National College of Ireland

Automated Detection of Dark Patterns in Website Design: Enhancing User Trust and Online Transparency

MSc Research Project -
Cybersecurity

Sundar Ayyappan Muthukumarasamy
Student ID: X23180749

School of Computing
National College of Ireland

Supervisor: Michael Pantridge

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sundar Ayyappan Muthukumarasamy
Student ID:	X23180749
Programme:	MSc in Cybersecurity **Year:** 2024
Module:	MSc research Practicum-II
Supervisor:	Michael Pantridge
Submission Due Date:	12/12/2024
Project Title:	Automated Detection of Dark Patterns in Website Design: Enhancing User Trust and Online Transparency
Word Count:	**6173** **Page Count:** **17**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sundar Ayyappan Muthukumarasamy

Date: 11/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Automated Detection of Dark Patterns in Website Design: Enhancing User Trust and Online Transparency

Sundar Ayyappan Muthukumarasamy
X23180749

Abstract

This project introduces a robust system for the automated detection of dark patterns on websites, aiming to enhance user protection and transparency in online interactions. Leveraging a Naive Bayes classifier trained on dark pattern categories such as Bait and Switch, Forced Continuity, Price Comparison Prevention, Hidden Costs, and Sneaking, the model achieves effective identification of deceptive design elements. The preprocessing of textual data involves employing the TFIDF vectorizer for feature extraction, optimizing the classifier's performance. Web scraping is facilitated through cloud scraping techniques and Beautiful Soup, enabling the extraction of relevant data for classification. The resulting model file is applied to classify scraped data, empowering users to make informed decisions while navigating online interfaces. This innovative approach addresses the ethical concerns associated with dark patterns and contributes to a safer and more transparent online environment.

Keywords: Dark pattern detection, Naive Bayes classifier, TFIDF vectorizer, Web scraping, Cloud scraper, User protection, Transparency, Deceptive design, Online ethics.

1 Introduction

The pervasive use of digital platforms for various activities has led to an increased prevalence of deceptive design strategies, commonly known as dark patterns, on websites. Dark patterns are user interface elements crafted to manipulate users into making decisions that may not align with their best interests. This project focuses on the development of a sophisticated system for the automated detection of dark patterns, enhancing user awareness and safeguarding online experiences. The methodology involves training a Naive Bayes classifier on distinct dark pattern categories, including Bait and Switch, Forced Continuity, Price Comparison Prevention, Hidden Costs, and Sneaking. To optimize the classification process, textual data is preprocessed using the TFIDF vectorizer, capturing the significance of terms within the dataset. Web scraping, facilitated by cloud scraping techniques and Beautiful Soup, enables the extraction of pertinent data from websites, creating a diverse dataset for model training. By combining machine learning, natural language processing, and web scraping, this project addresses the ethical concerns surrounding deceptive online practices. The resulting model empowers users by providing a means to identify and avoid

websites employing dark patterns, fostering a safer and more transparent digital ecosystem. The significance of this work lies in its potential to contribute to a user centric online environment, promoting trust and informed decision making in the digital realm.

2 Related Work

1.Title: Dark patterns in e-commerce: a dataset and its baseline evaluations
Authors: Yuki Yada,Jiaying Feng,Tsuneo Matsumoto,Nao Fukushima,Hayato Yamana,
Publication: 2022 IEEE International Conference on Big Data (Big Data)
Dark patterns mean designing the user interface of online services to make users conduct unwanted activities. Recently, dark patterns have been raised as an issue of privacy and fairness. Thus, a variety of studies on detection of dark patterns are expected. In this work, we provide a dataset for detecting dark patterns and prepare its baseline detection performance with the state-of-the-art machine learning approaches. First, the original dataset is from Mathur et al., 2019, which includes 1,818 instances of dark pattern texts collected from several e-commerce websites. Then, we added negative samples-namely, non-dark pattern texts-by scraping texts from the same online sources that were used to create the dataset of Mathur et al. We then showed that automatic detection performance works as a baseline with state-of-the-art machine learning approaches, including BERT, RoBERTa, ALBERT, and XLNet. Among them, the best performance is 0.975 when doing 5-fold cross-validation by RoBERTa. (Yada et al., 2022)

2.Title: AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces
Authors: S M Hasan Mansur,Sabiha Salma,Damilola Awofisayo,Kevin Moran,
Publication: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)

