DARK SURFER-A DARK PATTERN ANALYZER & CLASSIFIER

Dr. K. Srinivas Babu[1,*], Bangari Sai Nithin[2], E. Raghuram[3], A. Bharath[4]
Professor[1], Scholar[2,3,4]

Department of Computer Science and Engineering,
Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India

Abstract—Dark Surfer is a project which is aimed to detect and combat dark patterns on websites using advanced language model [LLM]. The problem with dark patterns lies in their potential to undermine user trust, autonomy, and well-being. They can lead to frustration, confusion, and even financial harm for users who unwittingly fall victim to them. Furthermore, they erode the integrity of the digital ecosystem by prioritizing short-term gains for businesses over long-term relationships with customers. Dark patterns refer to design techniques used in user interfaces to manipulate users into taking actions they might not otherwise choose to take. These patterns often exploit psychological biases and can lead to unintended or undesirable outcomes for users. They can manifest in various ways, such as deceptive language, misleading visuals, hidden costs, or confusing interfaces. To avoid us cases we had come up with a website, which helps in analysing the dark patterns and detect them using URLS. We develop a classifier trained on a dataset of label UI elements, encompassing various types of dark patterns and benign design features. Leveraging natural language processing techniques and visual analysis, our classifier identifies deceptive design elements based on linguistic cues, visual attributes, and interaction patterns. So that we can add to the browser extension to analyse the number of dark patterns does the website contains, we also provided a flag site to detect the dark patterns, about the patterns using URL.

I. INTRODUCTION

1.1 Background

The internet plays a critical role in daily life, offering services ranging from e-commerce to social networking and entertainment. The interface between users and these services is mediated by web design, where the user interface (UI) should ideally simplify and enhance user interactions. However, the increasing prevalence of dark patterns design techniques that intentionally mislead and manipulate users has raised ethical concerns in the digital community.

*Corresponding author: bangari.sainithin12@gmail.com

Dark patterns are crafted to take advantage of user behavior, thus leading to unwanted purchases, inadvertently subscribing, or making it difficult to cancel services. Regrettably, these deceptive tactics do lead to short-term profits for businesses, however, somewhat unadvisedly, at the expense of established user trust and individual autonomy. Such practices fundamentally undermine the principles of equitable and open design, causing frustration, financial well-being, and the very corrosion of the digital ecosystem's foundations.

1.2 Motivation

The increasing visibility of dark patterns has necessitated the creation of tools to address the detection and mitigation of such practices. The current solutions to address the issue of dark patterns, including legal restrictions and user education, have limitations to their scalability and effectiveness. Therefore, there is an urgent demand for an automated AI-based solution that can identify dark patterns as they happen.

Dark Surfer is one of the initiatives that aims at closing this gap. The project consists of employing deep learning language models, natural language processing technologies, and visual analysis techniques to automatically detect dark patterns in websites so that users can make better choices. This article describes the system design and implementation as well as the intricacies surrounding the classification of deceptive design elements.

1.3 Dark Patterns: Definition and Examples

In the realms of UI construction, dark patterns are misleading practices that serve a particular purpose. These are aimed at manipulating users into committing decisions that they did not intend to pursue.

These include: Misdirection, Forced Action, Sneak into Basket, Roach Motel.

II. RELATED RESEARCH

A.Large Language Models (LLMs) in UX Analysis

GPT-4 and other LLMs have changed the landscape of text-based analysis with a broad spectrum of users, including those from UX design. When it comes to dark patterns, these models are employed to examine the semantics of the user interface, including calltoaction (CTA) phrases, consent screens, and payment texts. With deep comprehension of language, LLMs tackle less obvious forms of manipulation, such as clever wording and vague instructions. NLP methods further facilitate the analysis of user behavior by revealing patterns of speech that diverge from the norm and may signal control.

