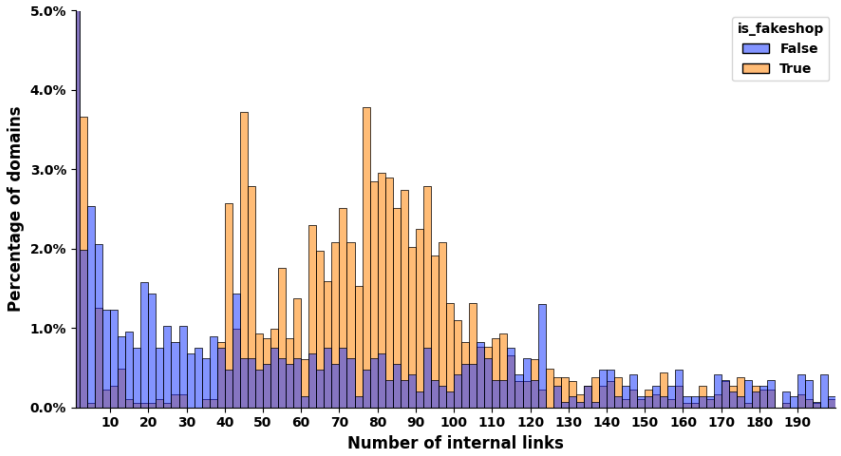

Figure 3.7: The compact distribution of the number of internal links for fake webshops hints at the presence of clusters of similar fake webshops. Outliers on the x-axis are excluded to enhance the clarity of the figure.

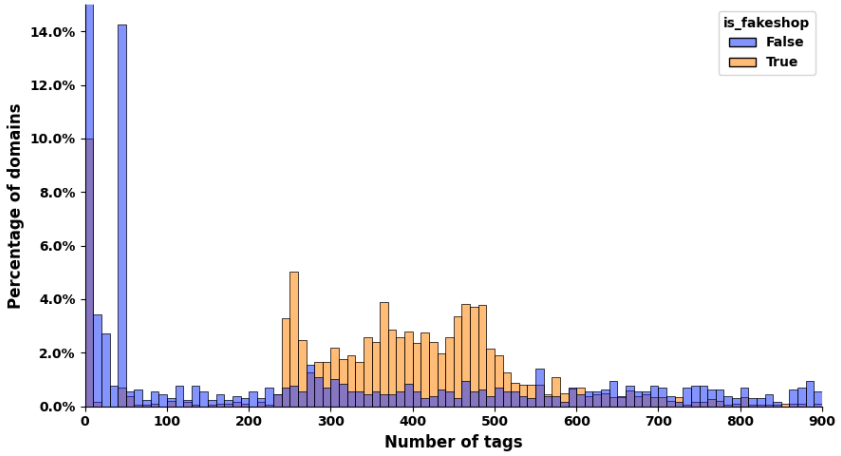

Figure 3.8: The compact distribution of the number of tags for fake webshops hints at the presence of clusters of similar fake webshops. Outliers on the x-axis are excluded to enhance the clarity of the figure.

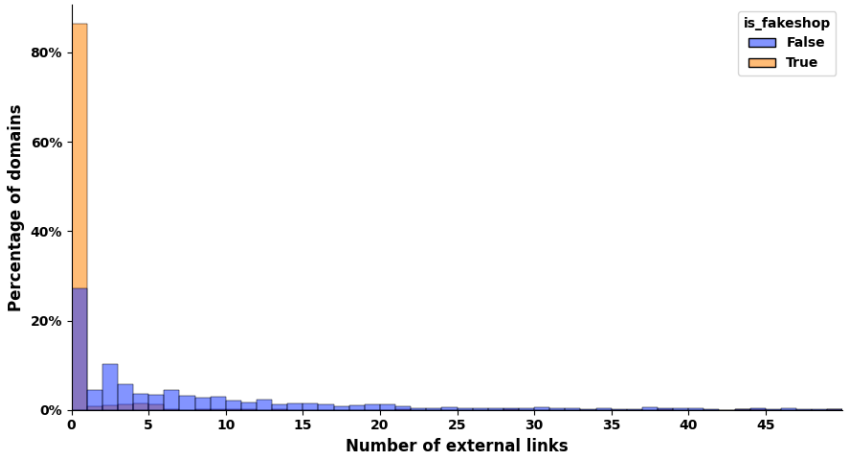

Figure 3.9: Fake webshops are often self-contained, in the sense that they do not link to other websites.

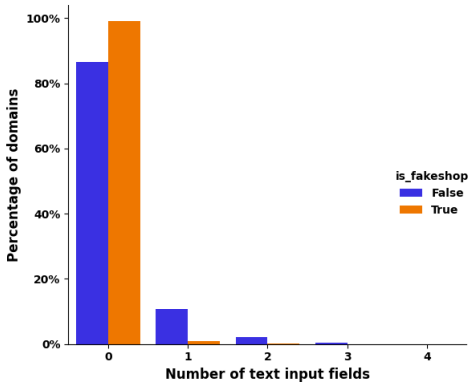

Figure 3.10: Text input fields are rare in both types of domains.

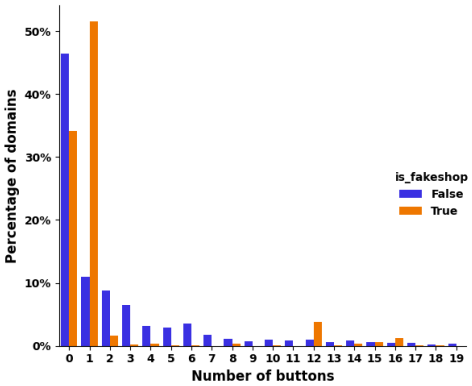

Figure 3.11: Most fake webshops display either one or no buttons on their homepage. The distribution for benign domains varies more.

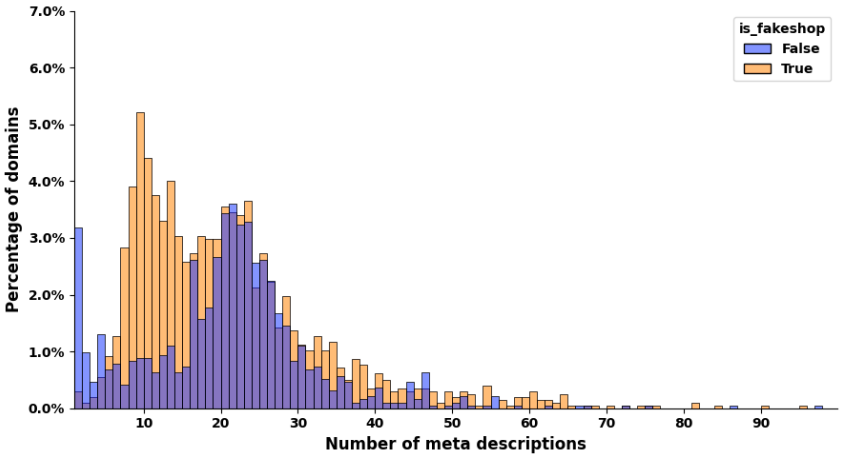

Figure 3.12: The distribution of the number of words in the meta description tag hints again at clusters of fake webshops.

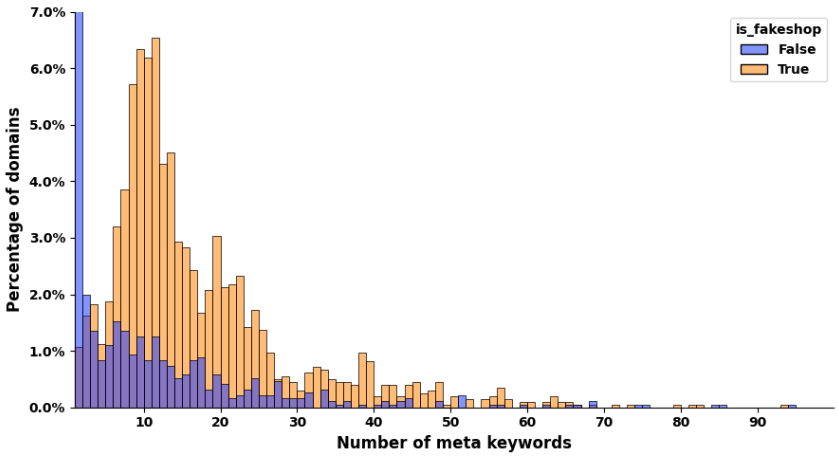

Figure 3.13: The distribution of the number of words in the meta keywords tag hints again at clusters of fake webshops. Most benign domains do not use the meta keywords tag.

3.2.4 Website-level features

The distribution of the AS of the hosting provider is less diverse for fake webshops than for benign domains (Figure 3.14). In particular, more than half of the fake webshops are hosted in three ASes. However, as we discussed in the previous section, DNS Belgium used the AS of known fraudulent shops to detect new ones. This may explain to some extent why certain ASes occur more often than others. To avoid overfitting on this feature, DNS Belgium did not include it during the training of the classifier. The same holds for the country code of the hosting provider (Figure 3.15). Some countries, like Turkey (TR) and the Republic of Seychelles (SC), seem to host only fake webshops, which is not necessarily true. Therefore also this feature was withheld from training.

It is very uncommon for fake webshops to configure a mail server (Figure 3.16). However, they hold TLS certificates almost as often as benign domains (Figure 3.17). While this may seem surprising, it is probably part of their strategy of resembling legitimate webshops as much as possible. Since some TLS certificates are available for free, e.g., from Let's Encrypt [24], cost is also not an issue.

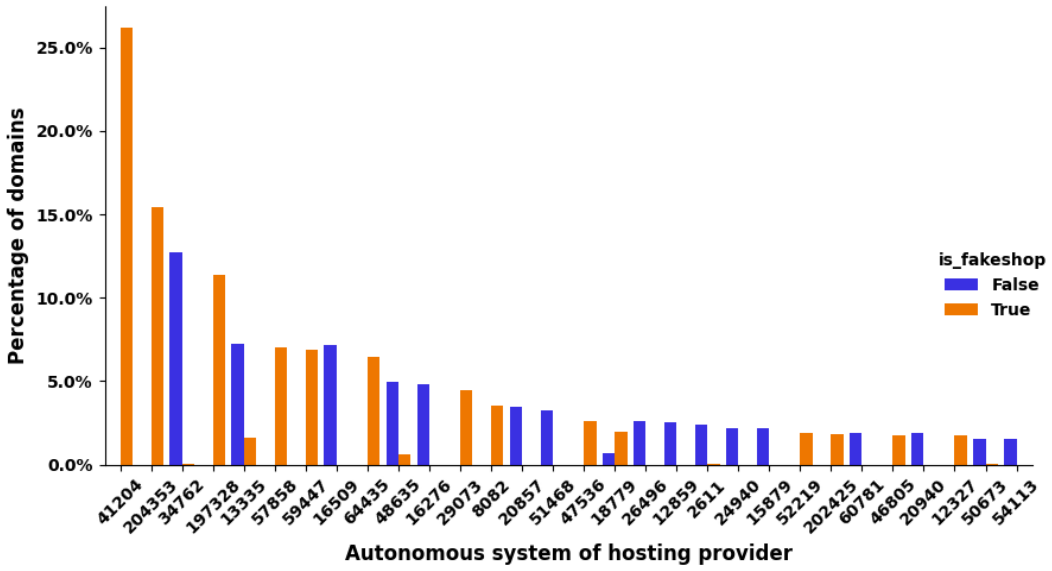

Figure 3.14: More than half of the fake webshops are hosted in only three ASes. For benign domains, the distribution is more diverse. For the sake of clarity, we only show the 30 most frequently occurring ASes.

3.3 Supervised classification

DNS Belgium's solution splits the labeled dataset into 70% training data and 30% test data. It then preprocesses the features as follows:

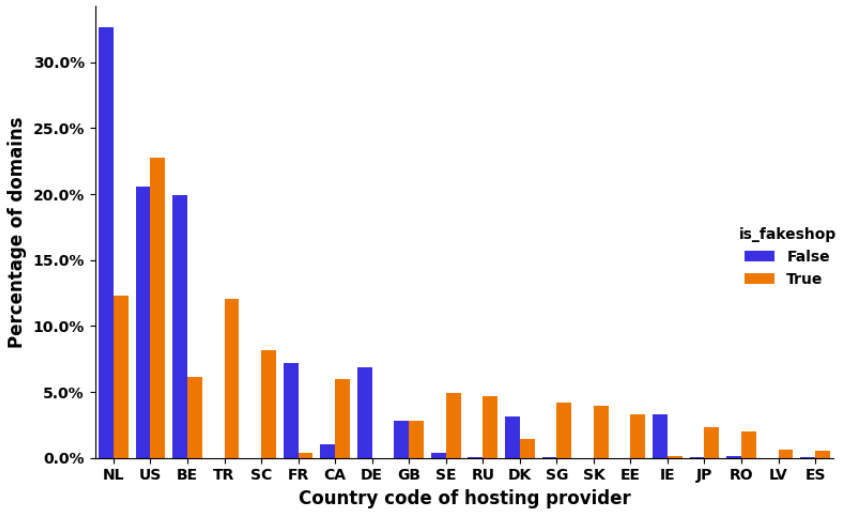

Figure 3.15: Many countries host only fake webshops, while benign domains are concentrated in a limited number of countries. For the sake of clarity, we show only the 20 most frequently occuring countries.