Past research has documented the prevalence of UI dark patterns, understood as interfaces likely to unwittingly nudge end-users into performing actions they had not originally intended to take. These misleading UI designs may arise either intentionally to serve the interests of an online service or as an unintended consequence of collaborative forms of design; in any case, they are apt to cause end-users harm either due to the over-selling of personal information or to financial loss. While there has been considerable research in developing taxonomies of dark patterns in many areas of software, both developers and end-users currently do not have support with regard to their detection, avoidance, and control, since such design features can sometimes be very subtle. The true automation of the detection of dark patterns is problematic, however, since any particular pattern type can manifest in a great many ways, leading to great variability. The contribution of this paper is an initial understanding of the extent to which common user interface dark patterns can be automatically detected within modern software applications. In this work, we present AidUI, the first fully automated approach that leverages computer vision in combination with natural language processing techniques for detecting a set of visual and textual features from application screenshots indicating ten kinds of UI dark pattern presence;

hence, detection, categorization, and localization. To validate our approach, we created ContextDP, by far the largest dataset of fully-localized UI dark patterns. Indeed, it's a large dataset comprising 301 dark pattern instances within 175 mobile and 83 web UI screenshots. The overall performance of our analysis result is such that AidUI reaches 0.66 general accuracy, 0.67 recall, and 0.65 F1-score on identifying the dark pattern, whereas it reports just a few false positives and with good localization for the detected patterns, as evidenced by an IoU score of 0.84. Moreover, a remarkable percentage of the dark patterns studied here can be detected reliably, with an F1 score higher than 0.82. Further work might help improve such detection for more patterns. The present study therefore provides evidence about the possibility of designing tools able to support developers in detecting and effectively mitigating the spread of deceptive user interface patterns. (Hasan Mansur et al., 2023)

3.Title: DarkDialogs: Automated detection of 10 dark patterns on cookie dialogs
Authors: Daniel Kirkman,Kami Vaniea,Daniel W. Woods,
Publication: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)

In theory, consent dialogs allow users to express privacy preferences regarding how a website and its partners process the useru2019s personal data. Reality is different, though. Dialogs often use subtle design techniques, so-called dark patterns, that nudge users to accept more data processing than users would otherwise do. Dark patterns undermine user autonomy and can result in breaches of privacy laws. We will introduce and implement an innovative system, DarkDialogs, aimed at the automatic extraction of any form of consent dialog appearing on different websites. This advanced system identifies in total 10 different dark patterns, which are often used within those dialogs. The authors evaluate DarkDialogs based on a hand-labeled dataset. Results show that it extracts dialogs with a very high accuracy of 98.7%, whereas 99% of the studied dark patterns are classified correctly. We deploy DarkDialogs on a sample of 10,992 websites where it successfully collects 2,417 consent dialogs and finds 3,744 different dark patterns appearing automatically on the consent dialogs. Then we test whether dark pattern prevalence is associated with each of: the websiteu2019s popularity, the presence of a third-party consent management provider, and the number of ID-like cookies (Kirkman et al., 2023)

4.Title: Cuteness as a u2018Dark Patternu2019 in Home Robots
Authors: Cherie Lacey,Catherine Caudwell,
Publication: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)

Dark patterns constitute a great and very current problem in the field of interaction design; they are an unnerving mixture of design patterns along with the principles of behavioral psychology that cooperate in deceitfully misleading users. However, the available literature about dark patterns mainly concerns digital interactions on screens. This current focus represents a critical gap because this academic work urgently needs to extend to cover implications about the dark patterns arising with home robots. In this paper, we reflect upon the concept of dark patterns within the context of the 'cute' aesthetic developed by home

robots, suggesting that their design incarnates a dark pattern in human-robot interaction, as it: 1) favors immediate reward over longer-term consideration; 2) reduces user conscious agency in the process of interaction; and 3) deliberately provokes an emotional response from users to facilitate the gathering of emotional data. The exploratory paper is of significant contribution to the already existing repository on Dark Patterns and their many uses, more so considering the focus on new technological interfaces coming to the fore in the area of home robotics. The effort is to lay the grounding necessary to facilitate ethical design practice particular to human-robot interaction.. (Lacey & Caudwell, 2019)

5.Title: Re-Designing Dark Patterns to Improve Privacy
Authors: Davide Maria Parrilli,Rodrigo Hernu00e1ndez-Ramu00edrez,
Publication: 2020 IEEE International Symposium on Technology and Society (ISTAS)

Dark patterns are extremely unethical mechanisms pursued in the field of digital design, most of the time for the purpose of collecting a good deal of deeply personal data about users without their prior consent. The thing with methodologies of dark patterns is that they are inherently prone for improvements which one day or another would serve as a means of strengthening and securing users' privacy, hence making them ethical tools based on recognition of users' rights. The present research, though at its end, tries to propose that such dark patterns can be provided thoughtfully in order to drive the users toward the choice of the most rigid settings of privacy, with the aim of keeping them in a digitally safer environment. The implication, therefore, is that, within the framing of ethics and practice, what was once held to be harmful and destructive now turns out to be well and truly building blocks which contribute positively and significantly towards the common good. (Parrilli & Hernandez-Ramirez, 2020)

6.Title: Identifying Dark Patterns in Social Robot Behavior
Authors: Elizabeth Dula,Andres Rosero,Elizabeth Phillips,
Publication: 2023 Systems and Information Engineering Design Symposium (SIEDS)

