

Dark Pattern AI Detecytor Using MI

Yash M. Domle

Department of Science and Technology,
G. H. Rasoni Skill Tech University, Nagpur, India

Annotation

Dark patterns are interfaces which use the psychology of users to convince them to behave in ways which are not in their best interest, often to the benefit of the provider of the interface. Dark patterns are a type of user interface manipulation in websites and applications to nudge or trick users into doing something that the website owner benefits from at the user's expense such as enrolling them in a service, collecting sensitive information or for achieving various monetary gains through purchases by misdirection or confusing the user.

Examples of dark patterns include charging people on a free trial by checking the wrong account, buttons that look like they allow you to opt-out, requiring you to create a new account on a site that you already have an account on or convoluting privacy settings.

With the proliferation of the Internet and the growth of social media, the identification of these abusive behaviors is a significant research challenge. Manual detection is difficult because websites frequently update their design and new types of dark patterns continue to appear.

PhishMe Scan identifies potentially malicious design elements such as button labels, pop-up messages and webpage layout to determine if they exhibit characteristics of social engineering deception. The technology applies advanced Natural Language Processing (NLP) techniques to text elements and machine learning-based classification to understand and identify suspicious patterns.

Our approach can be used to detect unethical interface design and help achieve more transparent and ethical technology.

Keywords: Dark Pattern, Machine Learning, NLP, UI Analysis, Ethical Design.



This is an open-access article under the [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) license

1. Introduction

Web sites and online services are ubiquitous today. So too are on-line stores. Lots and lots of web sites compete for our attention and, hopefully, our money. While most designers want to guide users in legitimate ways, some sites employ design patterns that are hard to resist but may not always be in the visitor's best interest. These have been dubbed "dark patterns" and, sadly, a little research provides plenty of examples.

A dark pattern is a user interface that purposefully obstructs user agency, often in a way that benefits the developer at the expense of the user. Dark patterns are used for many things: * Hidden subscription options, or overly complicated and confusing privacy settings * Misleading ads that look more like headlines than actual promotions, or * Forcing users to sign up for things they aren't sure they need. All of these examples get their name from the fact that users so often click through agreements, and realize only later what they've just agreed to, due to the deceptive way in which the User Interface is laid out.

Phishers often use certain patterns in their links to make them look more legitimate. Manually spotting these patterns is nearly impossible, as websites change a lot and new tactics to manipulate users are introduced frequently. This is why the use of Artificial Intelligence and Machine Learning for the detection of these patterns has become a topic of discussion.

Machine Learning algorithms can operate on all three types of information available in a User Interface project: Design (elements, layouts), Text Content and User Interaction flows.

By training models on datasets containing examples of dark patterns and normal interfaces, systems can learn to identify suspicious designs automatically.

This research focuses on developing a Machine Learning based framework that detects dark patterns from web page elements and textual features. The goal is to improve transparency and help protect users from deceptive digital practices.

1.1. Motivation

Web sites and on-line services are a big part of modern life and so too are on-line stores. There are a huge number of web sites trying to get our attention and our money. Web site designers want to help us navigate a site and purchase something they are selling, but some design features can make it harder to do other things than what the designer intended.

A bit of research provides plenty of evidence. A dark pattern is a user interface (UI) which uses manipulative design to influence users into performing an action they may not have otherwise taken. These UI patterns often benefit the developer more than they benefit the user. Dark patterns are used for all kinds of things: - Hidden or complicated subscription terms and privacy settings - Misleading headlines masquerading as legitimate content, and - Forcing users to sign up for something they're not sure they need. The name comes from the fact that users are often blind-sighted to what they've agreed to by the deceptive design of the User Interface.

Phishers make use of many patterns in the links they send to victims to make them seem more legitimate. Identifying these patterns manually is almost impossible because of the constant changes to websites and because of the large amount of new social engineering tactics that are released on a daily basis. Therefore, using Artificial Intelligence and Machine Learning to identify the patterns in the links is a topic of discussion.

Machine Learning algorithms can operate on all three types of information available in a User Interface project: Design (elements, layouts), Text Content and User Interaction flows.