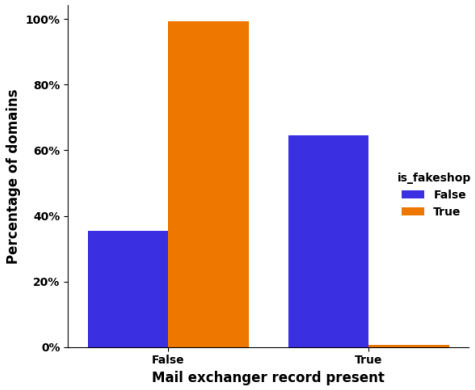

Figure 3.16: Almost none of the fake webshops configured a mail server. This is also not common practice for benign domains, but at least it is done more often.

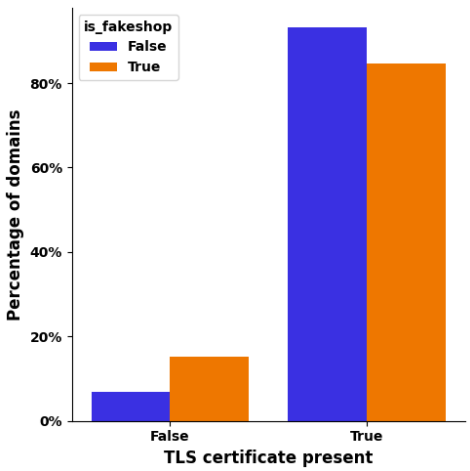

Figure 3.17: Fake webshops hold TLS certificates almost as often as benign domains.

- Missing boolean and numerical values are imputed with the mean.
- Numerical features are standardized, i.e., the mean is removed and they are scaled to unit variance.
- Words in `body_text` and `meta_text` that are at least three characters long, appear in at least 0.1% of the documents and at most 90% of the documents, are turned into TF-IDF features.

Next, a Random Forest Classifier is fitted to the preprocessed features. A grid of different parameters is tried and the best combination of parameters is determined using 5-fold cross-validation. To avoid data leakage, the preprocessing steps and the classifier are composed in a scikit-learn pipeline.

Table 3.3 presents the performance of the obtained classifier on the test set, as well as the performance of a tuned Gradient Boosting Classifier, which we implemented only for comparative reasons. Both classifiers achieve high precision and high recall, so at first sight, there is no need for a PU learning classifier. However, in practice, DNS Belgium's classifier does not detect that many fake webshops anymore and mainly predicts false positives. We determined the following possible causes for this problem:

1. Fake webshop operators know that DNS Belgium actively hunts fake webshops and prefer to register new domains in other zones.
2. The set of benign domains is not representative of the entire benign .be zone, causing the classifier to behave unexpectedly in underrepresented regions of the feature space.
3. The layout and strategy of fake webshops changes over time and the classifier is unable to generalize to new types of fake webshops. If the training set consists of clusters of similar fake webshops and these are spread out over the train, validation, and test sets, it could be that the classifier actually just remembers clusters and generalizes poorly.

The second and third causes seem more likely than the first. In the next chapter, we try to address them by learning from all the webshops in the .be zone and by clustering the fake webshops.

The most important features, as measured by the mean decrease in impurity (MDI) in the trees' nodes, are roughly the same for both classifiers (Figure 3.18). The MDI criterion is biased towards high cardinality features (such as `nb_tags`) [47], so we also computed the importance based on feature permutation. These importances revealed roughly the same important features, albeit with changes in the importance magnitudes.

Table 3.3: The Random Forest Classifier and Gradient Boosting Classifier achieve similar performances in terms of precision, recall and F1-score.

Classifier	Precision	Recall	F1
Random Forest	0.98	0.96	0.97
Gradient Boosting	0.98	0.97	0.98

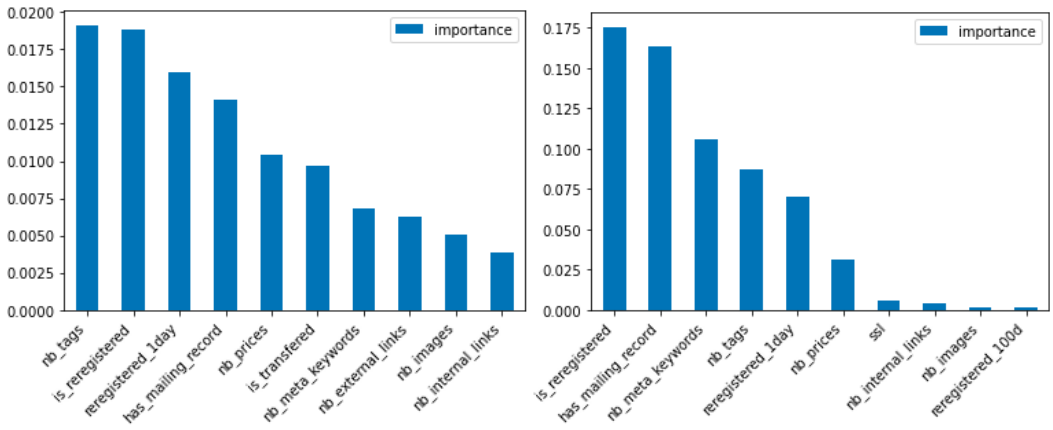

Figure 3.18: The most significant (non-TFIDF) features, based on the mean decrease in impurity, are similar for the Random Forest Classifier (left) and the Gradient Boosting Classifier (right).

Figure 3.19 visualizes the most important TF-IDF terms for each classifier. The Random Forest Classifier places high importance on currency names, e.g., (Australian / Canadian) dollar and (Norwegian / Danish) krone. Furthermore, we recognize words related to webshops, such as 'verzenden' (= shipping) and 'retouren' (= returning). The Gradient Boosting classifier's terms are less comprehensive, except for a few terms like 'cher' (French for expensive) and 'cad' (short for Canadian dollar). This is related to the fact that the Random Forest Classifier places much higher importance on the TF-IDF terms than the Gradient Boosting Classifier, as evidenced by the feature importance magnitudes in Figure 3.19.

Figure 3.19: The most significant TF-IDF terms for the Random Forest Classifier (left) are more comprehensive than those of the Gradient Boosting Classifier (right). This is related to the fact that the former classifier places higher importance on these terms. Word sizes are proportional to the word's importance for the classifier.

Chapter 4

Methodology

This chapter describes the different steps that were required to achieve a PU Learning-based classifier. First, it discusses the features that were added compared to the original classifier. Second, it describes how we collected the unlabeled data. Third, it reports how we clustered the fake webshops and used this clustering to define the cross-validation folds. Fourth, it provides more detail about the different PU classification methods we trained.

4.1 Implementation of additional features

This section recaps the features encountered during our literature study and discusses which of those we added to the existing classifier. We also investigate how they are distributed over the labeled dataset. Decisions about which features we implemented were based on the significance of those features in previous studies and on the ease of implementation in the available timespan.

4.1.1 Registration features

As evidenced by Table 2.1, registration features were often found to be significant in previous studies. Since DNS Belgium has the advantage of having registration information at its disposal, we wanted to include as many registration features as possible. More specifically, we tried to include the reported domains score, the email provider of the registrant, the ratio of lowercase characters in the registrant's name, and the registration hour.

Reported domains score. Some registrars may be more popular among fake webshop owners than others because they are cheaper or facilitate registration in bulk. To verify this, DNS Belgium determined for each registrar the number of domains that were taken down on suspicion of being a fake webshop. Of course, some registrars are just better known and are more popular in general, so they also determined the total number of new registrations per registrar (since 11/05/2018). The ratio of these two numbers confirms the suspicion that some registrars are used relatively

more often to register fake webshops than others (Figure 4.1). Furthermore, the three registrars with the largest reported domains score are also the registrars with the highest fake webshop count: NETIM, Gransy s.r.o. and Registrar.eu registered 1244, 153 and 1263 fake webshops respectively. Therefore, the reported domains score seems to be an informative feature.

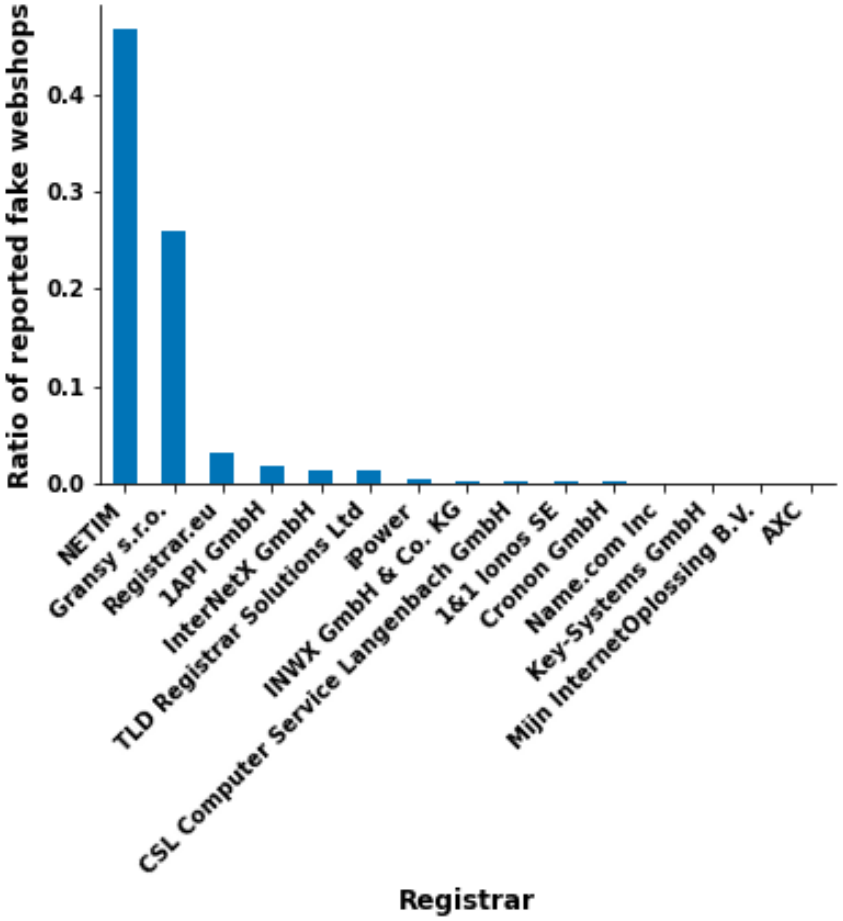

Figure 4.1: Some registrars are much more popular among fake webshops than others. For the sake of clarity, only the 15 registrars with the largest ratio are shown.

Email provider of the registrant. We think of two reasons why the email provider of the registrant can be informative. First, fake webshop owners do not necessarily live in Belgium and therefore do not always use an email provider that is popular in Belgium. The labeled dataset confirms this, as the most frequently encountered email providers are `hxmail.com` and `163.com` for the fake webshops and `gmail.com` and `hotmail.com` for the benign domains. To the best of our knowledge, the former two providers are rarely used in Belgium, while the latter two are encountered very often. Second, the fake webshop owner is not likely to reveal his true

identity and is therefore likely to use a fake email address. Again, the labeled dataset confirms this, as the fake webshops feature providers like `2kusr1cgwy6b.com` and `1wmldgby4bgb.com`, which are not legitimate email providers.

Similar to the reported domains score, we incorporate the email provider of the registrant through a trust score. Here, we calculate the score for a certain provider by dividing the number of occurrences of that provider in the labeled benign set by the total number of occurrences in the labeled data. Providers that were never encountered in the labeled dataset receive a trust score of 0. In retrospect, this method is not flawless, as there will also be benign webshops for which the email provider does not occur in the benign labeled set. We could improve the trust score by defining additional rules, stating for example that the provider should not contain more than two digits or more than four subsequent consonants. However, we did not implement such rules and used only the simple trust score.

Ratio of lowercase characters in registrant name. The absence of capital letters in a registrant's name could point to a carelessly constructed name and could therefore hint at a fake webshop. Although Table 2.1 did not indicate the ratio of lowercase characters in the registrant's name to be a significant feature, DNS Belgium stated they were under the impression that registrant names in complete lowercase characters were indeed more common in fake webshops than in benign domains. Figure 4.2 presents the distribution of this ratio over the labeled dataset and confirms this suspicion. The range $[0, 1)$ of this feature is not very informative, as the exact ratio will depend on the registrant's name. Therefore we binarize this ratio, such that the feature is only equal to 1 when the registrant's name is completely in lowercase characters (ignoring spaces).

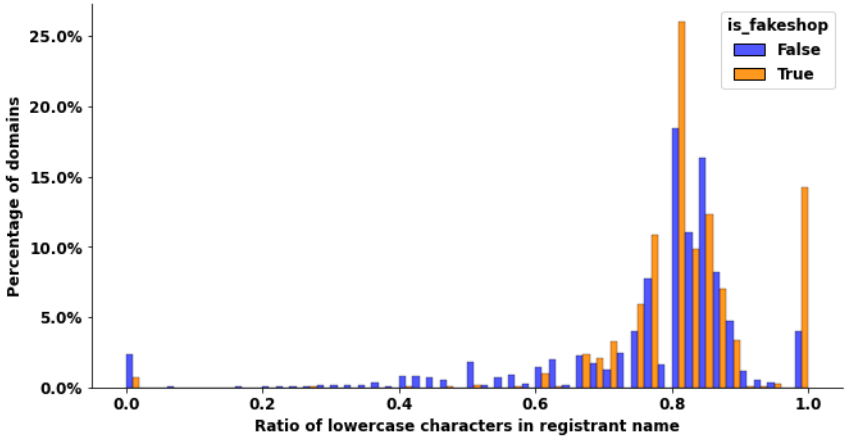

Figure 4.2: Fake webshops are often registered with a registrant name in complete lowercase characters.

