

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Hanzl** Jméno: **Petr** Osobní číslo: **420866**
Fakulta/ústav: **Fakulta informačních technologií**
Zadávající katedra/ústav:
Studijní program: **Informatika**
Studijní obor: **Znalostní inženýrství**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Detekce temných vzorů v českých internetových obchodech

Název diplomové práce anglicky:

Detection of Dark Patterns on Czech Webshops

Pokyny pro vypracování:

The goal of the thesis is to analyze content on selected Czech webshops in order to detect so called dark patterns.

1. Analyze and describe existing methods for dark patterns detection in the Czech Web environment as well as in the world.
2. Design a crawler to retrieve Czech Webshops content and identify relevant product pages.
3. Implement the crawler and a method for dark patterns detection on selected Webshops.
4. Evaluate and describe results of your method.

Seznam doporučené literatury:

Jméno a pracoviště vedoucí(ho) diplomové práce:

doc. Ing. Tomáš Vitvar, Ph.D., katedra softwarového inženýrství FIT

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **16.02.2021**

Termín odevzdání diplomové práce: _____

Platnost zadání diplomové práce: _____

doc. Ing. Tomáš Vitvar, Ph.D.
podpis vedoucí(ho) práce

podpis vedoucí(ho) ústavu/katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

Datum převzetí zadání

Podpis studenta

CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF SOFTWARE ENGINEERING



Master's thesis

Detection of Dark Patterns on Czech Webshops

Bc. Petr Hanzl

Supervisor: doc. Ing. Tomáš Vitvar, Ph.D.

5th of January, 2022

Acknowledgements

I wish to express my sincere thanks to my supervisor, doc. Ing. Tomáš Vitvar, Ph.D., for the continuous encouragement and advice given while writing this thesis.

Furthermore, I would like to thank my whole family, especially my parents, for raising me up and supporting me during my studies. Additionally, I would like to thank my friends. I would like to point out my friend David Whalan making me lose worries in spoken English and my other friend Ing. Tomáš Hodek for encouraging in my life decisions.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the “Work”), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work in any way (including for-profit purposes) that does not detract from its value. This authorization is not limited in terms of time, location and quantity.

In Prague on 5th of January, 2022

.....

Czech Technical University in Prague

Faculty of Information Technology

© 2022 Petr Hanzl. All rights reserved.

This thesis is a school work as defined by Copyright Act of the Czech Republic. It has been submitted at Czech Technical University in Prague, Faculty of Information Technology. The thesis is protected by the Copyright Act and its usage without author's permission is prohibited (with exceptions defined by the Copyright Act).

Citation of this thesis

HANZL, Petr. *Detection of Dark Patterns on Czech Webshops*. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2022. Available also from WWW: [⟨https://github.com/Lznah/DarkPatterns⟩](https://github.com/Lznah/DarkPatterns).

Abstrakt

Tato diplomová práce se zabývá vzory v uživatelské rozhraní, též známé jako temné vzory, které nutí uživatele dělat věci, nebo se rozhodovat jinak, než původně zamýšleli. Tato práce se zaměřuje na detekci temných vzorů použité webshopy na českém internetu a detekce probíhá ve velkém měřítku.

Práce vychází z již provedeného výzkumu z Princetonovy univerzity, který zkoumal temné vzory na anglických webshopech.

Bylo vytvořeno několik nástrojů pro získání značného počtu webshopů. Nástroje z původního výzkumu byly upravené tak, aby mohly být použity pro český jazyk.

Těmito nástroji bylo získáno několik datasetů mapující webshopy na českém internetu a temné vzory na nich použité.

Bylo zjištěno, že temné vzory jsou na českých webshopech hojně využívány.

Klíčová slova Temné vzory, Automatizované procházení webu, Interakce člověk-počítač, Shluková analýza, Webové obchody

Abstract

This thesis investigates patterns in user interfaces, also known as dark patterns, that force users to do things or make decisions differently than they originally intended. This thesis focuses on the detection of dark patterns used by webshops on the Czech Internet and the detection is done on a large scale.

This thesis builds on research already conducted at Princeton University that investigated dark patterns on English webshops.

Several tools were created to retrieve a significant number of webshops. Also the tools from the conducted research were modified to be applied to the Czech language.

These tools were used to obtain multiple datasets mapping webshops on the Czech Internet and the dark patterns used on them.

It was found that dark patterns are widely used on Czech webshops.

Keywords Dark patterns, Web crawling, Human-computer Interaction, Cluster analysis, Webshops

Contents

Introduction	1
1 State of the art	3
2 Dark Patterns	7
2.1 Definition	7
2.2 Taxonomy	8
2.3 Categories and types of Dark Patterns	10
3 Corpus Creation	23
3.1 Extracting webshops from Heureka	24
3.2 Retrieving true domain names	26
3.3 Cleansing of dataset	26
4 Data Collection	29
4.1 Discovering Product Page URLs	30
4.2 Discovering Textual Segments	33
5 Data Analysis	39
5.1 Preprocessing	39
5.2 Feature processing	40
5.3 Clustering	41
5.4 Analysis of output clusters	42
5.5 Results	43

Conclusion	53
Bibliography	55
A List of Acronyms	61
B Supplemental Material	63

List of Tables

List of Figures

1.1	Overview of the shopping website corpus creation, data collection using crawling, and data analysis, as proposed by Princeton University researchers.[21].	4
2.1	An example of Sneak into Basket dark pattern that was used on Alza.cz in 2018, the biggest Czech e-commerce website. The user added a power bank into his basket, and this webshop added a charger into the basket. Alza.cz claimed that users might need these additional buyings because users could not use the bought products without them [18].	10
2.2	An example of Hidden Cost dark pattern that appears at the very last step of the purchase flow on Mall.cz. This webshop adds a payment for insurance, which users may not notice. "Chci pojistit zásilku" can be translated as "I want to insure the shipment".	11
2.3	An example of Hidden Subscription that was used by Alza.cz in 2016[35]. Alza promoted 30 days free of its VIP membership. If users did not cancel their membership within those 30 days, Alza assumed that users were interested in continuing their membership and paying the fee.	12
2.4	An instance of 'Countdown Timers' dark pattern on Alza.cz's homepage. The caption "Nabídka končí za 14:08:24" can be translated as "The offer ends in 14:08:24". changes this offer for a different product every day.	13

2.5	Another instance of ‘Countdown Timers’ dark pattern, but this was found on CZC.cz homepage.	13
2.6	An instance of ‘Limited-time message’ dark pattern found on a product page of CZC.cz webshop. The red arrow is pointing at the caption ”pouze dnes!”, which can be translated as ”only today!”.	13
2.7	An instance of ‘Visual Interference’ dark pattern on Alza.cz. This instance appears in the last step of the buying process, where users fill in their payment information. Alza.cz steers users’ attention to the option with the green background ”Zaplatit 1 249 Kč a zapamatovat kartu pro příští nákupy” (English: ”Pay 1 249 CZK and save the card information for the future payments”) and hides the other option, by which users would not approve to save the card information.	14
2.8	An instance of ‘Trick Questions’ dark pattern on CZC.cz, where ”Nesouhlasím se zasíláním marketingových materiálů ...” can be translated as ”I do not agree with ...” While this sentence is not a question, it is certainly confusing because users must indicate their opposition to the newsletter subscription.	15
2.9	Instances of ‘Pressured selling’ and ‘Confirmshaming’ dark patterns found on Alza.cz again. Webshop offers additional services (a protective glass in this instance) for products in the basket, which is a cross-selling, that is defined as Pressured selling dark pattern. Also, webshop preselected an option ”Nebojím se odření displeje” (English: ”I am not worried about the scratches on the display”), which is intended to evoke worries and scare emotions in users, so it is considered as Confirmshaming dark pattern as well.	16
2.10	Another similar instance of ‘Pressured selling’ dark pattern. This instance was found on CZC.cz. ”Risknu to bez prodloužené záruky” can be translated as ”I will take my chances without the extended warranty.”	16

2.11	The third instance of ‘Pressured selling’ dark pattern. This instance is a modal window, that occasionally pops up right after the confirmation of the content of the basket. The headline says: ”Do not forget these important additional products.” The webshop preselects these additional products. In this example, there is ‘Visual Interference’ dark pattern as well. The styling of the acceptance button (the green button) tempts users to click on particular button.	16
2.12	An instance of ‘Activity Notifications’ dark pattern on flora-online.cz, where ”Dnes zakoupilo 31 zákazníků” can be translated as ”31 customers bought this product today”.	17
2.13	Another instance of ‘Activity Notification’ dark pattern. This instance was found on kytice-expres.cz. It shows the recently bought flowers by other users.”	17
2.14	An example of Testimonials of Uncertain Origins dark pattern found on kytice-expres.cz. The webshop claims 4381 rankings of a product (czech: ”Hodnoceno 4381x”) with an average score of four and a half stars out of five. There was no additional information on how the webshop obtained these references.	18
2.15	An instance of ‘Low-stock Message’ dark pattern on Alza.cz. The green text ”Poslední 1 kus” can be translated as ”The last 1 product left in stock”.	19
2.16	An instance of ‘High-demand Message’, that can be found in the basket of Alza.cz webshop. It can be translated as ”Dear stranger, hurry up! Some of the goods from your basket may disappear soon!”	19
2.17	Example of Hard to Cancel dark pattern in Alza Premium terms of service. If users want to cancel the auto-renewal service, they need to contact customer care via the contact form. The translation of the bottom paragraph from the terms of service is: ”Alza Premium membership can be cancelled at any time on the Alza website via the contact form.”	20
2.18	An example of Forced Enrollment that can be found on registration page of bestdrive.cz. By checking the first checkbox, users confirm their acceptance of the general terms of use. By checking the second checkbox, they agree to the processing of personal data, which also leads to the subscription.	21

3.1	All steps of corpus creation, which starts with a list of webshops available on Heureka.cz, paginated into 3,735 pages and ends with a list of 43 413 unique webshops' URLs in CSV format.	25
4.1	The workflow of discovering product page URLs. The crawler can be run in an unguided or a guided mode. The unguided crawl extracts Product Page URLs from a fraction of all Czech webshops. The output is manually labelled and creates a training dataset for the classification model. Further, this model guides the crawler with prioritizing possible Product Page URLs in its inner queue. Therefore, the crawling of a single page is rapidly speeded up. . . .	31
4.2	The workflow of discovering textual segments from the dataset of Product page URLs. OpenWPM framework creates multiple workers (Browser Managers) and serves them a sequence of tasks they follow. The task manager is capable of orchestrating the workers, that finished previous tasks and are ready for crawling a Product Page URL.	35
5.1	Distribution of webshops using at least one Dark Pattern over the ranking in Heureka's webshop list. Each bin is a size of two hundred webshops, representing a percentage prevalence of webshops containing dark patterns within the bin.	44
5.2	Distribution of different types of dark patterns over the ranking in Heureka's webshop list. Each bin is a size of two hundred webshops, representing a number of dark patterns of the type within the bin. .	46
5.3	Distributions of the five most used e-commerce solutions over Heureka's rank with a distribution of a sum of them all. Each bin is a size of two hundred webshops.	48
5.4	Distribution of webshops using Notifikuj.cz service of push notifications over the dataset of 10K highest-ranked webshops in Heureka's ranking.	50
5.5	An example of cross-selling as "Pressured Selling" dark pattern found on webshop beason.cz. This dark pattern appears in a pop-up window immediately after users add a product to a cart. "Ostatní zákazníci také nakoupili" can be translated into English as "Other customers also purchased".	51

5.6	An example of dark pattern "Pressured Selling" found on beason.cz. This dark pattern offers free shipping if a customer purchases for higher price. "Objednejte ještě za 900 Kč a budete mít dopravu ZDARMA" can be translated as "Order for another 900 CZK and you will get free shipping".	51
5.7	An example of "Trick Questions" dark pattern, which uses double negation in the sentence. The user may think that he is not giving his consent to the webshop for sending satisfaction surveys by not checking the checkbox. "Nesouhlasím se zasláním ..." can be translated as "I do not agree with ..."	52

Introduction

Dark patterns[12, 16, 33, 21] are ways of designing a user interface of websites, apps or any other computer system in a specific way to trick, confuse or coerce a user in doing unwanted actions like confirming to share more information than is needed to use the service, signing up for things that the user did not mean to, buying unwanted products and more.

Typically, when the user reads a website or uses an app, he does not read all the words and makes quick assumptions[12]. Dark patterns then trick the user by hiding information of unpleasant truth. The user also trusts in the experience that he has gained from using other websites or apps and expects specific actions to happen or not to happen by using a similar pattern in the user interface. The user is tricked here by expecting this user interface behaviour, but in reality, it does something more or less than what the user expects[33]. Dark patterns are not only able to take advantage of the user not paying enough attention. Another dark pattern uses psychological methods to make users feel bad and guilty for not doing what the dark pattern wants them to do[33].

Research into tricky user interface designs and deceptive practices has surprisingly much history, but it was neglected for many years. In 1999, Hanson and Kysar were the first who examined how companies abuse customers' cognitive limitations and profit from them. The rapid growth of the Internet and e-commerce increased more serious discussions and analyses of this topic. The term Dark Pattern itself was introduced by user interface expert Harry Brignull

in 2010 to create a library of different types of dark patterns and to shame websites using them[13].

In March 2021, the state of California added new regulation that now bans dark patterns that prevent users from opting out of the sale of their personal data[6]. Therefore, the topic of dark patterns becomes more and more relevant.

In 2019, a group of scientists from Princeton University introduced an automated approach that enables experts to identify dark patterns used on websites at scale[21].

This thesis's primary goal is to build on top of their research to analyse the prevalence of dark patterns on Czech webshops, also described in the Princeton study[21]. Their work and also this thesis focus on product pages and product purchase flow only because these are the most promising pages, where all the buying happens. Several subgoals need to be done to fulfil the primary goal:

- Create a dataset of Czech webshops.
- Adapt the published source codes from the prior research for the Czech language.
- Analyse gathered data.
- Evaluate and describe findings.