Social robots have become increasingly utilized in intimate environments where their roles can include caretakers for the elderly, general physical or emotional support, entertainment, and educators for children. To accommodate for these increasingly intimate relationships, robotics companies have begun employing robotics with the ability to identity emotions and respond with emotionality in return. This faux emotional relationship opens the door for potential user manipulation and exploitation through deceptive robot design. Dark patterns are deceptive design patterns used by websites or apps to manipulate users into actions the user did not intend. We argue that dark patterns can be programmed into social robotics to leverage these unidirectional human - robot emotional bonds to manipulate users, which could result in the exploitation of vulnerable populations like children and the elderly. Drawing from the dark pattern and social robotics literature, we suggest ways that dark

patterns can manifest themselves in these relationships. We also provide recommendations for ethical practices when designing emotional social robots. (Dula et al., 2023)

7.Title: Risk Analysis of Encountering Dark Patterns of UX E-commerce Applications Affecting Personal Data
Authors: Apichaya Nimkoompai,
Publication: 2022 6th International Conference on Information Technology (InCIT)
Much attention is paid to personal data protection on the digital world in the form of the PDPA, which concurs with attention to designs called nowadays as the u2018u2018Dark Patterns of UX especially in e-commerce. In this work, the focus is on mobile applications because the users are more convenient with mobile applications. The researcher collected data from potentiallyvulnerable, relevant samples, along with dark patterns data from various sources to derive a conclusion for future awareness campaign. The study revealed that 58.3% of the sampled group did not know dark patterns. The top misleading dark patterns that were commonly used by designers were Forced continuity (71.8%), Disguised ads (59.3%). The riskiest dark patterns was found to be risk of disclosing personal data in apps that force the user to pay prior to usage. The gathered data showed that many users were not familiar with dark patterns, which put them at risk of incurring damage or infringement of personal data. (Nimkoompai, 2022)

8.Title: Analysis of Dark Pattern-related Tweets from 2010
Authors: Jiaying Feng,Fan Mo,Yuki Yada,Tsuneo Matsumoto,Nao Fukushima,Fukuyo Kido,Hayato Yamana,
Publication: 2023 IEEE 8th International Conference on Big Data Analytics (ICBDA)
Dark patterns are defined as user interfaces that make users behave unintendedly, such as buying something or subscribing to some services. The use of dark patterns is considered an infringement of usersu2019 rights and privacy. In this study, we reveal the usersu2019 responses toward dark patterns by analyzing 12 years of tweets. Our findings include 1) users in countries in which dark patterns-related regulations have been implemented have a higher level of discussions about dark patterns; 2) tweets about dark patterns shifted from around 2017 from sharing their diversity to acting to resist them; 3) the commonly discussed dark pattern types of tweets are sneaking, obstruction, and interface interference, which are widely used in e-commerce sites. Our findings may help policymakers and regulators to promote our more secure internet use. (Feng et al., 2023)

9.Title: An Analysis of Differences between Dark Pattern and Anti-Pattern to Increase Efficiency Application Design
Authors: Pumarin Tiangpanich,Apichaya Nimkoompai,
Publication: 2022 7th International Conference on Business and Industrial Research (ICBIR)
As the technological growth that happens to be much more than expected, marketing has started shifting from offline marketing to social media marketing. This has therefore made marketers start using a mobile application to represent their brands and products to customers. It is here that UX, to make it more attractive, has to ensure that brands

communicate seamlessly with their consumers by using UI to portray the interaction in visible and comprehensible terms. The rule of thumb is that, in the modern context, information is one of the fundamentals of marketing. Lead to an organization trying to extract usersu2019 personal information without their consent, resulting in damage to the property of users via the Dark pattern of User Experience (UX). However, when designers design the application, they might notice that the design is having trouble or not having a good solution to deal with the problem that users might find. The design experience that can solve this problem is called Anti-patterns-in which approaches to common issues may seem obvious but are less than optimal in practice. While the dark pattern of user experience (UX) is deceitful, UX/UI design or inter-actions created with psychological knowledge is designed to mislead users to do something they did not intend to create value for the service that employs them. This will explain ways in which this research can help designers, with the right tools, create compelling artwork by understanding and applying anti-patterns and dark patterns. (Tiangpanich & Nimkoompai, 2022)

3 Problem statement

The prevalence of dark patterns in websites really threatens the general online user experience: misleading design strategies fool people into making decisions against their own best interests. Consequently, it is highly cumbersome to identify and mitigate such dark patterns manually, and this itself is a reason why there is a need for an automated system that makes use of machine learning, natural language processing, and web scraping in finding and categorizing those unethical design components. It satisfied the long-overdue requirement for comprehensive action protecting users from manipulative practices and helped nurture trust in the use of digital space.

4 Necessity

In the digital world today, dark patterns on the net have grown to a level where they've actually become a real hazard to internet users. Confidence and transparency, therefore, come into question. This is developing an increasing need to deliver a dark pattern automatic detection system that will help users of virtual interactions take protective measures against this kind of obtrusive design tactic. This kind of system will be fundamental for maintaining the users' autonomy, enabling them to make informed choices, and cultivating an ethical ground for a healthy digital environment while conducting user interface design.