Registration hour. The distribution of the time of registration is quite different for the fake webshops and the benign domains (Figure 4.3). The majority of the benign domains are registered during the Belgian daily working hours, while the fake webshops tend to be registered at night. This suggests that they are often registered by people in non-European countries. Due to a misunderstanding, we were under the impression that DNS Belgium's database no longer stored the registration hour for all domains. Therefore, we did not include it in our solution. This feature can still be extracted though and could be added in the future.

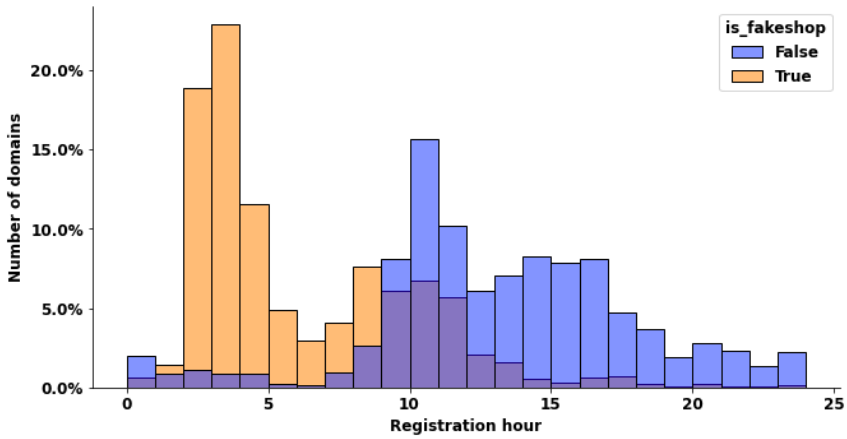

Figure 4.3: The registration pattern for benign domains resembles Belgian working hours, while fake webshops are mostly registered during Belgian nighttime.

4.1.2 URL-level features

So far, DNS Belgium's classifier did not include any URL-level features. Previous research searched for the presence of suspicious words or suspicious characters in the URL, but defining a list of such words or characters seems a rather subjective task. Furthermore, we already found that most fake webshops are re-registered domains, meaning the webshop operators did not choose the URL themselves and suspicious characters are unlikely. For the same reason, spelling mistakes in the domain name seem improbable. Additionally, domain names could be written in Dutch, French, German, or English, so we would have to spell-check in multiple languages.

Instead, we decided to calculate the distance between the domain name and the HTML title, similar to Cox and Haanen [15]. Keeping in mind that many fake webshops are re-registered, we expect a low similarity between them, while we expect a higher similarity for benign domains. We calculate two types of similarities. First, we wanted to determine the fraction of words in the domain name that appears in the HTML title. However, this requires splitting the domain name into words, which is non-trivial, especially since we do not know the language of the domain name. Instead, we turn it around and determine the fraction of words in the HTML title

that are contained in the domain name. This fraction is equal to 0 for almost all fake webshops and about half of the benign domains (Figure 4.4). Since fractions larger than 0 do not seem very informative, it makes sense to binarize this feature.

Of course, the HTML title may contain variations of words in the domain name instead of the exact same words, e.g., verbs instead of nouns or the other way around. Therefore, we also calculate the Levenshtein distance between the domain name and the HTML title. On average, we expect this distance to be smaller for benign domains than for malicious webshops, which is confirmed by Figure 4.5.

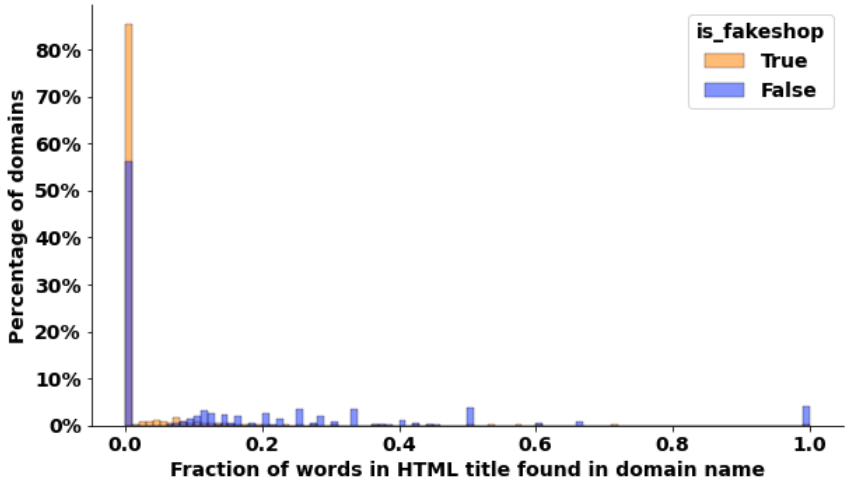

Figure 4.4: HTML title words for fake webshops almost never occur in the domain name.

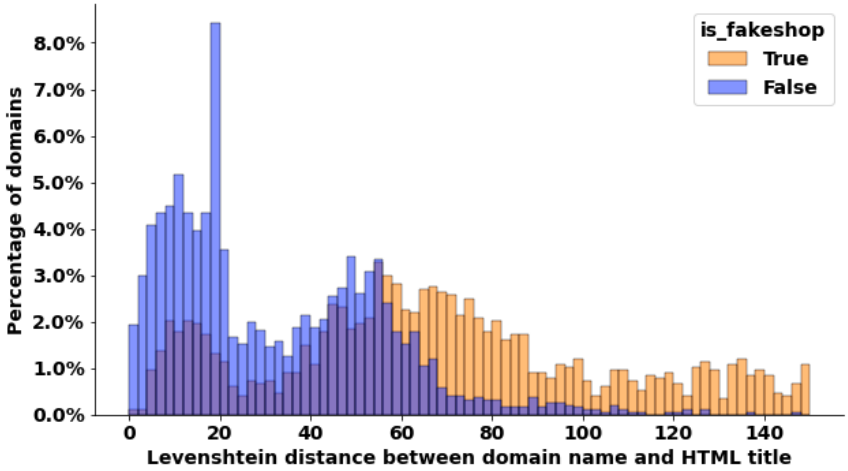

Figure 4.5: The Levenshtein distance between the HTML title and the domain name is larger on average for fake webshops than for benign domains.

Another interesting feature used by Cox and Haanen [15] is the semantic similarity between the domain name and the HTML title, calculated with the help of word embeddings. However, this would again require us to split the domain name into words of the correct language. Furthermore, we would have to use different trained embeddings depending on the language. Since this seemed hard to achieve in the available time span, we decided to not implement this feature.

4.1.3 Product-related features

So far, the only product-related feature of DNS Belgium's classifier was the number of numerical strings that resemble prices. Table 2.3 indicates that previous research tried many other features, such as the number of products, the average (percentage of) price reduction, and the percentage of discounted products. To be meaningful, these features require an accurate estimation of the number of products on display and a method to match prices to products. Since this is a non-trivial task and the research papers mentioning these features provided only a little information on their approach to these problems, we decided to not implement this kind of feature.

However, we did include another feature that was found to be significant in previous research: the number of currency mentionings in the HTML body. Currencies occur in different formats, so we track all of them [59]:

- Currency names, such as `euro` and `dollar`;
- Currency symbols, such as € and $;
- ISO codes, such as `EUR` and `USD`
- Fractional units, such as `cent`.

We make sure to also match expressions with preceding or trailing numbers, e.g., 25€ and €25. The number of currency occurrences acts as a proxy for the number of products on display, as there is usually a single price related to each product (unless it is discounted). For fake webshops, we expect a lot of discounted products on display, thus leading to a large currency count. Figure 4.6 illustrates that the homepages of fake webshops indeed display more currencies than those of benign domains. We should note that this is partially due to not all benign domains being webshops. However, even if we ignore the domains for which the currency count equals zero, it is clear that the distributions barely overlap. Therefore, we believe that the currency count is a useful feature. Appendix A provides a breakdown of the total currency count into the number of currency names, symbols, ISO codes, and units.

Furthermore, as discussed earlier, a fake webshop often serves multiple countries and displays prices in different currencies. Therefore, we also keep track of the number of different currencies encountered in the HTML. For each format, the distribution of the number of different currencies is given in Appendix A. Most notably, fake webshops display more different ISO codes and currency names.

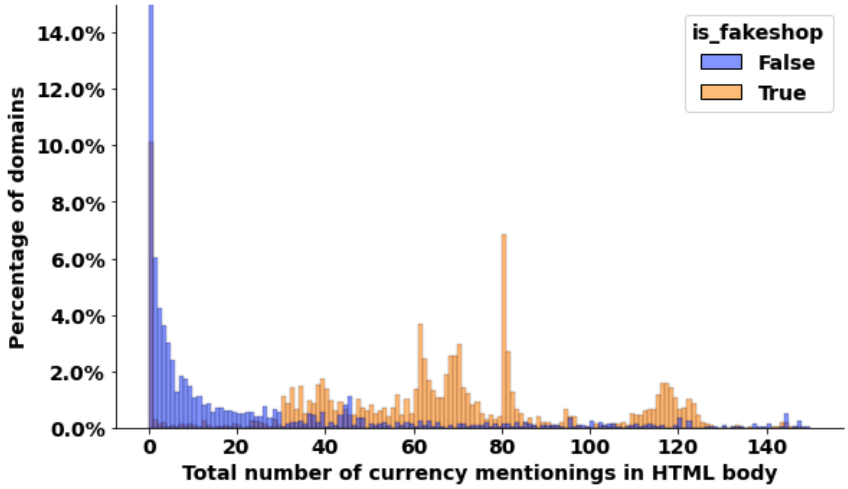

Figure 4.6: Currencies occur much more on homepages of fake webshops than on homepages of benign domains.

4.1.4 Merchant-related features

Table 2.4 made clear that the presence of information about the merchant can be a significant feature to discriminate between fake webshops and benign domains. Therefore, we added the number of (deep) links to social media and the presence of an address as features to our classifier.

Number of (deep) links to social media. As explained before, we do not expect fake webshops to be very active on social media. Indeed, Figure 4.7 confirms that we find less links to social media and in particular less deep links. However, some fake webshops did contain Facebook deep links, so we examined if these corresponded to actual facebook pages. It turns out that most of those deep links were of the form `www.facebook.com/sharer/sharer.php?u=http://nikebelgie.be`. This type of link allows to share the website on Facebook, but it is not a link to an actual Facebook page. As such, we subsequently excluded them from our deep link count. It is worth noting that these links were all found in a cluster of fake webshops selling shoes from different brands and with similar domain names, such as `nikeairmaxbelgie.be` and `filabelgie.be`. Appendix B provides an overview of the number of links and deep links to Facebook, Instagram, Twitter, and LinkedIn.

Presence of address. Previous research found that also the presence of an address on the webpage was a significant feature. However, fake webshops could easily display a false address to fool customers. Verifying the truthfulness of a displayed address seemed difficult, especially since this address could also be outside of Belgium. Now, assuming a legitimate webshop, we expect that the address provided during registration would match the address displayed on the website. While a fake webshop operator could also make sure these addresses match, he has no intrinsic motivation

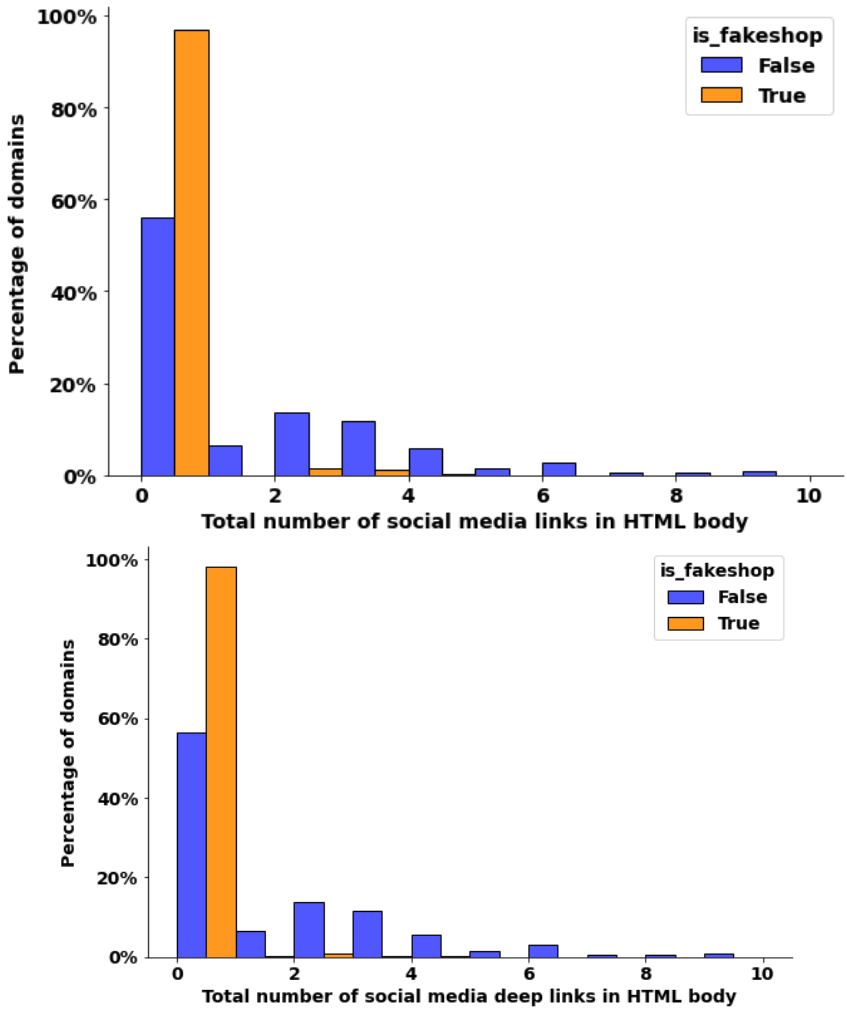

Figure 4.7: Fake webshops are very unlikely to display social media links on their homepage (top), and even less likely to display deep links (bottom).

to do so. Therefore, we calculate a simple score based on the presence of the domain's registration address in the HTML body. We award one point for the presence of the city name, one point for the street name, and one for the postal code. Most fake webshops score zero out of three on this metric and none of them scores higher than one out of three (Figure 4.8). Benign domains, on the other hand, do regularly score higher than one. Most of them still score zero, but the benign set also contains non-webshops, which may have no reason to display an address on their webpage. Furthermore, the address could be located on a page other than the homepage, although the same argument holds for the fake webshops.