This thesis does not aim to create a model capable of automated detection of dark patterns. Also, this thesis does not aim to study the prevalence of deceptive dark patterns that display transients values over time.

State of the art

Most studies[12, 16, 9] in the field of dark patterns have only described known existing types of dark patterns. Also, literature often proposes different dark pattern taxonomies. To find these patterns, scholars did manual research, analysing page by page.

In contrast to this approach, which requires much manual work, there is a study from Princeton University[21]. The researchers implemented mechanisms to reduce the manual work that needs to be done. They also propose an entirely new taxonomy. Furthermore, the researchers recategorised and made more accurate the currently known types from the literature, but they were able to find new types of dark patterns; thus, they extended the literature about these new types.

Princeton researchers focus their study only on textual information found on webshops. This limits the results of their work to only textual dark patterns.[21].

In an attempt to find these new types, researchers focused on product pages of webshops, because as they say, these pages are the most promising to contain dark patterns at any level of purchase flow[21]. Princeton Researchers did much work to find these dark patterns. Their work can be split into three steps, as can be seen in figure 1.1.

Corpus Creation is the first step; there are several scripts to get domain names of webshops. They gathered websites with the highest Alexa Rank

1. STATE OF THE ART

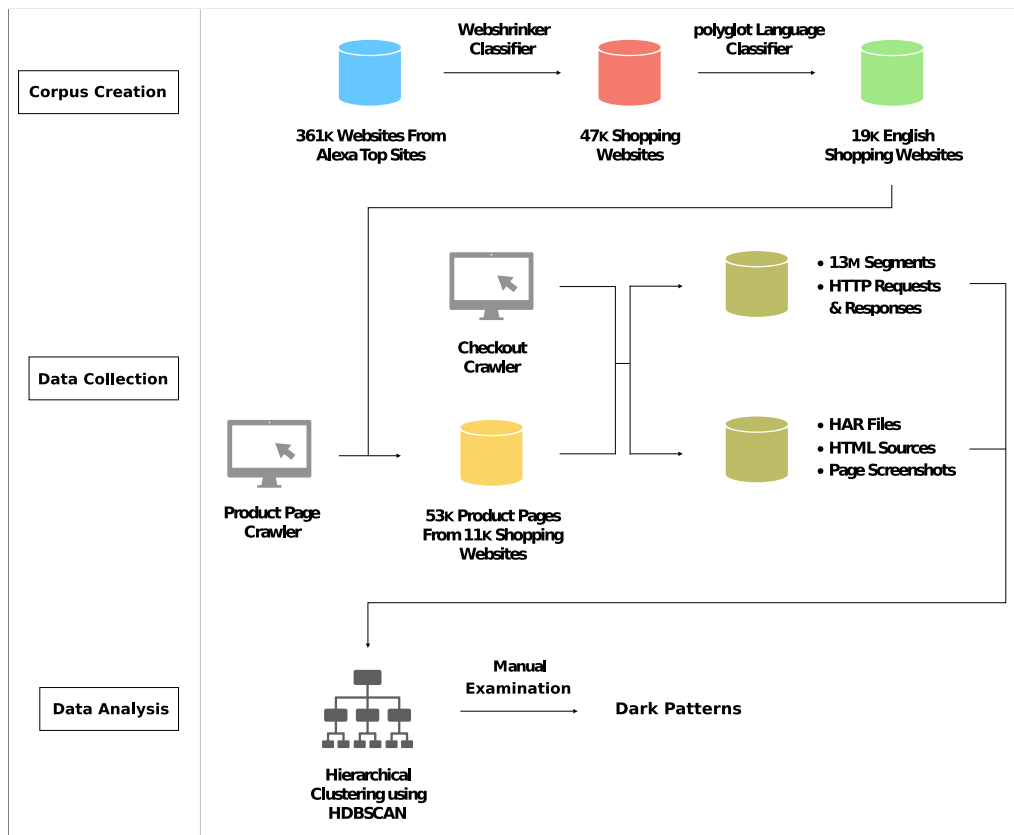


Figure 1.1: Overview of the shopping website corpus creation, data collection using crawling, and data analysis, as proposed by Princeton University researchers.[21].

via Alexa Rank API. Then, they used paid service Webshrinker to filter out only those websites that are webshops. The list of domains still contained non-English websites. They used a language classifier library Polyglot to filter them out of the list. Overall, researchers gathered a list of 19K English shopping websites[21].

Data Collection is the second step. It consists of two crawlers created by Princeton researchers. The first crawler is meant to find product links on a single website. To speed up the process of finding these product pages, they trained a classifier of Logistic Regression on a dataset of 1000 URL links manually labelled by the researchers. The first crawler found 53K pages in 11K domain names.

The second crawler, also referred to as a checkout crawler, is meant to simulate users' shopping flow. This ability to simulate users' flow means that the crawler follows the buying process steps, including selecting product options (e.g., size or colour), adding the product to the card, viewing the cart, and checking out. To evaluate whether or not this crawler can simulate users' shopping flow, the researchers randomly sampled 100 product pages and examined whether the crawler successfully reached the checkout page.

This crawler is built on OpenWPN, which is a web privacy measurement framework for privacy studies on a large set of websites. Princeton researchers implemented additional features to this framework. For example, they created a feature to store HAR files, which contain all the HTTP communication and Javascript calls. All these collected data are further utilised in an analysis phase by researchers. These data help researchers recognise whether or not a found pattern is one of the types of dark patterns.

The checkout crawler also divides visited pages into meaningful textual segments. Researchers define this textual segment and an algorithm to split the page's HTML code into these segments[21]. Consequently, the checkout crawler extracts data about the text and background colours, positions and dimensions of the segments and others. With this algorithm, they captured approximately 13 million segments across the previously noted 53K product URL pages.

Data Analysis is the last step of the research. It consists of data preprocessing, hierarchical clustering, examining and analysing the found clusters. The data cleansing phase reduced 90% of all segments to 1.3 million segments.

Data were transformed into a representation of Bag of Words (BoW)[37]. Then, Principal Component Analysis was performed on the BoW matrix. The outcome was three components, which together represented 95% of the variance in the data.

Researchers chose an algorithm called Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN)[7] to find clusters in data. They tried different hyperparameters of this clustering algorithm and picked the most promising results.

Then, they did two passes examining the clusters. In the first pass, they manually tagged clusters that can manifest as dark patterns. This pass reduced the number of clusters from 10,277 to 1,768. During the second examination, researchers manually examined which of these 1,768 clusters contain dark patterns[21].

Lastly, the researchers discussed the results, and they iteratively grouped the discovered dark patterns into types and categories. They revealed 15 types of dark patterns in 7 categories on 1,254 websites, representing 11,1% out of 10,277 webshops[21].

Dark Patterns

The ‘Dark Pattern’ is a relatively new term. This neologism was firstly used by Harry Brignull in 2010[15] when he registered a domain darkpatterns.org. In this domain, Brignull created an online library to share user interface patterns with deceptive characteristics that intentionally confuse and enrol users in unwanted situations. Another purpose of this online library is to shame websites that use dark patterns.

2.1 Definition

Brignull described dark patterns as so: ‘Dark Patterns are tricks used in websites and apps that make you do things that you did not mean to, like buying or signing up for something.’[12] Brignull’s definition is simplified to understand what dark patterns are with ease. However, it does not include all the dark patterns that Brignull describes. For example, there is a dark pattern that purposely focuses users attention on doing one action and distracts their attention from alternatives. Brignull’s definition does not imply this example.

A more accurate definition is the one used in the study made by Princeton researchers. They suggest this definition: ‘Dark patterns are user interface design choices that benefit an online service by coercing, steering, or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, they might not make.’ [21]

2.2 Taxonomy

Brignull also defined the first types of dark patterns. This list of types is continuously updated when a new type of dark pattern is found. In April 2021, there were twelve different types of dark patterns defined[14].

The researchers from Princeton University have redefined this list considering the results of their study. This list consists of fifteen types of dark patterns and seven broad categories. Their work also differs from the prior work[12, 5, 9] by the new proposed taxonomy. This new taxonomy focuses on the characteristics of dark patterns and cognitive biases that they exploit in users. They used their taxonomy to classify and describe discovered dark patterns.

This thesis uses the same taxonomy defined by Princeton researchers. This taxonomy consists of five dimensions:

Asymmetric

The user interface presents more alternatives to a user. It is an asymmetric characteristic of a dark pattern if the user interface requires less effort to continue with the alternative that might be disadvantageous for users. A typical example is buttons for accepting and rejecting cookies on websites. Usually, the rejecting button is less noticeable. Also, if users want to reject saving cookies, the user interface forces them to read much more text and click many buttons for every single cookie.

Covert

The user interface shows evidence of covert characteristics if users may fail to recognise the intended outcome of a specific action. Users have experience with other user interfaces, and they may predict a similar outcome from the interface that shows similar traits as a decoy to influence their decision-making process. For instance, most of the websites offer a subscription to a newsletter in the process of registration. Usually, this subscription to the newsletter is done by ticking a checkbox in the registration form. When users start to read a sentence mentioning the subscription, they automatically expect that not ticking the checkbox means not subscribing to the newsletter.

Deceptive

The user interface induces false beliefs in users by presenting them

misleading information. For instance, a website may offer a discount for a limited period of time, but in reality, the discount is permanent. Another example is a website that shows how many users are watching the given product and how many products are in stock. This information can take advantage of the deal by steering users into making quick decisions or inducing false beliefs of the product's exclusivity.

Hides Information

The user interface intentionally delay presenting necessary information in places or in time, where or when users do not expect them to be presented. For instance, a website may present extra fees for a bought product at the very last step of the checkout.

Restrictive

The user interface restricts the set of choices available to users and takes advantage of it. For example, a website may require signing up only with Facebook to collect additional personal information.

In addition to these dimensions, Princeton researchers define six different effects on users through exploiting different cognitive biases by specific dark patterns:

- **Anchoring Effect:** The tendency of users to over-rely on the first piece of information in the future decision-making process.
- **Bandwagon Effect:** The tendency of users to value more or believe in something simply because others do.
- **Default Effect:** The tendency of users to stick with default options.
- **Framing Effect:** The tendency of users to choose different options with knowledge of the same information, but with a different way of presenting the options.
- **Scarcity Bias:** The tendency of users to value more things that are more sparse.
- **Sunk Cost Fallacy:** The tendency of users to continue an action because they already invested time or other resources in it. Users tend to continue even if that action is capable of putting them in an even worse situation.

2.3 Categories and types of Dark Patterns

The types introduced in this section are the same defined in the paper from Princeton university[21], but with examples found on the Czech webshops. These types are based on the types firstly published by Harry Brignull[12]. Princeton researchers discovered 15 types of dark patterns in total, and they divided them into seven broader categories. The summarization of these types is in table 2.1 at the end of this section.

2.3.1 Sneaking

It is an attempt to hide, disguise, or delay information relevant to users. Users would likely change their move if they knew about this information. There are three types of dark patterns in this category: Sneak into Basket, Hidden Costs, and Hidden Subscription.

2.3.1.1 Sneak into Basket

This type of dark pattern adds additional products into the user's basket without their consent. Usually, he is not aware of this fact. The added products are bonuses or additional services – for example, an additional year of warranty or a gift card. The essential for these dark patterns is that it raises the total price, and users might not be aware of this fact.

This dark pattern exploits the *default effect* of cognitive bias in users that was described earlier in this thesis. The literature says that this dark pattern is not *covert* because users can see the added products in their baskets.

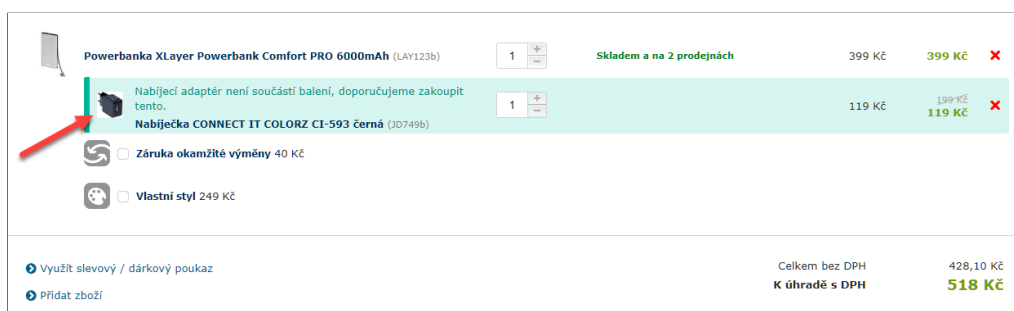


Figure 2.1: An example of Sneak into Basket dark pattern that was used on Alza.cz in 2018, the biggest Czech e-commerce website. The user added a power bank into his basket, and this webshop added a charger into the basket. Alza.cz claimed that users might need these additional buyings because users could not use the bought products without them [18].

2.3.1.2 Hidden Cost

This pattern is an attempt to add additional charges, typically at the end of the purchase process. Typical examples of this type of dark pattern are additional service fees or handling costs.

This type of dark pattern is also not *covert*, but it may be considered partially *deceptive* because the information is delayed from users. Also, this dark pattern can be classified into *hides information* dimension, as it attempts to hide information from users.

VYBERTE DOPRAVU	
<input checked="" type="radio"/> Doručení na adresu ve vámi vybraný den Upravit	678 Kč
Pardubická 1051, 53501 Přelouč +420777777777 Pátek 16.4. Čas doručení: 09-12 hod.	
<input type="checkbox"/> Chci pojistit zásilku Více info	49 Kč

Figure 2.2: An example of Hidden Cost dark pattern that appears at the very last step of the purchase flow on Mall.cz. This webshop adds a payment for insurance, which users may not notice. "Chci pojistit zásilku" can be translated as "I want to insure the shipment".

2.3.1.3 Hidden Subscription

This pattern signs up users into a subscription with a recurring fee. Users may not be aware of this subscription because the subscription is presented as a one-time payment or a free trial. This type of dark pattern usually appears together with another dark pattern named 'Hard to Cancel'.

