DETECTION, PREVALENCE AND USER PERCEPTION OF DARK PATTERNS IN E-COMMERCE OF BANGLADESH

YASIN SAZID
Exam Roll: 2323202
Session: 2022-23
Registration Number: 2017-515-171

A Thesis
Submitted to the Master of Science in Software Engineering Program Office
of the Institute of Information Technology, University of Dhaka
in Partial Fulfillment of the
Requirements for the Degree

MASTER OF SCIENCE IN SOFTWARE ENGINEERING

Institute of Information Technology
University of Dhaka
DHAKA, BANGLADESH

DETECTION, PREVALENCE AND USER PERCEPTION OF DARK PATTERNS IN E-COMMERCE OF BANGLADESH

YASIN SAZID

Approved:

Signature *Date*

31.07.2024

Supervisor: Dr. Kazi Muheymin-Us-Sakib

To every common man who fought for justice
For the bravery, the perseverance, and everything in between

Abstract

Dark patterns are deceptive user interface design tactics that abuse knowledge of human psychology to manipulate user decision-making, causing financial loss, oversharing of personal data, and induction of addiction. Most research on dark patterns has focused on Western contexts, but examining them in different cultural settings can produce novel insights. In developing countries like Bangladesh, rapid digitalization and low technological awareness make users particularly vulnerable to these tactics. Thus, analyzing dark patterns in such settings is crucial for informing local policy-making and safeguarding users in digital platforms.

This research investigates dark patterns in Bangladeshi e-commerce websites. A novel automated detection technique for dark patterns was developed first utilizing GPT-3's in-context learning, which successfully detects six out of seven common categories of dark patterns and demonstrates improved generalizability over other methods. This technique was then applied to extract dark pattern texts from 715 member websites of the E-Commerce Association of Bangladesh (e-CAB). Additionally, a user perception study surveyed Bangladeshi university students to assess their exposure, awareness, and concern regarding dark patterns.

The findings reveal that dark patterns are prevalent in approximately 18.3% of the analyzed websites, with 931 instances belonging to five common categories. The user perception survey indicates a general awareness and concern about dark patterns, particularly among students with a background in technology education. A novel categorization of dark patterns distinguishes between 'Passive Dark Pat-

terns' and 'Active Dark Patterns', each with distinct characteristics in terms of prevalence and user perception. These findings highlight the need for policy and regulatory measures to protect vulnerable users on digital platforms of Bangladesh.

Acknowledgments

I am grateful to my parents and sister for always supporting me through the worst and the best of times. My knowledge is an extension to your knowledge. Without you, nothing would have been possible.

I want to thank my wife for always making me feel confident and safe. You have helped in ways that even you yourself would not understand. You are the reason behind my sanity. I am proud of who you are; your kindness, generosity, sincereness, intellect, and most importantly your love.

I want to express my gratitude and utmost respect to my supervisor, Professor Dr. Kazi Muheymin-Us-Sakib, Institute of Information Technology, University of Dhaka. I have worked with you for so long that your impact on my life has become immeasurable. Without your support, guidance, and most importantly kindness, this thesis would not have seen the light of day.

I am thankful to my friends, seniors, and juniors from Institute of Information Technology, University of Dhaka. You have always treated me with kindness and given me confidence in my endeavours. I will always keep fighting to repay your blessings on me and strive to become the best version of myself.

I thank the Distributed Systems and Software Engineering (DSSE) research group of Institute of Information Technology, University of Dhaka. Special thanks to Noshin Tahsin Saaj and Nafis Fuad for their unconditional help and support.

I appreciate the researchers who previously worked on dark patterns. They are the reason I got exposed to such an interesting concept. This thesis is supported

by their research in every way possible. I do not know if I have done justice to you. But I will keep trying to improve myself.

Finally, I express my sincere gratitude to the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh; for granting me fellowship no.: 56.00.0000.052.33.001.23-09, dated 04.02.2024 and funding this research.

List of Publications

1. Y. Sazid, M. M. N. Fuad, and K. Sakib, "Automated Detection of Dark Patterns Using In-Context Learning Capabilities of GPT-3," in *Proceedings of the 30th Asia-Pacific Software Engineering Conference (APSEC)*, 2023, pp. 569-573.

2. Y. Sazid and K. Sakib, "Prevalence and User Perception of Dark Patterns: A Case Study on E-Commerce Websites of Bangladesh," in *Proceedings of the 19th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*, 2024, pp. 238-249.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dark patterns (DP) represent deceptive user interface design tactics that are strategically crafted to exploit human psychology and manipulate user decision-making processes [1]. Figure 1.1 illustrates such a dark pattern called 'Urgency', where a countdown timer is used to pressure users into accepting specific deals. These design practices, which abuse knowledge of psychology, have raised ethical questions regarding user welfare.

Dark patterns constitute a relatively new field of research and there are opposing perspectives about the ethical responsibilities of user interface designers. However, it is clear that discussions on dark patterns are important to safeguard the interests of digital platform users. As a result, researchers have worked on the definition and classification of dark patterns [1, 2], explored legal and policy-making perspectives [3], and studied user perception of dark patterns [4]. Detection techniques [5, 6] and data extraction efforts [2] regarding dark patterns have also been explored. Research and analysis on dark patterns have mostly occurred in Western regions like the United States and Europe [7]. However, cultural differences can have an impact on the analyses as dark patterns are related to designer strategies and human interactions. As a result, localized analysis of dark patterns is necessary to have a more generalized understanding of the phenomenon.

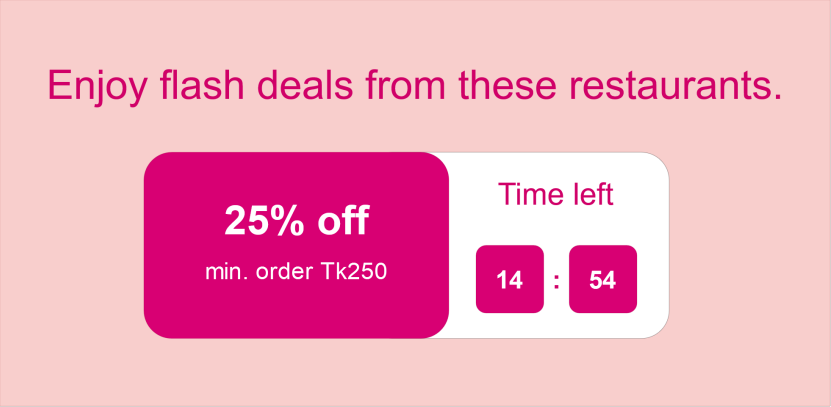


Figure 1.1: Countdown Timer as Urgency Dark Pattern

This research analyzes the landscape of dark patterns in the localized context of Bangladesh as it is a developing country experiencing rapid digitalization. It presents the prevalence of dark patterns in e-commerce websites of Bangladesh along with the related user perception. To extract dark patterns from Bangladeshi e-commerce websites, a novel automated detection technique is proposed, focusing on improved generalizability.

This chapter starts by discussing the motivation of the research. Research questions are then created and a plan for responding to them is provided. Subsequently, the contributions of the research are presented. Finally, the structure of this thesis is explained in order to provide readers with a clear guideline to go through it.

1.1 Motivation

Despite most end users being unaware of the term 'dark patterns', the phenomenon is more prevalent on digital platforms than commonly perceived. Evidence indicates that dark patterns are widespread across various digital environments, including social media platforms [8], e-commerce websites [2], mobile applications [9], and video games [10]. This prevalence raises significant concerns about the detrimental effects of dark patterns on users, particularly regarding their auton-

omy and well-being. These design tactics, intended to benefit digital platforms, often coerce, steer, or deceive users into making decisions they might not make if fully informed [2]. This can lead to feelings of frustration, confusion, and mistrust among users towards digital platforms. Most importantly, dark patterns can manipulate users into decisions that are not in their best interests, exploiting their cognitive vulnerabilities and leading to financial losses, excessive sharing of private data, and the fostering of addictive behaviours among both adults and children [2]. Digital platforms often use dark patterns to prompt users into disclosing more personal information than intended. Social media platforms, such as Facebook, also employ questionable tactics to induce digital addiction towards their platforms. E-commerce websites utilize dark patterns to entice and pressure users into overspending. Thus, investigating dark patterns is essential to mitigate their potential harm and safeguard user trust and welfare in the digital sphere.

Brignull et al. [11] first coined the term 'dark patterns'. They defined dark patterns as a set of "tricks used in websites that make a user do things that the user did not mean to". Researchers since have tried to define and classify dark patterns in various ways. Some researchers proposed taxonomies of dark patterns based on user experience [1, 2]. Some research efforts also focused on classifying dark patterns specifically based on privacy [9, 12]. Other researchers have also studied dark patterns found on social media [8], video streaming platforms [13], and proxemic interactions [14, 15]. Alongside definition and taxonomy, researchers have also tried to collect dark pattern data. They have collected dark patterns in various formats, including video recordings [4], screenshots [5], extracted features [16], and textual data [17] from user interfaces. As dark patterns abuse user psychology and manipulate their decision-making, researchers have also analyzed user perception and experience regarding the phenomenon. Users have been surveyed regarding their awareness [4, 18, 19], concern [18], ability to recognize [4, 18], and resist [18] dark patterns.

Majority of research on dark patterns has primarily been conducted in Europe and the United States. However, findings from these studies may not be universally applicable due to cultural and linguistic differences in other regions of the world [7]. Each country possesses distinct cultural characteristics and customs, which can influence the analysis of dark patterns. Variations in industry norms, user personality, impact of language on user emotions, ethical understandings, and availability of regulatory policies can all affect the manifestation and interpretation of dark patterns. Thus, further efforts are needed to examine dark patterns and related user perception from a localized perspective.

Bangladesh, a developing country in South Asia, has witnessed rapid digitalization in recent years, leading to significant growth in its e-commerce sector. However, technology-related awareness and concerns have not yet matured among the general population and regulatory bodies, primarily due to their inability to keep pace with the speed of digitalization. This situation makes Bangladeshi digital platform users, whose numbers are increasing rapidly, particularly vulnerable to the harmful effects of dark patterns. As a result, studying dark patterns within the Bangladeshi e-commerce industry is essential to understand the extent to which local users are exposed to these manipulative design tactics. Understanding local user perception regarding dark patterns is also an important area to explore. Such efforts are especially crucial for informing local policy-making to protect users from the detrimental consequences of dark patterns.

To understand the prevalence of dark patterns on Bangladeshi e-commerce websites, it is essential to extract relevant data from these platforms. The extraction of data related to dark patterns relies significantly on the successful detection of these deceptive tactics. Previous research by Mathur et al. [2] employed manual detection methods to extract dark patterns from over 11,000 e-commerce websites. Despite utilizing hierarchical clustering to ease the manual labeling process, this approach still demands substantial time and effort from researchers.

Automating the detection of dark patterns can significantly ease the data extraction process for researchers. Several studies have utilized machine learning techniques [16, 17] and rule-based methods [5] to detect dark patterns automatically. However, those techniques are either impractical for real-world settings or lack generalizability to be applied in the context of Bangladesh. Therefore, to facilitate the analysis of dark patterns in Bangladeshi e-commerce websites, it is important to develop a novel automated detection technique with a focus on improving generalizability.

1.2 Research Questions

This research aims to examine the prevalence of dark patterns in e-commerce websites of Bangladesh, alongside exploring related user perception. To achieve this objective, a novel dark pattern detection technique will be developed and subsequently applied to extract dark pattern data from Bangladeshi e-commerce websites. The formalization of this research effort can be expressed through the following research question.

RQ: How can dark patterns be effectively detected, classified, and analyzed in terms of their prevalence and user perception in Bangladeshi e-commerce websites?

To address this question, the following sub-questions (SQs) need to be explored.

- **SQ1:** How can dark patterns be effectively detected and classified in the context of Bangladesh?

  This sub-question aims to identify more effective methods for detecting and classifying dark patterns within the context of Bangladesh. To answer this question, existing automated detection techniques for dark patterns need to be explored first. An analysis of the weaknesses of these existing techniques has to be conducted to identify areas for improvement. Additionally, tech-

niques from other domains need to be examined to determine how they can improve the detection process in the context of Bangladesh. Based on these findings, a novel technique for the automated detection of dark patterns needs to be proposed.

- **SQ2:** How prevalent are dark patterns in Bangladeshi e-commerce websites, and how do users perceive them?

  This sub-question seeks to investigate the prevalence of dark patterns in Bangladeshi e-commerce websites. To answer this question, dark patterns need to be detected and extracted from Bangladeshi e-commerce websites. Subsequently, the collected data have to be analyzed to provide a comprehensive overview of the prevalence of dark patterns on these websites. To complement this prevalence study, an empirical study on user perception needs to be conducted, where participants have to be surveyed about their awareness and concern regarding dark patterns in the context of Bangladesh.

1.3 Contribution and Achievement

This research proposes a novel automated detection and classification technique for dark patterns and uses this technique to analyze the prevalence of dark patterns in e-commerce websites of Bangladesh. Local user perception of dark patterns is also analyzed in this research.

- **Automated Detection and Classification of Dark Patterns**

  Existing detection techniques for dark patterns are either not fully automated or suffer from a lack of generalizability. A novel approach for the automated detection of dark patterns is proposed in this research, focusing on improved generalizability. Large language models like GPT-3 have emergent abilities, such as in-context learning, which allows the model to learn

new tasks with contextual information and minimal demonstrations without extensive training or fine-tuning. Synthesized definitions of dark pattern categories are used as contextual information to utilize this emerging ability for detecting dark pattern texts.

This novel approach is validated by testing on the combined dark pattern textual dataset of Mathur et al. [2] and Yada et al. [17], which contains 1,178 dark pattern and 1,178 non dark pattern texts, totaling 2,356 texts from e-commerce websites. Experimental results using GPT-3 as the large language model achieve satisfactory performance for six out of the seven dark pattern categories considered in this research, as well as the non dark pattern category. This approach using in-context learning offers increased generalizability as the large language model is provided with contextual information through prompts, resulting in no modification of the underlying hyperparameters of the large language model. The model is not trained or fine-tuned with training data, making the approach less susceptible to data overfitting, and thus increasing generalizability. To further validate the generalizability of the approach, results using this approach are compared with the results of Yada et al. [17] on another dataset extracted from some e-commerce websites of Bangladesh. The proposed approach outperforms the results of Yada et al. on this test dataset.

- **Prevalence and User Perception of Dark Patterns in E-Commerce Websites of Bangladesh**

  There is no prior work regarding dark patterns in the context of Bangladesh. Thus, the prevalence of dark patterns in e-commerce websites of Bangladesh is analyzed in this research. More than 2,000 member companies of the E-Commerce Association of Bangladesh (e-CAB) are selected as a representative sampling frame for e-commerce websites of Bangladesh. After dis-

carding static websites and websites with no related data, 715 websites are analyzed and HTML contents from relevant webpages of those websites are extracted. A novel page segmentation algorithm is used to extract meaningful text segments from the HTML contents. More relaxed inclusion criteria for the algorithm are adopted compared to the page segmentation algorithm provided by Mathur et al. [2], resulting in an increased number of dark patterns in the final analysis. Detection and classification of dark patterns involve a combination of automated and manual approaches. As explained in the previous point, in-context learning capabilities of GPT-3 are used to detect dark patterns from the webpage texts. The automated detection step is used to cut down the number of texts to be considered for manual classification, which is carried out to remove the false positives and classify the texts into different dark pattern categories. A total of 931 instances of dark patterns are found in the analysis. 99 of these instances are unique and can be used as a dark pattern dataset for future research efforts. These dark pattern instances are classified into six common dark pattern categories. Dark patterns are found in 131 different websites, which is around 18.3% of the 715 e-commerce websites analyzed.

An empirical study on user perception of dark patterns is also conducted to complement the prevalence study. A survey of Bangladeshi university students is conducted. Participants belong to two groups - one with a background in technology education such as computer science, software engineering, etc., and the other without such an educational background. Participants are presented with pictures of user interface design instances containing different categories of dark patterns, followed by the collection of exposure, awareness, and concern ratings for each dark pattern category. The user perception survey indicates a general awareness and concern about dark patterns, particularly among students with a background in technol-

ogy education. A novel categorization of dark patterns distinguishes between 'Passive Dark Patterns' and 'Active Dark Patterns', each with distinct characteristics in terms of prevalence and user perception.

1.4 Organization of the Thesis

An overview of the chapters of this thesis is provided in this section. The chapters are organized as follows -

- **Chapter 2: Background Study**

  This chapter creates the knowledge base for understanding different terms used in this thesis. The chapter starts by explaining the taxonomy of dark patterns that is used in this study. This is followed by an introduction to large language models and in-context learning. A brief introduction of GPT-3 is also provided as it is used as the large language model for the experiments.