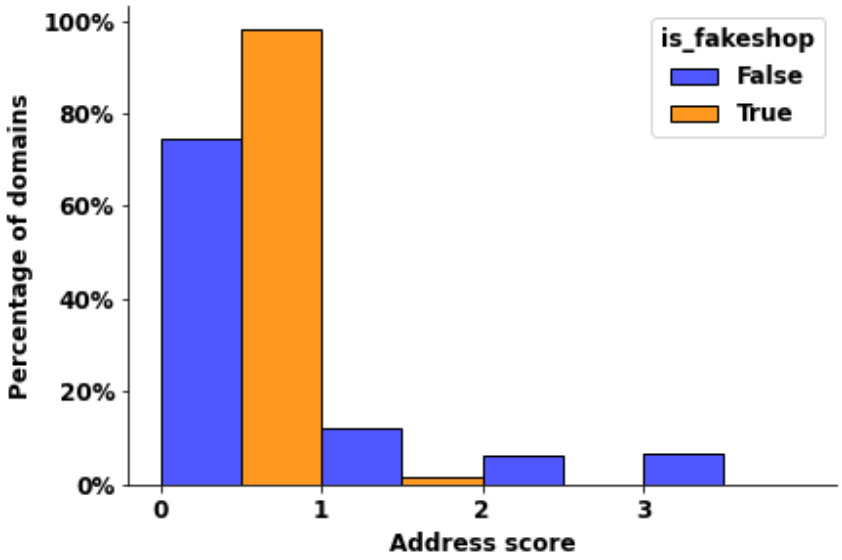

Figure 4.8: Fake webshops are very unlikely to display an address on their homepage.

4.2 Gathering of unlabeled data

Theoretically, we could learn from the entire .be zone, i.e., we could take every domain as an unlabeled example. However, we reasoned that the online ecosystem is very diverse and it may be hard to discriminate between fake webshops and any other type of domain. As such, we decided to discriminate first between webshops and non-webshops. While we could learn a classifier for this task, we believe a reasonable approximation is achieved by employing Wappalyzer's Technology Lookup API [57]. This API is useful to analyze the technology stack of websites and is used by DNS Belgium to inspect websites on .be domains.

Wappalyzer divides the individual technologies into categories, one of which is denoted as 'e-commerce'. Therefore, we select all domains in the .be zone with a technology belonging to the e-commerce category as our unlabeled dataset. This means that fake webshops could go unnoticed if they do not use an e-commerce technology recognized by Wappalyzer. However, we reason that fake webshops need to receive payments in some way to be profitable. Although they could write custom shopping cart software, this seems unlikely as it is very time-consuming and there exist free alternatives, like the WooCommerce plugin for WordPress. We therefore expect that the majority of currently active fake webshops will be contained in the unlabeled dataset.

DNS Belgium's database contains 110,995 records of domains that used an e-commerce technology at the moment they were crawled, which corresponds to about 10% of the .be domain names. We merged these records with the registration records and made sure the crawling timestamp lies between the start and end date of the

corresponding registration record.

We do not explicitly include the benign webshops of our labeled dataset in our unlabeled dataset. However, the ones that are still active are very likely to be included anyway, as they probably use some kind of e-commerce technology. On the other hand, the benign non-webshops are excluded.

4.3 Clustering for cross-validation

As discussed in the previous chapter, we suspect that the labeled set of fake webshops contains clusters of similar webshops and that DNS Belgium's classifier may remember these specific clusters instead of generalizing. In an attempt to reduce the risk of overfitting on these clusters, we now first cluster the fake webshops. During cross-validation, we then ensure that all domains belonging to the same cluster belong to the same fold. This way, we always train on a set of clusters and validate the performance of the model on another set of clusters. This method should provide a better estimate of the generalization capabilities than when the fake webshop clusters were dispersed over the training and validation folds.

Before clustering the fraudulent shops, we preprocess the features such that all feature distributions span the $[0, 1]$ range. Numerical features which are inherently bounded (i.e., the address score and the number of different currency names, symbols, ISO codes, and units) are transformed with scikit-learn's MinMaxScaler [48]. Other numerical features, for which outliers are possible, are processed with a QuantileTransformer [50], which transforms the feature distribution to a uniform distribution in the desired range. Boolean features and ratios are left unchanged. We ignore TF-IDF features for clustering since they would make the feature space very high-dimensional and render distances in that feature space meaningless. Furthermore, we removed the features containing the distances from the HTML title to the domain name, since many fake webshops are re-registered. Otherwise, two shops from the same cluster could look exactly the same, but be considered different because the original domain owners chose very different domain names. We also excluded the registration hour, since registrations at 11PM and 1AM are similar, yet would be considered very different.

To further reduce the negative impact of a high dimensional feature space on the feature performance, we compute the singular value decomposition and truncate it. We keep only the 15 largest singular values, which account for just over 90% of the variance in the data. Next, we perform an agglomerative clustering using the Ward criterion, which minimizes the variance of the clusters being merged [58]. We then selected the optimal number of clusters based on the silhouette coefficient [45]. From Figure 4.9, we found an optimal number of 152 clusters.

To verify the meaningfulness of the obtained clustering, we manually examined a

couple of clusters. The most convincing evidence of a good clustering was provided by cluster 12, which contains domains such as `filabelgie.be`, `conversewinkel.be` and `filagoedkoop.be`. All domain names contained brand names of shoes and furthermore, they all used the same type of Facebook deep links. Many other clusters contain domains that are registered in a relatively short time frame. For example, cluster 33 contains 45 domains registered in less than two months. Typically, multiple domains are registered over the course of a few minutes. Since we did not use the registration start date, end date, or registration hour for this clustering, we believe this is not a coincidence. However, the cluster of domains selling shoes did not exhibit this pattern, which indicates that not every cluster of fake webshops follows the same strategy.

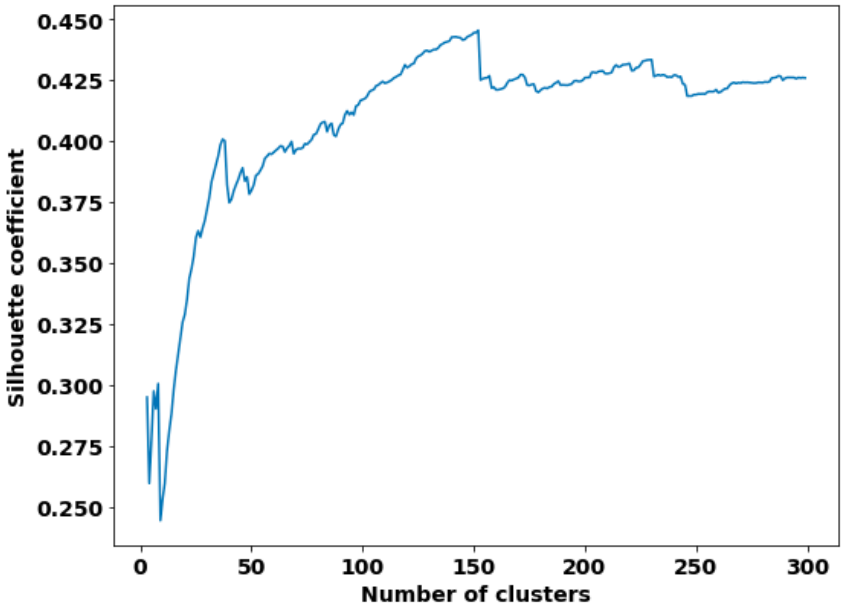

Figure 4.9: Silhouette score of the clustering as function of the number of clusters. The optimal number of clusters is 152.

During training, we would like to perform 10-fold cross-validation, but at this point, we have 152 clusters. Therefore, we require a method to group these clusters into ten folds. In the previous chapter, we hypothesized that the strategy of fake webshop operators might be changing over time. To test this, we plotted a timeline of the registration start dates for each cluster in Figure 4.10. Since the majority of the fake shops were registered in 2018, we can not detect a clear change in the observed clusters over time. Therefore, we believe it is not necessary to define the folds based on the registration dates of domains in the clusters. Instead, we use a bin-packing algorithm [35] to divide the clusters in 10 folds of roughly equal size. To enable cross-validation for the PU learning methods, we randomly assigned the unlabeled data to the ten folds.

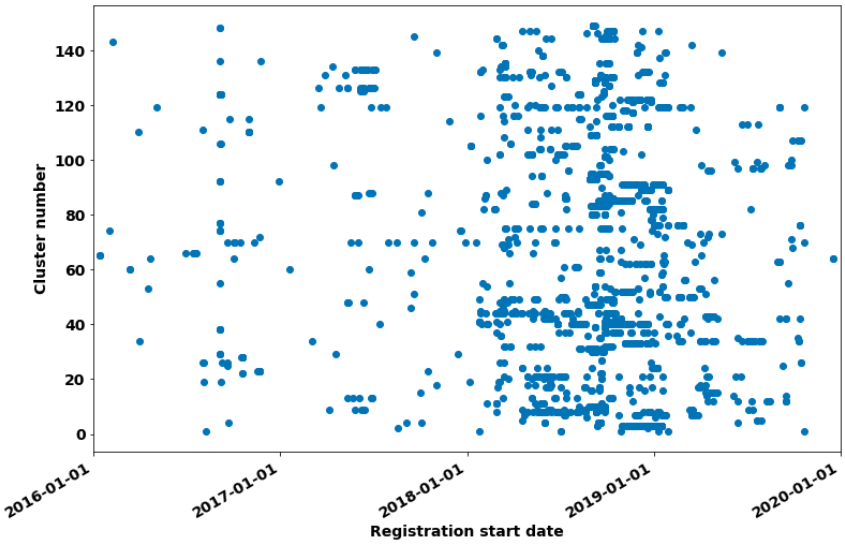

Figure 4.10: In general, the registration dates of different domains in the same cluster are spread out over time.

To examine the impact of this clustering on the generalization performance, we randomly divided the benign domains of the labeled dataset into 10 folds and then retrained DNS Belgium's classifier. We set apart three folds (i.e., 30%) for testing and kept seven folds (i.e., 70%) for training. With optimized parameter settings (tuned using 7-fold cross-validation), the recall on the test set dropped to 75% and 76% for the Random Forest Classifier and Gradient Boosting Classifier respectively, while the precision increased to 100% and 99%. This result confirms that changing fake webshop tactics could pose significant problems for DNS Belgium's current classifier.

4.4 Classifiers

Since PU learning has not been applied to fake webshop detection before, we were unsure about which PU classification methods would perform best. According to Bekker and Davis [4], two-step techniques should be preferred when the separability assumption holds and the class distributions lie not too close to each other. When the SCAR assumption holds or when the class distributions almost overlap, they recommend using biased learning or class prior incorporation methods. Given that we operate in an adversarial setting, where fake webshops try to resemble regular webshops as much as possible, it seems implausible that the classes will be easily separable. We therefore focused our efforts on biased learning and class prior incorporation methods. More specifically, we tried learning with Empirical Risk Minimization (ERM) methods and Robust Ensemble SVMs, as these techniques seemed most promising. We applied the same preprocessing steps as for clustering,

but also included the TF-IDF features and domain-title distances again.

Our dataset intuitively corresponds to the case-control scenario of PU learning, with a set of positives collected over time and an unlabeled dataset sampled from the distribution of webshops in the .be zone. However, the unlabeled dataset is not really a sample and the combination of both datasets corresponds roughly to the total distribution of webshops (with the caveat that the positives are no longer active). Therefore, we argue that we can also view this as a single training set scenario. In the sections that follow, we will assume a single training set scenario, as most developed PU learning techniques implicitly or explicitly assume such a scenario [4].