This dark pattern is classified to be partially *deceptive* because it may confuse and mislead users. Also, it can be said that this dark pattern *hides information* from users.

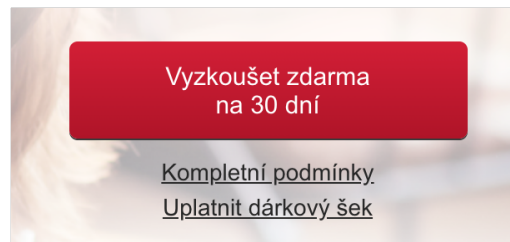


Figure 2.3: An example of Hidden Subscription that was used by Alza.cz in 2016[35]. Alza promoted 30 days free of its VIP membership. If users did not cancel their membership within those 30 days, Alza assumed that users were interested in continuing their membership and paying the fee.

2.3.2 Urgency

Dark patterns from this category speed-up users decision-making process by exploiting scarcity bias in users. For example, this can be done by showing more beneficial or time-limited discounts to users. As a result, users value products more than they would normally do. These dark patterns usually keep signalling that the special offer may be lost to users if they do not react promptly. This dark pattern is usually combined together with ‘Social Proof’ and ‘Scarcity’ types of dark patterns defined below.

2.3.2.1 Countdown Timers

This dark pattern is usually in the form of an indicator of a deadline, counting down to the end of the deadline.

This dark pattern is classified as partially *covert* because it evokes untrue feelings of immediacy in users and is sometimes classified as *deceptive* because the indicator sometimes shows false information. For example, the timer can reset every time it reaches the deadline.

2.3.2.2 Limited-time Messages

The ‘Limited-time message’ dark pattern differs from ‘Countdown Timer’ by static urgency message and not showing the exact time of the deadline.

With the taxonomy defined before, this dark pattern is classified as *covert* because of the same reason as ‘Countdown Timer’ dark pattern and *information hiding* because it does not show the deadline in its offers.

2.3. Categories and types of Dark Patterns



Figure 2.4: An instance of 'Countdown Timers' dark pattern on Alza.cz's homepage. The caption "Nabídka končí za 14:08:24" can be translated as "The offer ends in 14:08:24". changes this offer for a different product every day.

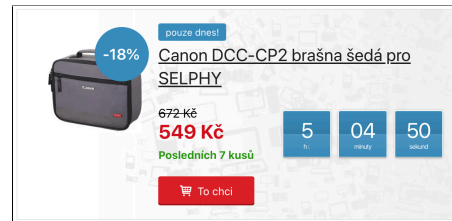


Figure 2.5: Another instance of 'Countdown Timers' dark pattern, but this was found on CZC.cz homepage.

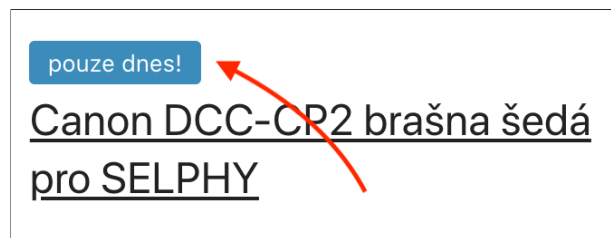


Figure 2.6: An instance of 'Limited-time message' dark pattern found on a product page of CZC.cz webshop. The red arrow is pointing at the caption "pouze dnes!", which can be translated as "only today!".

2.3.3 Misdirection

This category of dark patterns uses visuals and language to distract users' attention on other possible presented choices. Also, some types from this category use users' emotions to invoke bad feelings of being guilty or ashamed for not making a specific choice. Users trust or feel that the other choices are unavailable or less beneficial for them. Essential for this dark pattern is that other choices are not hidden. Users are aware of the other choices, but this category of dark patterns steers users away from the other choices. Princeton researchers discovered four types of dark patterns from this category: 'Confirmshaming', 'Visual Interference', 'Trick Questions', and 'Pressured Selling'.

2.3.3.1 Confirmshaming

The 'Confirmshaming' dark pattern uses language and emotions to focus the attention of users on one choice in order to distract attention on other choices. Researchers point out that this dark pattern usually appeared in popup

dialogues that asked for an email address in exchange for a discount. Some instances of this dark pattern evoke emotions of shame in users if they select an option that the webshop does not want them to select. Typical examples of such options are ‘No, I want to pay full price’ or ‘No thanks, I hate saving money’. This dark pattern exploits the framing effect of cognitive bias in users by presenting choices differently to users.

Thus, this dark pattern is classified as *asymmetric*. However, it is not *covert* since all the possible choices are presented to users.

2.3.3.2 Visual Interference

The ‘Visual Interference’ dark pattern uses different styles and visuals to draw users’ attention to certain choices - the choices that the website wants users to choose. A typical example of this dark pattern is two buttons in different styles for opting-in and opting-out for the website’s newsletter subscription. One of the buttons - the one that the website wants users to click on - looks more promising, more attractive to users’ eyes than the other one. The different styles steer users attraction to the opting-in choice.

By provided taxonomy, this type of dark pattern is partially classified as *asymmetric* because it sometimes unequally present choices to users. Users may not realise that the effect of the dark pattern influenced them. Because of this fact, this dark pattern is also classified as *covert*. Some instances can also be classified as *deceptive*, and Princeton researchers give an example of an option “lucky draw” among others that are deterministic and not random.

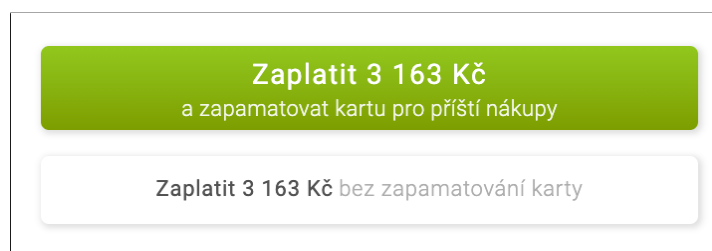
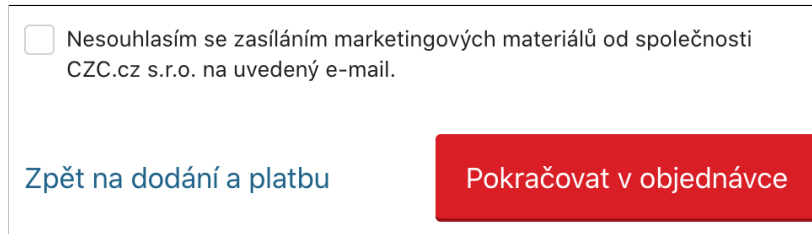


Figure 2.7: An instance of ‘Visual Interference’ dark pattern on Alza.cz. This instance appears in the last step of the buying process, where users fill in their payment information. Alza.cz steers users’ attention to the option with the green background “Zaplatit 1 249 Kč a zapamatovat kartu pro příští nákupy” (English: “Pay 1 249 CZK and save the card information for the future payments”) and hides the other option, by which users would not approve to save the card information.

2.3.3.3 Trick Questions

The ‘Trick Question’ dark pattern uses confusing language to confuse users and their ability to make decisions. A typical trick in the English language for this dark pattern is double negatives. For example, websites using this type of dark pattern invert the meaning of a check subscription checkbox, usually seen in registration forms, followed with confusing language ‘Uncheck this box if you prefer not to receive email updates’. Users need to pay more attention to properly understand which state of the checkbox means the subscription for the newsletter and which not. This type of dark pattern exploits the default effect in users, who erroneously believe that to them presented user interface follows traditional patterns. Also, this dark pattern exploits the framing effect by presenting the same information in a different, more confusing way to influence users in choosing different choices.

Therefore, Princeton researchers classify this type of dark pattern as *asymmetric* because opting out takes more effort than opting in. Also, researchers classify this dark pattern as *covert* because users may falsely understand the effect of their choice.



Nesouhlasím se zasíláním marketingových materiálů od společnosti CZC.cz s.r.o. na uvedený e-mail.

[Zpět na dodání a platbu](#) [Pokračovat v objednávce](#)

Figure 2.8: An instance of ‘Trick Questions’ dark pattern on CZC.cz, where “Nesouhlasím se zasíláním marketingových materiálů ...” can be translated as “I do not agree with ...” While this sentence is not a question, it is certainly confusing because users must indicate their opposition to the newsletter subscription.

2.3.3.4 Pressured Selling

Princeton researchers define the ‘Pressured Selling’ dark pattern as pre-selecting more expensive variations of the same product as default. Additionally, pressuring users into choosing the more expensive variations or buying related products is also considered as a tactic of this dark pattern. More cognitive biases are triggered and exploited by this dark pattern, such as the default effect, the anchoring effect (users may tend to overlook the other choices) and the scarcity bias (more expensive variations may seem to be more exclusive).

2. DARK PATTERNS

This dark pattern is for some instances classified as *asymmetric* (i.e., steering users and their acceptance towards more expensive options), and partially *covert* (users may fail to realise that the firstly shown price of the less expensive variation of the product is not the same price, as the more expensive default variation).

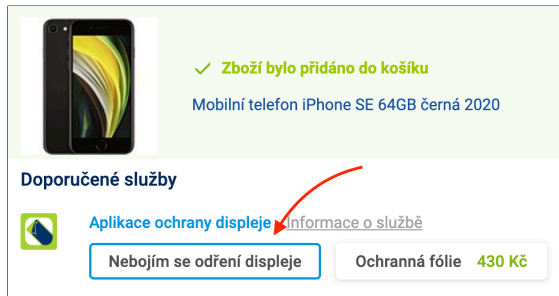


Figure 2.9: Instances of 'Pressured selling' and 'Confirmshaming' dark patterns found on Alza.cz again. Webshop offers additional services (a protective glass in this instance) for products in the basket, which is a cross-selling, that is defined as Pressured selling dark pattern. Also, webshop preselected an option "Nebojím se odření displeje" (English: "I am not worried about the scratches on the display"), which is intended to evoke worries and scare emotions in users, so it is considered as Confirmshaming dark pattern as well.

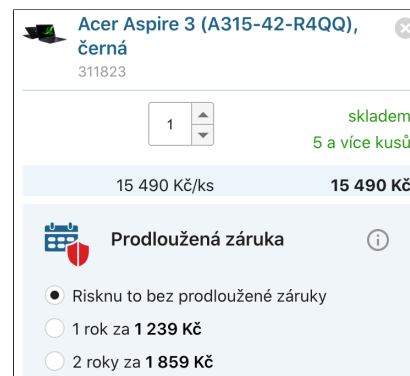


Figure 2.10: Another similar instance of 'Pressured selling' dark pattern. This instance was found on CZC.cz. "Risknu to bez prodloužené záruky" can be translated as "I will take my chances without the extended warranty."

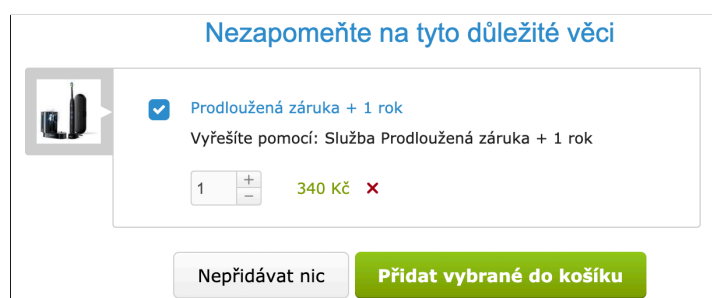


Figure 2.11: The third instance of 'Pressured selling' dark pattern. This instance is a modal window, that occasionally pops up right after the confirmation of the content of the basket. The headline says: "Do not forget these important additional products." The webshop preselects these additional products. In this example, there is 'Visual Interference' dark pattern as well. The styling of the acceptance button (the green button) tempts users to click on particular button.

2.3.4 Social Proof

The ‘Social Proof’ category of dark patterns is based on a social proof principle. Those hesitating individuals, who do not know what to do in a given situation, tend to observe others and mimic their moves, actions, and behaviour [8, 24]. This category of dark patterns misuses this behaviour of individuals, and it exploits the bandwagon effect of cognitive bias to its advantage. Princeton researchers define two types from this category: Activity Notifications and Testimonials of Uncertain Origin.

2.3.4.1 Activity Notifications

The ‘Activity Notifications’ dark pattern is information on product pages that indicate other users’ activity. The message can have different forms. It can be a number of other users watching the same product or a number of sold products to other users. Messages displaying recent purchases of other users (e.g., ‘User X just bought a product Y’) also count as ‘Activity Notifications’ dark pattern. Princeton researchers point out that some websites claim activity that is deceptive and not true. These websites use a misleading random number instead of factual information. This number also changes after some time, making it even more challenging to recognise as deceptive.



Figure 2.12: An instance of ‘Activity Notifications’ dark pattern on flora-online.cz, where “Dnes zakoupilo 31 zákazníků” can be translated as “31 customers bought this product today”.



Figure 2.13: Another instance of ‘Activity Notification’ dark pattern. This instance was found on kytice-expres.cz. It shows the recently bought flowers by other users.”

Some instances of this dark pattern can be classified as *covert* because users fail to understand that this dark pattern influences their decision-making process

in a way that they tend to buy a product, which is sold more often or is viewed more by other users. Also, some instances are classified as *deceptive* because they present made up untruthful information and users are not aware of this fact.

2.3.4.2 Testimonials of Uncertain Origin

This type of dark pattern refers to the use of customer testimonials whose origin is unclear and not sourced enough. The result of such testimonials is that users' decision-making process is influenced by untrue information, and they erroneously believe in the quality of products. In addition, a new directive will apply to all EU member states in 2022. This directive demands all e-shops to state how they ensure the authenticity of references[17].

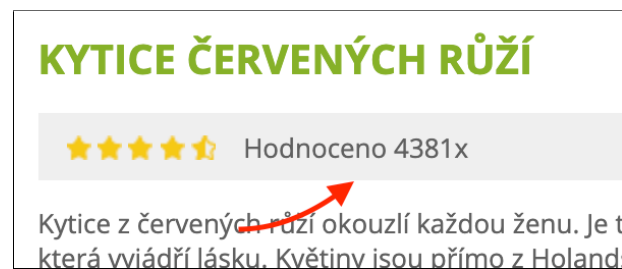


Figure 2.14: An example of Testimonials of Uncertain Origins dark pattern found on kytice-expres.cz. The webshop claims 4381 rankings of a product (czech: "Hodnoceno 4381x") with an average score of four and a half stars out of five. There was no additional information on how the webshop obtained these references.