- **Chapter 3: Literature Review**

  This chapter discusses the existing literature on various relevant areas concerning dark patterns. Studies that are directly related to this research are thoroughly examined. The chapter provides an in-depth discussion on automated detection, data extraction and analysis, and user perception of dark patterns. Alongside the methodologies of these studies, contributions and limitations are also presented.

- **Chapter 4: Automated Detection and Classification of Dark Patterns**

  This chapter presents a novel automated detection and classification technique for dark patterns using in-context learning capabilities of large lan-

guage models. Experimental results with GPT-3 as the large language model of choice are discussed, followed by threats to validity of the work.

- **Chapter 5: Prevalence and User Perception of Dark Patterns in E-Commerce Websites of Bangladesh**

  This chapter presents the prevalence of dark patterns in e-commerce websites of Bangladesh. Complementing the prevalence study, an empirical study is also presented regarding user perception of dark patterns. The methods of both studies along with the findings are discussed in detail, followed by threats to validity of the work.

- **Chapter 6: Conclusion**

  This chapter provides a summary of the entire thesis, shedding light on the key contributions and achievements of the research. It discusses the development of a novel dark pattern detection technique, the analysis of the prevalence of dark patterns on Bangladeshi e-commerce websites, and the exploration of related user perception. Furthermore, this chapter outlines potential avenues for future work, suggesting how the research could be expanded or refined to further contribute to the field of dark patterns.

Chapter 2

Background Study

This chapter provides a comprehensive overview of the fundamental terms essential to understanding the concepts of dark patterns and in-context learning. Initially, different ways of classifying dark patterns are discussed as documented in existing literature. Subsequently, the in-context learning ability exhibited by large language models is discussed, along with a general introduction to large language models and Generative Pre-trained Transformer (GPT).

2.1 Taxonomy of Dark Patterns

Academic discussions regarding dark patterns rest on uncertain grounds. Questions such as what constitutes a dark pattern, why such tactics are problematic, and if it is possible to classify them; are still not adequately answered. Researchers have tried to develop definitions and taxonomies on this topic to address such inquiries. Dark patterns have versatile representations in user interfaces, making it very challenging to create a holistic taxonomy. Brignull et al. [11] first defined the term 'dark patterns' as a set of "tricks used in websites that make a user do things that the user did not mean to". Researchers since have defined and classified dark patterns in various ways. According to the background study of this research, four different ways of classifying dark patterns can be found in existing literature.

- **Dark Patterns Based on Privacy**

  This type of classification specifically talks about dark patterns that foster user data collection. Jarovsky [12] proposed a taxonomy for dark patterns in personal data collection based on design strategies that can negatively impact user decision-making. The taxonomy is divided into four categories - 'Pressure', 'Hinder', 'Mislead', and 'Misrepresent'. Fritsch [20] identified three privacy dark patterns found in online services. These patterns include 'Fogging identification with security', 'Sweet seduction', and 'You can run but you can't hide'. These patterns have implications such as decreased privacy, increased profiling, and limited user control over their personal information. Bösch et al. [21] also described a list of privacy related dark design strategies including 'Maximize', 'Publish', 'Centralize', 'Preserve', 'Obscure', 'Deny', 'Violate', and 'Fake'. These dark patterns are not explained further, as this classification of dark patterns is not considered in this research.

- **Dark Patterns Based on Induction of Addiction**

  This type of classification specifically talks about dark patterns that capture user attention, foster addictive or compulsive behaviour, and make users waste their time. Roffarello et al. [8] coined the term 'attention-capture dark patterns'. They identified five such patterns - 'Recommendation', 'Autoplay', 'Pull-to-refresh', 'Infinite scrolling', and 'Social investment'. Their work emphasized the impact of these dark patterns on user attention and engagement. Zagal et al. [10] discussed three categories of dark patterns used in games - 'Temporal dark patterns', 'Monetary dark patterns', and 'Social capital-based dark patterns'. Chaudhary et al. [13] discussed dark patterns found in popular video streaming platforms from a user-centric perspective. The five identified patterns are 'Feature fog', 'Extreme countdown', 'Switch-off delay', 'Attention quicksand', and 'Bias grind'. These dark patterns are

not explained further, as this classification of dark patterns is not considered in this research.

- **Dark Patterns Based on Proxemic Interaction**

  This type of classification specifically talks about dark patterns in proxemic interactions, which are interactions between people and digital devices that involve physical closeness. Greenberg et al. [14] identified several proxemic dark patterns including 'The captive audience', 'The attention grabber', 'Bait and switch', 'Making personal information public', 'We never forget', 'Disguised data collection', 'The social network of proxemic contacts or unintended relationships', and 'The milk factor'. Lacey et al. [15] discussed 'Cuteness' as a dark pattern in the design of home robots. They identified three features of dark patterns - 'Producing the illusion of user sovereignty', 'Emphasizing short-term gains over long-term decisions', and 'Manipulating emotions'. By employing these strategies, home robots can use their cuteness to deceive users, prioritize immediate benefits, and collect emotional data without informed consent. This can lead to a loss of user agency and privacy, as well as the potential for emotional manipulation. These dark patterns are not explained further, as this classification of dark patterns is not considered in this research.

- **Dark Patterns Based on Generic Designer Strategy**

  This type of categorization is not specific to any particular domain. Gray et al. [1] defined and categorized dark patterns into five primary categories including 'Nagging', 'Obstruction', 'Sneaking', 'Interface Interference', and 'Forced Action'. Later, they [22] identified some additional design properties that could mislead or manipulate users even without the presence of their previously defined dark design strategies. These properties included 'Automating the user away', 'Two-faced', 'Controlling', 'Entrapping', 'Nickling-

and-diming', and 'Misrepresenting'. Even though Conti et al. [23] did not use the term 'dark patterns' in their work, they categorized different methods used in 'malicious interface design'. The described categories are 'Coercion', 'Confusion', 'Distraction', 'Exploiting errors', 'Forced work', 'Interruption', 'Manipulating navigation', 'Obfuscation', 'Restricting functionality', 'Shock', and 'Trick'. Mathur et al. [2] found seven broad categories of dark patterns in their work, following the taxonomies proposed by Gray et al. [1] and Brignull et al. [11]. This classification of dark patterns is followed in this research. Dark pattern categories under this classification are -

- Urgency
- Misdirection
- Social Proof
- Scarcity
- Obstruction
- Forced Action
- Sneaking

2.1.1 Urgency

'Urgency' refers to the category of dark patterns that impose deadlines on sales or deals to accelerate user decision-making and purchases. This exploits users' fear of missing out (FOMO), making discounts and offers more attractive. This category includes dark patterns such as 'Countdown Timers', which dynamically display a deadline, and 'Limited-time Messages', which are static urgency messages without an explicit deadline. Some examples of 'Urgency' collected by Mathur et al. are shown in Figure 2.1.

(a) Countdown Timer (b) Limited-time Message

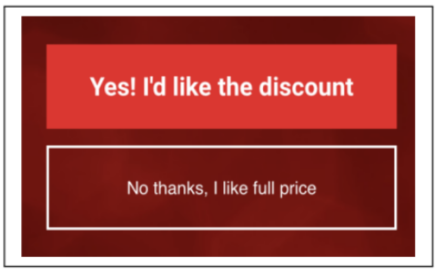
Figure 2.1: Dark Pattern: Urgency

2.1.2 Misdirection

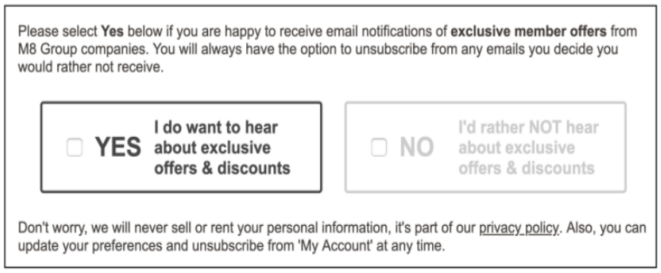
'Misdirection' refers to the category of dark patterns that employ visuals, language, and emotion to steer users toward or away from making a particular choice. It exploits affective mechanisms and cognitive biases in users without restricting their choices. This category includes dark patterns such as 'Confirmshaming', 'Trick Questions', 'Visual Interference', and 'Pressured Selling'. 'Confirmshaming' uses emotions like shame to steer users away from certain choices. 'Visual Interference' influences choices through style and presentation, while 'Trick Questions' uses confusing language like double negatives to manipulate user choices. 'Pressured Selling' employs high-pressure tactics to encourage purchasing a more expensive version of a product. Some examples of 'Misdirection' collected by Mathur et al. are shown in Figure 2.2.

2.1.3 Social Proof

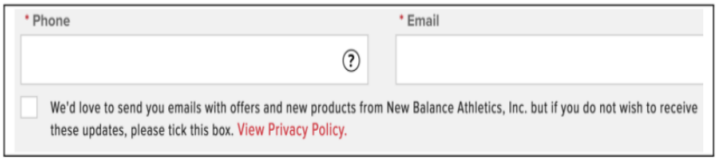
'Social Proof' refers to the category of dark patterns that accelerate user decision-making and purchases, by showing the actions and behaviours of other users. This category includes dark patterns such as 'Activity Notifications' and 'Testimonials of Uncertain Origin'. 'Activity Notification' dark pattern involves transient and attention-grabbing messages on product pages indicating the activity of other



(a) Confirmshaming



(b) Visual Interference



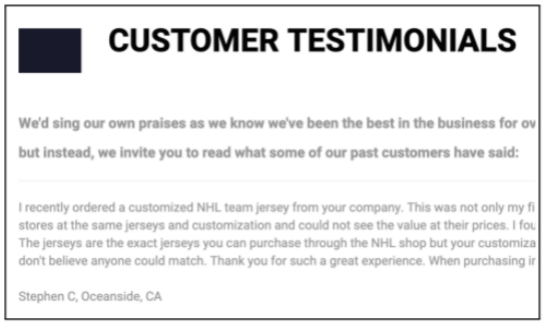
(c) Trick Questions

(d) Pressured Selling

Figure 2.2: Dark Pattern: Misdirection

users, such as recent purchases, items in carts, and product views. 'Testimonials of Uncertain Origin' dark pattern refers to customer testimonials with ambiguous sourcing information. Some examples of 'Social Proof' collected by Mathur et al. are shown in Figure 2.3.

(a) Activity Notification

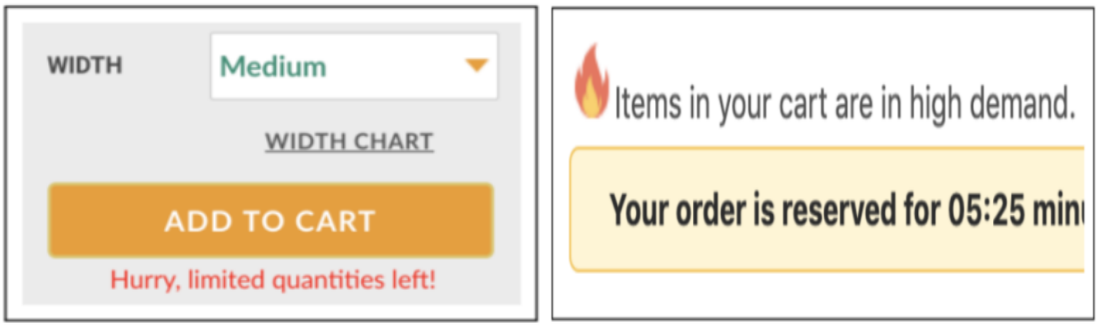


(b) Testimonials of Uncertain Origin

Figure 2.3: Dark Pattern: Social Proof

2.1.4 Scarcity

'Scarcity' refers to the category of dark patterns that indicate the limited availability or high demand of a product, thereby increasing its perceived value and desirability. This category entails dark patterns such as 'Low-stock Messages' and 'High-demand Messages'. 'Low-stock Messages' signal limited quantities of products while 'High-demand Messages' suggest that products are in high demand. Some examples of 'Scarcity' collected by Mathur et al. are shown in Figure 2.4.



(a) Low-stock Message (b) High-demand Message

Figure 2.4: Dark Pattern: Scarcity

2.1.5 Obstruction

'Obstruction' refers to the category of dark patterns that intentionally complicate a specific action beyond what is necessary to discourage users from taking that action. 'Hard to Cancel' is such a dark pattern that makes it easy for users to sign up for subscriptions but difficult for them to cancel later on, a detail often not initially disclosed on shopping websites. An example of 'Obstruction' collected by Mathur et al. is shown in Figure 2.5.

Figure 2.5: Dark Pattern: Obstruction

2.1.6 Forced Action

'Forced Action' refers to the category of dark patterns that require users to perform an additional action, which is often undesirable, to accomplish some other tasks. 'Forced Enrollment' is such a dark pattern that explicitly coerces users into creating accounts, or signing up for marketing communication. By using 'Forced Enrollment' dark pattern, online services and websites collect more information about their users than they might otherwise consent to, resulting from an all-or-nothing proposition. An example of 'Forced Action' collected by Mathur et al. is shown in Figure 2.6.

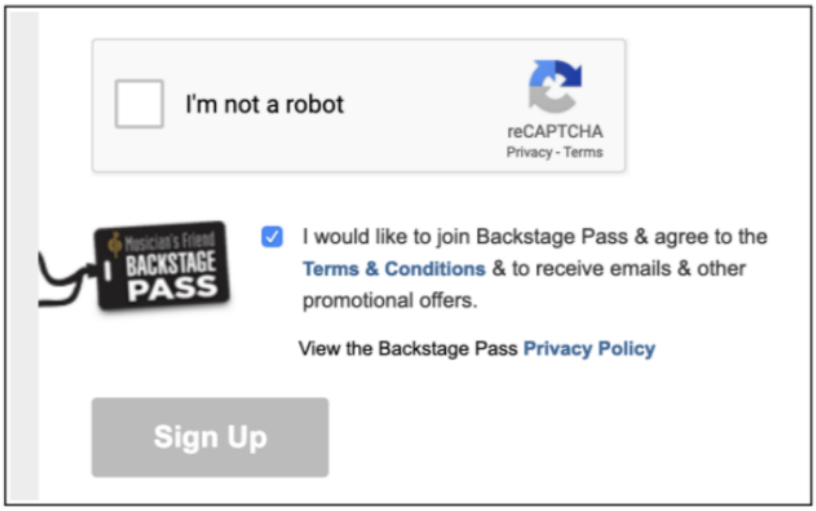
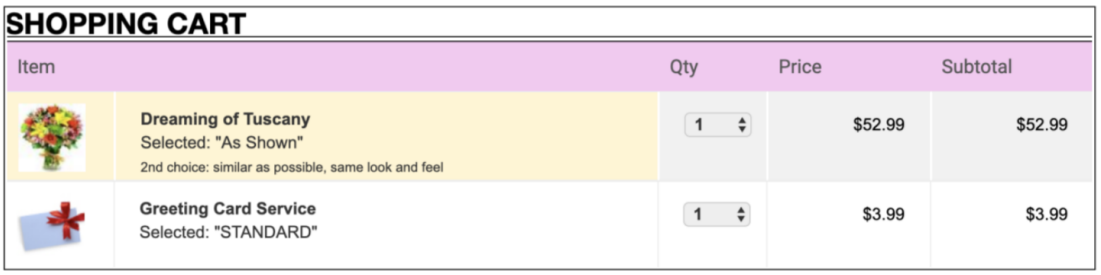


Figure 2.6: Dark Pattern: Forced Action

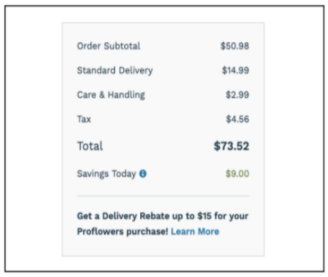
2.1.7 Sneaking

'Sneaking' refers to the category of dark patterns aimed at misrepresenting user actions or concealing/delaying information that users would probably oppose if they were aware of it. This category encompasses dark patterns such as 'Sneak into Basket', 'Hidden Costs', and 'Hidden Subscription'. 'Sneak into Basket' adds products to shopping carts without consent. 'Hidden Costs' reveals additional charges at the end of a purchase. Lastly, 'Hidden Subscription' charges users

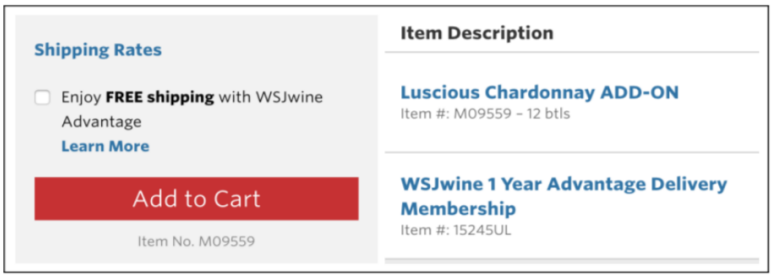
recurring fees under false pretenses of one-time fees or free trials. Some examples of 'Sneaking' collected by Mathur et al. are shown in Figure 2.7.