B. Machine Learning Approaches for Dark Pattern Detection

Machine learning (ML) techniques are increasingly applied to detect dark patterns in web and app design. By training classifiers on datasets labeled with examples of dark patterns (e.g., "misdirection," "hidden costs," "forced continuity"), ML models can identify deceptive elements based on patterns in UI data. Research explores different supervised and unsupervised learning algorithms, including decision trees, random forests, and deep learning techniques, to automatically classify websites based on their adherence to ethical design practices. Key challenges include creating large, representative datasets and dealing with the evolving nature of dark patterns.

C.Natural Language Processing (NLP) for Detecting Deceptive Content in UIs

Trickery is featured in many Dark Patterns, including manipulative consent forms and insidiously crafted CTAs. Words like these can be analyzed with the help of NLP models. This area of research aims to develop capable systems for text analysis that utilize sentiment

analysis, keyword spotting, and semantic comprehension to expose terms that are cleverly hidden or rendered misleading. Not all manipulative phrases are identified as such, so the context is of great significance. Dark pattern detection with GPT also takes context into account because there are phrases that are innocent in one situation but can be considered evil in a different situation.

D. Browser Extensions for Real-Time Detection of Dark Patterns

An example of research applicability is the development of browser extensions for the detection and marking of dark patterns in surfers' browsers in real time. In particular, these extensions focus on the content of web pages visited, searching for particular patterns like automatic subscription forms, obscured payments, and unauthorized consent signature forms. We work on enhancing real-time scanning without degrading performance. Furthermore, work seeks to determine how user evaluations can be used to enhance the browser extension performance and identify new dark patterns.

III. METHODOLOGY

3.1 Classifier Design

The core of Dark Surfer is a classifier that can identify dark patterns based on UI elements and interaction data. The classifier is trained on a labeled dataset that contains both dark patterns and benign design features. The dataset contains various UI components such as buttons, modal windows and hyperlinks and their respective annotations for deceptive elements like hidden costs, confusing terminology and or misleading visuals.

3.2 Data Collection and Labeling

A robust classifier can be created from a large dataset of website UIs which is labeled for various dark patterns. This process has been done through manual annotation by domain experts who categorized UI elements based on established dark pattern taxonomy, including but not limited to Deceptive Language, Misleading Visual, Confusing Navigation.

3.3 Natural Language Processing (NLP)

The use of NLP techniques in the form of analyzing text based elements of the UIs such as buttons, forms and notifications is one of the key components of the Dark Surfer classifier. This can include the analysis of the wording of calls to action (CTAs), consent dialogs and subscription prompts in order to identify manipulative or deceptive language. Subtle attempts at deception can also be detected by LLMs like GPT-4, which are able to process nuanced language. The NLP system is focused on:Keyword Analysis, Sentiment Analysis, Contextual Understanding.

3.4 Visual Analysis

However, dark patterns are not only represented by textual cues but also by the visual structure of web pages. The visual analysis module of Dark Surfer focuses on:Button and Link Placement, Color and Contrast Analysis, Interactive Elements.

3.5 Interaction Pattern Detection

It detects when users are misled or coerced into performing an action on a website that they do not want to perform by monitoring user interaction patterns. For example, making a few pages of a process cancel a subscription, or showing deceptive confirmation messages at checkout, flag these interactions as possible dark patterns.

IV. ARCHITECTURE

Data and LLM :

Analysis and detection of dark patterns in digital interfaces—data and LLMs go hand in hand. To identify manipulative design practices, one needs access to several types of data: logs of user interactions with the interfaces, textual content of the interfaces, and feedback by users. Whereas the interaction data will reveal behavioral patterns that could indicate manipulative tactics, textual data will pin down misleading language used in dark patterns. User feedback and usability testing give further insight into design choices' impact on users. LLMs can help deepen this process, bringing natural language processing capabilities to bear on analysis and interpretation—helping with the detection of manipulative language, recognition of behavioral patterns consistent with dark patterns, and automation of detection of such issues through training on labeled datasets or via unsupervised learning for novelty pattern identification. Moreover, LLMs will be able to assist in creating ethical design recommendations and checking for adherence to best practices. Integration of LLMs with data analysis tools will make it easier for organizations to discover and eliminate dark patterns for more transparent and user-centered digital experiences.