Using taxonomy defined by Princeton researchers, this dark pattern is classified as sometimes *deceptive*, and it depends on the truthfulness of the testimonials, which can be determined by scanning the website and looking for a submission form for sending testimonials.

2.3.5 Scarcity

The 'Scarcity' category contains such types of dark patterns that implement messages indicating limited availability or high demand for a product. Thus, the value of the product increases because of its exclusivity. This dark pattern forces users to make quicker decisions. Users may feel intimidated by losing the chance to buy this very desirable product because it could be sold out soon. Princeton researchers define two types of dark patterns: 'Low-stock Message' and 'High-demand Message'.

2.3.5.1 Low-stock Message

The 'Low-stock Message' dark pattern informs users about the limited availability of a product; thus, users want to prevent losing the chance to buy the product by making quicker decisions than they normally do. Some instances of this dark pattern show the exact quantities left on the stock. Others only show a message that stock is almost empty. This dark pattern exploits scarcity bias in users - making products more valuable only because it is low on stocks. Some websites use untruthful data to keep arousing the feelings of need in users all the time.



Figure 2.15: An instance of 'Low-stock Message' dark pattern on Alza.cz. The green text "Poslední 1 kus" can be translated as "The last 1 product left in stock".

Princeton researchers classify the 'Low-stock Message' dark pattern as partially *covert* because users fail to realise that these messages influenced their decision-making process. Some instances of this dark pattern are classified as *deceptive* for displaying false information to users about being low on stock, but it is not. Some other instances are classified as *information hiding* for hiding the exact quantities of the product on stock.

2.3.5.2 High-demand Message

The 'High-demand Message' dark pattern informs users that a product is in high demand and can be sold out soon.

Similarly to 'Low-stock Message' dark pattern, 'High-demand Message' is also classified as partially *covert*.

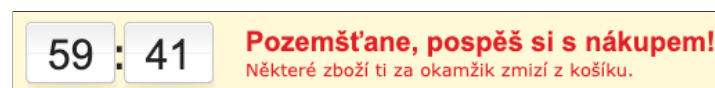


Figure 2.16: An instance of 'High-demand Message', that can be found in the basket of Alza.cz webshop. It can be translated as "Dear stranger, hurry up! Some of the goods from your basket may disappear soon!"

2.3.6 Obstruction

This ‘Obstruction’ category contains only one type of dark pattern, which is ‘Hard to Cancel’. This type of dark pattern refers to making specific actions harder to complete than other actions. For instance, signing up for a subscription to an annually paid service is often much more straightforward than cancelling the subscription [11]. Also, Princeton researchers mention examples when cancellation of a subscription is available only by calling customer service[21].

This dark pattern is sometimes classified as *restrictive* with the defined taxonomy because it restricts the available choices to cancel the previous subscriptions. The ‘Hard to Cancel’ dark pattern becomes *information hiding* when the website does not inform users how to cancel the subscription or about the fact that cancellation is not as easy as signing up.

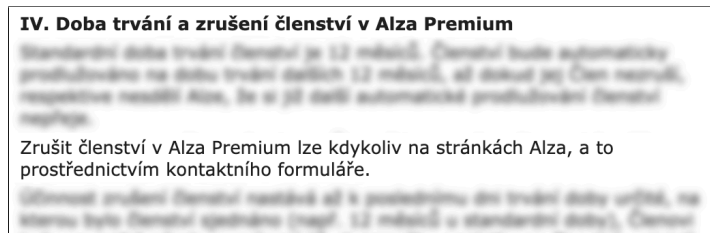


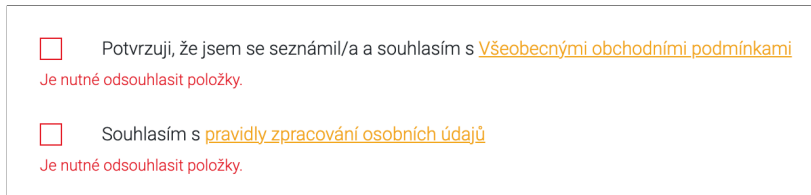
Figure 2.17: Example of Hard to Cancel dark pattern in Alza Premium terms of service. If users want to cancel the auto-renewal service, they need to contact customer care via the contact form. The translation of the bottom paragraph from the terms of service is: "Alza Premium membership can be cancelled at any time on the Alza website via the contact form."

2.3.7 Forced Action

The ‘Forced Action’ dark pattern category forces users to take additional action, even though they might not normally take it to finish their task. The ‘Forced Enrollment’ is the only type of dark pattern discovered and defined by Princeton researchers in this category. This type of dark pattern forces users (that want to use the service) into enrolling for a marketing newsletter or into creating accounts, which gives the website more information than is needed to use the service. Princeton researchers describe an example when users have to simultaneously sign up for a marketing newsletter alongside their consent to terms of service.

2.3. Categories and types of Dark Patterns

Princeton researchers define this type of dark pattern as *assymetric*, because of the requirement of the additional actions to complete users' tasks, which creates asymmetrically balanced choices, and *restrictive*, because it forces users into creating accounts and signing up for marketing newsletters.



Potvrzuji, že jsem se seznámil/a a souhlasím s [Všeobecnými obchodními podmínkami](#)
Je nutné odsouhlasit položky.

Souhlasím s [pravidly zpracování osobních údajů](#)
Je nutné odsouhlasit položky.

Figure 2.18: An example of Forced Enrollment that can be found on registration page of bestdrive.cz. By checking the first checkbox, users confirm their acceptance of the general terms of use. By checking the second checkbox, they agree to the processing of personal data, which also leads to the subscription.

2. DARK PATTERNS

Table 2.1: Summarisation of categories and types of dark patterns with their description, definition and cognitive biases they exploit [21].

Legend: ● = Always, ◐ = Sometimes, ○ = Never			Asymmetric?	Covert?	Deceptive?	Hides Info?	Restrictive?	Cognitive Biases
Category	Type	Description						
Sneaking	Sneak into Basket	Adding additional products to users' shopping carts without their consent	○	○	◐	●	○	Default Effect
	Hidden Costs	Revealing previously undisclosed charges to users right before they make a purchase	○	○	◐	●	○	Sunk Cost Fallacy
	Hidden Subscription	Charging users a recurring fee under the pretense of a one-time fee or a free trial	○	○	◐	●	○	None
Urgency	Countdown Timer	Indicating to users that a deal or discount will expire using a counting-down timer	○	◐	◐	○	○	Scarcity Bias
	Limited-time Message	Indicating to users that a deal or sale will expire will expire soon without specifying a deadline	○	◐	○	●	○	Scarcity Bias
Misdirection	Confirmshaming	Using language and emotion (shame) to steer users away from making a certain choice	●	○	○	○	○	Framing Effect
	Visual Interference	Using style and visual presentation to steer users to or away from certain choices	◐	●	◐	○	○	Anchoring & Framing Effect
	Trick Questions	Using confusing language to steer users into making certain choices	●	●	○	○	○	Default & Framing Effect
	Pressured Selling	Pre-selecting more expensive variations of a product, or pressuring the user to accept the more expensive variations of a product and related products	◐	◐	○	○	○	Anchoring & Default Effect, Scarcity Bias
Social Proof	Activity Message	Informing the user about the activity on the website (e.g., purchases, views, visits)	○	◐	◐	○	○	Bandwagon Effect
	Testimonials	Testimonials on a product page whose origin is unclear	○	○	◐	○	○	Bandwagon Effect
Scarcity	Low-stock Message	Indicating to users that limited quantities of a product are available, increasing its desirability	○	◐	◐	◐	○	Scarcity Bias
	High-demand Message	Indicating to users that a product is in high-demand and likely to sell out soon, increasing its desirability	○	◐	○	○	○	Scarcity Bias
Obstruction	Hard to Cancel	Making it easy for the user to sign up for a service but hard to cancel it	○	○	○	◐	●	None
Forced Action	Forced Enrollment	Coercing users to create accounts or share their information to complete their tasks	●	○	○	○	●	None

Corpus Creation

One of the steps of the analysis of the prevalence of the dark pattern on Czech webshops is to find webshops URLs on a large scale autonomously. The Princeton researchers used Alexa Rank[1] made by a web traffic analysis company, Alexa Internet. Alexa Rank is a measure of website popularity. Alexa Internet provides API to fetch a list of most popular websites by Alexa Rank. However, this list contains other types of websites as well, not only webshops that researchers focus on. Also, non-English websites are included as well. Because of that, researchers implemented a couple of mechanisms to cherry-pick English webshops only, discussed earlier in the state-of-the-art chapter of this thesis.

As said before, this thesis aims to analyse Czech webshops and because of that, using Alexa Rank is not efficient enough. Alexa API provides only the first five hundred thousand most popular websites, which is a reason why it contains a small number of Czech websites and even fewer Czech webshops. However, the Czech Internet (that means only websites in the Czech language) is relatively small compared to the English Internet. Also, the English Internet is under multiple jurisdictions, but the Czech Internet is not. Therefore, the Czech Internet is more consistent, and a result of it is that this environment allows creating companies that make Internet catalogues and comparison shopping websites (aggregators) that cover a significant portion of the Czech Internet.

These catalogues and tools also sometimes rank the listed websites by a measure that has connotations to popularity. For example, a number of testimonials are an excellent resource that reflects the popularity of the webshop. These catalogues and aggregators can be used to mine the URLs of Czech webshops from them instead of using Alexa Rank API. Also, if there is a similar measure as described above, the analysis results can be compared to Princeton's researcher analysis, revealing a correlation of dark patterns evidence on the website and the popularity of the website.

While searching the Internet, several such suitable sites were discovered that contain extensive lists of Czech webshops. Examples of the most suitable sites are Heureka.cz, Asociaceeshopu.cz and Shopy.cz.

Other facts that played a role and were considered in the selection of the only one website (that is later used for the creation of the list of Czech webshops) were the actual cover of the Czech Internet. Heureka has by far the highest number of webshops in their listings[29]. However, a few of the biggest webshops do not want to be listed on Heureka. Their reason is usually Heureka itself because it compares the prices of products on the enlisted webshops. Also, Heureka is a part of a business group that runs several competitive webshops. Because of that, the final list (made in this practical part of the study) of Czech webshops was manually checked if it contains the five biggest webshops (according to the list published on website peak.cz[30]), and it does.

Figure 3.1 shows steps of Corpus Creation with used technologies.

3.1 Extracting webshops from Heureka

Heureka provides a list of all registered webshops on their website. This list is paginated, where every page contains twenty records of webshops. However, Heureka does not provide the total number of these pages.

In February 2021, the total number of pages was 3,735. This number of pages was manually found by changing the query parameters from the URL until the page stopped returning error 404. At the same time, these 3,735 pages contained 74,698 webshops in total.

I implemented a web crawler to extract webshops' links and names from these pages. The crawler is written in Python 3, using Selenium framework with

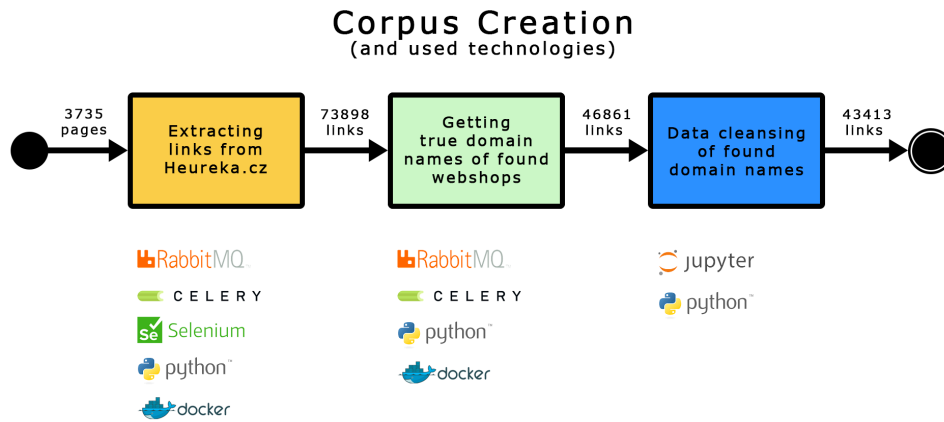


Figure 3.1: All steps of corpus creation, which starts with a list of webshops available on Heureka.cz, paginated into 3,735 pages and ends with a list of 43 413 unique webshops' URLs in CSV format.

Chrome browser in headless mode. Also, the crawler is parallelised to speed up this task using Celery asynchronous task queue.

Some of the crawled pages were not successfully downloaded because the Chrome browser occasionally failed to start or the webserver returned an empty page. The crawler was still able to successfully download 3,695 pages containing 73,898 webshops in total after scraping the HTML into 3,695 CSV files.

Such a high number of obtained webshops does not correspond with the number of webshops that Heureka claims to contain on its homepage (it claims to aggregate around 38,000 webshops). Also, the estimated number of webshops on the Czech Internet is around 41,000 according to a study made by Zbozi.cz and Shoptet.cz[29]. The cause of this is that the retrieved list of webshops contains many duplicities and already inactive webshops.

Another problem with this list is that the retrieved links are not the actual domain names of the webshops. These URLs are redirections, and they must be visited first to retrieve the actual domain name.

The two other steps of the Corpus Creation deal with these two problems.

3.2 Retrieving true domain names

As mentioned above, Heureka does not provide direct URLs to webshops in its listings. The provided links only redirect to the true URLs of webshops, and because of that, we implemented another crawler that follows these redirections and returns the true URLs. This crawler is also written in Python 3. This time, the task is only to retrieve the true URLs. Hence, Request library is used instead of Selenium, which is too complex for such a simple task. It remains parallelized using Celery.