(a) Sneak into Basket



(b) Hidden Costs



(c) Hidden Subscription

Figure 2.7: Dark Pattern: Sneaking

2.2 Automated Detection of Dark Patterns

The proposed approach for the automated detection of dark patterns utilizes the in-context learning capabilities of large language models. Therefore, this section provides an overview of in-context learning and large language models. GPT, the chosen large language model for the experiments conducted in this research, is also discussed here.

2.2.1 Large Language Models (LLMs)

In recent years, there has been a significant advancement in the field of artificial intelligence (AI) and natural language processing (NLP) with the development of large language models (LLMs). These models, such as GPT developed by OpenAI,

have revolutionized the way machines understand and generate human language. Large Language Models have become integral to various AI applications, including language translation, content generation, and conversational agents.

Large language models are neural network-based models that are trained on vast amounts of text data to understand and generate human language. These models utilize a technique called unsupervised learning, where they learn to predict the next word in a sentence based on the context of the previous words. This process allows the models to capture the complex patterns and structures of natural language, enabling them to generate coherent and contextually relevant texts.

The training data for these models often consists of diverse sources such as books, articles, websites, and other textual content from the internet. By leveraging such extensive and varied data, large language models can learn to understand and mimic human language, leading to impressive capabilities in natural language understanding and generation.

2.2.2 In-Context Learning (ICL)

Large language models have revolutionized natural language processing tasks by demonstrating remarkable performance on a wide range of applications. One key advancement in this domain is the concept of in-context learning (ICL). The idea of in-context learning was introduced in a seminal work by Brown et al. [24] in 2020, where they proposed a novel way for large language models to learn new tasks by conditioning on a few input-label pairs, also known as demonstrations. It is a prompt engineering strategy for large language models that do not require any training. Instead of training or fine-tuning models with large datasets, only a few examples are provided within the prompt as contextual information. As a result, models can learn tasks without updating underlying parameters [25]. This innovative approach has gained significant attention in recent years due to its

potential to revolutionize how machines understand and generate human language. Research has shown that large language models exhibit the capability to perform complex tasks via in-context learning, including solving mathematical reasoning problems [26]. These capabilities have been verified as emergent abilities for large language models [27].

One of the primary advantages of in-context learning is its capacity to generalize to new tasks without requiring extensive labeled data or task-specific training. By presenting the model with a few examples of input-output pairs, it can infer the underlying patterns and relationships within the data to generate accurate predictions on unseen instances. This flexibility and adaptability make in-context learning a promising approach for rapid task adaptation and deployment in real-world scenarios [28].

2.2.3 Generative Pre-trained Transformer (GPT)

GPT (Generative Pre-trained Transformer) is a series of large language models developed by OpenAI[1]. It stands as a seminal innovation in the landscape of natural language processing (NLP) and advanced machine learning. GPT leverages transformer architectures to achieve unprecedented levels of performance in various NLP tasks. At its core, GPT is distinguished by its pre-training strategy, wherein the model is exposed to vast amounts of text data to learn intricate patterns and dependencies inherent in language [29].

The architectural foundation of GPT lies in the transformer, a neural network architecture introduced by Vaswani et al. [30]. The transformer's key innovation is its attention mechanism, which enables the model to capture long-range dependencies within input sequences more effectively than traditional recurrent or convolutional architectures. This mechanism enables the parallel processing of input tokens, thereby enhancing computational efficiency and scalability.

[1]https://openai.com

Through extensive pre-training on large-scale text corpora, GPT acquires a robust understanding of language structure and semantics, enabling it to generate contextually relevant and grammatically coherent texts in a variety of contexts. The pre-training process involves training the model to predict the next token in a sequence based on the preceding tokens [29]. This self-supervised learning paradigm allows GPT to learn rich representations of language without the need for explicit annotations or supervision. GPT can achieve state-of-the-art performance across a wide range of NLP tasks including question answering, semantic similarity assessment, and document classification [29].

The versatility and efficacy of GPT have propelled it to the forefront of NLP research and applications in recent years, with multiple iterations developed since its initial release. Subsequent iterations, such as GPT-2 [31] and GPT-3 [24], have further expanded the model's capabilities, enabling it to generate longer and more coherent texts and exhibit a deeper understanding of context and semantics.

2.3 Summary

This chapter provides a comprehensive overview of the domain of dark patterns as well as the in-context learning capabilities of large language models (LLMs). It begins by explaining various ways of classifying dark patterns as documented in existing literature. The discussion includes classifications based on privacy, induction of addiction, proxemic interaction, and generic designer strategies. The absence of a holistic taxonomy due to the multifaceted nature of dark patterns is highlighted. Subsequently, in-context learning ability exhibited by large language models is discussed. An overview of LLMs is provided, showcasing their pivotal role in natural language processing (NLP) tasks and machine learning in general. In-context learning, a novel approach facilitated by LLMs, is explained as the ability to enable task adaptation without the need for fine-tuning. Generative Pre-

trained Transformer (GPT), a seminal innovation in the field of NLP developed by OpenAI, is also discussed. GPT's architectural foundation in transformer networks and its pre-training strategy are discussed, explaining its unparalleled performance in diverse NLP tasks.

Chapter 3

Literature Review

Dark patterns have attracted a lot of attention from researchers in recent times. Researchers have worked on the definition and taxonomy of dark patterns [1, 2] as well as automated detection techniques [5, 6]. Legal and policy-making perspectives [3], and user perception [4] about dark patterns have also been explored. This chapter provides an overview of past literature on the following topics.

- Automated Detection of Dark Patterns
- Dark Pattern Data Extraction and Analysis
- User Perception of Dark Patterns
- Localized Analysis of Dark Patterns

3.1 Automated Detection of Dark Patterns

Dark patterns lack a holistic definition and classification, making the task of detecting such design techniques very complex. Researchers have explored dark pattern detection techniques using machine learning [16, 17], and heuristic-based computer vision and natural language processing (NLP) [5, 32, 33]. This section presents an extensive review of literature related to the automated detection

of dark patterns. A synopsis of each of the important studies along with their contributions and limitations are discussed here.

3.1.1 Automated Machine Learning

Yada et al. [17] employed classical natural language processing (NLP) techniques with transformer-based pre-trained language models to establish baseline results for automated detection of dark pattern texts. Classical methods such as logistic regression, support vector machines (SVM), random forest, and gradient boosting were used alongside feature extraction techniques like bag-of-words representation. This approach involved a structured model training process with hyper-parameter optimization. Concurrently, transformer-based models, including BERT, RoBERTa, ALBERT, and XLNet, which are well-known for their effectiveness in NLP tasks, underwent fine-tuning on a textual dark pattern dataset. Hyper-parameter tuning of these deep learning models was refined iteratively through grid search.

Contributions: The study evaluated baseline detection performance using various machine learning methods, including classical NLP techniques and transformer-based pre-trained language models. It provided benchmark results for comparison with future algorithms.

Limitations: The models may suffer from overfitting, resulting in reduced generalization. Additionally, the study employed binary classification and did not classify dark patterns into specific categories.

3.1.2 Semi Automated Machine Learning

Soe et al. [16] employed supervised machine learning techniques to detect dark patterns in cookie banners. They utilized a dataset of cookie banners, represented by features including page location, option clarity, site functionality after cookie rejection, and the number of clicks required to reject all cookies. Various prediction

models were trained on this dataset to determine the presence of dark patterns in cookie banners.

Contributions: The study proposed a novel feature-based supervised machine learning technique for detecting dark patterns in cookie banners. It also discussed the challenges associated with the automated detection of dark patterns, offering directions for future research.

Limitations: The detection technique employed in this study is semi automated rather than fully automated. It necessitates the extraction of feature values from user interfaces, after which the method predicts the presence of dark patterns based on the feature set. While some features can be automatically extracted, others require manual intervention. As a result, this approach is impractical for real-world applications. Furthermore, the detection performance of this approach is not particularly high.

3.1.3 Heuristic-Based Computer Vision and NLP

Mansur et al. [5] introduced AidUI (Aid for detecting UI Dark Patterns), a comprehensive approach that integrates textual, iconographic, chromatic, and spatial analyses of user interfaces to facilitate the automated detection of dark patterns. The core of AidUI's methodology involves identifying various visual and textual cues whose co-occurrence indicates the presence of different dark patterns. To achieve this, AidUI employs a combination of computer vision and natural language template matching techniques for cue detection, effectively addressing the challenges posed by the variability of dark patterns. AidUI is a fully automated solution that needs only a screenshot as input, making it highly adaptable and practical for real-world applications.

Contributions: The study is a pioneering effort in automating the detection of dark patterns within screenshots of user interfaces. This marks a significant advancement in the field, as it is the first to achieve such automation successfully.

Limitations: The text analysis technique employed in the study relied exclusively on heuristically defined pattern-matching rules. This reliance poses significant challenges when applying the method to semantically complex texts, limiting its overall effectiveness.

3.2 Dark Pattern Data Extraction and Analysis

Dark pattern data extraction efforts rely heavily on the successful detection of such deceptive tactics in websites. However, as previously discussed, it is a challenging task, especially without a holistic definition and classification of dark patterns. This is also a relatively new field of research, and detection methods are yet to be matured. As a result, there have been very few efforts to extract dark pattern data on a large scale. Another challenge in such data extraction efforts is determining the appropriate format (image, video, text, etc.) to represent dark patterns. Researchers have extracted dark patterns in various formats, including video recordings [4], screenshots [5], extracted features [16], and textual data [17] from user interfaces.

Geronimo et al. [4] prepared a comprehensive video dataset comprising screen recordings of mobile application usage. The dataset contains 15 videos, accompanied by classifications of identified dark patterns at different timestamps within the videos. Mansur et al. [5] created a dark pattern dataset containing 501 instances by manually labelling user interface screenshots collected from different sources. Soe et al. [34] analyzed dark patterns in cookie banners of 300 news websites. For each cookie banner, they manually extracted a set of features in terms of dark patterns. However, data extraction and analysis efforts using video recordings, screenshots, or manually extracted features are quite challenging to carry out on a large scale. Alternatively, textual data extraction and analysis are more convenient in terms of scalability. A synopsis of each of the important litera-

ture related to textual data extraction and analysis along with their contributions and limitations are discussed in this section.

3.2.1 Dark Pattern Texts from E-Commerce Websites

Mathur et al. [2] collected the first large-scale textual dark pattern data. They crawled more than 11,000 e-commerce websites and extracted meaningful text segments using their page segmentation algorithm. These text segments were manually labelled with dark pattern categories. To ease the manual labelling of a large number of text segments, they used hierarchical clustering in their work.

Contributions: The study found 1,818 instances of dark patterns in 1,254 websites, which was roughly 11.1% of all analyzed websites in their work. The discovered dark pattern instances represent 7 broad categories.

Limitations: Even though they collected text segments using automated techniques, dark pattern detection and classification were carried out manually, which is a time consuming task.

3.2.2 Non Dark Pattern Texts from E-Commerce Websites

Yada et al. [17] expanded upon the dataset of dark patterns created by Mathur et al. by incorporating an additional 1,178 non dark pattern texts from the same e-commerce websites. This process involved collecting webpages from these websites and extracting texts from them. They modified the segmentation algorithm of Mathur et al. to facilitate the extraction of non dark pattern texts instead of dark pattern texts. Since the extracted texts could potentially contain dark patterns, they filtered out those identified by Mathur et al. as dark patterns. Subsequently, each remaining text was manually validated to ensure it did not contain dark patterns. From these validated texts, a random sample of 1,178 non dark pattern texts was selected and designated as negative samples for dataset construction. These negative samples were added with the positive samples collected by Mathur

et al., resulting in a combined textual dataset comprising 1,178 positive (dark pattern) texts and 1,178 negative (non dark pattern) texts, amounting to a total of 2,356 texts sourced from e-commerce websites.

Contributions: The study contributed a combined textual dark pattern dataset by augmenting the existing dark pattern texts extracted by Mathur et al. with non dark pattern texts extracted from the same websites. This combined dataset is designed to support supervised machine learning and other analytical purposes related to dark patterns.

Limitations: The study only extracted non dark patterns, but contributed nothing in terms of extracting dark patterns.

3.3 User Perception of Dark Patterns

As dark patterns abuse user psychology and manipulate their decision-making, user perception regarding the phenomenon is a significant area of research. Researchers have surveyed users regarding their awareness [4, 18, 19], concern [18], ability to recognize [4, 18], and resist [18] dark patterns. This section presents an extensive review of literature related to the user perception of dark patterns. A synopsis of each of the important studies along with their contributions and limitations are discussed here.

3.3.1 Awareness and Recognition of Dark Patterns

Geronimo et al. [4] investigated user awareness and their ability to recognize dark patterns within popular mobile applications. They conducted an online survey involving 589 participants from diverse backgrounds. Participants were asked to evaluate three user interface interaction videos - two containing dark patterns randomly selected from a pool of five, and one without dark patterns. The survey began with questions regarding familiarity with the applications and usage

frequency, followed by evaluations of the video interactions. Subsequently, participants were asked to identify any malicious designs within the videos without prior priming about dark patterns. Finally, they were shown screenshots of the identified dark patterns and asked to confirm their recognition.

Contributions: The study sheds light on DP-blindness, which refers to the inability of users to recognize dark patterns. Their findings revealed that while a substantial proportion of users failed to detect dark patterns, some displayed implicit distrust in the applications, indicating a level of awareness despite not identifying the expected dark patterns. The study also highlighted the necessity to improve user awareness mechanisms to counter dark patterns.

Limitations: The study asked participants if they recognized any 'malicious design'. Such a negative phrasing might have influenced participant responses.

3.3.2 Concern and Resistance to Dark Patterns

Bongard-Blanchy et al. [18] explored user awareness and their ability to recognize and resist dark patterns. They conducted an online survey in which participants were asked about their awareness of and concerns regarding manipulative online designs. Participants rated various statements on a Likert scale and provided examples of perceived influence and potential harm. Additionally, the survey assessed participants' engagement with online services and their perceptions of manipulation. Participants were also tested on their ability to identify dark patterns in online interfaces. The survey included 413 participants, who were representative of the UK population in terms of age, gender, and ethnicity.

Contributions: The study revealed that users can recognize dark patterns even though they are often unaware of the actual harmful consequences. They also identified that demographic factors such as age and educational background can influence user perception of dark patterns. Interestingly, participants who recognized manipulative designs more easily reported a lower likelihood of being

influenced by them. This indicates a potential link between the detection of dark patterns and the resistance against them.

Limitations: The sample was limited to the UK population, potentially limiting generalizability to other cultural contexts.

3.3.3 User Experience of Dark Patterns

Gray et al. [19] employed a mixed methods approach to investigate user experience and perception of manipulative technology interfaces. Data were collected through a survey study and follow-up interviews with interested respondents, including English-speaking and Mandarin-speaking users. The survey employed a qualitative design to capture users' daily interactions with manipulative interfaces, their emotional reactions to these manipulations, and their perceptions of the creators of such technologies. Follow-up interviews were conducted with a subset of American and Chinese participants, selected based on demographic characteristics and their experiences with technology. These interviews aimed to delve deeper into participants' experiences of manipulation and their perceptions of the designers responsible for these manipulative interfaces.

Contributions:

The study offered valuable insights into user perception of manipulative technology interfaces, focusing on research questions related to the triggers of manipulative experiences, the associated emotions, and user perception of those who create such manipulations. The findings indicated that users often sense that something is "off" or manipulative, even when they lack the precise terminology to articulate it.

Limitations: Language, cultural differences, and the limited scope of the sample could have impacted the accuracy and generalizability of the findings.

3.3.4 Normalization of Dark Patterns

Maier [35] investigated how users perceive, experience, and respond to dark patterns. He used focus groups and individual structured interviews to collect data from end users. Participants were selected based on availability and internet proficiency from various university departments to ensure diverse perspectives. The gathered data, including audio recordings and transcriptions, were analyzed to generate a comprehensive understanding of users' perceptions of dark patterns.

Contributions: The study suggests that while users may initially be angered by dark patterns, they often become accustomed to them over time. This leads to the normalization of these deceptive design practices. Maier's work also highlights that awareness mechanisms alone may not be sufficient to safeguard users from dark patterns.