5 Advantages

This automatic dark pattern detection system would be greatly advantageous: it would strongly protect users from manipulative practices, ensure transparency within digital interaction, and avoid the various treacherous one-way mirrors. The proposed system detects and categorizes dark patterns in a proactive way, leveraging machine learning and Web scraping in support of users' free actions online. This gives more trust not only to the whole

digital ecosystem but also contributes to the spread of best practices of ethical design methodologies among web developers and enterprises. In return, the automated detection mechanism smoothes the process of detection, saving time and effort invested in developing a user-centered online environment focused on ethical and transparent design principles.

The approach is scalable as long as it uses cloud scraping and preprocessing techniques while handling big datasets. It will also provide a user-friendly GUI interface able to provide real-time analysis, considering users input in terms of URLs to check for dark patterns. More than mere detection, ethical design is encouraged by making developers aware of the manipulative tactics to raise them toward transparency and user-centricity. Also, the modular and extensible architecture will allow adding new features in the future, including advanced machine learning algorithms and browser extension support. This work clearly furthers the development of new technologies to curb dark patterns and greatly supports trust and transparency in the virtual world.

6 Methodology:

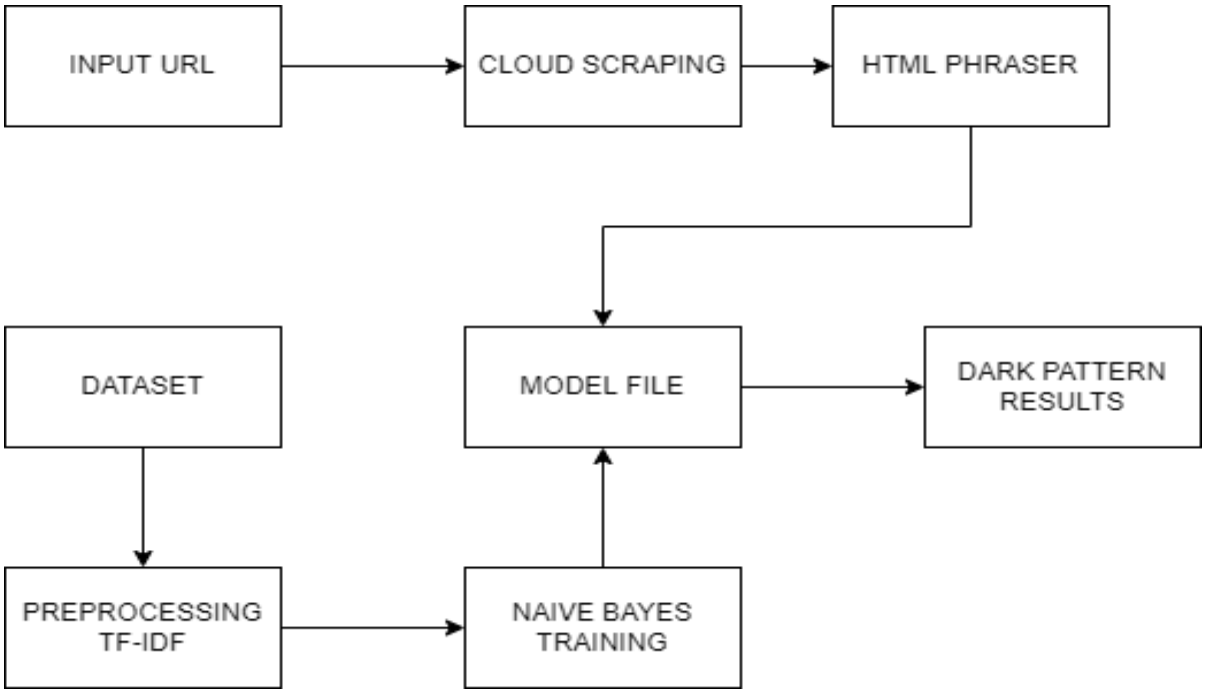

Fig 1: Implementation block diagram

6.1 BLOCK DIAGRAM DESCRIPTION:

The implementation consists of two main blocks: the Machine Learning Block and the Graphical User Interface (GUI) Block, each serving distinct yet interconnected functions.

6.2 Machine Learning Block:

In this section, several Python scripts are prepared using the scikit-learn library for developing and evaluating a Multinomial Naive Bayes classifier on the Dark Pattern detection task. First, the dataset has been cleaned using the Pandas library and split into training and test subsets. Text vectorization using the TF-IDF approach was conducted to improve feature extraction. The transformed training dataset is used for training the Naive Bayes model, whose performance is tested using the test set. Finally, the trained model along with the TF-IDF vectorizer, is serialized into files which can later be deserialized and used for real-time predictions.

6.3 GUI Block:

The following GUI block, which has been designed using the Tkinter library in such a manner that the user can interface the system of detection through it, contains one entry field for URL input, one button 'Analyze' to fire up the analysis, and a scrolled text widget for showing the result. This would interface with the current model and vectorizer in respect of fetching, parsing, and classification of information at the given URLs.
It will enable them to insert the URLs effortlessly, start the analysis, and get immediate feedback whether dark patterns exist or not, in a very practical and easy way to enhance online user protection.They are one system set together, which would cooperatively combine machine learning capabilities with an intuitive GUI for effectively carrying out user-centric dark pattern detection in web content.