This crawler adds an additional column to the dataset given from the previous crawler, which contains the true URL. If an exception occurred during the execution of the single task, its message is written there instead. If the web page returned a different status code than 200, the status code is also written there instead. The importance of this data about errors and exceptions are helpful in the validation of the whole task. Whether the whole task is successful or it returns too many errors and exceptions.

3.3 Cleansing of dataset

The given data from the previous crawler is further cleansed in a Jupyter notebook using primarily Pandas library.

In the first step, the dataset is split into two data frames of errors and true URLs. Dataset of errors contains 27,037 rows, and 23,238 of them are results of connection being refused after redirection. The first exception might be that Heureka implemented mechanisms to prevent the crawling of the redirections. This claim was refuted by **manually going through 100 random** links. None of these links redirects to an active webshop. The next most frequent errors were 404 errors with 1,953 occurrences, 403 errors with 750 occurrences and 503 errors with 504 occurrences. These are errors that indicate that the web store web page is no longer active. The other errors had an incidence of fewer than 200 occurrences.

The dataset of true URLs is further cleansed by filtering out other identified inactive webshops that were not identified in the error/rejection filtering. Firstly, such webshops URLs have a high frequency in the dataset because the webshops' URLs are often redirected to the webshops' hosting service website

after deactivation of the webshop. For example, many Czech webshops use Shoptet service, which allows users to rent a ready-to-use webshop solution for a monthly payment. After the users stop paying this fee, their webshop is inactivated (or deleted), making the original webshop be redirected to Shoptet’s custom web page informing visitors about the inactivation of the particular webshop. Secondly, many URLs of inactive webshops contain status codes in the URL without sending the actual status code in an HTTP response. Lastly, some domain names were inactive or resold and redirected to a new website (surprisingly redirected to porn websites in most of the cases). All this manual work led to creating a list of such URLs and filtering them out of the dataset. This shrank the dataset from 46,861 to 46,023 rows, removing another 838 rows.

The last step was to remove URI parts from the URLs and drop duplicate entries, which shrank the dataset to final 43,413 unique links to webshops, removing another 2,610 rows.

Links to download all the outputs and logfiles of crawling are on README page of GitHub repository of this thesis. The link to this repository is in Appendix B (Supplemental Material) of this thesis.

Table 3.1: A summary of URLs from 3,735 pages of Heureka’s webshop catalogue and how many and which type were removed in each step.

Extracted URLs	Accesible URLs	Non-accessible URLs
73,898	27,037	27,037
Dirty URLs	Duplicate URLs	Clean URLs
838	2,610	43,413

Data Collection

The second step of the analysis is to find the candidates for dark patterns. This thesis, like the paper on which it is based[21], focuses only on a textual representation of dark patterns in terms of finding candidates.

A random sample of one hundred records was drawn from the final dataset from the previous chapter. The URLs from this sample were manually visited, and it was found that the linked page was not a web store for six samples, and thirteen were already non-functional URLs. In addition, it was discovered that more of these non-compliant URLs mainly were located at lower positions in the dataset. This finding is not surprising, as large web stores last longer and thus are higher in the list of stores.

This sample is updated and used later in the data collection, where records with non-compliant URLs are replaced with compliant ones for a total of one hundred records.

Searching the candidates for dark patterns is divided into two steps or two crawlers, respectively. The goal of the first step is to find product pages on the webshops in the final list from the chapter Corpus Creation. The purpose of the second step is to capture textual candidates on the found product pages and save them into the SQLite database for further analysis.

These two crawlers are taken from the original research done at Princeton but modified for the purposes of this thesis—i.e. crawling of Czech webshops.

Websites can be client-site rendered, which means that it requires a client (web browser) to render the loaded Javascript scripts. Because of that, the proposed crawlers are based on Selenium. Navigation on the website is done with Javascript. Both crawlers are originally written in Python 2, which has been a deprecated version since January 2020. Only the *Product page crawler* is successfully rewritten to Python 3. The *Checkout crawler* is not because of the complex incompatibilities of libraries and technologies (Selenium, Geckodriver, Firefox) used by the crawler. However, this crawler and its installation had to be modified to work since some used libraries had already stopped supporting Python 2.

4.1 Discovering Product Page URLs

Discovering product URLs is a complex task for three main reasons. Firstly, a classification of whether a page is a product page or not is complex because product pages look different for different webshops, and there is no unified definition of how a web page should look. Also, the HTML source code of the product pages varies a lot.

Secondly, a single website contains many links, and only a tiny portion of them can be actual product pages. Crawling and classifying every page on the website would lead to unnecessary work. Lastly, the crawler must work in parallel on multiple processors to speed up processing the large dataset of webshops obtained in the Corpus Creation step.

The Princeton researchers built a crawler that contains a classifier capable of classifying product URLs from non-product URLs, and the crawler proposed in this thesis is hugely based on it.

However, the original crawler is built to discover product pages on English webshops. It had to be modified to work for Czech webshops. This includes adjusting the classifier to detect the product page URL and modifying the product page detection. Steps of building the classifier and steps of the crawler for discovering product page URL are shown in Figure 4.1

4.1.1 Product Page Detection

A page is classified as a product page if its HTML code contains only one "Add to Cart" button. The detection of such a button is more complicated than it may

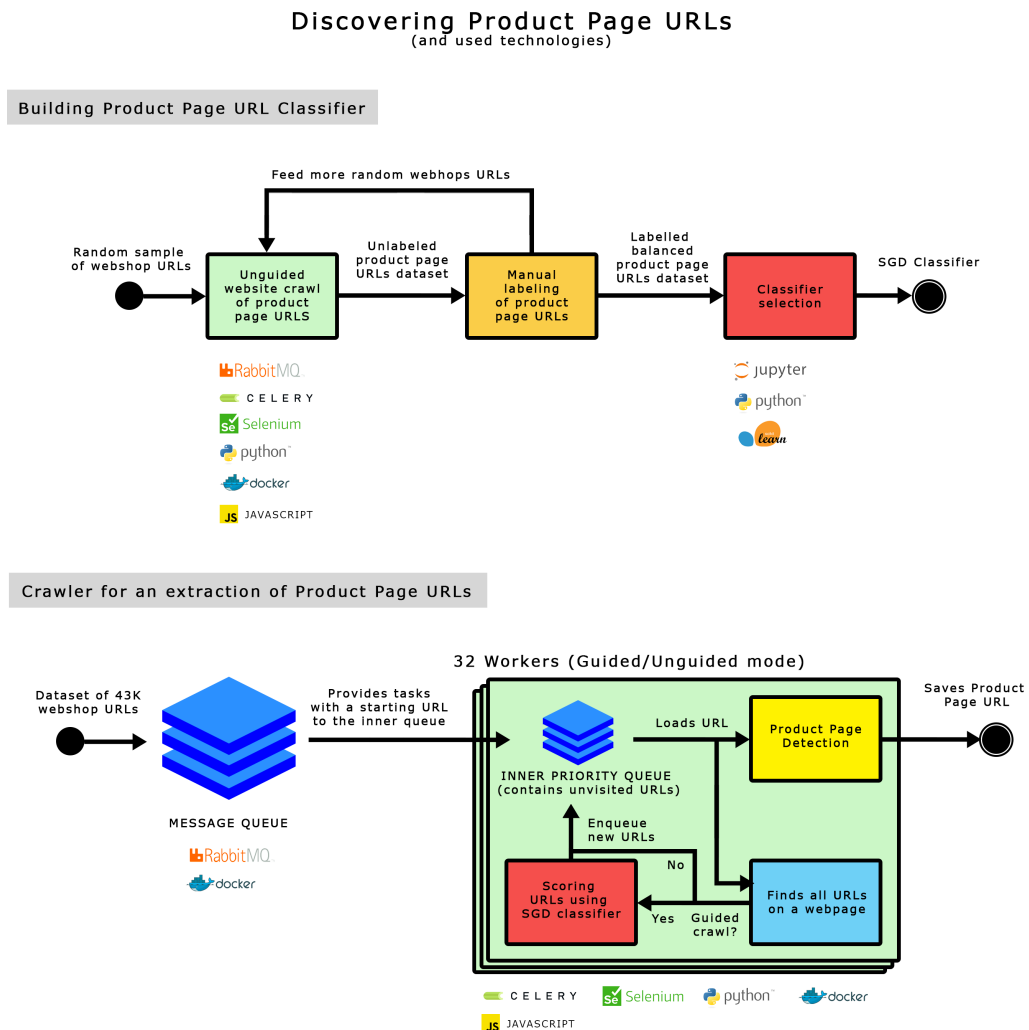


Figure 4.1: The workflow of discovering product page URLs. The crawler can be run in an unguided or a guided mode. The unguided crawl extracts Product Page URLs from a fraction of all Czech webshops. The output is manually labelled and creates a training dataset for the classification model. Further, this model guides the crawler with prioritizing possible Product Page URLs in its inner queue. Therefore, the crawling of a single page is rapidly speeded up.

seem. The researchers implemented complex scoring functions for the detection of such buttons. This includes that the candidates for the "Add to Cart" button are scored not only by the presence of the possible "Add to Cart" phrase (defined in a form of a regular expression) in the inner text or in its attributes but also by the button's size and a contrast ratio of button's colour to the background colour

of a body HTML element. To add support for the Czech language, a sample of 50 pages needed to be analysed for the use of "Add to Cart" phrases (see table 4.2). This analysis led to a modification of the used regular expression.

Table 4.1: "Add to Cart" phrases in the Czech language found on a random sample of 50 Czech webshops.

Phrase	#
Do košíku	19
Přidat do košíku	18
Koupit	7
Vložit do košíku	6

4.1.2 Unguided Crawl

The original classifier distinguishes product page URLs from non-product page URLs that are from English websites. The classifier was trained on a dataset of Czech URLs. This dataset was obtained by running the crawler on a random sample of 100 webshops (mentioned at the beginning of this chapter). The crawler was run to select random URLs to visit while spidering the website instead of predicting which URL was more likely to be a product page. In this random crawl, the crawler's detection marked 398 pages to be product pages. These pages were manually examined, and 308 were actual product pages (77% accuracy for a random crawl). Additional URLs (not marked as product pages) were iteratively added to the dataset and manually examined. Multiple iterations of additions and examination led to a balanced dataset of 377 product pages and 334 non-product pages.

4.1.3 Product page URL Classifier

The classifier is trained in a separate Jupyter Notebook, and it is a modified version of the notebook published by the researchers. This notebook differs from the original notebook in used feature variables, where it adds Czech equivalents to boolean features, representing a specific word in the URL, such as "category" and "product". The Czech counterparts are "kategorie" and "produkt". Czech webshops often use a word "detail" in their URL, which was

added as another boolean feature. The last added feature represents whether or not the URL contains a product ID. The rest of original features are a length of the URL, a number of hyphens and slashes and the longest number in the URL.

The dataset was split where 90% records are used for training and 10% for five-fold cross-validation. Tested classifiers were sklearn's Logistic Regression using L-BFGS solver[20] and Logistic Regression with Stochastic Gradient Descent learning [28]. Both classifiers had very similar results on average, but Stochastic Gradient Descent with 78 % accuracy was chosen as a classifier for the crawler because of its higher validation score (0.83 to 0.76).

However, none of the added feature variables significantly led to an increased accuracy compared to the original features.

4.1.4 Guided Crawl

Once the classifier was trained, the second crawl guided by this classifier was done on the full dataset of 46,023 webshops' URLs. The classifier helps the crawler to rank URLs on the page by likelihood of being product page URLs. The researchers set certain limits for the crawler from their observations, followed in this part as well. The crawler visits 100 pages or spends 15 minutes at maximum on a single website. It does not visit the same page more than two times. The crawling of a single website can be skipped if the crawler has already found five product pages.

A total number of 32 workers finished the crawling in 2 days, 7 hours, 20 minutes. During the crawling, 1,944,980 pages were visited from which 159,768 were identified as product page URLs on 43,411 different webshops. The remaining 2,612 pages were no longer accessible, redirected to a different domain or identified as not being in the Czech language.

4.2 Discovering Textual Segments

Since the guided crawl took a significant amount of time to complete, it was assumed that the crawler simulating purchase flow would take even more time to complete due to its higher complexity. During the manual browsing in the previous steps, it was found that the dataset of Czech webshops also contains websites that do not allow consumers to directly buy goods and they

just present goods. Obviously, the crawler cannot find product pages on such websites.

It can also be assumed that the smaller businesses do not have the funding for their own e-commerce software solution. Because of this, the vast majority of them use third-party solutions. These third-party solutions exhibit the same characteristics, and it cannot be expected from them to contain unique instances of dark patterns across multiple websites.

For these reasons, only the first 10K domains were selected from the list of 43,411 e-shops for the following crawling step. Ten thousand different domains are also comparable to the number of domains used by the Princeton researchers—i.e. 11,286 unique domains. The filtering of the list of given product URLs by domain name returned 33,782 product URLs.

The crawler used in this part of the study is almost identical to the crawler (in their study referred to as Checkout crawler) published by the Princeton researchers.

The crawler performs two concurrent tasks. It is Product Purchase Flow Simulation and Text Extraction using Page Segmentation. The workflow of this crawler is shown in Figure 4.2

4.2.1 Product Purchase Flow Simulation

By analyzing the code for the simulation, it was found that the crawling logic can remain the same because it is independent of the language used. Nevertheless, the original crawler does not support the Czech language. This is crucial in searching appropriate buttons for interactions with web pages. The modified crawler adds Czech language support.

It contains modified regular expressions that are used by scoring functions. The scoring functions are very similar to the one already described in Section 4.1.1. These functions allow the crawler to find the buttons on the website. Because of that, the crawler can run independently on many websites with different designs.