Limitations: The limited scope of the sample could have impacted the accuracy and generalizability of the findings.

3.4 Localized Analysis of Dark Patterns

Each region in the world has unique cultural traits that can impact the analysis of dark patterns. Variations in industry standards, user behaviours, language, ethical interpretations, and regulatory frameworks all play a role in dark pattern analyses. Research has examined dark patterns with data from various countries including the United Kingdom [18], the United States [19], China [19], Japan [7] as well as globally [2]. However, Hidaka et al. [7] argued that most research has been mainly conducted in Europe and the United States which they termed as coming from a 'Western context'. They highlighted differences between the manifestations of dark patterns according to geographic location such as fewer instances found in Japanese applications compared to American applications, emphasizing the need for localized examination of dark patterns and related user perceptions.

No such analysis of dark patterns has been carried out in the context of Bangladesh. Thus, it is important to analyze dark pattern data as well as user perception in the localized context of Bangladesh. Such efforts are especially crucial for local policy-making that can protect users from the harmful consequences of dark patterns.

3.5 Summary

This chapter provides a comprehensive overview of the literature surrounding dark patterns, dividing the research into four main areas - automated detection, data extraction and analysis, user perception, and localized analysis. In the realm of automated detection, several studies have utilized machine learning techniques and feature-based approaches to identify dark patterns, each with its own set of contributions and limitations. From classical methods to transformer-based models, researchers have made strides in benchmarking detection performance, although challenges like overfitting and impracticality in real-world settings persist. Data collection efforts have seen diverse data formats, with textual data collection being the most suitable in terms of scalability. While some studies have successfully collected textual dark pattern related datasets from a large sample of websites, the manual nature of dark pattern detection remains a bottleneck, highlighting the need for more scalable approaches. User perception of dark patterns has been a key area of investigation, with studies revealing varying levels of awareness and resistance among participants. Cultural and demographic factors have also been shown to influence users' ability to recognize manipulative designs, underscoring the importance of localized research efforts. Such studies are vital for informing policy-making and safeguarding users from the harmful effects of dark patterns.

Chapter 4

Automated Detection and Classification of Dark Patterns

Dark patterns manifest in user interfaces through various representations, ranging from subtle nudges to explicit coercion. Automated detection of dark patterns is challenging due to this diverse nature of representation. Existing detection techniques either necessitate manual intervention or suffer from overfitting, resulting in poor generalization across different scenarios. To address these challenges, this study proposes an automated detection technique for dark patterns that aims for improved generalization. For this purpose, inclusive definitions of dark pattern categories are first synthesized from existing literature. This contextual information is then prioritized using the in-context learning capabilities of GPT-3 to detect and classify dark pattern texts without the need for any fine-tuning. Experimental results show that the proposed technique offers satisfactory performance for six out of the seven dark pattern categories explored in this study, with the highest overall accuracy of 92.57%. The improved generalization of the approach is also validated by outperforming an existing baseline model on a test dataset.

4.1 Introduction

The deceptive nature of dark patterns often makes it difficult for users to identify the exact problems they face, even when they experience negative emotions because of these design strategies. This confusion and frustration can trap users in a cycle that prevents them from recognizing and avoiding manipulative design tactics. Automated detection of dark patterns can help users identify and avoid these tactics, while also facilitating the large-scale extraction of dark patterns. However, such an automated process faces significant challenges due to the diverse manifestations of dark patterns within user interfaces. Effective detection techniques must exhibit a high degree of adaptability and generalizability to adequately serve real-world users. To address these challenges, it is essential to develop a mechanism that is capable of processing contextual information in a human-like manner. Such an approach is crucial for effectively addressing the dynamic and multifaceted nature of dark patterns.

Various approaches exist for the automated detection of dark patterns in user interfaces. One such method involves processing images or screenshots of user interfaces. This technique uses pixel correlations to detect dark patterns. However, it overlooks the textual information in the user interface, which is crucial for identifying certain types of dark patterns [16]. Alternatively, user interface texts can be processed using machine learning (ML) or heuristic-based mechanisms. Heuristic-based mechanisms, while useful, often lack generalizability as they are designed for specific cases, making them less effective in diverse scenarios. In contrast, Yada et al. [17] proposed an ML-based technique to detect dark pattern texts by training models on a labeled textual dataset. Although this method offers greater flexibility compared to heuristic-based systems, it remains vulnerable to data overfitting. This issue occurs when the training dataset lacks sufficient diversity, resulting in a classification model that performs poorly on unseen data.

This study proposes a novel approach for the automated detection of dark patterns, focusing on improved generalization. To accomplish this, comprehensive definitions of dark pattern categories are first synthesized. These definitions provide essential contextual information to large language models (LLMs), along with zero, one, or two examples per category. The in-context learning capabilities of LLMs, such as GPT-3, enable these models to learn new tasks using contextual information and only a few examples, without the need for extensive training or fine-tuning. Engineered prompts are utilized to take advantage of this capability for detecting and classifying dark pattern texts.

The proposed approach has been validated through testing on a combined dataset compiled by Mathur et al. [2] and Yada et al. [17], consisting of dark pattern and non dark pattern texts. The method demonstrates satisfactory classification accuracy for six out of seven dark pattern categories, as well as for the non dark pattern category, with the highest overall accuracy of 92.57%. The utilization of at most two examples per category effectively mitigates the risk of overfitting, thereby improving the generalizability of the approach. To further demonstrate its generalization across diverse datasets, the results of this method have been compared with those obtained by Yada et al. [17] on a manually labeled test dataset. The proposed approach outperforms the results of Yada et al. on this test dataset.

4.2 Methodology

This study introduces a novel approach to the challenging task of automated dark pattern detection, emphasizing generalizability. The approach begins by synthesizing comprehensive definitions of dark pattern categories, which are then prioritized as contextual information to engineer prompts for large language models. GPT-3 is such a large language model that has in-context learning capabilities, which is

used for detecting dark pattern texts with a reduced risk of overfitting. Figure 4.1 provides an overview of the automated dark pattern detection methodology employed in this study.

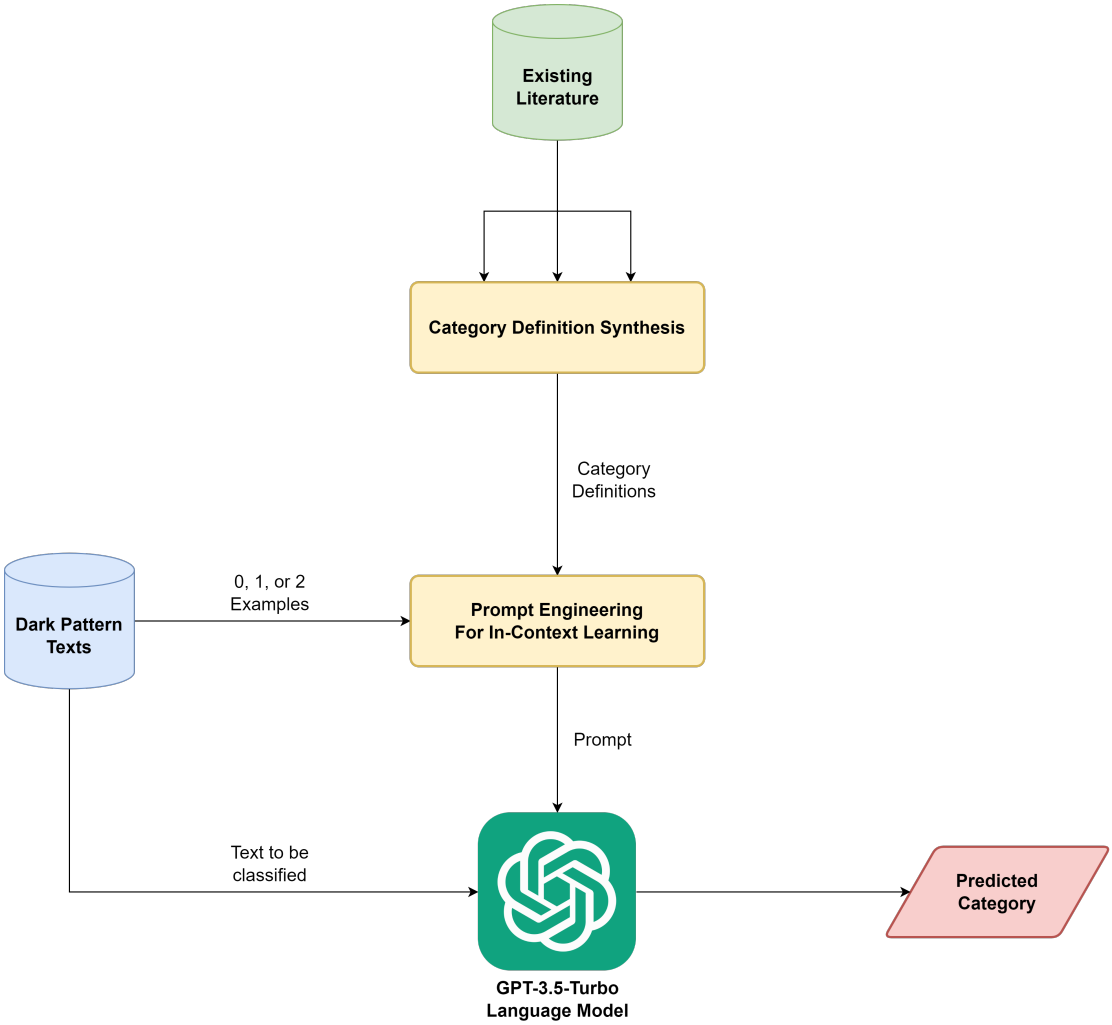


Figure 4.1: Automated Dark Pattern Detection Using In-Context Learning

4.2.1 Textual Dark Pattern Dataset

This study utilizes a comprehensive dataset compiled by Mathur et al. [2] and Yada et al. [17], which includes 1,178 dark pattern texts and 1,178 non dark pattern texts, totaling 2,356 texts extracted from e-commerce websites. The dark pattern texts in this dataset are classified into seven broad categories - 'Misdi-

rection', 'Urgency', 'Scarcity', 'Social Proof', 'Obstruction', 'Forced Action', and 'Sneaking'. These categories have already been discussed in Chapter 2. The non dark pattern texts are labeled under the category 'Not Dark Pattern'. Exhaustive definitions for these eight categories are formulated to facilitate in-context learning, a methodology that is further detailed in the following subsection. Out of the 2,356 instances in the dataset, 2,342 are exclusively reserved for validating the classification models. The remaining 14 texts are allocated for demonstration or validation purposes, depending on the prompting techniques employed for GPT-3.

4.2.2 Dark Pattern Category Definition Synthesis

To classify texts into dark pattern categories, these categories must be defined first. Given the importance of this contextual information in the classification task, the precise meaning and wording of category definitions can significantly influence the results. Each dark pattern category may have multiple variants or types, as discussed in Chapter 2. For example, 'Countdown Timers' and 'Limited-time Messages' are variants of the 'Urgency' category. Instead of utilizing high-level definitions, category definitions are formulated using the definitions of their corresponding variants. For this purpose, relevant information from the literature [1, 2, 5, 11] has been synthesized to create inclusive definitions for the corresponding variants of each category. These variant definitions, enclosed in single quotes, are then enumerated to formulate comprehensive definitions for the categories. Figure 4.2 illustrates the methodology employed in synthesizing these category definitions, which are listed in TABLE 4.1. Additionally, a definition for the overarching term 'dark patterns' is synthesized from the same sources, defined as "user interface design techniques that use knowledge of human behavior to benefit the service provider by coercing, steering, tricking or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, users might not make" [1, 2, 11].

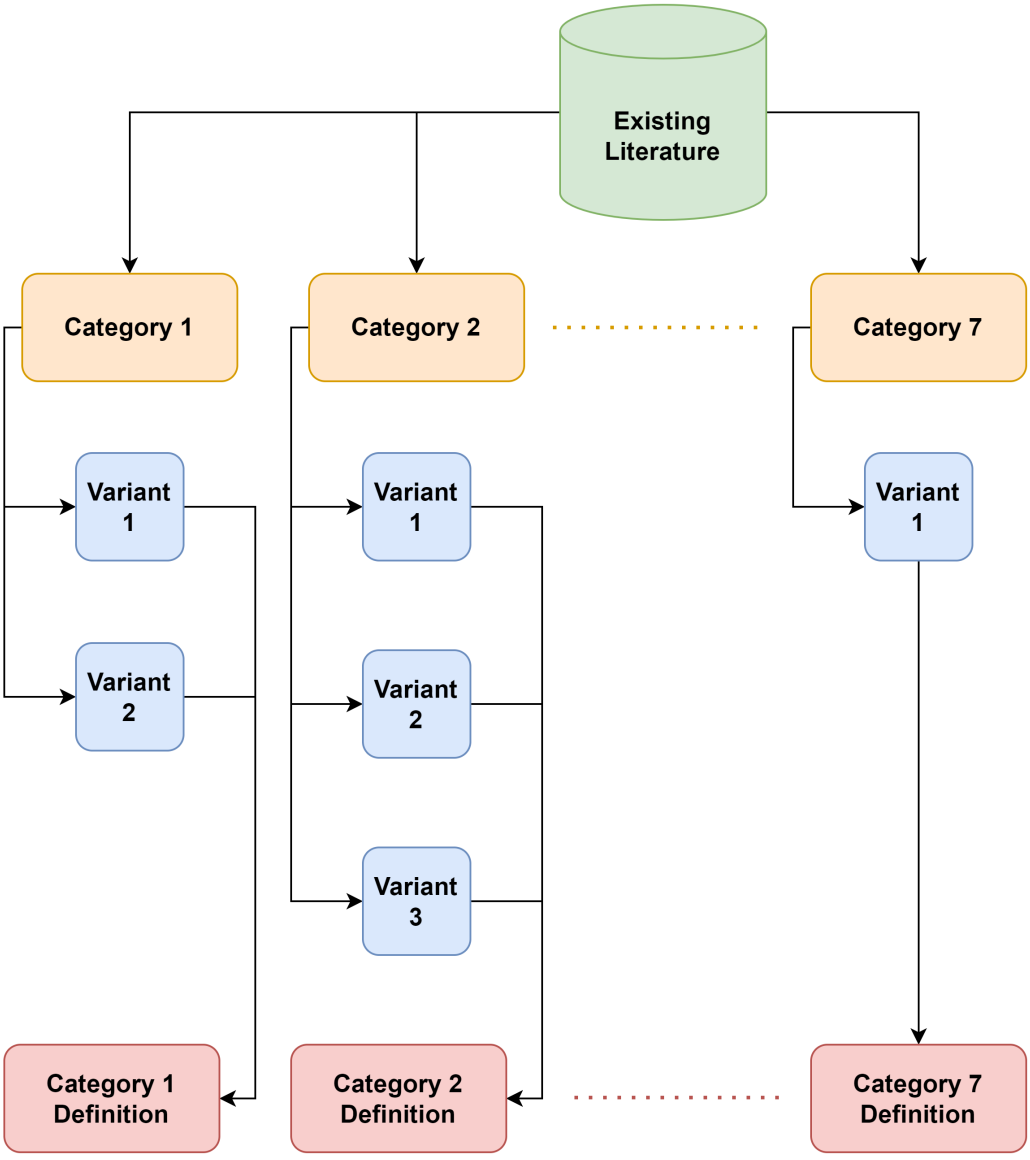


Figure 4.2: Dark Pattern Category Definition Synthesis

4.2.3 Classification Using GPT-3

'GPT for Sheets™ and Docs™' extension available in 'Google Sheets' is used to classify all the texts in the valiation dataset. This extension provides convenient GPT functions designed for integration within spreadsheets. Specifically, the 'GPT()' function is utilized, which sends a prompt to GPT and retrieves the generated response. This function has a mandatory parameter 'prompt', along with three

Table 4.1: Synthesized Definitions of Dark Pattern Categories

Dark Pattern Category	**Definition**
Misdirection	'Misdirection' includes 'attracting the user's attention to one thing in order to distract their attention from another' and 'use of visuals, confusing language, and emotion (shame) to intrigue users to or away from particular choices'.
Urgency	'Urgency' means 'pressuring the user to accept deals or discounts with time limitations, count-down timers, or sale deadlines'.
Scarcity	'Scarcity' means 'pressuring the user to accept deals or discounts with indications of low stock, limited supply, high demand, or low availability'.
Social Proof	'Social Proof' includes 'informing the user about the activity on the website (e.g., purchases, views, visits) to make the product more credible' and 'testimonials on a product page whose origin is unclear'.
Obstruction	'Obstruction' includes 'making a process more difficult than it needs to be, with the intent of obstructing certain actions' and 'making it easy for the user to sign up or subscribe for a service but hard to cancel it'.
Forced Action	'Forced Action' means 'requiring the user to perform an undesirable action (e.g., creating accounts or sharing information) to access certain functionality'.
Sneaking	'Sneaking' includes 'hiding or delaying the relevant information from the user', 'adding additional products to shopping carts without their consent', 'revealing previously undisclosed charges to users right before the checkout' and 'enrolling users in a recurring subscription or payment plan without clear disclosure or their explicit consent'.
Not Dark Pattern	If a text does not fall into the other 7 categories, it is defined as 'Not Dark Pattern'.

optional parameters - ‘value’, ‘temperature’, and ‘model’. By adjusting these parameters, the function can be customized to achieve the desired outcomes.