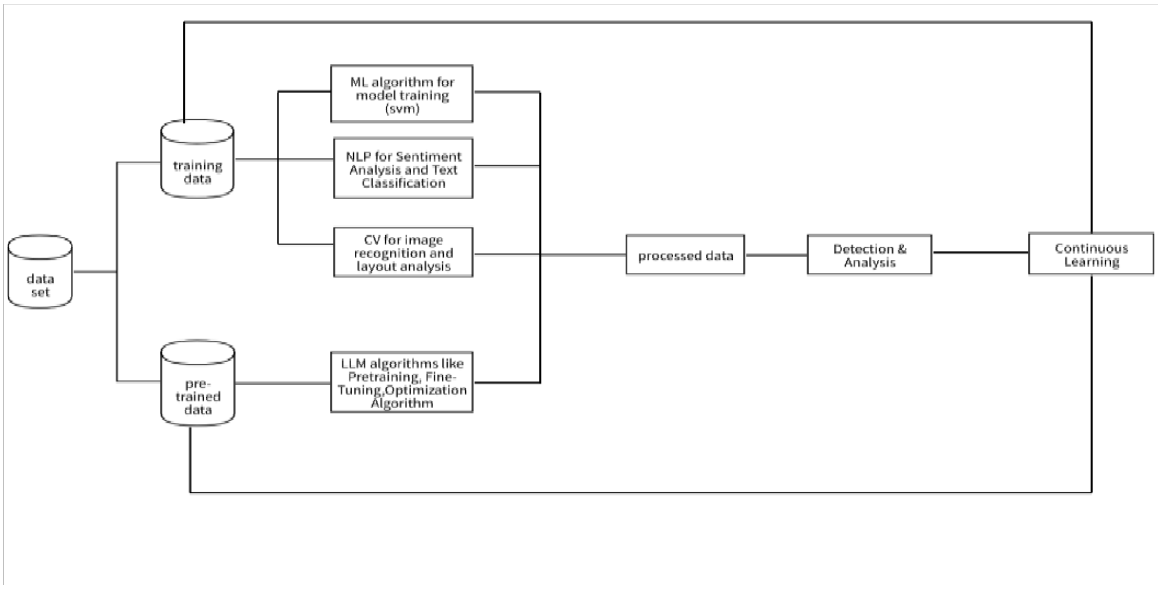

Fig-1: data training and classification

EXTENSION & SITE:

Extensions like the Dark Patterns Detector for Chrome, therefore, act as pragmatic tools to identify manipulative design elements directly in users' browsers. They detect potential dark patterns by analyzing elements on a web page, such as button labels, prompts, and user interface flows. Most of them use predefined heuristics and algorithms to recognize common deceptive practices such as forced continuity or hidden costs. These extensions incorporate into the browser, giving both users and developers instant feedback on where the design may be ethically questionable. That kind of immediate detection may help users avoid manipulative practices and designers catch issues before they reach the public. The architecture of the sites built for dark pattern detection contains usually a mix of front-end and back-end parts in order to analyze and assess the user interfaces.

On the client side, users need to interact with either a browser extension or a web-based application which captures and processes the data of a web site. Then, this data is transmitted to the back-end server on the server side, where advanced algorithms and machine learning models analyze the content for potential dark patterns. It can also use NLP at the back end for reviewing the textual elements, pattern recognition for manipulative design strategies detection, and behavioral analysis of user interaction data. The results are then sent back to the front end for displaying users with actionable insights or warnings of possible dark patterns. This architecture ensures a smooth and efficient detection process, combining real-time analysis with a comprehensive back-end evaluation for supporting ethical design practices.