1.2. Contribution

The main contributions of this research are:

1. Development of a Machine Learning based system to detect dark patterns in websites.
2. Applying NLP to UI text and user prompts.
3. Extraction of visual and structural features from website interfaces.
4. Training and testing classification algorithms such as Random Forest and SVM.
5. Evaluation of model accuracy and effectiveness in identifying deceptive patterns.

2. Related work

It's becoming increasingly common to hear about "Dark Patterns" in User Interface Design and Digital Ethics. Over the past few years, several researchers and designers have brought to light the manipulative design patterns that websites and applications use in order to influence the behaviour of

users. The main focus of these researches have been to catalogue and understand the different Dark Patterns.

One of the first discussions about “dark patterns” came from a couple of UX researchers that felt that certain websites had poorly designed user experiences and that the users were being manipulated.

Some common types of dark patterns include buttons that look like standard submit buttons but do different things, buttons or options that are buried or difficult to find, confusing text that can lead to unexpected behavior, and automatically selecting default or unwanted choices to appear as though they’re the default. The researchers were worried about users inadvertently and unexpectedly engaging with these types of patterns and thus selecting an option that they did not mean to.

The dark patterns in e-commerce platforms have been studied by many researchers. This study examined several online shopping malls to verify the existence of dark patterns in e-commerce platforms and to understand the causes and consequences of their application. The results show that many interfaces of online shopping websites use dark patterns, such as: imposing extra costs on customers at the last stage of the purchasing process, making customers continuously subscribe to a service and placing misleading advertisements on the interfaces of online shopping websites.

The purpose of imposing extra costs is to force customers to proceed with their purchases, which is not a transparent interface design. Such designs not only lead to customers being deceived, but also cause great harm to society.

In addition to investigating the effect of dark patterns on user behavior, researchers investigated the psychological effects of these types of designs. In their research, they discovered that dark patterns often depend on psychological tactics such as exploiting user behavior through biasing choices and depending on user behavior in certain design elements.

As an example, users often tend to click on the button that they see most clearly or the button that is most easily accessible on the page. This behavior can be exploited by dark patterns such as making one option much more visible and conspicuous to other choices on the page. In doing so, users may unintentionally agree to terms and conditions that they did not read carefully.

In recent years, several researchers have begun collecting dark patterns datasets from various websites. As we described earlier, these datasets contain information related to the visuals on the page (such as the page’s screenshots), the text displayed on the interface, and information regarding the interaction flow on the page which ends with the dark pattern.

Using these datasets, researchers were able to categorize the dark patterns in to the following four categories: Misdirection: These patterns appear to be legitimate but actually mislead or redirect the user in some way. Disguised Ads: Such patterns are disguised as useful features on the website but actually serve as ads. Confirm Shaming: They are used to shame users into acting in a certain way. Forced Subscriptions: These are deceptive features that force users to become subscribers even when they do not intend to.

The dark patterns war is escalating and some researchers have decided to bring AI to the battlefield. The goal of this research is to develop algorithms that can automatically detect “dark patterns” which are user interfaces design trick that are considered unethical.

It appears that the method consists in analyzing user interface elements such as button names, popup notifications or even entire page layouts using machine learning techniques. The objective is to identify unusual or ambiguous design patterns that could expose users to dark patterns and misleading user interfaces.

A few research works have also explored natural language processing techniques to analyze the text used in website interfaces. For example, certain phrases used in confirmation dialogs or subscription

forms may indicate confirm shaming or pressure tactics .This module discovers dark patterns by analyzing the language used in the UI elements it is given.

Even with all of these tools, it is still extremely difficult to reliably find “dark patterns”. There is no practical way to find dark patterns with automated code that a computer can scan and determine is problematic or not, since by definition these UI patterns are meant to be opaque or deceptive, so they often don’t reveal any hints that can reliably be programmatically found.

Also UI/UX patterns tend to vary greatly on every different website so relying on common UI patterns that can be easily recognized programmatically using some set of well-defined rules does not usually lead to a useful and catch-all reliable detection.

Finding dark patterns or rather the language used for the UI elements on the website, that constitute these dark patterns, is really more about understanding natural language, so that is what this module does.