The parameter ‘temperature’ determines whether the language model prioritizes creativity or precision in its responses. It ranges from 0 to 1, with lower values favouring precision and higher values favouring creativity. For classification into dark pattern categories, the temperature is set to 0 to ensure precise answers. The parameter ‘model’ specifies the language model to be used. The ‘gpt-3.5-turbo’ model is used in this study for its advanced natural language processing capabilities. The mandatory parameter ‘prompt’ supplies GPT with the contextual information needed for dark pattern classification. The optional parameter ‘value’ denotes the value appended to the end of the prompt. To classify one single user interface text, that text is provided in this ‘value’ parameter. In this way, the ‘GPT()’ function is called for all the user interface texts in the dataset.

4.2.4 Prompt Engineering

Prompts are engineered for the large language model of GPT-3 in a way so that it can classify texts into dark pattern categories. Engineered prompts used in this study consist of three essential components - ‘Context’, ‘Input’, and ‘Output’. The ‘Input’ section serves to inform the model about the expected input and guides it on how to process it effectively, ensuring that the model understands the task at hand. Meanwhile, the ‘Output’ section describes the desired format for the model’s answers, enabling the systematic extraction and interpretation of the responses. The heart of the prompt engineering lies in the ‘Context’ section, where in-context learning is applied for the following prompting techniques.

- Zero-Shot Prompting
- One-Shot Prompting
- Few-Shot Prompting

4.2.4.1 Zero-Shot Prompting

In zero-shot prompting, no example for any of the classification categories is used. Definition and classification of dark patterns are provided as contextual information, along with the definitions of all classification categories. A template for zero-shot prompting is illustrated in Figure 4.3.

Figure 4.3: Zero Shot Prompting Template

4.2.4.2 One-Shot Prompting

In one-shot prompting, a single example is provided for each of the seven dark pattern categories, along with the other contextual information mentioned in zero-shot prompting. Since the 'Not Dark Pattern' category encompasses all texts that do not fit into the other seven categories, providing a specific example for this category could bias the language model toward that particular example. Therefore, no example is provided for the 'Not Dark Pattern' category. Each <Definition of Category X> is followed by <Example 1 of Category X>, except for the last category, 'Not Dark Pattern'. The first example for each category shown in TABLE 4.2 is used in one-shot prompting. The seven example texts are randomly selected and removed from the validation dataset. A template for one-shot prompting is illustrated in Figure 4.4.

4.2.4.3 Few-Shot Prompting

Few-shot prompting follows a similar approach to one-shot prompting but includes more than one example per category in prompts. To ensure that the approach remains as less dependent on the data as possible, only two examples are provided per category, along with the other contextual information. As with the one-shot learning scenario, no example is provided for the 'Not Dark Pattern' category to avoid bias. Each <Definition of Category X> is followed by <Example 1 of Category X> and <Example 2 of Category X>, except for the last category, 'Not Dark Pattern'. The first set of seven example texts used for the seven dark pattern categories are the same as those used in one-shot prompting, while the additional seven example texts are randomly selected. All fourteen texts are detailed in TABLE 4.2 and are excluded from the validation dataset. A template for few-shot prompting is illustrated in Figure 4.5.

Table 4.2: Example Texts Used for Dark Pattern Categories

Dark Pattern Category	**Examples**
Misdirection	1. No, I'll rather pay full price 2. No thanks, I dont want a discount
Urgency	1. Items reserved for 15:00 2. 1 day 08:15:25
Scarcity	1. 1 LEFT 2. 87% offers claimed. Hurry up!
Social Proof	1. IN 43 PEOPLE'S SHOPPING BAG 2. Armin Dinovic bought 10M Runescape 3 Gold Order total: 3,70€ About 10 seconds ago
Obstruction	1. You may change the items in your order, or cancel the Smartship at anytime, up until 3 days prior to the scheduled ship date of your Smartship by calling Customer Service at 1-800-518-0284 2. We may also disclose your information to third parties who may contact you with details of other products and services which may be of interest. If you do not want your name and mailing details made available in this way please email opt-out@nextdirect.com
Forced Action	1. I would like to join Backstage Pass & agree to the Terms & Conditions & to receive emails & other promotional offers 2. I agree to receive marketing emails from Natural Life and agree to our Privacy Policy and terms of use
Sneaking	1. Order Subtotal $19.99 Standard Delivery $12.99 Care & Handling $2.99 Tax $2.38 Total $38.35 Savings Today $10.00 2. Purchase protection added

Context
<Definition of Dark Patterns>
<Classification of Dark Patterns>

<Definition of Sneaking>
<Example 1 of Sneaking>

<Definition of Urgency>
<Example 1 of Urgency>

<Definition of Misdirection>
<Example 1 of Misdirection>

<Definition of Social Proof>
<Example 1 of Social Proof>

<Definition of Scarcity>
<Example 1 of Scarcity>

<Definition of Obstruction>
<Example 1 of Obstruction>

<Definition of Forced Action>
<Example 1 of Forced Action>

<Definition of Not Dark Pattern>

Input
<Input Format>
<Processing Instruction>

Output
<Output Format>
<Text to be classified>

<Predicted Category>

Figure 4.4: One Shot Prompting Template

Context
<Definition of Dark Patterns>
<Classification of Dark Patterns>

<Definition of Sneaking>
<Example 1 of Sneaking>
<Example 2 of Sneaking>

<Definition of Urgency>
<Example 1 of Urgency>
<Example 2 of Urgency>

<Definition of Misdirection>
<Example 1 of Misdirection>
<Example 2 of Misdirection>

<Definition of Social Proof>
<Example 1 of Social Proof>
<Example 2 of Social Proof>

<Definition of Scarcity>
<Example 1 of Scarcity>
<Example 2 of Scarcity>

<Definition of Obstruction>
<Example 1 of Obstruction>
<Example 2 of Obstruction>

<Definition of Forced Action>
<Example 1 of Forced Action>
<Example 2 of Forced Action>

<Definition of Not Dark Pattern>

Input
<Input Format>
<Processing Instruction>

Output
<Output Format>
<Text to be classified>

<Predicted Category>

Figure 4.5: Few Shot Prompting Template

4.3 Evaluation

Precision, recall, and F1 score are calculated for each category to assess the performance of the dark pattern classification, along with overall accuracy. However, recall gives a better understanding of how the proposed technique performs in classifying texts from a particular category. This is because the precision (and thus, F1 score) of a category is a concern for the performance of other categories (impacting their recall), not the category in concern. These results are detailed in TABLE 4.3. The generalizability of the detection technique is further validated using an additional test dataset of dark pattern texts.

4.3.1 Result Analysis

The entire dataset of 2,356 texts is subjected to classification using zero-shot prompting. The model accurately classifies 1,969 texts, while 387 texts are misclassified. This translates to an overall accuracy of 83.57%. The model performs significantly better in classifying texts from categories such as 'Forced Action', 'Urgency', 'Scarcity', 'Social Proof', and 'Not Dark Pattern'. In contrast, it performs very poorly when dealing with categories like 'Misdirection', 'Sneaking', and 'Obstruction', indicating the limitations of relying solely on definitions for classification, without any examples.

2,349 texts are classified using one-shot prompting. The model accurately classifies 2,137 texts, while 212 texts are misclassified. This translates to an overall accuracy of 90.97%. Like the zero-shot model, the one-shot model also performs significantly better in classifying texts from the same 5 categories. Even though the model performs poorly in the 'Sneaking' category, it performs significantly better than the zero-shot model in classifying texts into 'Misdirection' and 'Obstruction' categories. Recall of the 'Misdirection' category rises from 6.67% in zero-shot to 53.61% in one-shot, and the 'Obstruction' category rises from 11.11% in zero-shot

Table 4.3: Experimental Results of Automated Detection Using GPT-3

Prompting Technique	**Category**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Zero Shot	Misdirection	83.57%	44.83%	6.67%	11.61%
	Urgency		73.13%	93.33%	82.01%
	Scarcity		87.73%	92.34%	89.98%
	Social Proof		98.3%	92.63%	95.38%
	Obstruction		75%	11.11%	19.35%
	Forced Action		6.56%	100%	12.31%
	Sneaking		0%	0%	0%
	Not Dark Pattern		85.62%	91.51%	88.47%
One Shot	Misdirection	90.97%	97.2%	53.61%	69.1%
	Urgency		84.65%	92.34%	88.33%
	Scarcity		93.68%	95.92%	94.79%
	Social Proof		98.68%	96.46%	97.56%
	Obstruction		78.57%	84.62%	81.48%
	Forced Action		6.12%	100%	11.54%
	Sneaking		0%	0%	0%
	Not Dark Pattern		92.45%	94.65%	93.54%
Few Shot	Misdirection	92.57%	97.69%	65.8%	78.64%
	Urgency		90.65%	93.27%	91.94%
	Scarcity		92.34%	98.56%	95.35%
	Social Proof		99.32%	94.84%	97.03%
	Obstruction		82.14%	92%	86.79%
	Forced Action		4.35%	100%	8.33%
	Sneaking		0%	0%	0%
	Not Dark Pattern		94.59%	94.91%	94.75%

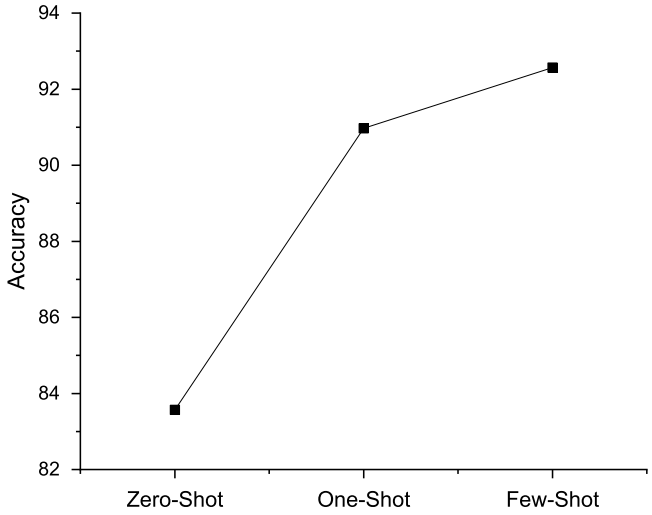


Figure 4.6: Overall Accuracy Across Prompting Techniques

to 84.62% in one-shot. This insight indicates that providing more examples for these categories might improve performance. One possible reason behind this is the wide variation in semantics for the instances of these categories.

Few-shot prompting is tested on 2,342 texts. The model accurately classifies 2,168 texts, while 174 texts are misclassified, achieving an overall accuracy of 92.57%. Like the other two models, this model also can not classify texts from the 'Sneaking' category. This category mostly leverages website logic and policy, rather than individual webpage texts. As a result, this category of dark patterns is challenging to detect only with textual information. The definition of this category provided in TABLE 4.1 also conforms to the findings. Few-shot prompting performs well for all the other seven categories of texts. It classifies texts from the 'Obstruction' category with a recall of 92%, which is a huge increase from 11.11% in zero-shot and is higher than 84.62% in one-shot. It further increases the results of the one-shot model for the 'Misdirection' category from 53.61% to 65.8%.

Figure 4.6 demonstrates that increasing the number of examples in prompts improves the overall accuracy of the proposed approach. The few-shot model with

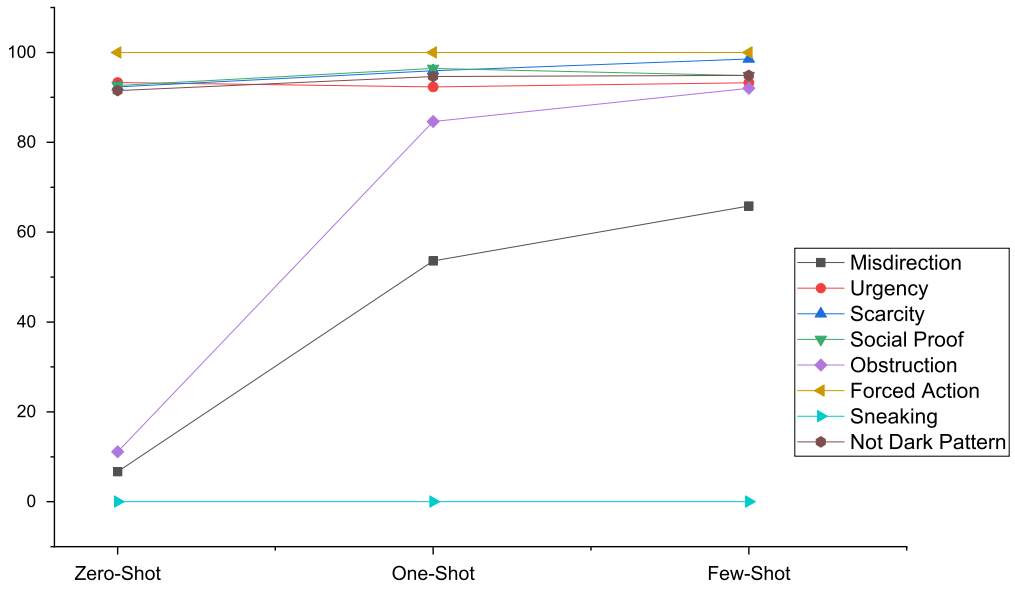


Figure 4.7: Category-Wise Performance Across Prompting Techniques

two examples per category is the best-performing model. However, providing more examples has the potential to further improve the performance of the approach.

Figure 4.7 illustrates that GPT-3 performs well in five of the eight classification categories for all three prompting techniques. Providing more examples in cases of 'Misdirection' and 'Obstruction' categories improves the classification by significant margins. The best-performing model, the few-shot model, performs well in seven of the eight classification categories. However, the proposed approach fails in the 'Sneaking' category for all three prompting techniques.

4.3.2 Generalization

A maximum of two examples per class are used to achieve the reported results, providing evidence that the proposed technique is less prone to overfitting and can generalize well. To further validate this, the proposed approach is compared with the approach used by Yada et al. [17] on a test dataset. As Yada et al. used an e-commerce dataset in their work, a comparable test dataset is created

by extracting and manually labeling dark pattern texts from 30 Bangladeshi e-commerce websites. The proposed technique is then compared with the SVM model used by them. Figure 4.8 shows that while the SVM model used by Yada et al. detected 92.2% of dark patterns in their own dataset, its performance drops to 42.8% on the new test dataset. In contrast, the proposed model using few-shot prompting performs similarly to the SVM model on the dataset of Yada et al., but outperforms it on the new test dataset, detecting 58.67% of dark patterns.

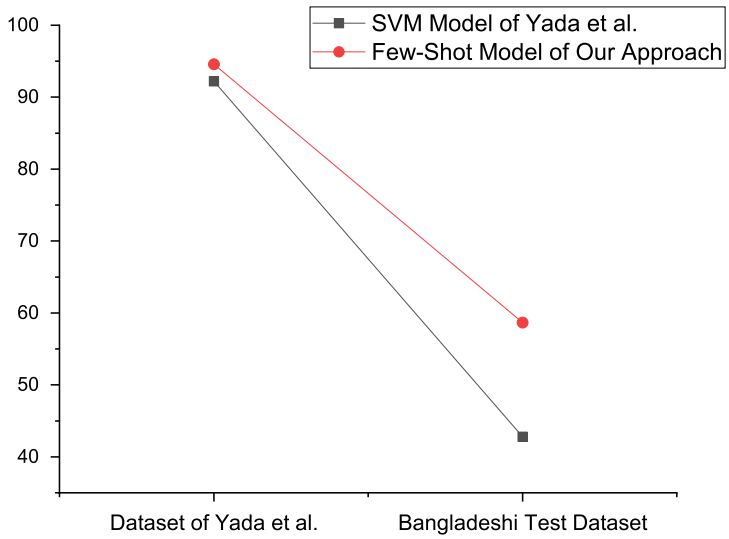


Figure 4.8: Performance Comparison with Yada et al.

4.4 Threats to Validity

This section discusses the potential threats that may affect the validity of this study and its findings.