4.4.1 Empirical Risk Minimization [4]

Recall that ERM methods aim to reweight the data such that the expected empirical risk of the weighted dataset equals that of a fully labeled dataset. More formally, we define the risk of a classifier g as

$$R(g) = \alpha\, \mathbb{E}_{f+}\left[L^+(g(x))\right] + (1-\alpha)\, \mathbb{E}_{f-}\left[L^-(g(x))\right], \tag{4.1}$$

where $\mathbb{E}$ denotes the expectation operator and L^+ and L^- are the loss functions for positive and negative examples respectively. The corresponding empirical loss in a supervised setting is then equal to

$$\hat{R}(g \mid \mathbf{x}, \mathbf{y}) = \alpha \frac{1}{|\mathbf{y}=\mathbf{1}|} \sum_{x:\mathbf{x}|\mathbf{y}=\mathbf{1}} L^+(g(x)) + (1-\alpha)\frac{1}{|\mathbf{y}=\mathbf{0}|} \sum_{x:\mathbf{x}|\mathbf{y}=\mathbf{0}} L^-(g(x)). \tag{4.2}$$

In the PU learning setting, we cannot observe the labels $\mathbf{y}$ though, and we have to rewrite the risk in terms of the labeled and unlabeled data. In the single training set scenario, one can show that it is possible to rewrite the risk as

$$R(g) = \alpha c\, \mathbb{E}_{f_l}\left[\frac{1}{e(x)} L^+(g(x)) + \left(1 - \frac{1}{e(x)}\right) L^-(g(x))\right] + (1-\alpha c)\, \mathbb{E}_{f_u}\left[L^-(g(x))\right]. \tag{4.3}$$

The empirical risk then reduces to

$$\hat{R}(g \mid \mathbf{x}, \mathbf{y}) = \frac{\alpha c}{|\mathbf{s}=\mathbf{1}|} \sum_{x:\mathbf{x}|\mathbf{s}=\mathbf{1}} \left(\frac{1}{e(x)} L^+(g(x)) + \left(1 - \frac{1}{e(x)}\right) L^-(g(x))\right) + \frac{1-\alpha c}{|\mathbf{s}=\mathbf{0}|} \sum_{x:\mathbf{x}|\mathbf{s}=\mathbf{0}} \left(L^-(g(x))\right) \tag{4.4}$$

$$= \frac{1}{|\mathbf{s}|}\left(\sum_{x:\mathbf{x}|\mathbf{s}=\mathbf{1}} \left(\frac{1}{e(x)} L^+(g(x)) + \left(1 - \frac{1}{e(x)}\right) L^-(g(x))\right) + \sum_{x:\mathbf{x}|\mathbf{s}=\mathbf{0}} \left(L^-(g(x))\right)\right), \tag{4.5}$$

where we applied the equalities $|\mathbf{s}=\mathbf{1}| = \alpha c|\mathbf{s}|$ and $|\mathbf{s}=\mathbf{0}| = (1-\alpha c)|\mathbf{s}|$. From this result, it follows that we should replicate all unlabeled examples as negatives, while we should replicate the positives once as a positive with weight $\frac{1}{e(x)}$ and once as a

negative with weight $(1 - \frac{1}{e(x)})$. This reweighted dataset is expected to yield the same empirical risk as a fully labeled dataset.

Under the SCAR assumption, the propensity scores are constant and we have that $e(x) = c$. The labeling frequency (or equivalently, the class prior α) is unknown though, and we have to estimate it from our data. Several methods have been proposed in the literature, but Bekker and Davis [3] found that the kernel embedding method KM2 [42] and their own decision tree induction approach TiCe [3] performed best on a small benchmark. Since the estimation of the class prior is not the main subject of this thesis, we refer to their respective papers for more details.

Under the SAR assumption, we require a method to estimate the propensity scores. Since examples can be unlabeled either because the class probability is low or because the propensity score is low, additional assumptions are necessary to enable learning. Bekker et al. [5] assume that only a subset of the attributes, called the propensity attributes x_e, influences the propensity score, i.e., $e(x) = e(x_e)$. This assumption is sufficient to enable learning in the SAR setting. They propose an Expectation-Maximization algorithm to simultaneously train the classification model and the propensity score model. After convergence, they retrain the classifier using Equation 4.5 with the obtained propensity score.

Bekker and Davis [5] published the code accompanying their paper, which compares their SAR learning method with, among others, ERM under the SCAR assumption using class prior estimation by KM2 and TiCe. They used logistic regression for each approach, such that the loss functions L^+ and L^- correspond to log losses. We extended their code to allow using our custom cross-validation method. Furthermore, we combined the preprocessing and training steps in a *pipeline* from scikit-learn [49]. The only parameter we tuned for SCAR is (the inverse of) the regularization strength C. For the SAR approach, we considered different regularization strengths for the classifier (C_{class}) and the propensity score model (C_{prop}). The propensity attributes for the SAR method consist of the AS and country code of the hosting provider and the TFIDF features. The reasoning behind this choice is that DNS Belgium tracked down many fake webshops based on the AS and country code of their hosting provider, so these attributes definitely influence the probability of a fake webshop being labeled. Furthermore, the TFIDF features reflect the type of goods sold on the website and therefore also implicitly a price category. As we discussed earlier, we suspect that customers may be more prone to report a website when their financial loss is more substantial. By including the TFIDF features in the propensity attributes, we hope to account for this possible bias.

The AS and country code were originally excluded from the data because they were so heavily used for labeling, but for the SAR approach, we include them again. Among the fake webshops, there were 66 missing AS values and 19 missing country codes. Since we wanted to keep all available fake shops for training, we imputed these values by sampling from the distribution of these attributes over the fake webshops.

There were also a few missing values among the unlabeled domains, but this amount was negligible compared to the number of unlabeled domains, so we dropped the corresponding records.

4.4.2 Robust Ensemble SVM

The most advanced biased learning algorithm we discussed in Chapter 2 is the Robust Ensemble SVM method of Claesen et al. [12], which has five parameters to tune. First, the number of base models n_{models} influences the stability of the ensemble, but more stability comes at the cost of increased computation time. Next, we have the number of positive examples n_{pos} and unlabeled examples n_{unl} sampled for each base model. Small numbers result in high variability in the base models and therefore require more base models, while larger numbers yield more stable base models and decrease the number of required base models. Finally, we should tune the misclassification penalty for unlabeled examples C_U and the weighting parameter w_{pos}, which influences the misclassification penalty for positive examples as follows:

$$C_P = C_u \times w_{pos} \times \frac{n_{unl}}{n_{pos}}. \tag{4.6}$$

While it is possible to change the kernel function as well, this would introduce even more parameters, so we decided to use only a linear kernel.

The RESVM classifier does not provide probabilities of an example being positive, but reports the fraction $v(x)$ of base models that classify the example as positive. If the vote is unanimous, we sum the decision values of the individual base models to allow ranking of the domains. Denoting the decision value of base model i for test domain x with $\psi^{(i)}(x)$, the decision value of the RESVM classifier can be calculated as

$$d(x) = \begin{cases} v(x) & \text{if } 0 \leq v(x) < 1 \\ \sum_{i=1}^{n_{models}} \psi^{(i)}(x) & \text{if v(x) = 0} \\ 1 + \sum_{i=1}^{n_{models}} \psi^{(i)}(x) & \text{if v(x)=1} \end{cases}$$

We based our implementation on the code accompanying the paper on RESVM and adapted it to allow using our custom cross-validation method. Furthermore, their code expects preprocessed data and does not support the use of scikit-learn's pipeline. We should not preprocess the data beforehand though, as this would introduce data leakage into the model. More specifically, this would imply that we incorporated knowledge about the validation folds during the training procedure. To remedy this, we make sure to preprocess the data using only the domains available in the training folds and then transform the validation fold accordingly.

Chapter 5

Experimental evaluation

This chapter covers the experimental evaluation of the classifiers described in Section 4.4. The first section describes their performance as measured by our custom 10-fold cross-validation method. It also elaborates on an issue we encountered for the ERM methods. The second section defines the procedure we followed to evaluate the classifiers empirically. The third section describes the empirical results, while the fourth and final section compares our best classifier with DNS Belgium's Random Forest Classifier.

5.1 Classifier performance

We started our evaluation by tuning the parameters for each PU learning method using our custom 10-fold cross-validation. For the methods relying on the SCAR assumption, we did so based on the F_1' score. Soon, we faced the issue of very small scores for the ERM methods (ERM-SCAR-KM2 and ERM-SCAR-TiCe in Table 5.1). They achieved very high recall, but only because they predicted that more than half of all the unlabeled domains were fake webshops. However, DNS Belgium believes they have already detected at least one out of two fake webshops in the .be zone. Since ERM methods should return unbiased estimates of the probabilities of webshops being fake [4], this result was unexpected. We were able to pinpoint the issue to the estimation of the labeling frequency by KM2 and TiCe. Both methods predict labeling frequencies around 2%, which is equivalent to estimating that 49 out of 50 webshops are undetected yet. We varied the attributes used for estimation and experimented with parameter settings for both algorithms, but the impact on the estimated label frequency was only minimal. Therefore, the most logical conclusion is that the assumptions underlying KM2 and TiCe do not hold for our dataset. Bekker et al. [5] state that all class prior estimation methods to date, including KM2 and TiCe, attribute the absence of class labels rather to a low labeling frequency than to a low class prior. Since we operate in a setting with a low class prior, we expect this is what causes the poor performance.

Since class prior estimation by KM2 and TiCe yields poorly performing classifiers,

Table 5.1: Performance of tuned PU learning methods relying on the SCAR assumption, as estimated with 10-fold crossvalidation on the training set.

Performance metric	ERM-SCAR-KM2	ERM-SCAR-TiCe	ERM-SCAR-C	RESVM
F_1'	1.596	1.215	37.219	43.365
Recall	0.999	1	0.879	0.9749

we decided to pick a value for c based on background knowledge. More specifically, we conservatively estimate a labeling frequency of 50%, which results in far better F_1' scores (EMR-SCAR-C in Table 5.1). RESVM yields even better performance though.

To the best of our knowledge, no evaluation metrics have been proposed to tune methods based on the SAR assumption. Therefore, we tried out a grid of parameters and computed the F_1' score and recall to gain an impression of the quality of the model. We then inspected the probability estimates to gain insight into overfitting (i.e., probabilities close to 0 and 1) and underfitting issues (i.e., nearly constant probabilities). Among the models that achieved high F_1' scores, i.e., similar to ERM-SCAR-C and RESVM, we selected the one that showed the most gradual decrease from high to low probabilities.

5.2 Empirical evaluation procedure

Since the unlabeled dataset contains more than 100,000 domains and manual verification of domains is a time-intensive process, it is impossible to obtain the correct labels for all domains. Instead, we first split the unlabeled data into roughly 20% used for testing and 80% used for training. Following the reasoning that fake webshop tactics may change over time, we pick the 20,000 domains that were registered most recently as test set, while we use the older domains for training.

However, 20,000 domains are still too much to evaluate manually. Therefore, we decided to only verify the 500 domains with the highest score for each (tuned) classifier. This comes down to 2.5% of the test data, which is a little more than the percentage of fake webshops in the training set (2.15%). During the manual verification, we mainly looked for the following clues:

- Presence of a VAT number, either on a contact page or in the Terms&Conditions. We verified the correctness of Belgian VAT numbers via `btw-opzoeken.be`, while we used `kvk.nl` for companies from the Netherlands. These websites mention the supposed main activity of the company, so we made sure these were in accordance with the activity of the domain;
- Presence of social media accounts and the number of likes/followers. When in doubt, we also checked whether the social media accounts were actively used;

- Presence of Terms&Conditions, return and shipping policies, privacy policy... We made sure these documents were not generic documents (without information about the company) or copies of documents from other companies;
- Presence of a contact page with an address, phone number, and/or email address;
- Online reviews;
- Peculiarities in the registration history.

5.3 Empirical results

Often it was hard to differentiate between a legitimate, but poorly constructed webshop and a truly fake webshop. Young companies that are just starting out often lack good policies, a VAT number, and/or online reviews, but are strongly present on social media. On the other hand, some older companies have websites, but are not present on social media. It seems they also forget to extend their registration sometimes, leading to peculiar gaps in their registration timeline. Only in rare cases we can state with certainty that a certain webshop is fake. Therefore, the counts of fake webshops in the remainder of this section refer to webshops we have serious doubts about. Further investigation by DNS Belgium will tell whether these are truly fake webshops or not.