The crawler is able to handle edge cases to simulate user interaction, such as dismissing pop-up dialogues and identifying interfaces for selecting product attributes. Also, the crawler takes a screenshot every second and saves only the

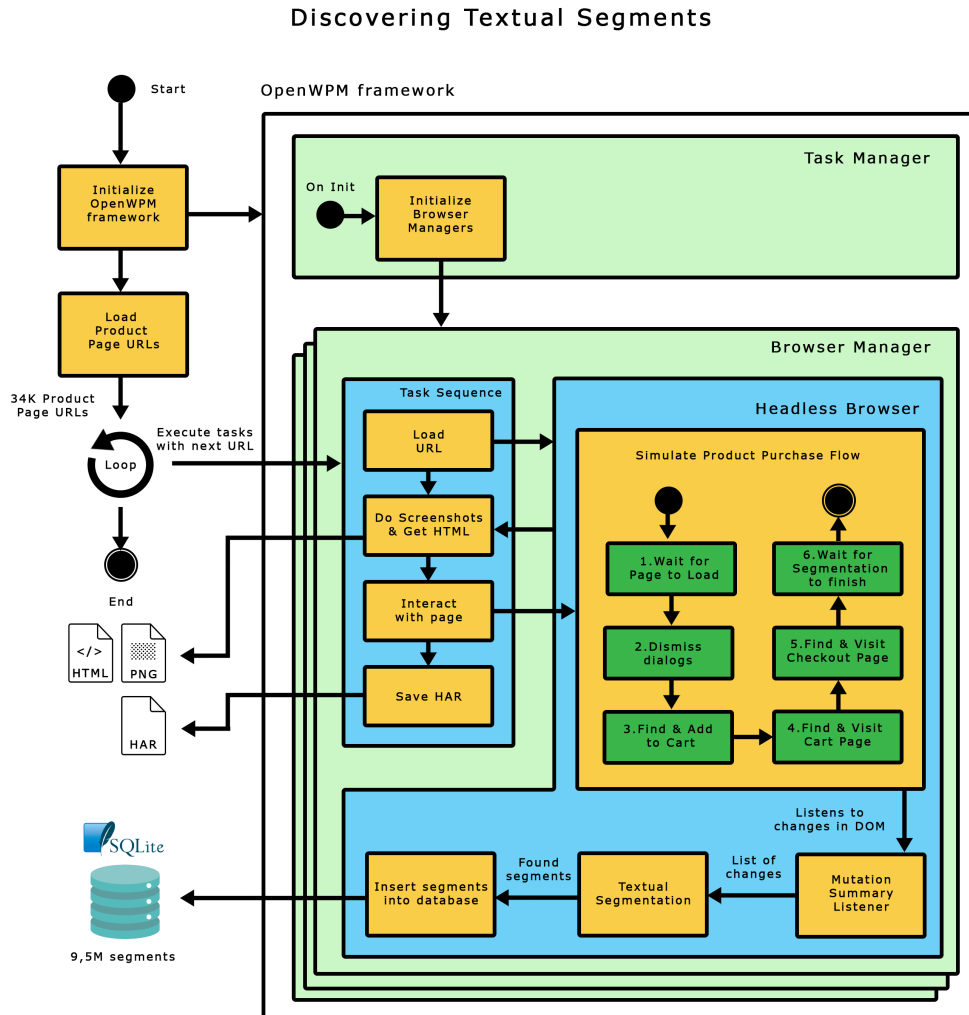


Figure 4.2: The workflow of discovering textual segments from the dataset of Product page URLs. OpenWPM framework creates multiple workers (Browser Managers) and serves them a sequence of tasks they follow. The task manager is capable of orchestrating the workers, that finished previous tasks and are ready for crawling a Product Page URL.

screenshots that differ in base64 form [2] from a previously taken screenshot. The whole communication, every interaction is saved in HTTP Archive (HAR) but is omitted from the analysis done in this thesis.

The product purchase flow consists of four steps:

4.2.1.1 Adding Product to Cart

Firstly, the crawler searches the interface for selecting product variants and selects one of them. This is not a very complex algorithm. It searches for every select HTML element on the page, and the crawler tries to set it to a random value. This was not modified from the original crawler because it is not dependent on the used language. After that, the crawler searches for the "Add to Cart" button. The regular expression defining the possible values was modified to support Czech languages, which is made and described earlier in the section 4.1.1. Once the button is found, the crawler clicks it to add the product to the cart.

4.2.1.2 Searching View Cart

Secondly, the crawler searches for a "View Cart" button in a similar way as it searches for the "Add to Cart" button. The regular expression defining the possible phrases was also modified to support Czech equivalents of "View Cart" and its synonyms. Results from the analysis of one hundred websites have shown that one hundred per cent of Czech e-shops uses the word "košík" (or its declination).

4.2.1.3 Searching Checkout Button

Thirdly, once the product is added to the cart, the crawler locates a button that leads to the final Checkout page.

On a random sample of one hundred product URLs, the following keywords from phrases used to proceed to the Checkout page were found:

Table 4.2: Most common Czech words inside buttons that proceed to the Checkout page. Occurrences were found on a random sample of one hundred Czech webshops.

Phrase	#	Phrase	#
Pokračovat	51	Pokladna	7
Objednávka	25	Přejít	6
Doprava	14	Objednat	5
Platba	13	Dále	1

4.2.1.4 Waiting on Checkout Page

The crawler does no interaction on the checkout page. It only waits 10 seconds, so the page segmentation defined below has enough time to finish.

4.2.2 Text Extraction using Page Segmentation

Since the page segmentation is language-independent, it is used in its original form as it was published by the researchers.

While crawling the website, the crawler extracts page segments of every visited page and from every change that can occur during the crawl. Researchers describe page segments as building blocks of a website, representing smaller sections. These segments shape a dataset later used in the data analysis and clustering. Researchers define[21] a single page segment as a visible HTML element that does not contain block-level element[4] and contains at least one text element[25].

The crawler waits for the web page to load completely before the page segmentation starts. Even after the complete loading of the web page, the browser can render additional content into the DOM. The user's interaction may trigger these changes in the web page, or pop-up windows may appear. To capture all these changes, researchers integrated Mutation Summary [26] library into the crawler. This library observes changes in the DOM and emits events that retrieve all the changes in a form of an array, which is again processed by the page segmentation.

The researchers included a pseudocode of the page segmentation algorithm and an illustration of the algorithm's output in their work[21].

Each segment is stored as one record in the SQLite database, where each document contains the segment's HTML Element type, innerText, dimensions, coordinates on the page and styles (CSS colour and background-color).

From the crawling of the 10K Czech webshops, the crawler visited 33,782 web pages, where it captured 9,5M segments. It can be said that this number of captured segments corresponds to the number of segments the Princeton researchers extracted from the same number of web pages. They extracted over 13M segments from 53K web pages.

Data Analysis

Analyzing millions of segments is not optimal for an expert analyst. The work of this expert is expensive and time-consuming. Therefore, methods that reduce the number of segments and thus the expert's work need to be used to make the analysis manageable for the expert. The output is a list of text segments that contain dark patterns.

The methods used in this section follow the work of the Princeton researchers. At the same time, the results of this work are again compared with the results of their work.

The data analysis can be divided into four steps:

5.1 Preprocessing

The SQLite database, which has 9.5 million segments, contains many duplicate segments across multiple websites. For example, "Add to Cart" buttons, various unified headings such as "Product Description", and others. Since only the text of the segments is analyzed, only those segments that have unique text across a single domain are selected from the dataset for further processing. Also, all the numbers in the dataset have been replaced with placeholders, thus reducing the dataset even more.

The output of this preprocessing is a reduction in the number of segments from 9.5M segments to 805K. Thus, this approach led to a 92% reduction in the number

of segments. Again, the results are similar to those of the Princeton researchers who achieved a 90% reduction.

5.2 Feature processing

In order to be able to use clustering in the next step, the texts of the segments must first be transformed into a representation for which the similarities between the segments can be expressed mathematically (hereafter, the document means the internal text of the segment). For this purpose, the Bag-of-Words model is used here. This model is a type of word embedding that represents a document as a string of the number of occurrences of words from a dictionary of all words used across all documents.

However, many words do not have only one base form. Especially in the English language, a single word can have many forms due to inflections such as declension and conjugation. The basis of the Bag-of-Words model is the previously mentioned dictionary. If that dictionary contained all the occurrences of the different forms of words, the dictionary would be unnecessarily large and inefficient. For example, the distances of two very similar documents could be disproportionately large simply by rewriting them in a different tense.

This mischief can be avoided by stemming, or lemmatisation, where stemming returns the roots of words. Lemmatization produces the basic forms of words (infinitive for verbs and first-person singular for nouns, adjectives, pronouns and numerals). Lemmatization also considers the context of the word and is, therefore, more accurate[**stemming-and-lemmatisation**]. On the other hand, it is slower than stemming.

The Princeton researchers used stemming from the NLTK Python library[3]. Still, because both methods (stemming and lemmatisation) depend on the language and because the NLTK library does not support the Czech language, another library had to be chosen.

Such a library is UDPipe[32] by the Institute of Formal and Applied Linguistics at Charles University. Also, one of the functionalities of this library is tokenisation in the Czech language, which is needed to split the documents into individual words (also referred to as tokens)[**tokenisation**].

Each document is tokenised during the dictionary creation process, producing a list of tokens for which lemmas are obtained and then added to the dictionary. Also, stop words from the Czech language and punctuation are filtered out of these lists.

The vocabulary after all the described steps above had a size of 269K tokens. However, this vocabulary still contained tokens, which did not have enough occurrences in the documents.

Furthermore, only those that appeared in the documents at least 100 times were selected. There were only 188 such tokens. The Count Vectorizer[10] was used to create the BoW matrix, which counts the number of token occurrences in a document.

Using Principal Component Analysis (PCA) with three retained components on the BoW matrix led to a dimensional reduction which captured 95% of the variance in the data.

5.3 Clustering

The goal of clustering is to group data together. In this case, it means clustering segments into clusters based on similarity. The expert then evaluates the resulting clusters, which makes the expert's job of manual passes easier.

The clustering method used was HDBSCAN (Hierarchical Density-based Spatial Clustering of Applications with Noise)[7]. According to the Princeton researchers, they selected this clustering algorithm because it is robust to noise and, in particular, allows to choose the minimum size of the output clusters.

In total, HDBSCAN was performed for four different hyperparameter settings. The number of output clusters and the size of the noise cluster was analysed. The metric used and the minimum cluster size mentioned earlier were the hyperparameters varied. The metrics used were L_1 and L_2 norms, also known as Manhattan and Euclidean distance. The hyperparameter of the minimum cluster sizes selected was 5 and 10 segments, which keep the size of noise small and prevents two or more clusters (that are separable) from forming only one.

The analysis showed the number of clusters is significantly lower for the models with a minimum cluster size of 10 segments. Similarly, as for the results from Princeton researchers, the difference between selected metric distances was not very significant for data. As expected, models with a larger minimum cluster size have a larger noise cluster size. However, this noise cluster is slightly less than 50% larger, while the number of all clusters is twice as small. Therefore, a model with a minimum cluster size of 5 segments was selected using the Manhattan distance as the metric with 4,248 clusters (one cluster is the noise cluster). The table 5.1 summarised the number of clusters and size of noise for the given hyperparameters.

Table 5.1: Number of clusters and size of noise cluster for different distance metrics and minimum size of a cluster.

Minimum cluster size	5		10	
	L1	L2	L1	L2
Distance metric				
Number of clusters	9,040	9,088	4,249	4,265
Size of noise cluster	80,980	80,083	98,436	97,651

5.4 Analysis of output clusters

The clusters that were obtained in the previous step are manually scanned in two steps.

In both passes, I put myself in the role of an expert who evaluates what is and what is not a dark pattern. I used the knowledge I gained from writing the Dark patterns section. I also used available literature [14][16][19][5][9]. In uncertainty, I also used the Internet to find out examples what is and what is not a dark pattern, to keep my decisions even more objective. However, the subjective component could still play a role in the decision making process.

In the first pass, I selected those clusters for which any segment could manifest a dark pattern. For example, the selected clusters were commonly countdowns, total cart prices, user references, notifications, product options, logins and registrations. Only the text components of the segments were checked, not

how the segment actually looks on the page. This pass resulted in the number of clusters being reduced from 4,249 to 477.

In pass two, I investigate these 477 clusters by directly visiting the website where the dark pattern is searched. If the page no longer exists or does not match the segment, then I investigated screenshots that were obtained during the simulated purchase flow instead. I extended this search by manually going through the entire shopping process directly on the web page and manually searching for all dark patterns.

Lastly, this output dataset of found dark patterns is examined and cleaned from duplicities.

5.5 Results

At first, it is important to mention the limitations of the analysis and the evaluation.

Limitations

As mentioned in the previous chapter, deciding what is still a harmless user interface pattern and what is already a dark pattern is a complex task. It always depends on the subjective opinion of the expert. At the same time, different types of dark patterns affect each customer differently. Several steps were done as described earlier in Analysis of output clusters in order to make the analysis to be more objective. Also, more experts analysing the clusters would lead only to more objective results.

Another limitation of this work is that it considers only the textual segments on the page and ignores the appearance of these segments. Thus, it cannot find dark patterns in images, for example.

During the analysis, it was also found that some webshops use dark patterns to increase their number of references on Heureka.cz. This behaviour could significantly affect their position in the overall ranking, especially for the lower-ranked webshops, where a small number of references can make a significant change in the rank.

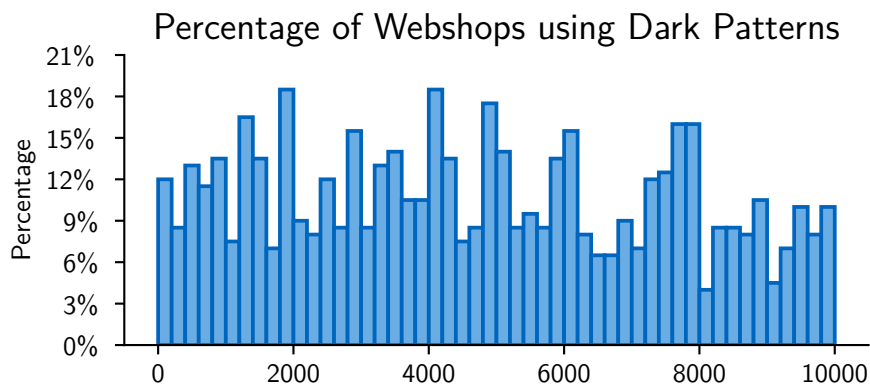


Figure 5.1: Distribution of webshops using at least one Dark Pattern over the ranking in Heureka's webshop list. Each bin is a size of two hundred webshops, representing a percentage prevalence of webshops containing dark patterns within the bin.