- **Threats to External Validity**

  Evaluation on only two datasets poses an external validity threat as results may vary based on test data sources. However, contextual information is prioritized in prompt engineering to mitigate this threat. Use of GPT-3 as

the large language model of choice may also affect external validity as other LLMs may generate different results.

- **Threats to Internal Validity**

  Random selection of example texts may affect the internal validity. There is scope to incorporate other sampling techniques that may change the results. The order of dark pattern categories used in prompt engineering also poses a threat to internal validity.

4.5 Summary

The proposed automated approach for detection of dark pattern texts uses in-context learning capabilities of GPT-3 to address the diverse nature of dark patterns. Results show that the technique detects six out of the seven dark pattern categories successfully and can be generalized to other sources of dark pattern data as well. Using in-context learning allows the detection technique to reduce the risk of overfitting by emphasizing the contextual information regarding dark patterns. As a result, this technique can be as much generalized as the contextual information provided during prompting.

Chapter 5

Prevalence and User Perception of Dark Patterns in E-Commerce Websites of Bangladesh

Researchers have explored dark patterns from many different perspectives. However, most of the work related to dark patterns have been conducted in Western countries. This study investigates dark patterns within the localized context of Bangladesh, emphasizing their prevalence and user perception. Bangladeshi e-commerce websites are examined for dark patterns using a combination of automated methods, including the in-context learning capabilities of GPT-3 and a novel segmentation algorithm, alongside manual validation to mitigate the limitations of automated techniques. An empirical study is also conducted by surveying Bangladeshi university students about exposure, awareness, and concern regarding dark patterns. Based on the findings of both explorations, six dark pattern categories are ranked and subsequently divided into two novel groups with distinct characteristics - 'Active Dark Patterns' and 'Passive Dark Patterns'. The findings also reveal that educational background in technology makes users more aware and concerned about dark patterns. 18.3% of the websites analyzed in this

study contain dark patterns, indicating the prevalence of such design practices in e-commerce websites of Bangladesh.

5.1 Introduction

Existing research on dark patterns and related user perception has predominantly focused on Europe and the United States. However, research efforts in other localized contexts may generate novel insights because of cultural and linguistic differences [7]. Thus, further research is needed to analyze dark pattern data and related user perception in a localized context like Bangladesh.

As a developing country, Bangladesh has rapidly embraced digitalization, resulting in significant growth in its e-commerce sector in recent years. However, this growth also raises the risk of dark patterns that can negatively impact the increasing number of online users. Studying dark patterns in the Bangladeshi e-commerce industry can provide valuable insights into the extent of local user exposure to these manipulative design tactics. Exploring local user perception regarding dark patterns is also a crucial area of investigation.

Dark patterns have varying representations in user interfaces, making it difficult to identify and analyze related data on a large scale. As identification methods of dark patterns are still in the early stages, there have been only a few efforts to extract and analyze dark pattern data on a large scale. Mathur et al. [2] conducted the first exploration of textual dark pattern data in a large sample of 11,000 shopping websites, extracting 1,818 instances of dark pattern texts. Yada et al. [17] extended their exploration to extract 1,178 non dark pattern texts from the same websites.

Understanding user perception of dark patterns is another relevant area of research, as these design tactics abuse user psychology. Geronimo et al. [4] found that users struggle to recognize dark patterns, emphasizing the need for improved

awareness mechanisms. Bongard-Blanchy et al. [18] discovered that users can identify dark patterns but are often unaware of their harmful consequences, with demographic factors influencing user perception as well. Gray et al. [19] revealed that users can sense manipulation, even without precise terminology to describe such experience. Maier [35] found that users may initially react negatively to dark patterns but eventually get normalized to them, suggesting awareness alone might not be enough as a countermeasure. Hidaka et al. [7] revealed that Japanese applications contain fewer instances of dark patterns compared to the United States. This finding emphasizes the necessity to analyze dark pattern data and user perception in different local contexts. No such exploration has been done yet in the local context of Bangladesh.

This study examines the prevalence of dark patterns in e-commerce websites of Bangladesh. Member companies of the E-Commerce Association of Bangladesh (e-CAB) are selected for analysis to demonstrate the prevalence of dark patterns. 715 websites are analyzed and HTML contents from relevant webpages of those websites are extracted. A novel page segmentation algorithm is used to identify candidate dark pattern texts from the HTML contents. It is developed with more relaxed inclusion criteria for candidate dark patterns compared to the segmentation algorithm developed by Mathur et al. [2], resulting in improved detection of dark patterns. Detection and classification of dark patterns involves a combination of automated and manual techniques. In-context learning capabilities of GPT-3 are used to detect dark patterns from the candidate texts according to the automated detection technique explained in Chapter 4. A manual review is carried out to remove the false positives and classify the texts into different dark pattern categories.

The investigation into the prevalence of dark patterns is complemented with an empirical study on user perception of dark patterns. A survey is conducted with Bangladeshi university students belonging to two groups - one with a background

in technology education and the other without such a background. Participants are presented with user interface design instances of different dark pattern categories and then exposure, awareness, and concern ratings are collected for each dark pattern category.

931 instances of dark patterns are found in the analysis, 99 of which are unique instances. These dark pattern instances are classified into six common dark pattern categories, including 'Misdirection', 'Urgency', 'Scarcity', 'Social Proof', 'Obstruction', and 'Forced Action'. The 931 dark pattern instances belong to 131 different websites, which is roughly 18.3% of the 715 e-commerce websites analyzed in this study. Average ratings of awareness and concern across all dark pattern categories reveal that users are aware of the influences of dark patterns and are concerned about such design tactics in general. Based on the findings from the user perception survey and the prevalence data across dark pattern categories, the categories are divided into two distinct groups - 'Passive Dark Patterns' containing 'Social Proof', 'Urgency', and 'Scarcity' categories and 'Active Dark Patterns' containing 'Forced Action', 'Obstruction', and 'Misdirection' categories. Distinct characteristics of these two groups are found in terms of prevalence, user awareness, and concern. User perception survey also reveals that users with a background in technology education are more aware and concerned about dark patterns than users with no such background. The main contributions of this study can be summarized as follows -

- Analysis of the prevalence of dark patterns in e-commerce of Bangladesh
- Novel page segmentation algorithm that extracts candidate dark pattern texts from HTML contents of webpages
- Collection of a dark pattern dataset containing 99 unique text instances labelled with six dark pattern categories
- Ranking dark pattern categories by user exposure, awareness, and concern

- Novel grouping of dark pattern categories according to their approach of abusing user psychology
- Revealing the influence of technology education (computer science or related fields) on user awareness and concern regarding dark patterns

5.2 Prevalence of Dark Patterns in E-Commerce Websites of Bangladesh

In this study, dark patterns from e-commerce websites of Bangladesh are extracted to analyze the local context and show the prevalence of such design tactics in local e-commerce websites. A sampling frame is defined first for local e-commerce websites, followed by automated crawling of relevant website data. The crawled textual data is then processed using a combination of automated and manual methods to retrieve dark pattern texts across different categories. Figure 5.1 provides an overview of the different stages of data extraction carried out for this purpose. The resulting dataset of dark patterns and other technical artifacts associated with this study are publicly available on GitHub[1].

5.2.1 Sampling Procedure

The sampling frame for this study is derived from the membership list of the E-Commerce Association of Bangladesh (e-CAB)[2], which is a Bangladeshi local organization that has more than 2,000 e-commerce companies as members. The members' list of e-CAB is the largest available representative sample of Bangladeshi e-commerce companies. Thus, this list of ECAB members is considered as the sampling frame for this study.

[1]https://github.com/bsse1006/dark-patterns-bangladesh
[2]https://e-cab.net

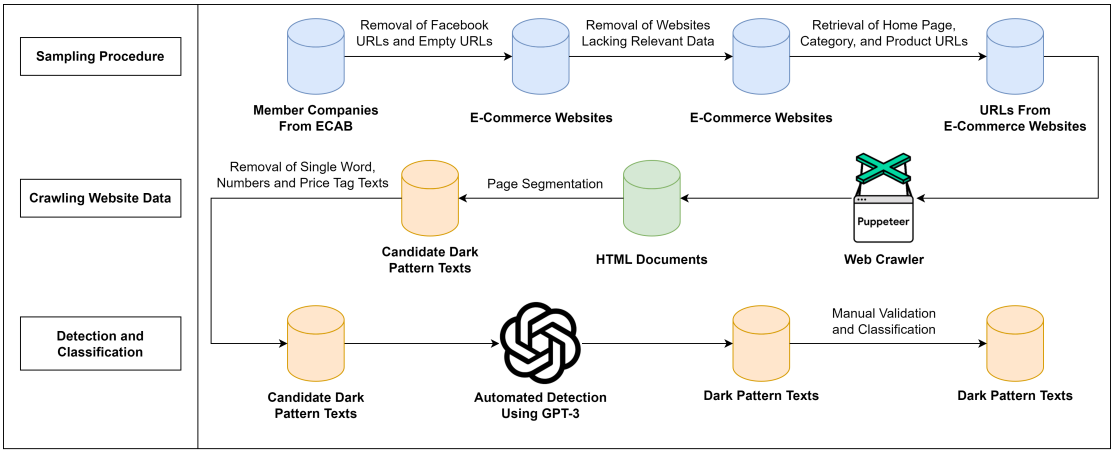


Figure 5.1: Overview of Dark Pattern Data Extraction

Upon inspecting the members' list of e-CAB available online at 'https://e-cab.net/member-list', a general pattern is found in webpage URLs associated with each e-CAB member profile - 'https://e-cab.net/company-profile/<member id>', where <member id> refers to the numbers '0001' to '2238' which are distinct member id numbers associated with each e-CAB member companies. The Javascript library 'puppeteer' is used to crawl the 'company name' and 'website URL' data for each e-CAB member company from their associated member profile pages using this general URL pattern. Excluding the member profile pages that are unavailable, data related to 2,216 member companies of e-CAB is collected. Two exclusion criteria is applied to these 2,216 instances of data - (1) 104 companies are removed because their associated 'website URL' data contain URLs of Facebook pages, not of company websites, and (2) 57 companies are removed for not having any specified 'website URL' at all. After excluding these companies, a total of 2,055 e-commerce websites are considered for further inspection.

5.2.2 Crawling Website Data

The 2,055 e-commerce websites are crawled to extract text segments from them. All the websites are manually examined first to collect and store relevant URLs, followed by automated crawling of the HTML contents from those URLs. Sub-

sequently, a page segmentation algorithm is applied on the HTML contents to identify text segments for further inspection.

5.2.2.1 URL Retrieval

The home or landing page of all of the 2,055 company websites are manually visited. However, 650 websites are unreachable. The remaining 1,405 websites are inspected to collect two additional URLs alongside the home page URL from each website. For product-based companies, one category page and one product page per website are navigated. On the other hand, for service-based companies, the offered services page and the portfolio page are navigated. In this way, all the websites are manually inspected to collect URLs of the home page, one category/services page, and one product/portfolio page. Two exclusion criteria are applied - (1) websites that have static pages with no further navigation available, and (2) websites that lack relevant information. These exclusion criteria reduce the number of websites to 715, from which 2,119 URLs are collected.

5.2.2.2 HTML Data Retrieval

The Javascript library 'puppeteer' is used again to crawl the 2,119 URLs and retrieve the whole HTML content associated with each URL. Some websites detect the 'puppeteer' crawler as a bot and restrict it from accessing the original HTML content. The 'puppeteer-extra-plugin-stealth' library is used to bypass this bot detection. This library conceals certain properties from outgoing HTTP requests that can flag the requests as coming from a bot. HTML contents of 51 URLs are not retrieved because of connection and navigation problems. The HTML contents of the remaining 2,068 URLs are collected.

5.2.2.3 Candidate Dark Pattern Extraction

Meaningful text segments need to be extracted from HTML contents that are called 'candidate dark patterns'. For this purpose, a segmentation algorithm is needed to analyze webpages and identify candidate dark patterns. Mathur et al. [2] proposed such a segmentation algorithm, where segments are defined as visible HTML elements that contain no other block-level elements and at least one text element. Using this algorithm leads to the discovery of very few dark patterns in the HTML contents. Upon exploring the causes, it is found that the algorithm is too restrictive, discarding potential dark pattern texts.

The segmentation algorithm is modified to make the inclusion criteria as relaxed as possible. Segments are defined as any visible HTML element that is a leaf (in other words, does not contain any child elements) and contains text content. However, paragraph elements are excluded from this definition, as they usually represent longer blocks of descriptive text that are not relevant to the analysis. Details of the new page segmentation algorithm are illustrated in Algorithm 1.

Algorithm 1: Page Segmentation

Input: HTML Document
Output: List of Text Segments

```
1  Function extractTextSegments(page):
2      ignoredElements ← ["p", "script", "style", "noscript", "br", "hr"];
3      segmentsList ← [];
4      foreach element in page do
5          if element is not in ignoredElements and element is visible and
             element does not contain any child elements then
6              innerText ← element.innerText;
7              segmentsList.append(innerText);
8          end
9      end
10     return segmentsList;
```

This version of the segmentation algorithm results in the discovery of both, increased candidate dark patterns and subsequently increased dark patterns. So this algorithm is used to extract candidate dark pattern texts from the HTML

contents obtained in the previous step. 770,570 texts are extracted from the 2,068 HTML contents under analysis.

5.2.3 Detection and Classification

It is a time-consuming task to detect dark patterns by manually reviewing the large set of 770,570 texts. Thus, a combination of automated and manual approaches is employed to detect and classify dark pattern texts from the candidate texts. Dark pattern texts are first detected using an automated approach leveraging machine learning. Then a manual review of each text and associated URLs is carried out to remove the false positives and classify the remaining texts into different dark pattern categories.

5.2.3.1 Data Preprocessing

Candidate dark pattern texts that consist of only numbers or a single word are removed. Price tag texts containing the local currency sign are also removed using regular expressions. This preprocessing reduces the set of candidate dark pattern texts by 43% to 435,536 texts. Removing the duplicates would reduce the set even further. However, multiple websites or multiple webpages within the same website can contain the same dark pattern text. Thus, removing duplicates would misrepresent the actual prevalence of dark patterns in e-commerce websites of Bangladesh. As a result, the duplicates are not removed in this step.

5.2.3.2 Automated Dark Pattern Detection

Machine learning is used to automatically classify the candidate texts as 'Dark Pattern' or 'Not Dark Pattern'. Yada et al. [17] provided some baseline machine learning models that could be used for this purpose. Initially, their best-performing model ($RoBERTa_{large}$) is used for automated detection, trained on 1,178 dark pattern and the same number of non dark pattern texts provided in

their dataset. The hyper-parameters are set according to their work. However, this approach results in a large number of false positives. As a result, the subsequent manual verification would be time-consuming, undermining the purpose of automated detection. Upon inspecting the causes behind high false positives, it is found that the model is vulnerable to data overfitting.

An automated detection technique is developed in Chapter 4 that can detect dark patterns without the need for training data, thus eliminating the concern regarding data overfitting. This approach utilizes the in-context learning capabilities of GPT-3 to detect dark pattern texts. The best-performing model from this study is the few-shot model where two example texts per category are used along with the category definitions to engineer prompts for GPT-3. This approach is able to detect six common categories of dark pattern texts successfully - 'Misdirection', 'Urgency', 'Scarcity', 'Social Proof', 'Obstruction', and 'Forced Action'. As the approach using in-context learning results in a very low rate of false negatives (under 10%) according to experimental results discussed in Chapter 4, this approach is used to detect dark pattern texts from the candidate texts obtained in the previous step. To optimize resource usage, only the unique texts are considered in this step. Texts with numerical variations (for example, 'Only 4 left in stock' and 'Only 1 left in stock') are treated as duplicates, ensuring that similar texts are not processed by GPT-3 separately. There are 91,585 such unique texts among the set of 435,536 candidate texts. The automated approach detects 2,185 out of those 91,585 texts as dark patterns.

5.2.3.3 Manual Dark Pattern Classification

Each of the 2,185 dark pattern texts is manually reviewed to check whether they are really instances of dark patterns or not. For this purpose, the whole set of texts are manually labelled by three researchers of our team separately. Conflicts in labelling are then resolved through exhaustive discussions between the three re-

searchers. During this manual labelling process, dark pattern categories explained by Mathur et al. [2] are followed, which has been presented in Chapter 2. After this manual labelling of texts, 2,086 instances are discarded as they do not belong to any dark pattern category considered in this study. The remaining 99 instances are classified into their respective dark pattern categories following the same process. During the manual review, if the textual data is insufficient for labelling, the associated URL is visited to gather context about the text. This results in a dataset of dark patterns containing 99 unique text instances labelled with their respective dark pattern categories. As only the unique texts are considered in the previous step, the set of 435,536 candidate texts is revisited to extract and label the duplicates matching the 99 unique dark pattern texts. Numerical variations are also treated as similar texts in the previous step. Thus, all such groups of texts are labelled with the same dark pattern category. After considering the duplicates and numerical variations, the final set of dark patterns consists of 931 instances.