During the evaluation, we did not only encounter fake webshops and normal webshops. Instead we regularly came across domains that we can divide into the following categories:

- Domains for which the webshop seems to be under construction. First, this category includes domains that display a page from e.g., Shopify stating there is one step left to finish setting up the webshop. Second, it contains domains that were registered relatively recently, have a non-operational shop and contain mostly *Lorem ipsum* text;
- Domains that are currently offline;
- Domains that are offered for sale, e.g., on `dan.com` or `godaddy.com`;
- Domains for which the connection times out and we receive a 502 Bad Gateway error;
- Domains for which our browser displays a warning message that the connection is not private and attackers might try to steal our information;
- Domains that redirect to porn websites;

- Domains re-registered by Chinese or Eastern European companies in the business of heavy machinery. This type of website is known by DNS Belgium and they are unsure about its purpose. Since they do not sell any goods online, they are surely not fake webshops though;
- Domains that redirect to the domain `thinkeos.com`, which claims to be a Finance and Business Intelligence company. For some unknown reason, unrelated domains like `pandemic.be`, `relatiezoeken.be` and `wavrecapital.be` redirect to it. The contact information on their website can also be found on `solarpanels.be`, `mijndrukker.be` and others. While the website has something suspicious, we cannot classify it as a fake webshop as it does not sell anything;
- Domains that redirect to the domain `hardwarecity.nl`, which is a Dutch webshop selling laptop-related goods. They registered domain names like `laptopbatteries.be` and `laptop-keyboard.be`. The company is legitimate, but since we encountered so many domains redirecting to its domain, we classify them in a separate category.

For the domains that were offline, for sale, or under construction, we wondered whether they were any different at the time of crawling. As DNS Belgium takes screenshots during their crawls, we were able to verify this. For each of these three categories, we took a small, random sample of ten domains and compared the homepage at crawl time with the current homepage. As it turns out, none of the domains that are currently offline were offline at crawl time. Instead, we found another Chinese heavy machinery domain and several domains in foreign languages. At first sight, they did not resemble fake webshops, but they were definitely not normal webshops. We assume that most of these domains were taken down by DNS Belgium. Nine out of ten domains that are currently on sale were also on sale at crawl time, while the tenth was just a weird domain. Eight out of ten domains under construction were the same at crawl time, while the other two were likely fake webshops.

We now consider the results obtained by the different PU learning techniques. It turns out that the ERM and RESVM methods perform quite differently in terms of the categories they detect (Table 5.2). First and foremost, RESVM yields the most suspected fake webshops. The ERM-based methods discover more webshops under construction and offline domains, which could have contained fake webshops at the moment of crawling. However, based on our previous analysis, we do not expect this to change the number of fake webshops significantly. Apart from suspected fake shops, RESVM mainly detects webshops and picks up only a limited amount of unrelated categories. On the other hand, both ERM techniques seem to give high scores to side categories like heavy machinery websites. Furthermore, during evaluation, it felt like the normal webshops receiving high scores from RESVM were more difficult to discriminate from fake webshops than normal webshops receiving high scores from the ERM-based methods. Based on these observations, the RESVM

Table 5.2: Distribution of 500 highest scoring domains over the identified categories. ERM-based methods detect a lot of side-categories, while RESVM mainly detects true webshops. Numbers in boldface indicate which PU learning method detected the most domains in each category.

Category	ERM-SCAR-C	ERM-SAR	RESVM
Suspected fake webshop	38	25	**58**
Normal webshop	180	150	**334**
Webshop under construction	34	**109**	53
Domain offline	**51**	50	33
Domain for sale	**70**	60	1
Connection time-out	**8**	**8**	0
Connection not private	7	**23**	6
Redirect to porn	**20**	19	11
Heavy machinery	10	**11**	1
`thinkeos.com`	**32**	**32**	0
`hardwarecity.nl`	**50**	13	0

classifier is the method of our choice.

Table 5.2 also suggests that using the SAR assumption does not improve the performance of ERM compared to using the SCAR assumption. On the contrary, the number of suspected fake webshops and normal webshops even decreases. However, all fake shops detected by either method occur in the top 10% highest-scoring domains for both methods, and most even in the top 5%. We therefore believe that using the SAR assumption does not necessarily degrade performance, although it does certainly not improve it.

It is interesting to look at the overlap between the suspected fake webshops detected by different classifiers (Figure 5.1). There is a large overlap between the suspected fake webshops detected by the ERM-based methods, while they show little overlap with the suspected domains detected by RESVM. Furthermore, ERM-based methods often give low scores to suspected fake shops detected by RESVM and vice versa. This could imply that neither of these methods is able to identify all fake webshops. However, as mentioned before, we can currently not state with certainty which webshops are fake and which are not. DNS Belgium will now examine the contact information provided by the suspicious domains and suspend them if this information is invalid. If the provided contact information seems correct, but they still have serious doubts about the legitimacy of the webshops, they may contact the FPS Economy to undertake legal action. Only after these evaluations will it be possible to draw definite conclusions about the effectiveness of the considered PU

learning techniques. Nonetheless, the classifier performance from Section 5.1 and our own empirical evaluation suggest that RESVM is the preferred classifier. We should note, however, that the comparison is not entirely fair, as the voting system in RESVM yields a non-linear model, while the ERM methods are based on linear models.

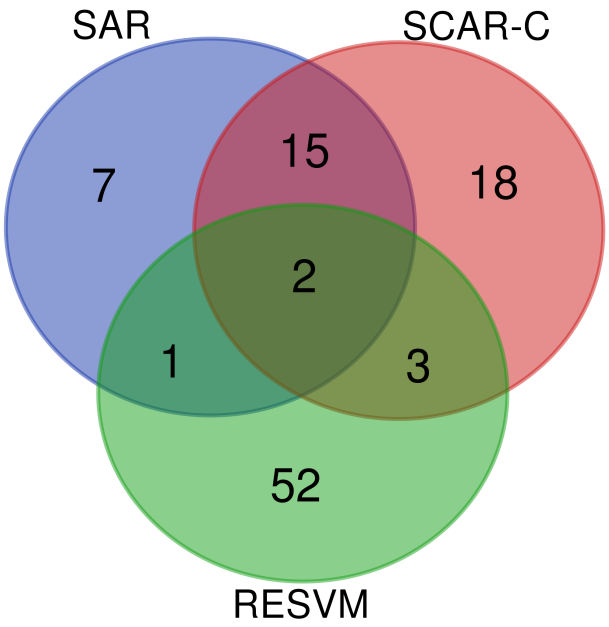

Figure 5.1: ERM-based methods show a large overlap in the fake webshops they detect, while there is little overlap with RESVM. Combined, the PU learning methods detected 98 potential fake webshops.

5.4 Comparison of PU learning with supervised learning

The goal of this thesis was to assess the applicability of PU learning to the detection of fake webshops. Therefore, we now compare the strengths and weaknesses of DNS Belgium's Random Forest Classifier and the best PU classifier, i.e., the RESVM classifier.

First, the PU classifier has the advantage of not requiring negative labels. On the one hand, these are already acquired, so this is not a real problem. On the other hand, if we want to extend the supervised classifier with newly discovered webshops, new negative labels should be gathered to keep the training set balanced. For the PU classifier, new fake webshops can be added to the training data without an additional cost. Furthermore, we can regularly update the unlabeled dataset with the new webshops in the .be zone.

Second, it appears that both classifiers achieve a similar precision. DNS Belgium provided us with an overview of the number of domains that were marked suspicious by their classifier and the number of domains that were efficiently revoked. They report having taken down 260 domains out of 2005 suspicious domains, which is equivalent to a precision of 12.97%. These numbers are not exact, as they also include some heavy machinery websites and potentially some other types of malicious domains. Nevertheless, this score is a reasonable estimate of the true precision. If we assume that the 58 suspicious shops detected by RESVM are all fake webshops, the precision of this classifier would be 11.60%, which is comparable to the Random Forest Classifier's precision. The true precision may be somewhat smaller if not all suspicious domains turn out to be truly fake webshops, or somewhat larger if some of the offline domains used to be fake webshops. We conclude that the PU learning classifier neither performs significantly better nor performs significantly worse than the supervised classifier. However, it is important to note that the precision of the Random Forest Classifier is decreasing over time.

Third, we tried to get an estimate of how both classifiers perform in terms of recall. If the 58 webshops detected by RESVM are truly fake, they were probably missed by the Random Forest Classifier. However, we should also consider it the other way around and assess whether RESVM can detect the fake webshops detected by the Random Forest Classifier. To do so, DNS Belgium provided us with a set of recent predictions of their classifier, which were manually verified. There were 51 suspicious domains, of which some were revoked and some are still online. Since DNS Belgium only has the authority to suspend domains based on the contact information and not based on content, the fact that a domain is still online does not imply it is not a fake webshop. We analyzed the prediction scores given to these domains by the RESVM classifier and came to the following conclusions. First, 20 domains would probably have been detected, as they resided in the top 2.5% of predictions in either the test or unlabeled training set. Second, 14 domains received low prediction scores, meaning they would have gone unnoticed by the PU classifier. Seven out of these 14 domains belonged to the test set, and we found that only one of these resided in the top 500 of one of the other PU learning techniques. Third, 17 domains belonged neither to the training nor to the test set. The most likely explanation is that those webshops did not use an e-commerce technology recognized by Wappalyzer. This seems the major drawback of our proposed approach, as the classifier will never be able to detect these domains.

Chapter 6

Conclusion

Our contribution. This thesis started out with an extensive analysis of the features of DNS Belgium's supervised classifier, which led to several insights. First, we found that fake webshops were often re-registered using a drop-catch mechanism. Second, it is very unlikely to encounter telephone or email address links on the homepage of a fake webshop. Third, the distributions of the number of images, internal links and tags suggest the labeled dataset contains clusters of similar fake webshops.

Next, we examined the performance of the Random Forest Classifier implemented by DNS Belgium. While the classifier achieves both high recall and high precision, DNS Belgium reports that their classifier does no longer detect a lot of fake webshops, but mainly generates false positives. We came up with three potential explanations. First, it could be that fake webshop operators stopped using .be domain names, as they know they are being hunted. Second, the labeled set of benign domains could be unrepresentative of the entire benign .be zone. Third, fake webshop operators could be changing tactics, which may pose problems if the classifier does not generalize well.

Prior to addressing these issues, we came up with additional features inspired by recent research in the domain of fake webshop detection. We incorporated trust scores for the registrar and the registrant's email provider, and looked at the ratio of lowercase characters in the registrant's name. By capturing the discrepancy between the domain name and the HTML title, we pick up on re-registered and hacked domain names that now operate fake webshops. We also integrate features that capture the presence of currencies in order to detect fake webshops displaying products with high discounts or in multiple currencies. Finally, we check on the merchant's identity by investigating the number of (deep) links to social media and by comparing the address provided during registration with the address displayed on the website.

We then attempted to improve the existing classifier by switching from a supervised approach to a PU learning approach. This allowed us to incorporate domains with unknown labels, thus mitigating the potential issue of an underrepresented benign set, as we could now learn from every domain in the .be zone. However,

we reasoned that it would probably be easier to discriminate fake webshops from legitimate webshops than from any kind of benign domain. Therefore, we first used Wappalyzer's Technology Lookup API to identify all websites using e-commerce technology and designated these as our unlabeled dataset.

Next, we addressed the potential problem of overfitting on the clusters. We first clustered the fake webshops and found convincing evidence of a meaningful clustering. Then, we proposed a custom cross-validation method to get a better estimate of the classifier's capabilities to generalize to new clusters. As it turned out, the classifier's recall decreased significantly, indicating that changing tactics of fake webshops may not be detected by the current classifier.

We then trained two PU learning methods that assume an unbiased labeling mechanism and one that assumes a bias in this mechanism. We achieved the most promising results with the Robust Ensemble SVM method, which yielded 58 suspicious domains out of the 500 evaluated ones. Since these did not strongly resemble the fake webshops from the training data and since they were not detected by DNS Belgium's Random Forest Classifier, RESVM seems to display superior generalization capabilities. Furthermore, it achieves similar precision. The major drawback seems to be that some fake webshops are not included in the unlabeled data, because they do not use an e-commerce technology recognized by Wappalyzer.

Future work. To alleviate the issue of fake webshops not appearing in the unlabeled data, DNS Belgium could try to learn from the entire .be zone instead of only from the webshops. The potential drawback is that the prior probability of being a fake webshop would decrease even further, and we are unsure about the implications this may have. Alternatively, we could try to run predictions on the entire .be zone based on the current model, but this seems not recommended, as there would be a discrepancy between the training set distribution and the test set distribution.

We believe the greatest improvements could be achieved by incorporating more fake webshops that were detected in the past. Since the amount of labeled fake webshops is limited, we believe any additional example could improve the classifier. Especially if the fake webshop strategy is changing, it is advisable to always include the most recent detected fake webshops.

Appendix A

Appendix A: currency occurrences in HTML body

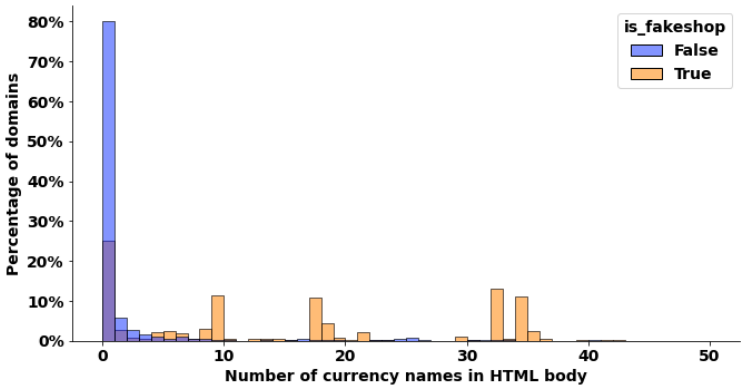

Figure A.1: Distribution of the occurrence of currency names in the HTML body.