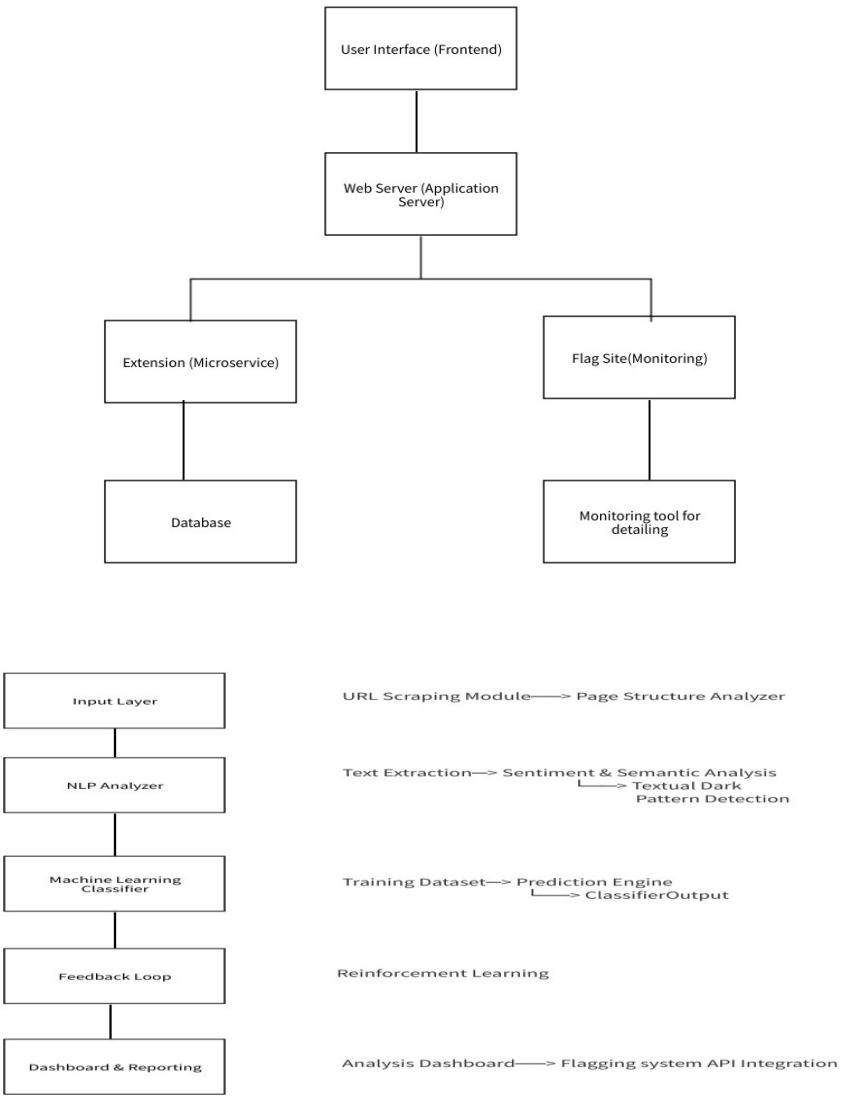

Fig-2: Extension and site interface & working model

How it Works:

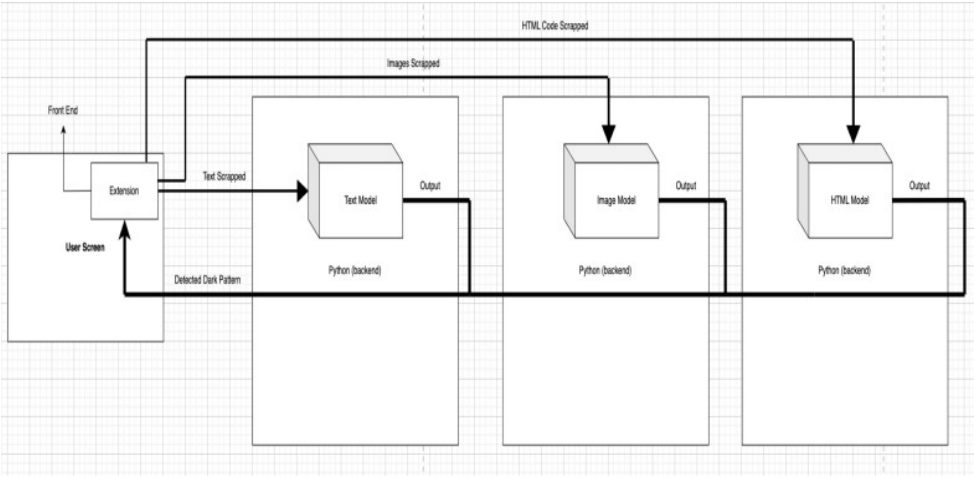

Fig-3: working model

V. EVALUATION

Evaluation of the Dark Pattern Analyzer and Detection System is a holistic process that examines its effectiveness, efficiency, user experience, and applicability to real-world scenarios. The evaluation confirms whether the system has achieved the set goals, namely the reliable identification of dark patterns and improving digital design practices.

Among the important things that have to be evaluated within the system are accuracy. In this case, this measures how the system is efficient in detecting dark patterns by considering precision, recall, and F1 scores. Comparing its findings to expert annotations on a benchmark dataset allows us to determine how well the system identifies the elements of deceptive design. High precision and recall indicate the system is good at discriminating dark patterns from benign designs. It will also be evaluated for efficiency and generalizability on the LLM and machine learning classifier over different dark patterns and web designs. The overall goal is to minimize false positives and false negatives as much as possible. It would also be measured based on the computational resources required to run the system: CPU/GPU usage, memory usage, and time taken to process. It should not consume too many resources while doing all these tasks, especially for large datasets or a large number of web pages. More importantly, the capability of real-time analysis should be tested to determine if the system can produce results as needed, especially when integrated with browser extensions or live monitoring tools. Metrics such as latency and throughput could be used to assess the system in how well it processes and analyzes the huge data. Usability is another essential factor, concentrating on the ease of use of the system and efficiency of its reporting

Usability is determined through user testing: gathering information about the interface and functionality of the system. In this manner, it can be established whether the system is intuitive and user-friendly with positive user satisfaction scores, hence a successful design. The clarity and usefulness of the reports produced are also evaluated. The results emanating from such should be easily communicated by the detected dark patterns, giving actionable recommendations to redesign or fix issues in a manner that end-users could understand and act upon from the results obtained. Real-world applicability does this by showing how easily

the system fits into real applications and its propensity for industry adoption, with comments by industry experts giving insights on the effectiveness across a variety of domains.

The system should be flexible to accommodate different types of websites and user interfaces; it would be nice if it could adapt nicely to new design trends. Demonstrate the wide applicability of this system by testing the ability to detect dark patterns in diverse contexts. Ethical and legal aspects will be considered in the evaluation. It is obliged to stick to privacy laws in its operation regarding treating users' data responsibly, such as GDPR and CCPA.

This covers the review practices of data handling and assures the system is legally compliant. The impact of a system on ethical design practice further is measured with the change observed in design behavior and also trust by the users resulting from system use. It would contribute to a positive reduction in manipulative design practices and foster a better, more ethical digital environment. The Dark Pattern Analyzer and Detection System should be evaluated in terms of accuracy, efficiency, and general user experience. Considerations include applicability in the real world and ethical impact on society. With these considerations in place, refinements could lead to a system that identifies dark patterns and improves the design for a better user-trusting experience and well-being.