5.3 User Perception of Dark Patterns

An empirical study is conducted to analyze user perception regarding the presence of dark patterns in e-commerce websites of Bangladesh. This study aims to evaluate and rank dark pattern categories based on users' exposure, awareness, and concern. Additionally, the analysis investigates potential differences in user perception based on participants' educational background in technology.

5.3.1 Study Design

An open survey is conducted, targeting primarily Bangladeshi university students. Initially, participants are queried about their general understanding of the capacity of user interface designs to influence user decision-making, as well as the frequency of their use of e-commerce websites and applications. Later, participants are

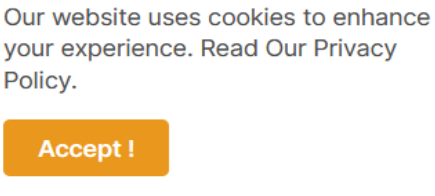
presented with six user interface design instances containing dark patterns from the six categories previously mentioned. Presenting only dark pattern instances could bias participants in favour of a negative perception regarding the design instances. To counteract that possibility, four non dark pattern design instances are also presented in the survey. These ten user interface design instances are illustrated in Figure 5.2. Participants are presented with these design instances in a random order and three ratings are measured for each design. 'Exposure' rating is measured based on whether the participant has ever encountered such a design or not (0 = never encountered, 1 = encountered). To measure 'Awareness' and 'Concern' ratings, the following two statements are presented respectively.

- This kind of design can influence users to do something that users normally would not do.
- I am worried about the influence of this kind of design on users' decision-making and choices.

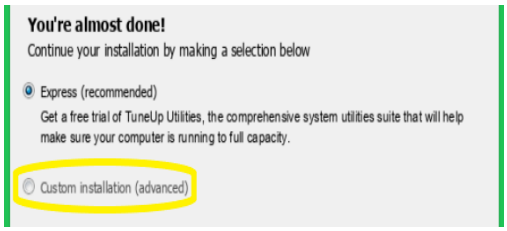
Participants are instructed to rate their agreement on a 5-point Likert scale (from 1 = strongly disagree to 5 = strongly agree). In addition, demographic information (age, sex, and educational background) is collected from participants.

5.3.2 Demographics of Participants

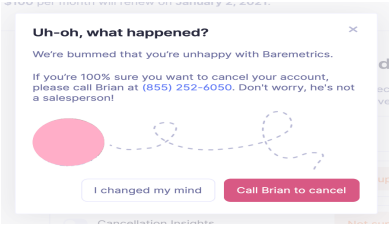
Survey responses are collected from 68 participants. The demographics of the participants are as follows - 45 male, 23 female. Their age range from 19 to 27 years (mean 22.57, SD 1.47). Participants are divided into two groups based on their educational background - 36 participants have an educational background in computer science or related fields and the other 32 have no such background in technology education. The rationale behind this grouping is that participants with an educational background in technology may have different perceptions of dark patterns compared to those without such a background.



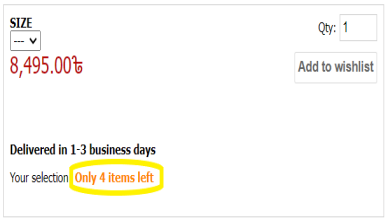
(a) Forced Action



(b) Misdirection

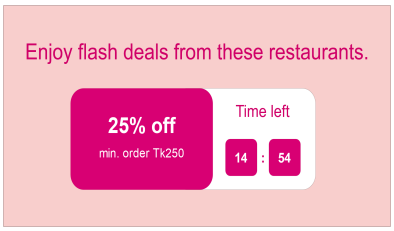


(c) Obstruction



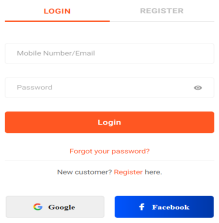
(d) Scarcity

(e) Social Proof

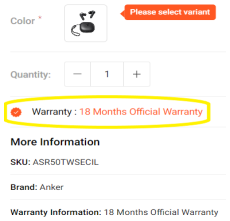


(f) Urgency

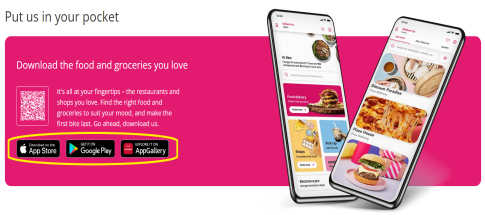
(g) Not Dark Pattern



(h) Not Dark Pattern



(i) Not Dark Pattern



(j) Not Dark Pattern

Figure 5.2: User Interface Design Instances Used in the Survey

5.4 Result Analysis

This section discusses the findings regarding the prevalence of dark patterns in e-commerce websites of Bangladesh as well as the user perception survey.

5.4.1 Prevalence of Dark Patterns

The extracted 931 dark pattern texts are analyzed to showcase the prevalence of dark patterns in e-commerce websites of Bangladesh. These dark pattern texts are found across 131 websites, constituting approximately 18.3% of the 715 analyzed e-commerce websites. This is a much higher rate compared to the work of Mathur et al. [2], who found dark patterns in 11.1% of their analyzed websites. The increased rate can be credited to the use of a more relaxed page segmentation algorithm than theirs, resulting in improved detection of dark patterns. It is important to note that this study exclusively analyzes the home page, one category/services page, and one product/portfolio page per website and the analysis only focuses on the texts of these pages. Given the scope of the exploration, this finding serves as a conservative estimate. As a result, it can be said that the actual prevalence of dark patterns in the e-commerce websites of Bangladesh is even higher.

Figure 5.3 illustrates the frequency of each dark pattern category. 'Social Proof' and 'Urgency' have the highest number of instances among all the categories with 466 and 271 occurrences respectively. 165 instances belong to the 'Scarcity' category while 'Forced Action' and 'Misdirection' have 23 and 6 occurrences respectively.

Figure 5.4 illustrates the number of unique instances across each dark pattern category. 'Urgency' tops the other categories in this regard, with 41 unique instances. 'Social Proof' and 'Scarcity' have 26 and 24 unique instances respectively while 'Forced Action' and 'Misdirection' have only 6 and 2 respectively. No instance of the 'Obstruction' category is found in the study. This prevalence

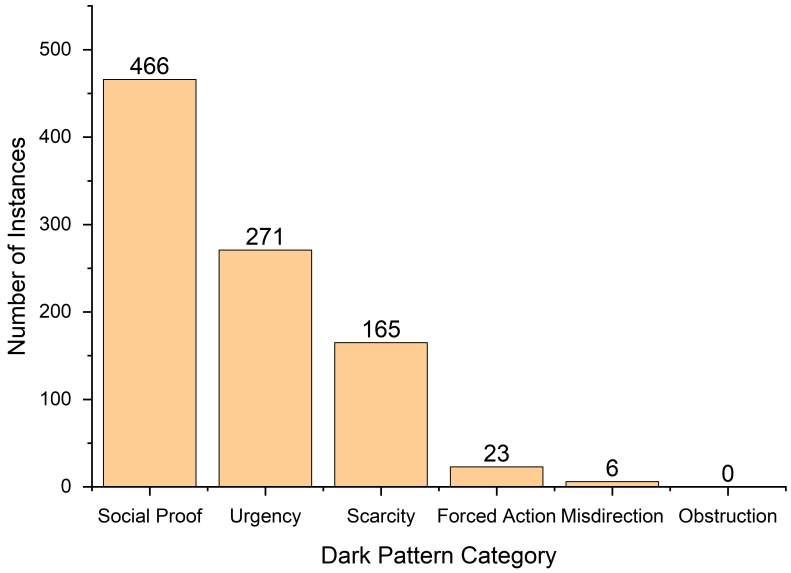


Figure 5.3: Number of Instances Across Dark Pattern Categories

of unique dark pattern texts is compared across categories with the findings of Mathur et al. [2], who also gathered unique instances of dark pattern texts in their work. While the rankings of dark pattern categories in terms of prevalence do not perfectly align between the two studies, two distinct groups emerge. Both analyses indicate that 'Urgency', 'Social Proof', and 'Scarcity' (Group A) had the highest number of unique instances, while 'Forced Action', 'Misdirection', and 'Obstruction' (Group B) had relatively fewer unique instances. As a result, it can be said that Group A dark patterns are more prevalent in the analysis than Group B dark patterns.

Figure 5.5 also confirms such a grouping of dark pattern categories in terms of prevalence across websites. 'Urgency', 'Scarcity', and 'Social Proof' have the more prominent presence, with occurrences on 78, 28, and 26 websites respectively. 'Forced Action', 'Misdirection', and 'Obstruction' are much less prominent, with occurrences on only 7, 2, and 0 websites respectively. Interestingly, 121 of the 131 websites are associated with a single dark pattern category, with the remaining 10 being associated with at most two dark pattern categories.

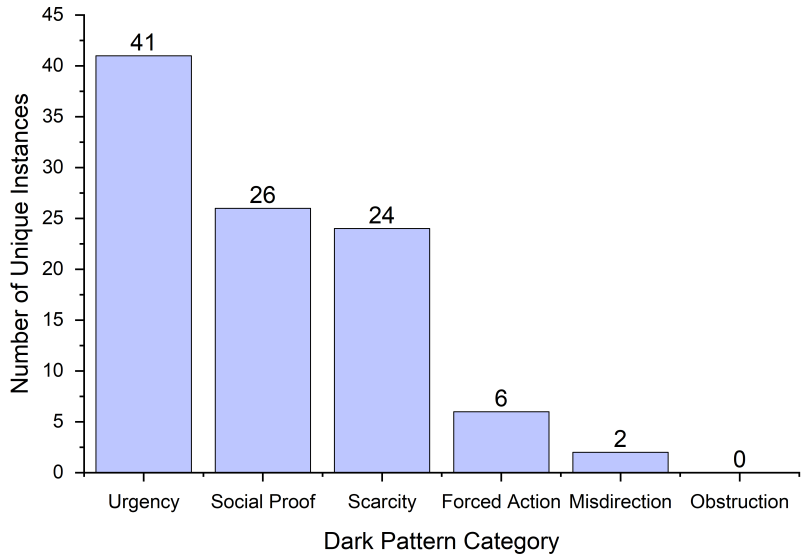


Figure 5.4: Number of Unique Instances Across Dark Pattern Categories

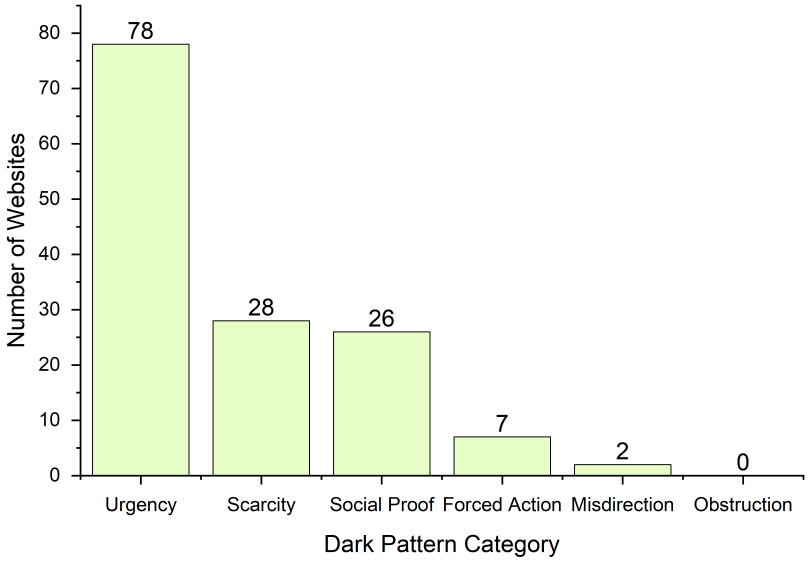


Figure 5.5: Number of Websites Across Dark Pattern Categories

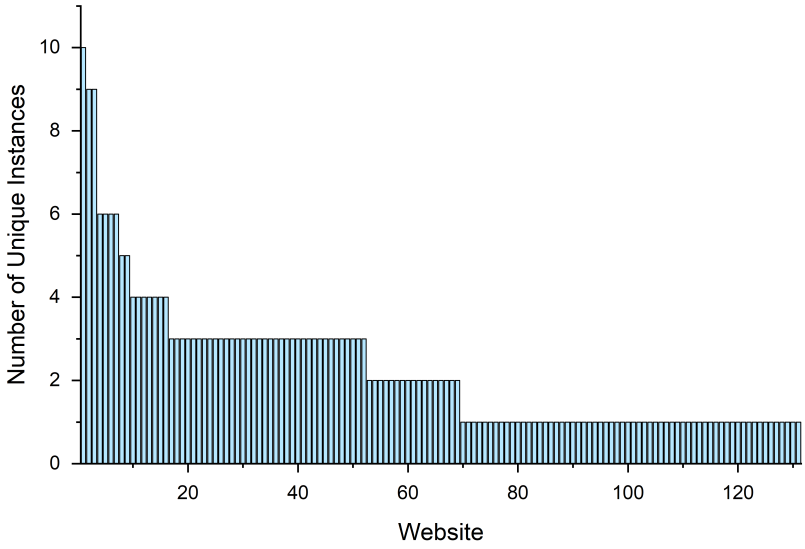


Figure 5.6: Number of Unique Instances Across Websites

Figure 5.6 illustrates the frequency of dark patterns across the 131 websites. Most of the websites present only 1, 2, or 3 unique dark pattern texts, with a few presenting larger quantities. The highest number of unique instances on a single website is 10.

5.4.2 User Perception of Dark Patterns

Based on the survey responses, the dark pattern categories are ranked and grouped according to levels of exposure, awareness, and concern. Additionally, a Mann-Whitney U test is conducted to determine if there are significant differences in user perception based on educational background in technology.

5.4.2.1 Ranking Dark Pattern Categories

Survey data are analyzed to rank different categories of dark patterns based on users' ratings of exposure, awareness of their influence, and concern regarding them. Rankings are determined by calculating the mean ratings of exposure, awareness, and concern for each dark pattern category by averaging all participant

responses. The resulting rankings are provided in Table 5.1. Notably, participants consistently rate their awareness and concern regarding dark patterns above the neutral value of 3, indicating a general awareness of the potential influences of dark pattern design instances and significant concern about them. However, the category rankings of awareness and concern do not perfectly align with each other, suggesting that participants are not necessarily more concerned about a design simply because they perceive it to be more influential, and vice versa.

5.4.2.2 Grouping Dark Pattern Categories

Exposure, awareness, and concern ratings produce further evidence in favour of the grouping of dark pattern categories in the previous section. There is a clear distinction in the ratings between the two groups of dark patterns. Participants are more exposed to Group A dark patterns than Group B, except for 'Forced Action'. Such a contrast in the exposure ratings makes more sense as they align directly with the prevalence of these two groups explained in the previous section. Participants find Group A containing 'Urgency', 'Scarcity', and 'Social Proof' categories more influential than Group B containing 'Misdirection', 'Obstruction', and 'Forced Action'. The reverse tendency is found in concern ratings as participants find Group B more concerning than Group A, except for 'Misdirection'.

The contrast between awareness and concern ratings across the two groups of dark pattern categories can be explained by how these design instances manifest themselves in user interfaces. Group A dark patterns get users' attention to certain actions using the attraction of limited sales or discounts, and the rarity or credibility of a product. Even though they can create a sense of hurry among users, the user experience is not necessarily unpleasant, resulting in less concern. On the other hand, Group B dark patterns make users feel uncomfortable as they get controlled, forced, or deceived by the user interface. As a result, users are more concerned about this group of dark pattern categories. According to this

Table 5.1: Dark Pattern Category Rankings Based on User Perception

Variable	**Rank**	**Dark Pattern Category**	**Mean Rating**
Exposure	1	Forced Action	94.12%
	2	Urgency	88.24%
	3	Scarcity	86.76%
	4	Social Proof	77.94%
	5	Misdirection	73.53%
	6	Obstruction	30.88%
Awareness	1	Urgency	4.21
	2	Scarcity	3.99
	3	Social Proof	3.96
	4	Obstruction	3.96
	5	Forced Action	3.81
	6	Misdirection	3.65
Concern	1	Obstruction	3.81
	2	Forced Action	3.78
	3	Urgency	3.56
	4	Scarcity	3.44
	5	Social Proof	3.22
	6	Misdirection	3.17

finding, Group A dark patterns are named as 'Passive Dark Patterns' and Group B dark patterns are named as 'Active Dark Patterns'.