7 Module description and evaluation:

1. Dataset Loading and Preprocessing:

Step 1: Load the dataset from 'dataset.csv' using Pandas.

Step 2: Split the dataset into training and testing sets, with 80% for training and 20% for testing.

2. Text Vectorization using TFIDF:

Step 3: Utilize the TFIDF vectorizer (with a maximum of 5000 features) to convert text data into numerical vectors.

Step 4: Transform both the training and testing sets.

3. Training the Naive Bayes Model:

Step 5: Employ a Multinomial Naive Bayes classifier for training on the TFIDF transformed training data.

Step 6: Generate a model that learns the patterns associated with different dark pattern categories.

4. Model Evaluation:

Step 7: Evaluate the trained model using the testing set.

Step 8: Calculate and print the accuracy score and a classification report.

Bait and switch	0.88	1.00	0.94	81
Forced continuity	0.96	0.66	0.78	38
Hidden costs	0.97	0.74	0.84	47
Not Dark Pattern	0.90	0.98	0.94	244
Sneaking	0.94	0.81	0.87	62
accuracy			0.91	472
macro avg	0.93	0.84	0.87	472
weighted avg	0.92	0.91	0.91	472

Fig 2 :Prediction performace report

The above figure summarizes the prediction performances of the machine learning model in terms of precision, recall, F1-score, and support over the following classes: "Bait and switch," "Forced continuity," among others. Precision tells how many of the model's positive predictions were correct. Recall tells us how well a model captures all actual positives. F1-score balances Precision and Recall. Support refers to the actual number of occurrences for each class within the given dataset.

Principal results:

The model's accuracy is high, 91%, and in general performs quite well, especially in "Not Dark Pattern" and "Bait and Switch." Classes such as "Forced Continuity" have lower recall, with a value of 0.66, indicating a miss of true instances for such classes. While the macro averages consider all classes to be of equal importance, the weighted average considers class support; hence, it balances between classes in the dataset.

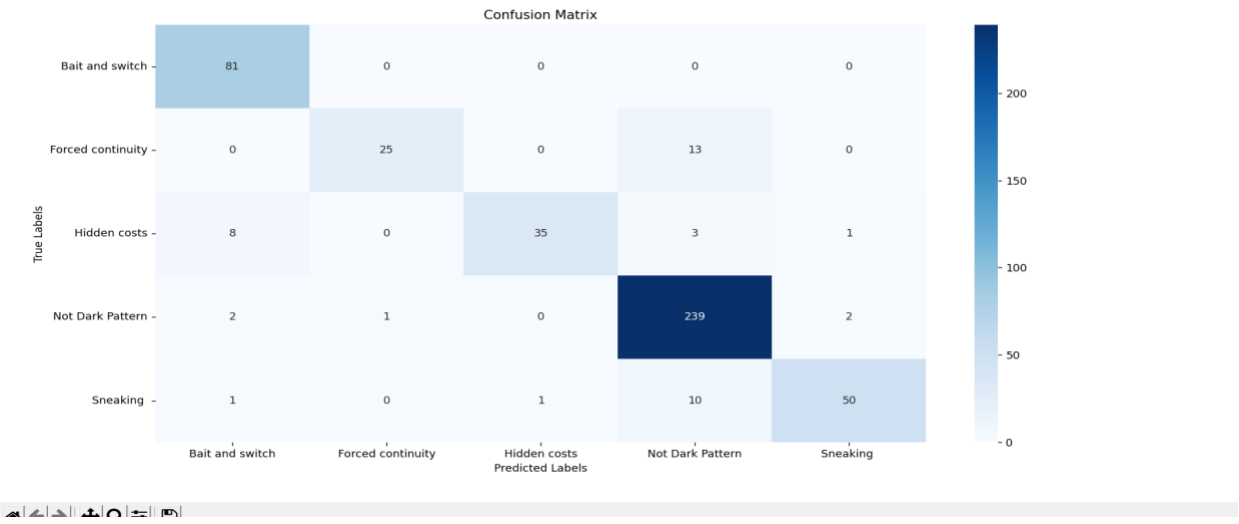

Fig 3 : Confusion matrix

The confusion matrix checks how well the classification model's predictions match the true labels. The diagonal cells show correct classifications, and the off-diagonal cells show misclassifications. The model does well overall, with most predictions found along the diagonal. It correctly predicted "Bait and switch," at 81 times. "Forced continuity" had correct predictions of 25 but was wrong in being labeled as "Not Dark Pattern" 13 times. "Hidden costs" had 35 correct identifications but were mainly misclassified for "Bait and switch," at 8, and for "Not Dark Pattern," at 3. The "Not Dark Pattern" class performed pretty well, with 239 correct classifications. Minor misclassifications are observed, namely for "Sneaking" there are 2 errors and for "Forced continuity" there is 1 error. More precisely, for the "Sneaking" category, out of a total of 50, there were 10 marked as "Not Dark Pattern" and 1 marked as "Hidden costs.". On the whole, this model should work fine. There is apparent much confusion between some classes, such as "Forced continuity" vs. "Not Dark Pattern", or "Hidden costs" with others. These could also do better with a better model or more balanced data.