Lastly, not every webpage was a part of the analysis due to simulation of the checkout flow only. Homepage, listings of products, registration page, login page, and payment page were omitted in the flow.

5.5.1 Webshops using Dark patterns prevalence

A total number of 1,419 dark patterns were found on 1,081 webshops from a total of 10K webshops, which makes 10.81% of all webshops to contain at least one instance of Dark Pattern. No dark patterns of the Hidden Costs, Confirmshaming and Hidden Subscription types were found during the manual passes. The found instances of Dark Pattern are divided into categories and types shown in table 5.2.

It can be seen in the figure 5.1 that the position in the Heureka's ranking of Czech webshops has a slightly negative correlation to the number of webshops in each bin, using at least one dark pattern. Using a Spearman's ρ to test the monotonicity of a probability of finding a dark pattern on a webshop is considered as statistically dependent on the position in the Heureka's ranking—i.e webshops higher in the ranking are more likely to use dark patterns (Spearman's $\rho = -0.291$, p-value = $0.040 < 0.05$).

Also, a number of instances over the ranking for every defined dark pattern type is shown in figure 5.2. The monotonicities of the dark pattern types were not tested because of a low number of instances for some types of dark patterns.

Table 5.2: Number of Dark patterns instances found on Czech webshops, divided into categories and types.

Category	Type	# Instances
Sneaking	Sneak into Basket	2
	Hidden Costs	0
	Hidden Subscription	0
Urgency	Countdown Timer	23
	Limited-time Message	17
Misdirection	Confirmshaming	0
	Visual Interference	28
	Trick Questions	68
	Pressured Selling	924
Social Proof	Activity Message	223
	Testimonials	19
Scarcity	Low-stock Message	38
	High-demand Message	7
Obstruction	Hard to Cancel	6
Forced Action	Forced Enrollment	75

However, it can be seen that all of the dark pattern types were used over the whole spectrum of Heureka's ranking, and none type was used only by the top webshops.

5.5.2 Used E-commerce solutions

Utilizing the segments gathered from the page segmentation, stored in the SQLite database, it is possible to select webshops that use an e-commerce solution for the Czech market. Usually, the companies providing such solutions include their names in the footer of the websites. The biggest and very frequently used

Instances of Dark Pattern types over the Heureka's ranking

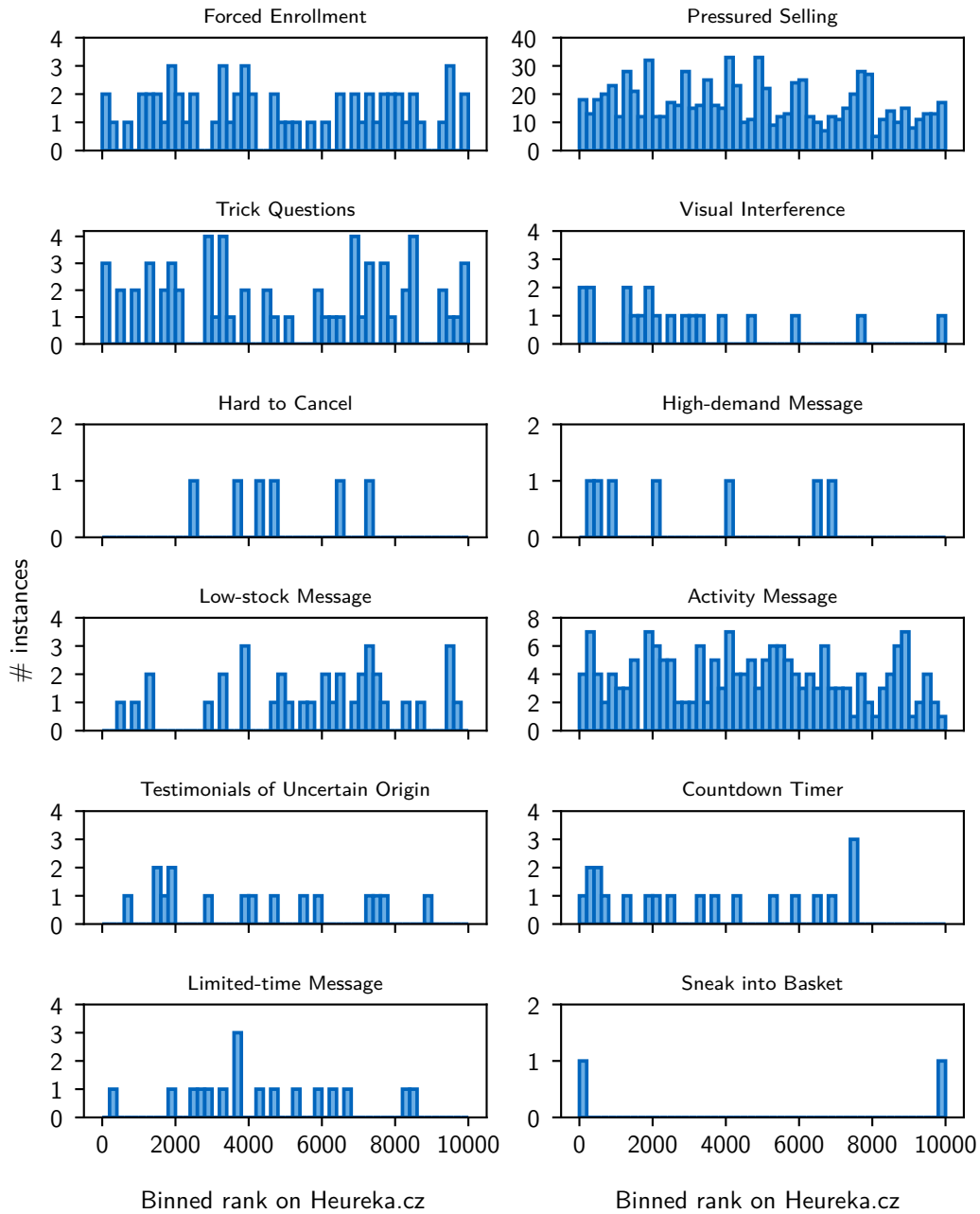


Figure 5.2: Distribution of different types of dark patterns over the ranking in Heureka's webshop list. Each bin is a size of two hundred webshops, representing a number of dark patterns of the type within the bin.

solutions are recommended in multiple articles on the Internet[31][34][23][22][27]. By reading these articles, five such solutions were chosen for testing their market share in the top 10K webshops in Heureka's ranking. Table 5.4 shows total numbers of webshops using these solutions. Figure 5.3 shows the distribution of these five e-commerce solutions over Heureka's rank.

A total number of 2.37K webshops use one of the defined e-commerce solutions, which is 23.7% out of all tested webshops. As expected, Shoptet.cz is the most used one with a share of 68%.

Table 5.3: Numbers and percentages of five selected e-commerce solutions used by webshops. The numbers refer to the first 10K webshops selected from Heureka's ranking.

Solution	# of websites	Share [%]
Shoptet.cz	1,618	68.2
Eshop-rychle.cz	570	24.0
FASTCentrik.cz	103	4.3
Upgates.cz	60	2.5
Webnode.cz	22	0.9
Total	2373	100

5.5.3 Notifikuj.cz as a third-party dark pattern provider

Compared to the English Internet, which has many providers of Dark Patterns as a Social Proof notification[21], the Czech Internet has only one such provider. The provider's name is Notifikuj.cz. During the analysis, no other provider was found to be used; even searching the Internet did not lead to discovering any other provider.

Notifikuj.cz provides dark patterns in the form of push notifications on the e-shop website. By analyzing these notifications obtained from the segment database and studying the Notifikuj.cz website, it was found that the provider allows inserting a total of five notification patterns. Four of these notifications are Activity Message dark patterns. The last notification shows ratings and

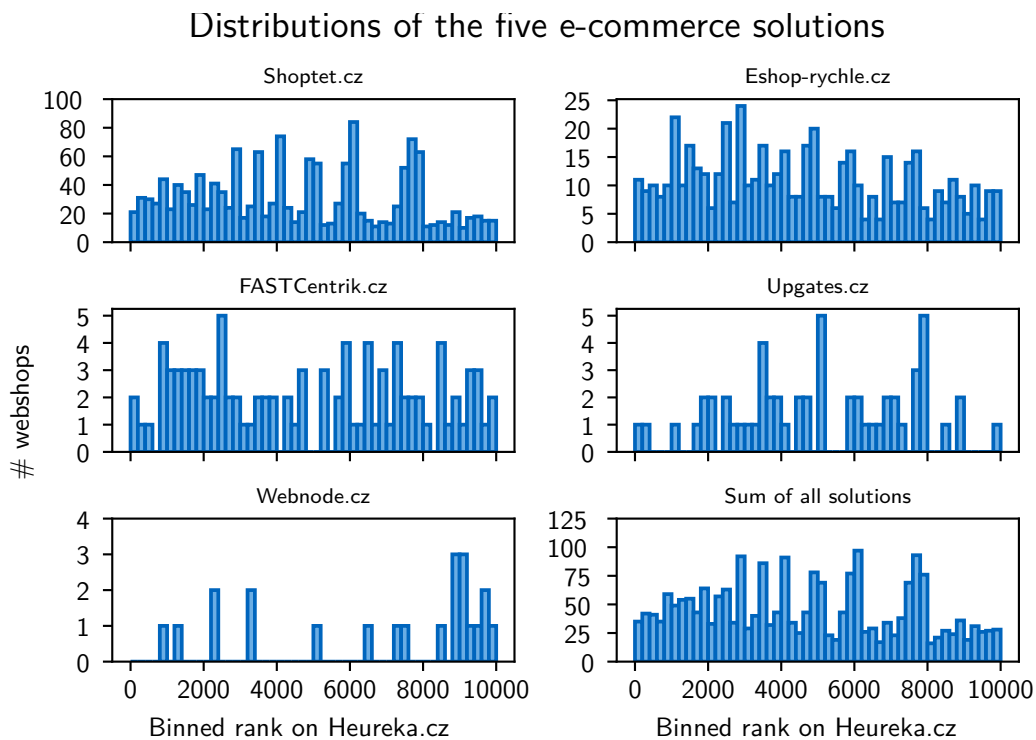


Figure 5.3: Distributions of the five most used e-commerce solutions over Heureka's rank with a distribution of a sum of them all. Each bin is a size of two hundred webshops.

references obtained on Heureka.cz and Zbozi.cz. It should be noted that the displayed ratings and references are accurate and not deceptive. Notifikuj.cz aggregates real data from Heureka.cz and Zbozi.cz, which can be found directly from these sites. Therefore, it is not a type of dark pattern Testimonials of Uncertain Origin.

It was found that 55 webshops use service Notifikuj.cz. As it is expected, more webshops that rank higher use Notifikuj.cz more frequently than lower-ranked webshops. This can be seen in figure 5.4. However, this claim is not supported by strong evidence because there is simply not enough data.

5.5.4 Frequently used Dark Patterns on Czech Internet

During the analysis, three techniques manifesting dark patterns occurred very frequently on the webshops. Their prevalence in the resulting dataset of 1419

Table 5.4: Types of discovered Dark Patterns on the five most-used Czech e-commerce solutions.

	Shoptet.cz	Eshop-rychle.cz	ASTCentrik.cz	Upgates.cz	Webnode.cz
Sneak into Basket	✗	✗	✗	✗	✗
Hidden Costs	✗	✗	✗	✗	✗
Hidden Subscription	✗	✗	✗	✗	✗
Countdown Timer	✓	✗	✗	✗	✗
Limited-time Message	✓	✗	✗	✗	✗
Confirmshaming	✗	✗	✗	✗	✗
Visual Interference	✓	✗	✗	✗	✗
Trick Questions	✓	✗	✓	✓	✗
Pressured Selling	✓	✓	✓	✓	✗
Activity Message	✓	✓	✓	✓	✗
Testimonials	✗	✗	✗	✗	✗
Low-stock Message	✓	✗	✗	✗	✗
High-demand Message	✓	✗	✗	✗	✗
Hard to Cancel	✗	✗	✗	✗	✗
Forced Enrollment	✓	✓	✓	✓	✗

dark patterns was examined. It was found that these three dark patterns are represented in the dataset by 894 instances, which makes 63% of the dataset.

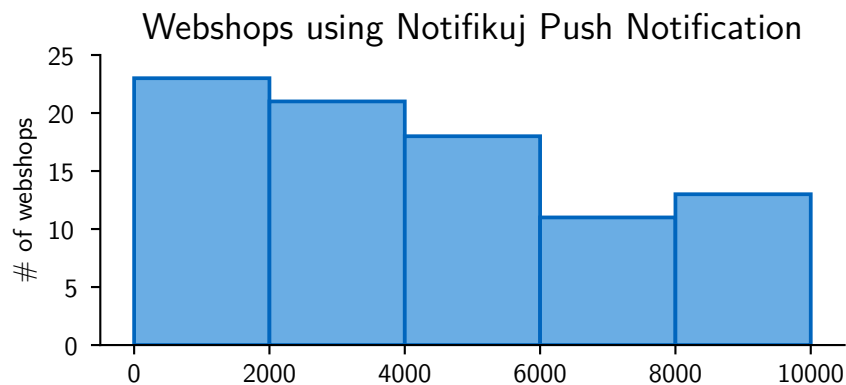


Figure 5.4: Distribution of webshops using Notifikuj.cz service of push notifications over the dataset of 10K highest-ranked webshops in Heureka's ranking.

This section further describes the prevalence of these four dark patterns in the dataset. Examples are shown in figures.