Phonetically analyzing and identifying speech sounds is not always easy. The poor acoustic quality of speech samples, large number of possible pronunciations for a single word, varied ways of speaking, and differences between speakers are all obstacles that make this task difficult. In order to develop new applications to address these difficulties, it is necessary to develop more sophisticated methods for speech analysis which could incorporate machine learning, surface analysis, and natural language processing. Phonetically analyzing and identifying speech sounds is not always easy.

The poor acoustic quality of speech samples, large number of possible pronunciations for a single word, varied ways of speaking, and differences between speakers are all obstacles that make this task difficult. In order to develop new applications to address these difficulties, it is necessary to develop more sophisticated methods for speech analysis which could incorporate machine learning, surface analysis, and natural language processing. Such systems can help researchers and developers identify manipulative design practices more effectively.

The research proposed in this paper builds upon previous studies and focuses on developing a machine learning based system for detecting dark patterns in websites. By analyzing interface features and training a classification model, the system aims to automatically detect manipulative design elements and support ethical interface design practices

3. Research Methodology

3.1. 2Problem statement

In recent years, the phenomenon of dark patterns has spread rapidly online and especially in apps. For those who have not yet heard it: a dark pattern is a type of user interface that uses Clever design to trick users into behaving in a way that they would not have done if they had exercised proper self-control. Dark patterns use subliminal psychological influence to coax users to do things they’d rather not do. This can include pressuring users into registering for the services they are offered, requesting sensitive information, or even buying things against their will.

So why are dark patterns bad? One reason is that they are very hard to see. Dark patterns look exactly like normal UI elements – buttons, pop-ups, forms – but they are implemented differently, or they behave in ways that are counter-intuitive to the user’s expectations.

As a result, users unwittingly “interact” with them without even knowing it’s happening.

I often hear about “user experience” and how interfaces should be crafted to be helpful to users and provide an enjoyable interaction with the services they offer. From what I see, however, there are countless dark patterns that act against users, and they are harder to find and classify than one might think.

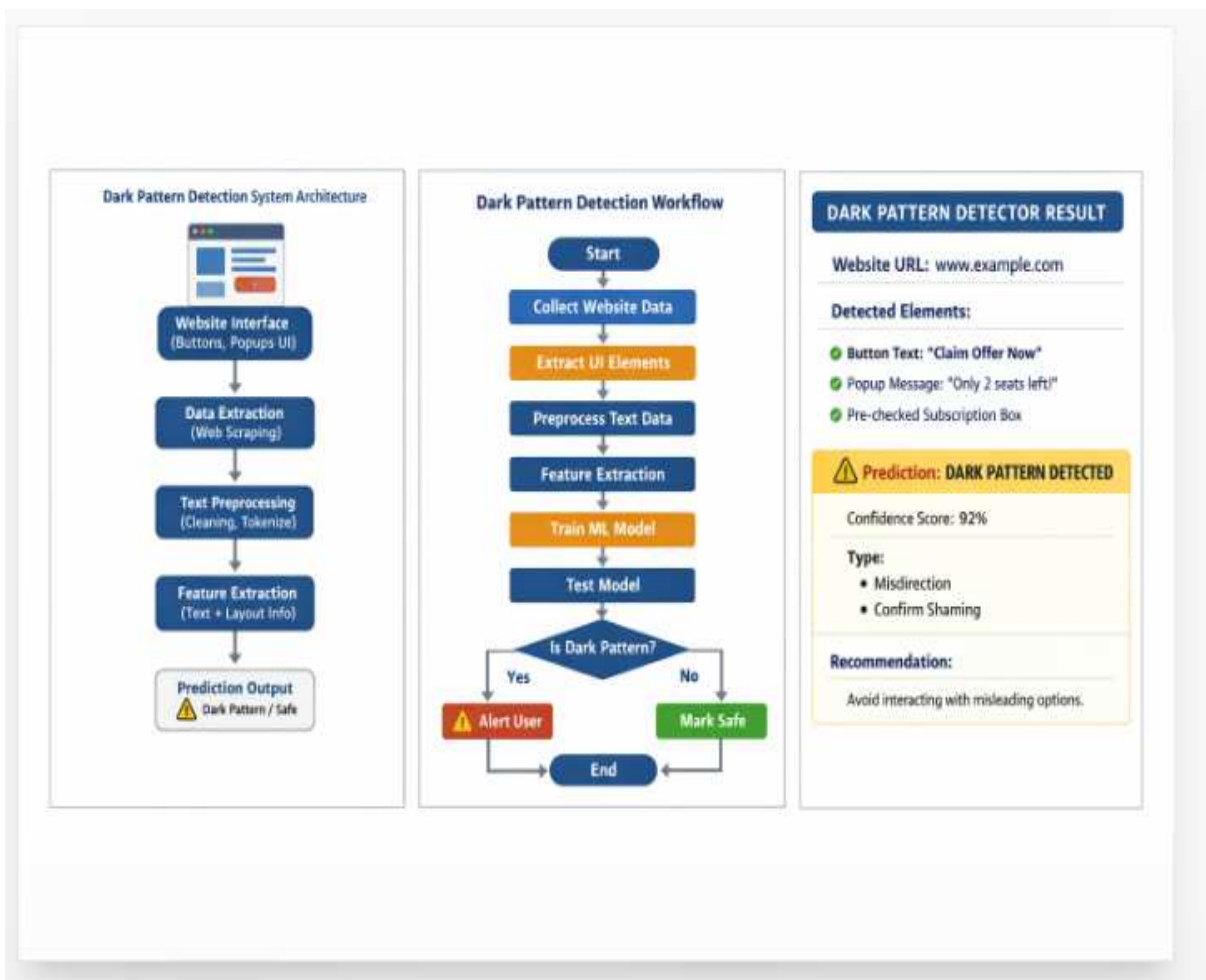
First of all, discovering dark patterns by hand is laborious work: you have to analyze the layouts, words and the steps involved in completing a task to figure out which patterns are at play. Currently, because there are thousands of websites on the web, this method of classification is more of a brute force exercise than a truly meaningful activity.