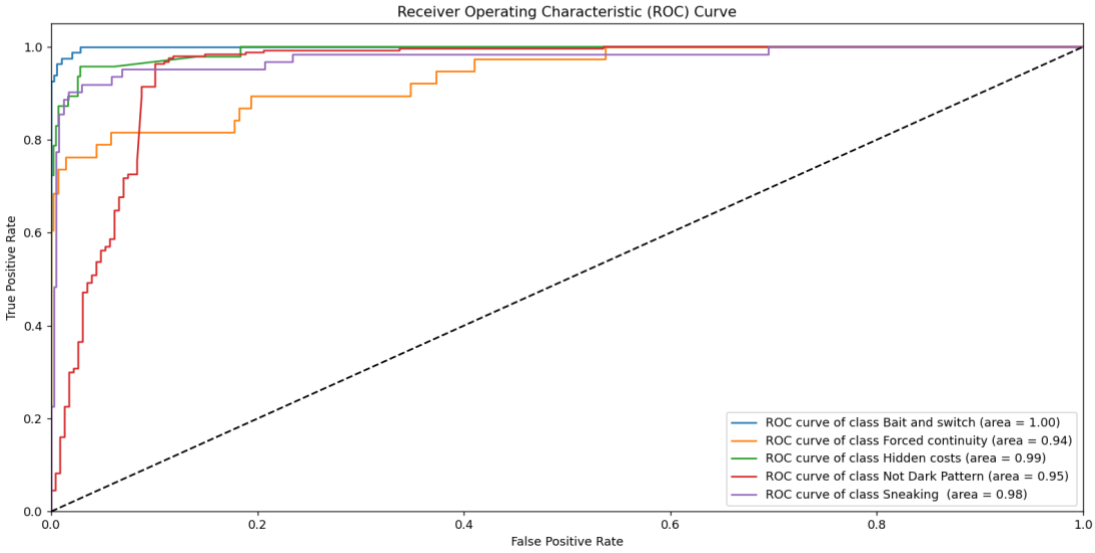

Fig 4 : ROC curve evaluation

The above ROC curve evaluates the model's performance for different classes, with high AUC values indicating strong classification ability. The model performs exceptionally for "Bait and switch" (AUC = 1.00) and "Hidden costs" (AUC = 0.99) and maintains robust performance for "Sneaking" (AUC = 0.98) and "Not Dark Pattern" (AUC = 0.95). "Forced continuity" has the lowest AUC (0.94), reflecting some misclassification issues observed earlier. Overall, the curves demonstrate that the model reliably distinguishes between classes, with only slight weaknesses in handling "Forced continuity."

5. Model and Vectorizer Serialization:

Step 9: Save the trained Naive Bayes model and TFIDF vectorizer using the joblib library.

Step 10: Create 'dark_pattern_detection_model_naive_bayes.pkl' and 'tfidf_vectorizer.pkl' files.

6. Web Scraping and Classification:

Step 11: Utilize the trained model and vectorizer in a Tkinterbased GUI application for realtime analysis of URLs.

Step 12: Use cloudscraper and BeautifulSoup to fetch and parse HTML content.

Step 13: Clean the extracted text by removing unnecessary elements and save it in 'sc.txt'.

Step 14: Classify the cleaned text into dark pattern categories, segmenting text by paragraphs.

7. GUI Implementation:

Step 15: Develop a Tkinterbased graphical user interface (GUI) with input for entering URLs.

Step 16: Include a button to trigger the analysis and a scrolled text widget to display the results.

Step 17: The 'Analyze' button invokes the scrape_and_classify function, providing predictions for each paragraph in the URL's content.

8. Tkinter Event Loop:

Step 18: Launch the Tkinter event loop to ensure the continuous functioning of the GUI.

Step 19: Users can input URLs, trigger the analysis, and receive real-time predictions on the presence of dark patterns.

This methodology seamlessly integrates machine learning, web scraping, and GUI development to create a comprehensive system for detecting dark patterns in website content. The trained model and vectorizer enable efficient real-time analysis of URLs, offering users insights into the potential presence of deceptive design elements.