Cross-selling

The most prevalent is cross-selling, a technique when a webshop recommends additional products to purchase. Webshops often claim that other users have also bought these products or additional products could come in handy. The dataset of dark patterns contains 695 instances of cross-selling.

These instances are of pattern type "Pressured Selling". It is important to mention that during the analysis, an instance of cross-selling was flagged as manifesting the dark pattern when the instance was pushing to purchase additional products. For example, this was usually done via pop-up windows or by recommending additional products in further steps of the purchase process. Figure 5.5

Free Shipping

Many webshops allow free shipping after certain criteria are met. Such a criterion is the price of the purchase. This can affect a customer's judgment of spending more money on products he initially did not want, only to have free shipping (Some webshops even offer gifts for free). In addition, webshops make users very aware of the need to purchase more products to get free shipping. In total, 132 instances of this dark pattern are present in the dataset of dark patterns.

Ostatní zákazníci také
nakoupili



436 Kč

> Do košíku

Figure 5.5: An example of cross-selling as "Pressured Selling" dark pattern found on webshop beason.cz. This dark pattern appears in a pop-up window immediately after users add a product to a cart. "Ostatní zákazníci také nakoupili" can be translated into English as "Other customers also purchased".

As the cross-selling above, this is also a type of "Pressured Selling" dark pattern, and both are often used together. Figure 5.6 shows an instance of this dark dark pattern.



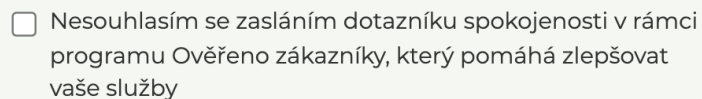
Objednejte ještě za **900 Kč** a budete mít dopravu **ZDARMA**.

NÁKUPNÍ KOŠÍK

Figure 5.6: An example of dark pattern "Pressured Selling" found on beason.cz. This dark pattern offers free shipping if a customer purchases for higher price. "Objednejte ještě za 900 Kč a budete mít dopravu ZDARMA" can be translated as "Order for another 900 CZK and you will get free shipping".

Heureka's Satisfaction Survey

High credibility is very important for webshops[36]. For Czech webshops, good reviews on the Heureka portal can earn a certain amount of credibility. In addition, if a webshop is involved in the "Approved by Customers" program (which adds even more credibility), the review can be uploaded to the portal only by accessing it via a special link. This link is sent to the customer's e-mail if he/she gives his/her consent. However, this consent is often unconscious, as it manifests the "Trick Questions" dark pattern. In this case, it uses a double negation in the sentence, as can be seen in Figure 5.7.



Nesouhlasím se zasláním dotazníku spokojenosti v rámci programu Ověřeno zákazníky, který pomáhá zlepšovat vaše služby

Figure 5.7: An example of "Trick Questions" dark pattern, which uses double negation in the sentence. The user may think that he is not giving his consent to the webshop for sending satisfaction surveys by not checking the checkbox. "Nesouhlasím se zasláním ..." can be translated as "I do not agree with ..."

Conclusion

This thesis described known dark patterns, and examples were found from the Czech Internet. The taxonomy of these dark patterns was also described and the effects that dark patterns use in users' cognitive biases.

Automated web crawlers were created to mine the Heureka page. This created a dataset that contains a large fraction of Czech webshops and their locations, which corresponds to the approximate size and popularity of the webshop. This dataset was also cleaned of no longer active webshops and duplicates.

The original crawlers developed by Princeton researchers to retrieve product pages and to simulate the shopping process were modified to work for Czech webshops.

For the first crawler, it was necessary to manually crawl Czech webshops and create a balanced dataset of URLs that are and are not product pages. This data was then used during the learning phase of the classification model, which was used to find even more product pages.

In the case of the second crawler, it was again necessary to manually go through the shopping process on the webshops. The most common Czech phrases in the buttons used to navigate this shopping process's steps were extracted. These phrases were used to modify the crawler, which can now simulate the shopping process on Czech webshops. Many screenshots, HTML and HAR files of individual pages were saved. All page segments of the web pages were also saved.

CONCLUSION

These extracted segments were first clustered using machine learning methods. This reduced millions of segments into thousands of clusters. These clusters were then manually crawled in two passes. The first pass selected those clusters that were suspicious of the possibility of a dark pattern. This approach reduced the number of clusters to hundreds. During the second pass, websites from these clusters were directly visited or screenshots previously obtained were examined.

In total, 1,419 dark patterns were discovered on 1,081 of the 10K webshops crawled. Thus, at least one instance of a dark pattern was found on approximately 10.81% of all webshops. The found instances were categorized in the types of dark patterns. It was also found that larger webshops use dark patterns more often.

The evaluation also included whether the webshops were built on one of the five largest Czech e-commerce solutions for webshop development. This revealed that approximately 23.7% of webshops are built on one of these five solutions. It was also found out which solutions actively use which types of dark patterns.

Another finding was the analysis of a service that provides dark patterns in the form of push notifications. This service was found on 0.55% of all webshops, and apparently, larger webshops use this service more often.

Lastly, three frequently used dark patterns found on Czech webshops were described, and their examples were shown. Instances of these three dark patterns make 63% of the whole dataset of all instances.

Much of the original code has been rewritten to Python 3, making it easier to use in the future.

All the datasets, screenshots, HTML and HAR files obtained can be further researched. The model for classifying production pages can also be used in other crawlers. For example, it can be used by a crawler to easily find product pages of competing webshops or for a prices aggregator. Also, the extracted dataset of dark patterns can be used to create a web browser add-on that will alert users to dark patterns on a page.

Bibliography

1. ALEXA INTERNET, Inc. *The top 500 sites on the web* [online]. 2021 [visited on 2021-05-10]. Available from: <https://www.alexa.com/topsites>.
2. *Base64* [online]. 2021 [visited on 2021-10-15]. Available from: <https://developer.mozilla.org/en-US/docs/Glossary/Base64>.
3. BIRD, Steven; KLEIN, Ewan; LOPER, Edward. *Natural Language Processing with Python*. O'Reilly Media, 2009.
4. *Block-level elements* [online]. 2021 [visited on 2021-10-16]. Available from: https://developer.mozilla.org/en-US/docs/Web/HTML/Block-level_elements.
5. BÖSCH, Christoph; ERB, Benjamin; KARGL, Frank; KOPP, Henning; PFATTHEICHER, Stefan. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies*. 2016, vol. 2016, no. 4, pp. 237–254. Available from DOI: 10.1515/popets-2016-0038.
6. *California bans 'dark patterns' that trick users into giving away their personal data* [online]. 2021 [visited on 2021-03-30]. Available from: <https://www.theverge.com/2021/3/16/22333506/california-bans-dark-patterns-opt-out-selling-data>.
7. CAMPELLO, Ricardo J. G. B.; MOULAVI, Davoud; SANDER, Joerg. Density-Based Clustering Based on Hierarchical Density Estimates. In: PEI, Jian; TSENG, Vincent S.; CAO, Longbing; MOTODA, Hiroshi; XU, Guandong

BIBLIOGRAPHY

- (eds.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172. ISBN 978-3-642-37456-2.
8. CIALDINI, Robert B. *Influence: Science and practice*. Pearson education Boston, MA, 2009.
 9. CONTI, Gregory; SOBIESK, Edward. Malicious Interface Design: Exploiting the User. In: *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 271–280. WWW '10. ISBN 9781605587998. Available from DOI: 10.1145/1772690.1772719.
 10. *CountVectorizer* [online]. 2021 [visited on 2021-12-16]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.
 11. CRESTODINA, Andy. *The Long Goodbye: 7 Sites That Make It Hard to Unsubscribe* [online]. 2016 [visited on 2021-04-10]. Available from: <https://unbounce.com/conversion-rate-optimization/when-friction-is-good/>.
 12. *Dark Patterns* [online]. 2010 [visited on 2021-03-28]. Available from: <https://www.darkpatterns.org/>.
 13. *Dark Patterns - About us* [online]. 2010 [visited on 2021-04-01]. Available from: <https://www.darkpatterns.org/about-us>.
 14. *Dark Patterns - Types of Dark Pattern* [online]. 2010 [visited on 2021-04-05]. Available from: <https://www.darkpatterns.org/types-of-dark-pattern>.
 15. GRAUER, Yael. *Dark Patterns are designed to trick you (and they're all over the Web)* [online]. 2016 [visited on 2021-04-05]. Available from: <https://arstechnica.com/information-technology/2016/07/dark-patterns-are-designed-to-trick-you-and-theyre-all-over-the-web/>.
 16. GRAY, Colin M.; KOU, Yubo; BATTLES, Bryan; HOGGATT, Joseph; TOOMBS, Austin L. The Dark (Patterns) Side of UX Design. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. CHI '18. ISBN 9781450356206. Available from DOI: 10.1145/3173574.3174108.

17. HEJNÁ, Veronika. *Konec falešných hodnocení. E-shopy musí ukázat, proč jim věřit* [online]. 2020 [visited on 2021-04-20]. Available from: <https://www.penize.cz/ochrana-spotrebitele/414983-konec-falesnych-hodnoceni-e-shopy-musi-ukazat-proc-jim-verit>.
18. KASÍK, Pavel. *Alza to zase zkouší. Zákazníkům do košíku „tajně“ přihodí nechtěné věci* [online]. 2018 [visited on 2021-04-16]. Available from: https://www.idnes.cz/technet/internet/nenapande-prihazovani-zbozi-do-kosiku.A181210_135725_sw_internet_pka.
19. KYSAR, Douglas; HANSON, Jon. Taking Behavioralism Seriously: A Response to Market Manipulation. *SSRN Electronic Journal*. 2001, vol. 6. Available from DOI: 10.2139/ssrn.265656.
20. *LogisticRegression* [online]. 2020 [visited on 2021-06-03]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
21. MATHUR, Arunesh; ACAR, Gunes; FRIEDMAN, Michael; LUCHERINI, Elena; MAYER, Jonathan; CHETTY, Marshini; NARAYANAN, Arvind. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 2019, vol. 1, no. CSCW.
22. *Nejlepší e-shop řešení 2021 - recenze (pronájem eshopu)* [online] [visited on 2021-12-23]. Available from: <https://entuzio.cz/e-shop-platformy/>.
23. *Nejlepší eshopová řešení v roce 2021* [online] [visited on 2021-12-23]. Available from: <https://www.nastrojeproweb.cz/clanky/eshopova-reseni>.
24. NODDER, C. *Evil by Design: Interaction Design to Lead Us into Temptation*. Wiley, 2013. ISBN 9781118422144. Available also from: https://books.google.cz/books?id=ytwgZ%5C_QELT4C.
25. *Node.nodeType* [online]. 2021 [visited on 2021-10-16]. Available from: <https://developer.mozilla.org/en-US/docs/Web/API/Node/nodeType>.
26. *rafaelw/mutation-summary* [online]. 2015 [visited on 2021-12-14]. Available from: <https://github.com/rafaelw/mutation-summary>.
27. *Recenze e-shopových řešení* [online] [visited on 2021-12-23]. Available from: https://compari.cz/eshopova-reseni/#3_nejlepsi_e-shopove_reseni.

BIBLIOGRAPHY

28. *SGDClassifier* [online]. 2020 [visited on 2021-06-03]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html.
29. SHOPTET, a.s. *Stav e-commerce v ČR v roce 2021* [online]. 2021 [visited on 2021-05-12]. Available from: <https://www.ceska-ecommerce.cz/>.
30. SKLENSKÝ, Martin. *Česká e-commerce stále roste. Vládne jí pětice obřích e-shopů* [online]. 2018 [visited on 2021-05-13]. Available from: <https://www.peak.cz/ceska-e-commerce-stale-roste-vladne-petice-obrich-e-shopu/2492/>.
31. *Srovnání 5 nejlepších nástrojů na tvorbu e-shopu* [online] [visited on 2021-12-23]. Available from: <https://www.5nej.cz/srovnani-e-shopovych-reseni/>.
32. STRAKA, Milan; STRAKOVÁ, Jana. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 88–99. Available also from: <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>.
33. *The Year Dark Patterns Won* [online]. 2010 [visited on 2021-03-28]. Available from: <https://www.fastcompany.com/3066586/the-year-dark-patterns-won>.
34. *Velké srovnání TOP e-shopových řešení!* [Online] [visited on 2021-12-23]. Available from: <https://www.lupa.cz/pr-clanky/velke-srovnani-top-e-shopovych-reseni/>.
35. *Wayback Machine* [online]. 2016 [visited on 2021-04-23]. Available from: <https://web.archive.org/web/20161129023949/https://www.alza.cz/alzapremium>.
36. *Why are Customer Reviews so Important for Your Online Shop?* [Online] [visited on 2022-01-02]. Available from: <https://business.trustedshops.com/blog/customer-reviews>.

37. ZHANG, Yin; JIN, Rong; ZHOU, Zhi-Hua. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*. 2010, vol. 1, pp. 43–52. Available from DOI: 10 . 1007 / s13042-010-0001-0.

List of Acronyms

DP	Dark Pattern
HTTP	Hyper Text Transfer Protocol
HAR	HTTP Archive
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheets
API	Application Programming Interface
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Application with Noise
BoW	Bag of Words
PCA	Principal Component Analysis
URL	Uniform Resource Locator
CSV	Comma-separated Values
DOM	Document Object Model
L-BFGS	Limited Memory Broyden–Fletcher–Goldfarb– Shanno algorithm
SGD	Stochastic Gradient Descent

Supplemental Material

The source code of the thesis and the implementation can be found on the attached medium or online at GitHub.

Github Repository <https://github.com/Lznah/DarkPatterns>

	README.md	a brief contents description	
	MT_Petr_Hanzl_2022.pdf	thesis text in PDF format	
	data/	folder with gathered datasets	
	docker/	folder with supporting files for Docker	
	src/	source code of the practical part of this thesis	
		classifier/	folder with the trained SGD classifier
		crawler/	folder with all crawlers
		analysis/	folder with Jupyter notebooks used during analysis
	thesis/	the directory of \LaTeX source codes of the thesis	

Directory structure B.1: Contents of the attached medium