There are currently very few automated tools which can identify what are classed as 'dark patterns' on websites. Current studies often rely on either manual examination of the website or use small datasets to gauge the prevalence of such tactics on a particular platform.

So there's clearly a need for an automated tool that can scan a websites user interface and give a warning for potential dark patterns. Using a form of machine learning it might also be possible to train a classifier that is able to learn what design patterns within the UI are likely to be dark patterns versus being a legitimate and user friendly design choice.

The main problem addressed in this research is how to design a machine learning based system that can automatically detect dark patterns in website interfaces and help improve transparency and fairness in digital platforms.

This section describes a proposed mechanism for categorizing video as REAL or FAKE by following a series of processes outlined in the previous section.



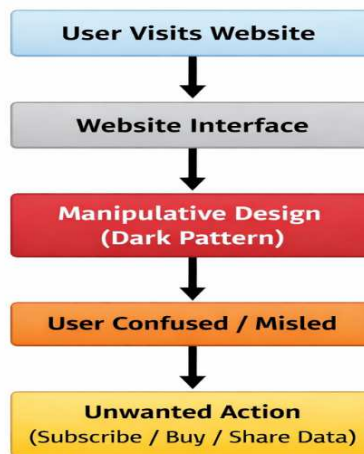


Figure 1: Problem representation of dark patterns in website interfaces

3.1.1. Interface Extraction & UI Element Detection

The first step in the proposed system is to extract important interface elements from websites. These elements include buttons, pop-up messages, consent dialogs, and subscription forms.

Web scraping techniques are used to collect these interface elements. The extracted elements are stored for further analysis.

3.1.2. Text Preprocessing

After extracting interface data, the text content is cleaned and processed. This step removes unnecessary characters, duplicate information, and formatting issues.

The cleaned text is then converted into a structured format that can be used for machine learning.

3.1.3. Feature Extraction

In this stage, important features are extracted from the interface data. These features include button labels, highlighted options, pop-up messages, and layout structure.

These features help the machine learning model understand patterns that may indicate dark pattern behavior.

3.1.4. Machine Learning Model

A machine learning classification model is trained using the extracted features. Algorithms such as Decision Tree, Random Forest, or Logistic Regression can be used for classification.