7.1 MACHINE LEARNING:

At the core of the proposed framework for detecting dark patterns, machine learning provides a rich foundation and offers key enhancements that will make the solution markedly more effective and scalable. (Ferreira, n.d.)An implementation of the Naive Bayes classifier, trained on these examples of dark pattern classification-which would be Bait and Switch, Forced Continuity, Price Comparison Prevention, Hidden Costs, and Sneaking-allows the automated detection of such deceptive design features. (Shen et al., 2018)It learns from the previous data, adapts to the change in pattern, and becomes more precise with time. Inclusion of a TFIDF vectorizer at the pre-processing stage enhances the quality of feature extraction by encapsulating semantic importance of terms and thereby fine-tunes the model for improved recognition of textual data patterns.(Lingadeu, n.d.) Machine learning allows the system to be dynamically adapted to the ever-changing characteristics of web content in order to ensure its efficiency in detecting new, emerging dark patterns. This trained model file encapsulates knowledge learned from various datasets and will serve as a flexible tool for classifying scraped information coming from different websites. This also allows the system to adapt to the continuous evolution of the deceiving strategies in use and to increase the generalization ability across a wide range of platforms and contexts. Machine learning

automation ensures that identifications of dark patterns occur in a timely enough manner so that protection of users' online interactions can be provided. (*How to Use Tf-Idf with Naive Bayes?*, n.d.; Minaee et al., 2020)It is expected that, through this initiative, machine learning capabilities will be leveraged to continue to develop a more robust and proactive strategy toward mitigating the problems brought about by misleading design practices on the internet toward ensuring a safe and reliable online environment.(Ferreira, n.d.)

7.2 ALGORITHM CONTRIBUTION:

The implemented algorithm will apply the Naive Bayes classifier for the purposes of detecting dark patterns appearing on websites. The Naive Bayes algorithm is a statistical machine learning algorithm based on Bayes' theorem, assuming that the features are independent of each other. (Ferreira, n.d.)For the dark pattern detection, features were provided from the textual data by the TF-IDF vectorizer, which captures the importance of the terms in the dataset. It is a methodology that turns textual data into numerical vectors; hence, the algorithm can effectively analyze and classify the patterns. (*How to Use Tf-Idf with Naive Bayes?*, n.d.)For that purpose, one can train it by providing the algorithm with a labeled dataset wherein each of the data points is marked with their class of dark pattern: Bait and Switch, Forced Continuity, Hidden Costs, Not Dark Pattern, and Sneaking. Therefore, the Naive Bayes model would have learned a probability of features for each class and, hence, a backing for its future prediction.(Otten & Otten, n.d.)

It utilizes the learned probabilities in the process of classification regarding how a certain text belongs to each category. It designates the category with the highest probability for a text-in other words, the existence of a certain dark pattern. It shows a classification report with the algorithm's efficacy: precision and recall for each category of distinct pattern type-including F1-score. Enhanced with TF-IDF vectorization, Naive Bayes did quite well in picking out the patterns of devious design on the websites, based on the probabilistic analysis of text, to help build a systematic and reliable system for detection.(Minaee et al., 2020; *Naive Bayes Classifier Tutorial: With Python Scikit-Learn*, n.d.; Otten & Otten, n.d.)

8 Discussions

The investigation of automatic detection of dark patterns using a Naive Bayes classifier yielded many findings of considerable importance that were further supported by extensive experimental evaluations. The overall performance of the system was at an accuracy rate of 91%, demonstrating that it was reliable in identifying misleading design features from target categories. Necessary metrics for evaluating the performance of the system included precision, recall, F1-score, and ROC-AUC values that indicated various strengths and vulnerabilities that had arisen.

8.1 Model Performance

The Naive Bayes classifier developed on the textual data processed by the TFIDF vectorizer works well, especially for classes such as Bait and Switch, Hidden Costs, and Sneaking. In evidence, the high level of classification metrics, especially in these classes, means the system could reliably identify patterns in this data. More importantly, "Not Dark Pattern" was also presenting outstanding results, hence confirming the capability of the system in distinguishing true manipulative content from innocuous design elements. The evaluation of the Forced Continuity pattern indicated a poor performance about correct classification that the model achieved. Classes are inaccurately represented, as evidenced by lower recall and precision in the confusion matrix and ROC-AUC values. This could be because some classes had too much spillover from the features of other classes, or not enough representation of this class in the dataset. Such deficiencies in the data and application of higher-order modeling techniques need to be addressed to improve the performance of classification.

Category	Precision	Recall	F1-Score	Support
Bait and Switch	0.94	0.88	0.91	81
Forced Continuity	0.66	0.74	0.7	47
Hidden Costs	0.78	0.98	0.87	244
Not Dark Pattern	0.94	0.9	0.92	472
Sneaking	0.81	0.84	0.82	62
Weighted Average	**0.91**	**0.91**	**0.91**	**-**

8.2 Analytical Observations

The important features were highly salient with the TFIDF vectorizer and hence enhanced the classifier to extract the fine-grained pattern from the text data. It played an indispensable role in the whole system performance; great generalization capability can be observed in the test sets from the model. The model's performance was evaluated with some prediction metrics, and support showed the balance between classes in the data, although certain classes are imbalenced, impacting the classification accuracy of a category.

8.3 Analysis of Confusion Matrix

The confusion matrix gave enough details on the strengths and weaknesses of the model. "Hidden Costs," for example, was sometimes incorrectly put into the category of "Bait and Switch," showing features common in both classes. Similarly, instances of Forced Continuity were mistakenly labeled as "Not Dark Pattern," illustrating how there is definitely room for an even finer approach to feature extraction. These results indicate that increasing specificity in the feature set and supplementing extra contextual analysis might strongly benefit the model.