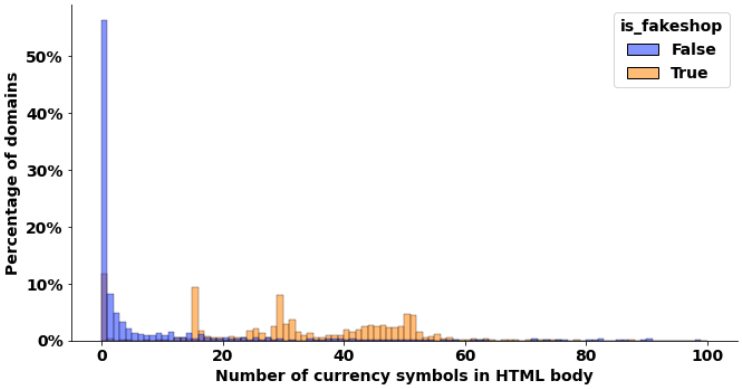

Figure A.2: Distribution of the occurrence of currency symbols in the HTML body.

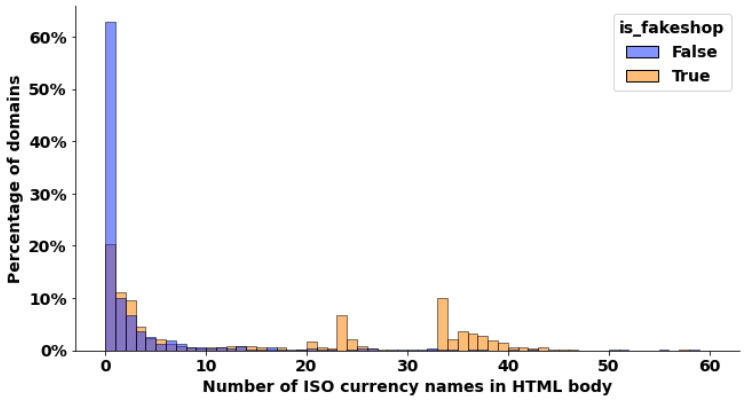

Figure A.3: Distribution of the occurrence of ISO codes in the HTML body.

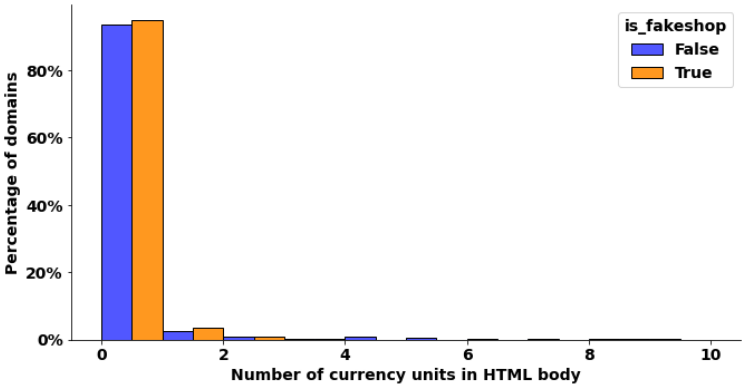

Figure A.4: Distribution of the occurrence of fractional currency units in the HTML body.

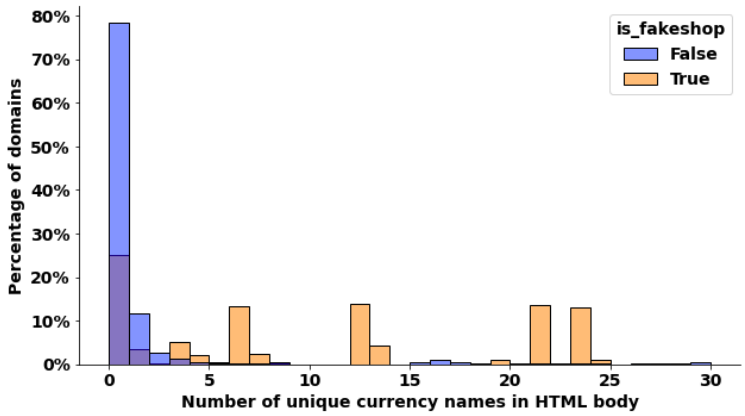

Figure A.5: Distribution of the number of different currency names in the HTML body.

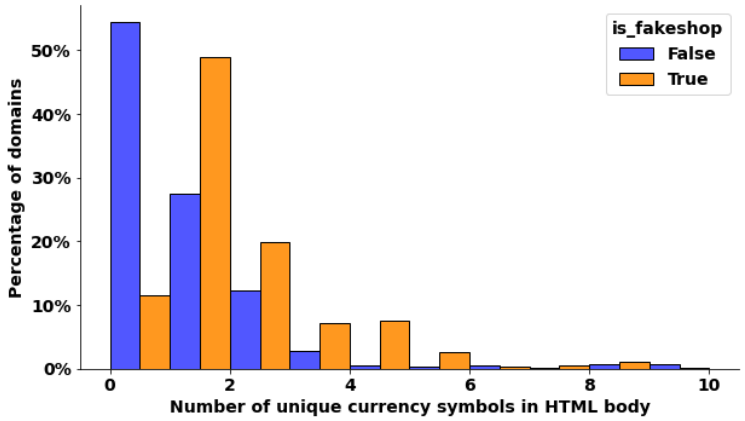

Figure A.6: Distribution of the number of different currency symbols in the HTML body.

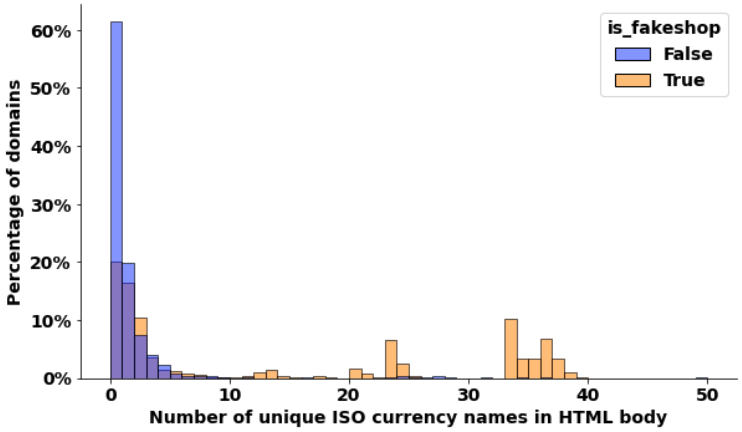

Figure A.7: Distribution of the number of different ISO codes in the HTML body.

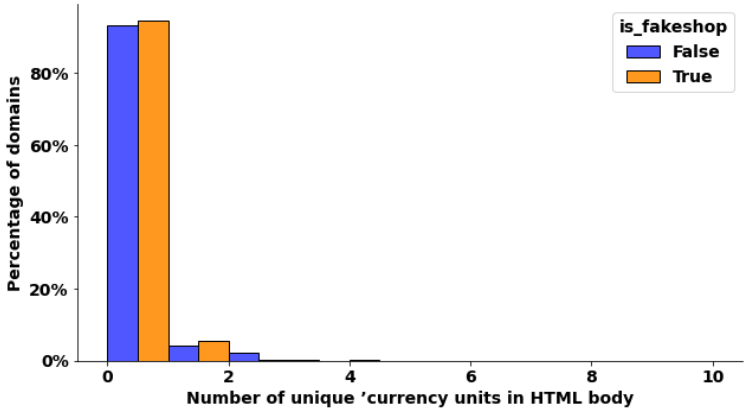

Figure A.8: Distribution of the number of different fractional currency units in the HTML body.

Appendix B

Appendix B: presence of (deep) links to social media

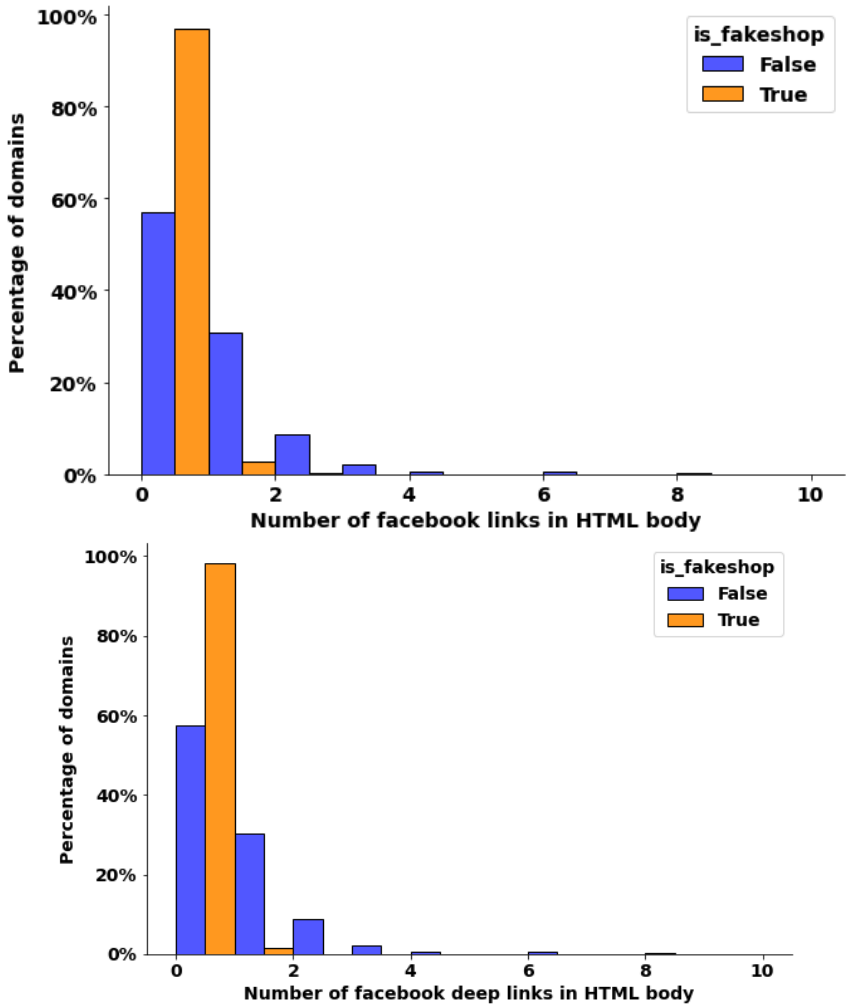

Figure B.1: Distribution of the number of links to Facebook (top) and the number of deep links to Facebook (bottom).

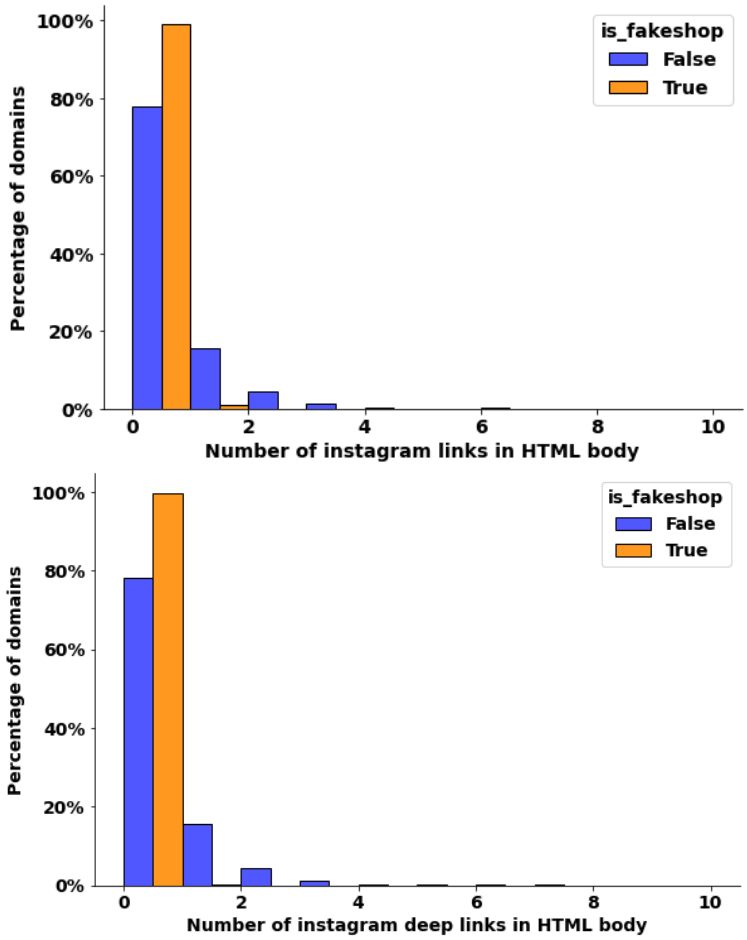

Figure B.2: Distribution of the number of links to Instagram (top) and the number of deep links to Instagram (bottom).