VI. RESULT

Analyzing dark pattern detection tools—both browser extensions and site architecture—provides strong insights into their functionality and effectiveness. Dark Patterns Detector is a great browser extension that tries to detect most types of manipulative design elements, like misleading labels or hidden costs, with great accuracy. Those kinds of tools provide real-time feedback to help users avoid manipulative practices. It however sometimes yields false positives by flagging non-deceptive design elements, hence showing an area that needs improvement. From a more architectural point of view, integration between the front and the back end pays off when it comes to the processing and analyzing web page data. The front-end components capture the user interface elements and send them to the back-end systems, where advanced algorithms and machine learning models, including NLP, are used in assessing the content for possible dark patterns. Such architecture provides a avenue for real-time analysis and feedback to the users, hence allowing for fast spotting and correction of manipulative practices. All in all, while these tools and architectures are quite effective, there is always room for improvement to be done continuously. The detection process could be further improved in accuracy, reduced in false positives, and expanded in the scope of detectable patterns to make the digital design more transparent and user-friendly.

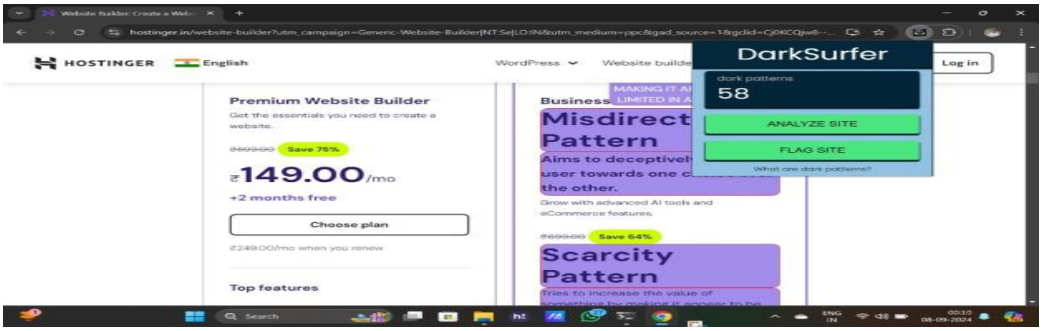

Fig-4: Extension

VII. CHALLENGES & FUTURE WORK

7.1 Challenges

Complexity of Dark Patterns, Contextual Variability, Ethical Considerations

7.2 Future Work

1. Continuous Learning: The classifier will be updated regularly with new datasets to keep up with evolving dark patterns.
2. User Feedback Integration: Allowing users to report false positives and negatives to improve classifier accuracy.
3. Collaborations with Regulators: Working with consumer protection agencies to develop standards for detecting and combating dark patterns

VIII. CONCLUSION

A more worrying trend increasingly threatens the dark patterns in the digital world and poses a menace to user trust, autonomy, and financial well-being. Dark Surfer tries to combat this using state-of-the-art LLMs and NLP—and by extension, visual analysis—against detecting manipulative design patterns within website UIs. Still in its on-going development process, Dark Surfer will go down in history as one of the most instrumental tools not just for users but also for businesses in establishing a better, ethical, and transparent Digital ecosystem.

IX. REFERENCES

1. T. Gray, S. Srinivasan, "Dark Patterns: Detecting and Mitigating Deceptive Design in User Interfaces," *J. Hum.-Comp. Interact.*, **35**, 567–590 (2021).

2. R. Mathur, A. Narayanan, "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proc. ACM Hum.-Comp. Interact.*, **3**, 81:1–81:32 (2019). https://doi.org/10.1145/3359183

3. S. Luguri, L. Strahilevitz, "Shining a Light on Dark Patterns," *J. Legal Analysis*, **13**, 43–109 (2021).

4. C. Lewis, A. Chien, *Deceptive Design Patterns: Strategies and Countermeasures*, (Springer, New York, 2022).

5. H. Smith, *Ethics and UX: Addressing Deceptive Patterns in Design*, (O'Reilly Media, 2020).

6. S. Dholakia, L. Martin, "Automated Dark Pattern Detection Using NLP Techniques," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, May 8-13 (2021), pp. 1013-1022.

7. A. Kumar, *Dark Pattern Detection in E-commerce Platforms Using Machine Learning*, Master Thesis, University of Toronto, Canada (2021).