- **Passive Dark Patterns**

  This group of dark patterns consists of 'Urgency', 'Social Proof', and 'Scarcity' categories. All of these techniques use users' 'fear of missing out' (FOMO) and create a sense of attraction in users to take certain actions. They rely on social conformity and the perception of limited availability (time or quantity) to motivate users to take action quickly. The main identifying factor of these dark patterns is that they attract users, rather than forcing them. Because of their subtle presentation in user interfaces, they do not seem abnormal to users, even when users think they have a high influence on their decision-making. As passive dark patterns have been 'normalized' to users, finding countermeasures for this group of dark patterns is more difficult. As a result, safeguarding against these strategies necessitates legal and policy conversations regarding marketing and advertisement strategies in general.

- **Active Dark Patterns**

  This group consists of 'Forced Action', 'Misdirection', and 'Obstruction' categories. These techniques manipulate user behavior through deception, obstacles, or coercion. These design instances forcefully drive users to act in ways that are against their wishes or interests. The main identifying factor of these dark patterns is that they mislead or force users to make certain choices. As a result, users often feel uncomfortable while encountering these dark patterns. These techniques undermine user autonomy and can lead to a negative user experience. Thus, this group of dark patterns does not fall into traditional marketing or advertisement strategies. Legal and policy regulations over design principles of online platforms can have a significant impact on counteracting these dark patterns.

Two deviations are visible in Table 5.1 from the analysis regarding the grouping of dark pattern categories - 'Forced Action' in the case of exposure ratings and 'Misdirection' in the case of concern ratings. These deviations are due to participant bias or misunderstanding about the particular design instances used in both cases. As shown in Figure 5.2a, participants are presented with a cookie banner featuring solely an 'Accept' button without a corresponding 'Reject' button, as an example of 'Forced Action'. The heightened frequency of responses indicating high exposure to 'Forced Action' could stem from participants' familiarity with cookie banners rather than the dark pattern design featuring only an 'Accept' button and lacking a 'Reject' button. As shown in Figure 5.2b, participants are presented with an installer interface containing two choices where one is recommended and the other is blurred, as an example of 'Misdirection'. Participants could have perceived such a recommendation as a helpful thing and not considered situations where such preselection and design interference could go against their interests; leading to less concern about this category.

5.4.2.3 Technology Education and User Perception of Dark Patterns

A Mann-Whitney U test is used to investigate differences in user perception between participants with an educational background in technology and those without such a background. For this purpose, ratings of exposure, awareness, and concern across all dark pattern categories are averaged to get mean exposure, awareness, and concern ratings for each participant, which are the three variables used in hypothesis testing. For each of the three variables, the null and alternative hypotheses are as follows (CS = participants with an educational background in technology, NCS = participants with no educational background in technology) -

- H_0: Ratings from CS are lower than or equal to that of NCS
- H_a: Ratings from CS are greater than that of NCS

Table 5.2 presents the results from the one-tailed Mann-Whitney U test (level of significance, $\alpha = 0.05$). While the null hypothesis is not rejected for exposure, it is rejected for awareness and concern; indicating statistically significant evidence that participants with an educational background in technology are more aware and concerned about dark patterns than participants without such a background.

Table 5.2: Results of Mann-Whitney U Test ($n1$=36, $n2$=32)

Variable	**Mann-Whitney U Statistic**	**P Value**	H_0 **Rejected**
Exposure	614	0.3184	No
Awareness	790	0.0042	Yes
Concern	739	0.0225	Yes

5.5 Threats to Validity

This section discusses the potential threats that may affect the validity of the study and its findings.

- **Threats to External Validity**

  The sampling frame of the study poses an external validity threat as not all e-commerce websites of Bangladesh are members of e-CAB. However, this is the largest representative sample of e-commerce websites originated in Bangladesh. Analyzing a few URLs per website also poses a threat, as other URLs may produce different results. However, e-commerce websites generally have similar user interface designs across category/services pages and product/portfolio pages. That is why one instance of each of these webpage types is considered, along with the homepage. The survey participants in the empirical study are university students. Thus, the findings may differ for other age groups and educational levels.

- **Threats to Internal Validity**

  Survey participants could have made assumptions about which responses are socially desirable for this study, posing a threat to the internal validity. However, this bias is mitigated by communicating that there are no 'right' answers. Random ordering of the user interface design instances also poses a threat to the internal validity.

- **Threats to Construct Validity**

  The detection and classification techniques used in the study may affect the construct validity. Automated detection may have missed some dark pattern data as false negatives. However, the manual classification is carried out carefully to reduce the remaining threats as much as possible. Survey participants always can misunderstand questions or appropriate contexts. To mitigate this threat, a physical survey is carried out instead of a remote one, allowing clear explanations to be provided to the participants.

5.6 Summary

This study marks the first exploration into dark patterns in the context of Bangladesh. Automated and manual methods are combined to extract dark patterns from the websites of member companies of the E-Commerce Association of Bangladesh (e-CAB). 715 websites are analyzed and dark patterns are exposed in 18.3% of those websites, which is 7.2% higher than in previous research using a similar approach. 68 university students are also surveyed about their perception of dark patterns. Based on the findings from both explorations, dark pattern categories are divided into two distinct groups - 'Passive Dark Patterns' and 'Active Dark Patterns'. Most of the dark pattern instances found in this study belonged to 'Passive Dark Patterns'. Analysis also reveals that users with a background in technology education are more aware and concerned about dark patterns than other users.

Chapter 6

Conclusion

Dark patterns — deceptive design tactics used to manipulate user decision-making — can lead to financial losses, unnecessary sharing of personal information, and the fostering of addictive behaviours among users. As a result, there has been a growing research interest in analyzing these design strategies. Researchers have focused on defining and classifying dark patterns, as well as developing automated detection techniques. Additionally, policy implications and user perception of dark patterns have been explored. Researchers have also extracted dark pattern related data in various formats, including video recordings, screenshots, extracted features, and textual data. Most of this research has been conducted in Europe and the United States, likely due to greater consumer protection awareness in these developed regions. However, findings from such analyses may vary in different parts of the world due to cultural and linguistic differences. In developing countries like Bangladesh, where rapid digitalization is occurring without sufficient consumer protection policies, users may be more vulnerable to dark patterns. Thus, this research investigates the prevalence of dark patterns in e-commerce websites of Bangladesh, alongside related user perception. Such an analysis necessitates effective extraction of dark pattern data, which relies heavily on the successful detection of these deceptive tactics on websites. Previous studies employed ma-

chine learning and heuristic-based techniques for the automated detection of dark patterns. However, existing methods often prove impractical or lack the generalizability required for real-world application. To address these limitations, this research proposes a novel automated detection technique for dark patterns as well.

6.1 Automated Detection and Classification of Dark Patterns

For automated detection of dark patterns, a novel approach is proposed using in-context learning capabilities of large language models. The main objective of this approach is to increase the generalizability of automated detection. To achieve this, definitions of dark pattern categories are inclusively synthesized from existing literature and used as contextual information to leverage the in-context learning capabilities of large language models. Zero, one, or two examples per dark pattern category are provided for prompt engineering along with the synthesized definitions. No usage of training or fine-tuning in this approach increases the generalizability of the automated detection of dark patterns.

The proposed approach is validated on the textual dark pattern dataset provided by Mathur et al. [2] and Yada et al. [17]. On this validation dataset, experimental results with GPT-3 as the large language model, achieve satisfactory performance for six of the seven dark pattern categories considered, as well as the non dark pattern category. Even more than the classification performance, it is important to note that this performance is achieved without any training or fine-tuning. Thus, this approach is less susceptible to overfitting, increasing its generalizability. To further validate this, experimental results of this approach are compared with the results of Yada et al. [17] on a test dataset. The proposed approach performs better on this dataset, highlighting the applicability of in-context learning in improving the generalizability of automated dark pattern detection.

6.2 Prevalence and User Perception of Dark Patterns in E-Commerce Websites of Bangladesh

A case study is executed on e-commerce websites of Bangladesh to expose the prevalence of dark patterns in such websites. More than 2000 member company websites of the E-Commerce Association of Bangladesh (ECAB) are considered for analysis. After excluding static websites and websites with no relevant data, 715 websites are analyzed and HTML contents from relevant webpages of those websites are retrieved. A novel page segmentation algorithm is used to identify meaningful text segments from the HTML contents. The algorithm is developed with more relaxed inclusion criteria compared to the segmentation algorithm developed by Mathur et al. [2], resulting in improved detection of dark patterns. Detection of dark patterns is carried out with in-context learning capabilities of GPT-3, while a manual review is carried out to remove the false positives and classify the texts into different dark pattern categories. 931 dark pattern instances are found in the analysis, belonging to 131 different websites, which is approximately 18.3% of the analyzed sample of 715 e-commerce websites. 99 of these dark pattern instances are unique, which can be used as a textual dark pattern dataset for future research efforts.

The study on the prevalence of dark patterns is complemented by an empirical study on user perception of dark patterns. A survey is conducted with Bangladeshi university students. Participants in the survey are divided into two groups - one with a background in technology education and the other without such a background. Participants are presented with user interface design instances of different dark pattern categories and then exposure, awareness, and concern ratings are collected for each dark pattern category. According to average ratings of awareness and concern across all dark pattern categories, users are aware of the influences of dark patterns and are concerned about them in general. Based on the findings

from the user perception survey and the prevalence data, dark pattern categories can be divided into two distinct groups - 'Passive Dark Patterns' comprising 'Social Proof', 'Urgency', and 'Scarcity' categories and 'Active Dark Patterns' comprising 'Forced Action', 'Obstruction', and 'Misdirection' categories. These two groups express distinct characteristics in terms of prevalence, user awareness, and concern. It is also revealed in the user perception survey that users with a background in technology education are more aware and concerned about dark patterns than users with no such background.

6.3 Future Work

This research is a primary exploration of the automated detection of dark patterns, the prevalence of such strategies in e-commerce websites of Bangladesh, and related user perception. From the processes and findings of this research, extensions can be studied in the following directions.

- **Developer Perception of Dark Patterns**

  This research carries out an empirical study on user perception of dark patterns. To complement this, it is also important to analyze developer or designer perception regarding dark patterns. Future research efforts can explore such analyses, and compare user and developer perceptions. Analyzing developer perception of dark patterns can bring novel insights to light, for example, whether developers are aware of these tactics, whether and why they think such practices to be unethical or not, what their intentions are in employing such strategies, etc. Such insights can help in formulating what really constitutes a dark pattern, creating countermeasures for dark patterns, creating developer or designer guidelines, and comparing developer and user perceptions of dark patterns.

- **Legal and Regulatory Frameworks for Bangladesh**

  An important reason behind the efforts to expose the prevalence of dark patterns in Bangladesh is to motivate potential legal and regulatory frameworks. Such frameworks should be the ultimate goal in safeguarding Bangladeshi online platform users against such deceptive design tactics. However, this necessitates further research with the help of collaboration between the fields of Bangladeshi law and human-computer interaction.

- **Automated Detection Tool for Dark Patterns**

  This research presents a novel approach for textual dark pattern detection as well as previous works on the topic. Future efforts should try to incorporate this approach with previous works to implement an automated dark pattern detection tool. Such a tool can be used in many real-life scenarios, for example, helping users detect dark patterns in user interfaces, helping researchers extract dark pattern related data from large numbers of webpages, etc.

- **Cultural and Linguistic Differences in Dark Patterns**

  This research presents a localized analysis of dark patterns in the context of Bangladesh. This serves as a starting point that can motivate future research focusing on more specific cultural and linguistic factors related to dark patterns. Especially, the usage of the Bengali language in user interfaces needs to be analyzed in terms of automated dark pattern detection. It is also important to explore novel types of dark patterns in Bangladeshi websites.

Bibliography

[1] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of UX design," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, apr 2018.

[2] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark patterns at scale," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1–32, nov 2019.

[3] J. Luguri and L. Strahilevitz, "Shining a light on dark patterns," *SSRN Electronic Journal*, 2019.

[4] L. D. Geronimo, L. Braz, E. Fregnan, F. Palomba, and A. Bacchelli, "UI dark patterns and where to find them," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, apr 2020.

[5] S. M. H. Mansur, S. Salma, D. Awofisayo, and K. Moran, "AidUI: Toward automated recognition of dark patterns in user interfaces," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, IEEE, may 2023.

[6] I. Stavrakakis, A. Curley, D. O'Sullivan, D. Gordon, and B. Tierney, "A framework of web-based dark patterns that can be detected manually or automatically," 2021.

[7] S. Hidaka, S. Kobuki, M. Watanabe, and K. Seaborn, "Linguistic dead-ends and alphabet soup: Finding dark patterns in japanese apps," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2023.

[8] A. M. Roffarello and L. D. Russis, "Towards understanding the dark patterns that steal our attention," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, ACM, apr 2022.

[9] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher, "Tales from the dark side: privacy dark strategies and privacy dark patterns.," *Proc. Priv. Enhancing Technol.*, vol. 2016, no. 4, pp. 237–254, 2016.

[10] J. P. Zagal, S. Björk, and C. Lewis, "Dark patterns in the design of games," 2013.

[11] H. Brignull, M. Leiser, C. Santos, and K. Doshi, "Deceptive patterns – user interfaces designed to trick you." `https://www.deceptive.design/`, 2023.

[12] L. Jarovsky, "Dark patterns in personal data collection: Definition, taxonomy and lawfulness," *SSRN Electronic Journal*, 2022.

[13] A. Chaudhary, J. Saroha, K. Monteiro, A. G. Forbes, and A. Parnami, ""are you still watching?": Exploring unintended user behaviors and dark patterns on video streaming platforms," in *Designing Interactive Systems Conference*, ACM, jun 2022.

[14] S. Greenberg, S. Boring, J. Vermeulen, and J. Dostal, "Dark patterns in proxemic interactions," in *Proceedings of the 2014 conference on Designing interactive systems*, ACM, jun 2014.

[15] C. Lacey and C. Caudwell, "Cuteness as a 'dark pattern' in home robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, mar 2019.

[16] T. H. Soe, C. T. Santos, and M. Slavkovik, "Automated detection of dark patterns in cookie banners: how to do it poorly and why it is hard to do it any other way," *arXiv preprint arXiv:2204.11836*, 2022.

[17] Y. Yada, J. Feng, T. Matsumoto, N. Fukushima, F. Kido, and H. Yamana, "Dark patterns in e-commerce: a dataset and its baseline evaluations," in *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, dec 2022.

[18] K. Bongard-Blanchy, A. Rossi, S. Rivas, S. Doublet, V. Koenig, and G. Lenzini, "" i am definitely manipulated, even when i am aware of it. it's ridiculous!"-dark patterns from the end-user perspective," in *Designing Interactive Systems Conference 2021*, pp. 763–776, 2021.

[19] C. M. Gray, J. Chen, S. S. Chivukula, and L. Qu, "End user accounts of dark patterns as felt manipulation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–25, 2021.

[20] L. Fritsch, "Privacy dark patterns in identity management," in *Open Identity Summit 2017 : Proceedings*, no. 277 in Lecture Notes in Informatics, pp. 93–104, 2017.

[21] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher, "Tales from the dark side: Privacy dark strategies and privacy dark patterns," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, pp. 237–254, jul 2016.

[22] C. M. Gray, S. S. Chivukula, and A. Lee, "What kind of work do "asshole designers" create? describing properties of ethical concern on reddit," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, ACM, jul 2020.

[23] G. Conti and E. Sobiesk, "Malicious interface design," in *Proceedings of the 19th international conference on World wide web*, ACM, apr 2010.

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[25] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "Metaicl: Learning to learn in context," *arXiv preprint arXiv:2110.15943*, 2021.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[27] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[28] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[29] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[32] J. Chen, J. Sun, S. Feng, Z. Xing, Q. Lu, X. Xu, and C. Chen, "Unveiling the tricks: Automated detection of dark patterns in mobile applications," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–20, 2023.

[33] D. Kirkman, K. Vaniea, and D. W. Woods, "Darkdialogs: Automated detection of 10 dark patterns on cookie dialogs," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 847–867, IEEE, 2023.

[34] T. H. Soe, O. E. Nordberg, F. Guribye, and M. Slavkovik, "Circumvention by design-dark patterns in cookie consent for online news outlets," in *Proceedings of the 11th nordic conference on human-computer interaction: Shaping experiences, shaping society*, pp. 1–12, 2020.

[35] M. Maier, "Dark patterns–an end user perspective," 2019.