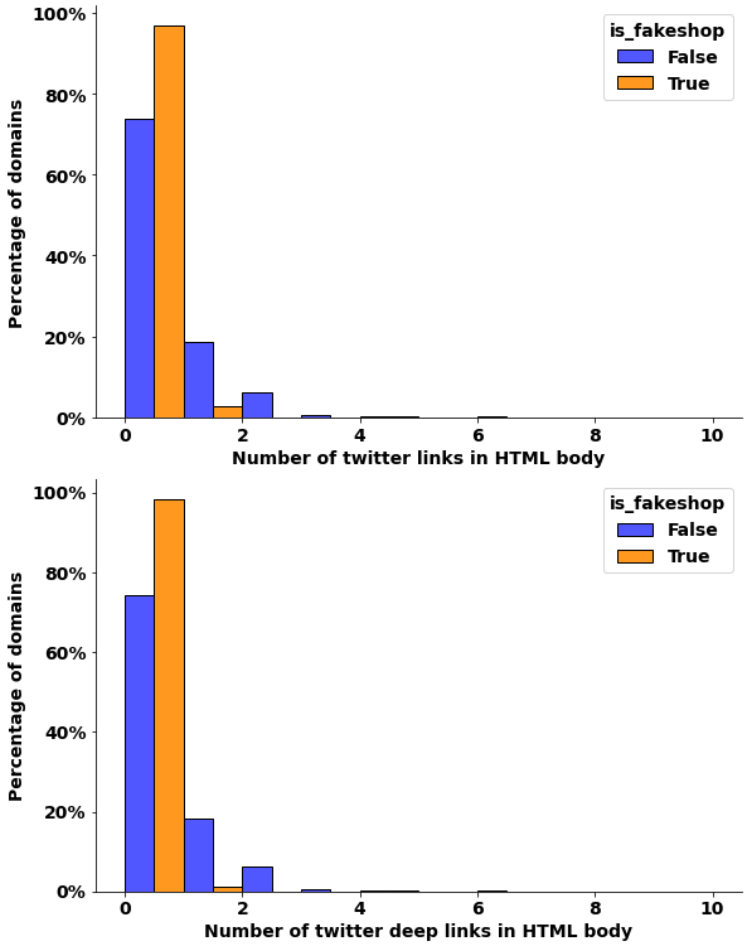

Figure B.3: Distribution of the number of links to Twitter (top) and the number of deep links to Twitter (bottom).

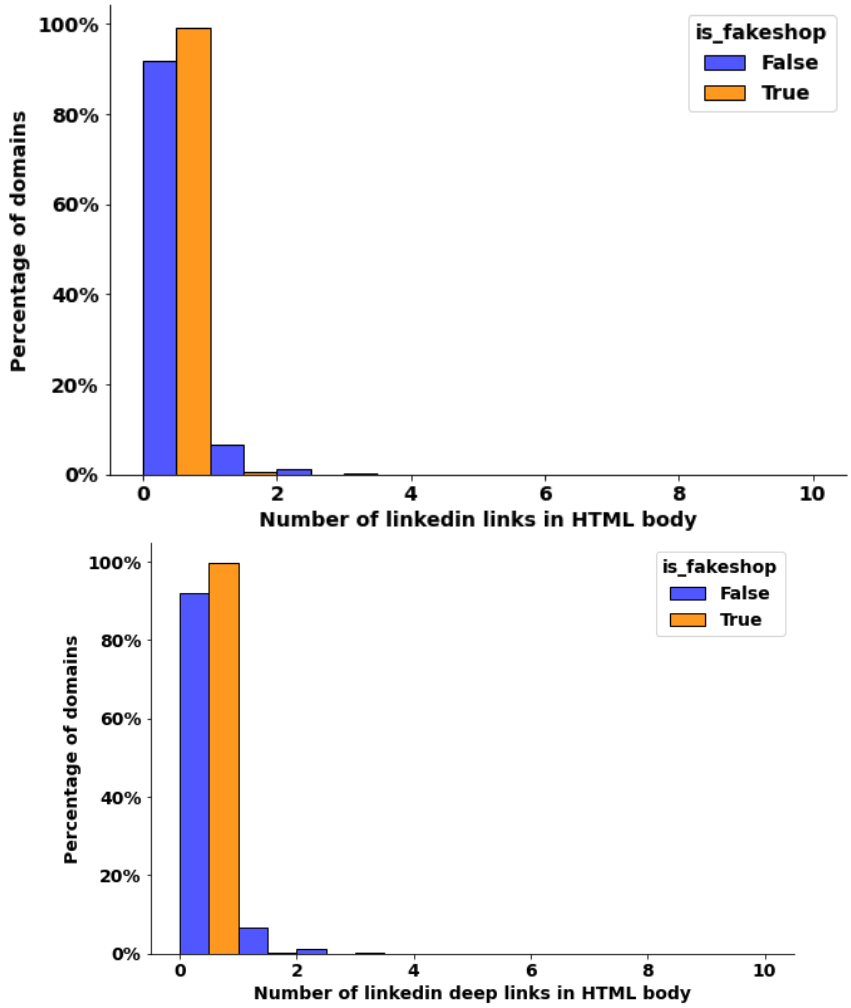

Figure B.4: Distribution of the number of links to LinkedIn (top) and the number of deep links to LinkedIn (bottom).

Bibliography

[1] S. N. Bannur, L. K. Saul, and S. Savage. Judging a site by its content: Learning the textual, structural, and visual features of malicious web pages. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, pages 1–10, New York, NY, USA, 2011. Association for Computing Machinery.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer Vision-ECCV 2006*, volume 3951, pages 404–417, Jul 2006.

[3] J. Bekker and J. Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, Feb. 2018.

[4] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. In *Machine Learning*, volume 109, pages 719–760, 2020.

[5] J. Bekker, P. Robberechts, and J. Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 71–85, Cham, 2020. Springer International Publishing.

[6] E. Boogert. Tien procent nederlandse webshops frauduleus, 2017. Accessed on 23-12-2020 via https://www.emerce.nl/nieuws/tien-procent-nederlandse-webshops-frauduleus.

[7] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[9] D. Canali, M. Cova, G. Vigna, and C. Kruegel. Prophiler: A fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 197–206, New York, NY, USA, 2011. Association for Computing Machinery.

[10] C. Carpineto and G. Romano. Learning to detect and measure fake ecommerce websites in search-engine results. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 403–410, New York, NY, USA, 2017. Association for Computing Machinery.

[11] Centraal Bureau voor de Statistiek. Cybersecuritymonitor 2017: Een eerste verkenning van dreigingen, incidenten en maatregelen, 2017. Accessed on 23-12-2020 via https://www.cbs.nl/nl-nl/publicatie/2017/06/cybersecuritymonitor-2017.

[12] M. Claesen, F. De Smet, J. Suykens, and B. De Moor. A robust ensemble approach to learn from positive and unlabeled data using svm base models. *Neurocomputing*, 160:73–84, 07 2015.

[13] Comeos VZW. E-commerce belgium 2019, 2019. Accessed on 23-12-2020 via https://static.comeos.be/E-commerce_Belgium_2019__2.pdf.

[14] Comeos VZW. E-commerce survey 2020, 2020. Accessed on 23-12-2020 via https://static.comeos.be/E-Commerce_Belgium_10.2020.pdf.

[15] M. Cox and S. Haanen. Content-based classification of fraudulent webshops. Master's thesis, University of Amsterdam, 2018. Accessed on 21-12-2020 via https://delaat.net/rp/2017-2018/p30/report.pdf.

[16] L. Daigle. Whois protocol specification. RFC 3912, RFC Editor, Sep 2004.

[17] DNS Belgium. 5733 domain names taken offline in 2019, 2019. Accessed on 23-12-2020 via https://www.dnsbelgium.be/en/news/domainnames-offline.

[18] DNS Belgium. About dns belgium: who we are and what we stand for, 2021. Accessed on 17-02-2021 via https://www.dnsbelgium.be/en/about-dns-belgium.

[19] M. C. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1386–1394. JMLR.org, 2015.

[20] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[21] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. Association for Computing Machinery.

[22] Europol. How to detect fraudulent sites selling fakes. Accessed on 23-12-2020 via https://www.europol.europa.eu/activities-services/public-awareness-and-prevention-guides/how-to-detect-fraudulent-sites-selling-fakes.

[23] Europol. Biggest hit against online piracy: over 20520 internet domain names seized for selling counterfeits, Sep 2017. Accessed on 23-12-2020 via https://www.europol.europa.eu/newsroom/news/biggest-hit-against-online-piracy-over-20-520-internet-domain-names-seized-for-sellin

[24] I. S. R. Group. A nonprofit certificate authority providing tls certificates to 260 million websites, 2021. Accessed on 03-06-2021 via https://letsencrypt.org/.

[25] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck. Understanding the domain registration behavior of spammers. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 63–76, New York, NY, USA, 2013. Association for Computing Machinery.

[26] F. He, T. Liu, G. I. Webb, and D. Tao. Instance-dependent PU learning by bayesian optimal relabeling. *CoRR*, abs/1808.02180, 2018.

[27] D. Hernandez, R. Guzman-Cabrera, M. Montes, and P. Rosso. Using pu-learning to detect deceptive opinion spam. *WASSA 2013*, page 38, 01 2013.

[28] K. Jaskie and A. Spanias. Positive and unlabeled learning algorithms and applications: A survey. pages 1–8, 07 2019.

[29] H. Kazemian and S. Ahmed. Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications*, 42(3):1166 – 1177, 2015.

[30] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015.

[31] W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 448–455. AAAI Press, 2003.

[32] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. Domain-z: 28 registrations later measuring the exploitation of residual trust in domains. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 691–706, 2016.

[33] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, page 179, USA, 2003. IEEE Computer Society.

[34] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–, Nov 2004.

[35] B. F. Maier. Binpacking 1.4.5, 2021. Accessed on 20-05-2020 via https://pypi.org/project/binpacking/.

[36] MarkMonitor. White paper: Fighting counterfeit sales online. Technical report, MarkMonitor, 2017. Accessed on 23-12-2020 via https://www.markmonitor.com/download/wp/wp-Fighting_Counterfeit_Sales.pdf.

[37] D. McCoy, A. Pitsillidis, J. Grant, N. Weaver, C. Kreibich, B. Krebs, G. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *21st USENIX Security Symposium (USENIX Security 12)*, pages 1–16, Bellevue, WA, Aug. 2012. USENIX Association.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. Accessed on 23-12-2020 via http://arxiv.org/abs/1301.3781.

[39] F. Mordelet and J.-P. Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014. Partially Supervised Learning for Pattern Recognition.

[40] W. Mostard, B. Zijlema, and M. Wiering. Combining visual and contextual information for fraudulent online store classification. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, pages 84–90, New York, NY, USA, 2019. Association for Computing Machinery.

[41] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All your iframes point to us. In *Proceedings of the 17th Conference on Security Symposium*, SS'08, pages 1–15, USA, 2008. USENIX Association.

[42] H. G. Ramaswamy, C. Scott, and A. Tewari. Mixture proportion estimation via kernel embedding of distributions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2052–2060. JMLR.org, 2016.

[43] Y. Ren, D. Ji, and H. Zhang. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 488–498, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[44] G. Rodriguez. Generalized linear models, 2014. Accessed on 23-12-2020 via https://data.princeton.edu/r/glms.

[45] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[46] R. E. Schapire. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[47] Scikit-learn. Feature importances with a forest of trees, 2020. Accessed on 21-05-2021 via https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html.

[48] Scikit-learn. Minmaxscaler, 2020. Accessed on 21-05-2021 via https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[49] Scikit-learn. Pipeline, 2020. Accessed on 21-05-2021 via https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html#sklearn.pipeline.Pipeline.

[50] Scikit-learn. Quantiletransformer, 2020. Accessed on 21-05-2021 via https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html?highlight=quantiletransformer#sklearn.preprocessing.QuantileTransformer.

[51] J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, Jun 1999.

[52] Test Aankoop. Bestel niets bij deze valse webshops, Sep 2018. Accessed on 23-12-2020 via https://www.test-aankoop.be/hightech/internet/nieuws/800-valse-webshops-ontmaskerd.

[53] H. Tian, S. M. Gaffigan, D. S. West, and D. McCoy. Bullet-proof payment processors. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–11, 2018.

[54] T. Wabeke, G. C. M. Moura, N. Franken, and C. Hesselman. Counterfighting Counterfeit: detecting and taking down fraudulent webshops at a ccTLD. In *Proceedings of the Passive and Active Measurement Workshop*, Eugene, OR, USA, 2020.

[55] J. Wadleigh, J. Drew, and T. Moore. The e-commerce market for "lemons": Identification and analysis of websites selling counterfeit goods. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1188–1197, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[56] D. Y. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. M. Voelker. Search + seizure: The effectiveness of interventions on seo campaigns. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 359–372, New York, NY, USA, 2014. Association for Computing Machinery.

[57] Wappalyzer. Apis: Automate technology lookups. https://www.wappalyzer.com/api/. Accessed: 11-05-2021.

[58] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[59] Wikipedia. List of circulating currencies, 2021. Accessed on 20-05-2020 via https://en.wikipedia.org/wiki/List_of_circulating_currencies.

[60] H. Yu, J. Han, and K.-C. Chang. Pebl: Web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):70–81, 2004.

[61] Y.-L. Zhang, L. Li, J. Zhou, X. Li, Y. Liu, Y. Zhang, and Z.-H. Zhou. Poster: A pu learning based system for potential malicious url detection. CCS '17, pages 2599–2601, New York, NY, USA, 2017. Association for Computing Machinery.