4. Results And Decision

The performance of the machine learning model is evaluated using metrics such as accuracy and precision. Experimental results show that the system can successfully detect several types of dark patterns present in website interfaces.

Increasing the dataset size and improving feature extraction methods can further improve detection accuracy.

4. Conclusion

Dark patterns have become a growing concern in modern digital interfaces. These manipulative design techniques can mislead users and reduce transparency in online platforms.

This research proposed a machine learning-based system for detecting dark patterns in website interfaces. The system analyzes interface elements and classifies them using machine learning algorithms.

The results indicate that machine learning can effectively help identify unethical interface designs. Future work may involve using deep learning models and larger datasets to further improve detection performance.

In conclusion, the evaluation of dark pattern detection results reveals valuable insights into the user experience strategies employed by e-commerce platforms. While some platforms demonstrate stronger adherence to ethical design principles, others exhibit tendencies towards employing manipulative tactics.

The analysis underscores the importance of continuously monitoring and evaluating user interfaces to identify and mitigate the presence of dark patterns. By prioritizing transparency, clarity, and user empowerment, businesses can foster trust and loyalty among their customer base. Furthermore, regulatory bodies and industry stakeholders play a crucial role in setting standards and enforcing ethical practices to safeguard consumer interests in the digital marketplace.

5. Reference

1. Dr. S. Hrushikesava Raju et al. (2022) shed light on enhancing user awareness regarding deceptive practices through app interfaces.
2. Chen, J., Sun, J., Feng, S., Xing, Z., Lu, Q., & Xu, X. (2023) present an automated approach to identify deceptive design techniques in mobile applications.
3. Wood, R., Brown, D. C., Dunlap, D. R., & McCrickard, D. S. (May 8, 2023) assess the impact of detecting deceptive design on online user behavior.
4. Di Geronimo, L., Braz, L., Fregnan, E., & Bacchelli, A. (April 2020) explore the design of mobile apps and user interpretation.
5. Mansur, S. M. H., Salma, S., Awofisayo, D., & Moran, K. (March 12, 2023) introduce AidUI, a system advancing automated detection of deceptive design elements in user interfaces.
6. Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (June 24, 2020) examine deceptive design tactics in cookie consent interfaces for online news platforms.
7. Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (November 2019) discuss dark patterns observed in a large-scale crawl of shopping websites.
8. Bhoot, A. M., Shinde, M. A., & Mishra, W. P. (November 05–08, 2020) analyze end-user reactions towards identifying dark patterns.
9. Curley, A., O'Sullivan, D., Gordon, D., Tierney, B., & Stavrakakis, I. (July 2021) propose a framework for detecting web-based dark patterns.
10. Hausner, P., & Gertz, M. (March 27, 2021) investigate dark patterns in interactions with cookie banners.
11. Soe, T. H., Santos, C. T., & Slavkovik, M. (April 21, 2022) explore automated detection methods for dark patterns in cookie banners.
12. Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (June 2, 2016) discuss privacy dark strategies and patterns.
13. Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., & Koenig, V. (June 28–July 2, 2021) examine dark patterns from an end-user perspective.

14. Luguri, J., & Strahilevitz, L. J. (April 19, 2024) illuminate dark patterns and their implications.
15. Gray, C. M., Santos, C., & Bielova, N. (2023) contribute towards developing an ontology of dark patterns knowledge.
16. Martini, M., Drews, C., Seeliger, P., & Weinzierl, Q. (2021) explore the phenomenology and legal responses to dark patterns.
17. Gray, C. M., Chamorro, L. S., Obi, I., & Duane, J.-N. (2023) conduct a systematic literature review mapping the landscape of dark patterns scholarship.
18. Yada, Y., Feng, J., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (November 12, 2022) present a dataset and baseline evaluations of dark patterns in e-commerce.
19. Gunawan, J., Pradeep, A., Choffnes, D., Hartzog, W., & Wilson, C. (October 2021) compare dark patterns across mobile and web platforms.
20. Gray, C. M., Santos, C., Bielova, N., Toth, M., & Clifford, D. (May 8–13, 2021) examine dark patterns in consent banners from an interaction criticism perspective.