True Labels	Bait and Switch	Forced Continuity	Hidden Costs	Not Dark Pattern
Bait and Switch	81	0	8	0
Forced Continuity	13	25	1	10
Hidden Costs	0	0	35	3
Not Dark Pattern	0	0	0	239
Sneaking	0	0	1	10
Weighted Average	**0.91**	**0.91**	**0.91**	**-**

8.4 Real-World Applicability

The idea was to make the system more practically applicable by incorporating machine learning into a user-friendly GUI capable of real-time analyses. In this respect, users can provide website URLs and get instant feedback on the detected dark patterns, thus helping with decision-making. But out there, in the wild, there is still darkness, and its techniques continue to diversify, which is an issue. From time to time, the model and dataset need to change because, over time, the efficiency could be lost by newly emerging patterns. In future, The system already showed quite a lot of satisfactory potential, but several paths can still be pursued as far as improvements are concerned. Furthermore, visualization-integrated GUIs and cooperation with browser extension developers may further broaden the system's reach and usability. In this way, the experiments conducted during this research prove the efficiency of the suggested system while pointing out some aspects to be further developed. Addressing the acknowledged limitations and making use of advances in machine learning, the system has the potential to evolve into a comprehensive tool to combat deceptive online practices and further enhance a safer and more ethical online environment.

9 Conclusion and Future Work

In conclusion, the Naive Bayes classifier, as evidenced by the classification report, has demonstrated commendable performance in detecting various dark patterns on websites. This research, therefore, designed and consequently applied the all-encompassing scheme for dark pattern detection in websites. Detection was based on the Naive Bayes classifier, significantly

augmented with techniques such as TFIDF vectorization and web scraping. So far, it has an accuracy of about 91%, hence quality performance in the recognition of different forms of deceptive practices such as Bait and Switch, Hidden Costs, and Sneaking. This will mean that in the case of picking out misleading design features, the system will be efficient and robust in sending out signals. The classification report, in conjunction with all the metrics, reflects the fact that the model can make subtle differences among a wide variety of different patterns with very high reliability. This forms a very important capability in ensuring safe and transparent navigation in the digital environment.

It combines machine learning with user-friendly GUI design so that users can comfortably crunch the contents of a website for dark patterns, building trust and hence assuring ethics in digital interaction. This will further cement the capability of automated systems to solve some of the age-old challenges in the detection of manipulative online practices that ensure user autonomy and digital transparency. From a pragmatic perspective, the collaboration of the concerned industry players, like web developers and regulatory agencies, might convincingly adopt this system as a normative tool of design ethics. By giving recommendations on how the system should be embraced to guide website development practices, the study contributes to lowering the occurrences of dark patterns while transforming the online world into a safer and more user-focused space

This, provides a very good basis for the automatic detection of dark patterns and, at the same time, underlines great opportunities that are currently opening within machine learning and natural language processing regarding how to tackle ethical issues related to online design. After further refinement and development, this proposed system may change how users interact with online services, thus creating for the individual a much more open, trustworthy, and fair virtual space.

FUTURE SCOPE:

In the future, this might be further refined by the incorporation of even more sophisticated machine learning algorithms such as deep learning models that will offer even higher accuracy and generalization. With an increased diversity of examples of dark patterns in the dataset, improvements can also be made using state-of-the-art natural language processing methods. Real-time updates, put alongside the capability for adaptation to changing patterns of online design tactics, make the system relevantly efficient. This is also where improvements might be afforded to the GUI by adding in visualization and explanations of the detected pattern, which would enhance user understanding.

Because of that, it can easily be integrated into a browser extension or some security applications, and it would greatly extend the current scope and scale of it. That would mean the users will have much wider access, and it will be much easier to apply it in practice to the everyday browsing, further consolidating the system's position of a building block in the development of the principles of ethical design. That means great security for users, as well as an opportunity to develop a safer, more secure cyber space for all participants.

References

Dula, E., Rosero, A., & Phillips, E. (2023). Identifying Dark Patterns in Social Robot Behavior. 2023 Systems and Information Engineering Design Symposium, SIEDS 2023, 7–12. https://doi.org/10.1109/SIEDS58326.2023.10137912

Feng, J., Mo, F., Yada, Y., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (2023). Analysis of Dark Pattern-related Tweets from 2010. 2023 IEEE 8th International Conference on Big Data Analytics, ICBDA 2023, 100–106. https://doi.org/10.1109/ICBDA57405.2023.10104855

Ferreira, H. (n.d.). Basics of Machine Learning and a simple implementation of the Naive Bayes algorithm. Https://Medium.Com/Hugo-Ferreiras-Blog/Basics-of-Machine-

Learning-and-a-Simple-Implementation-of-the-Naive-Bayes-Algorithm-80c1e67a2e8a.

Hasan Mansur, S. M., Salma, S., Awofisayo, D., & Moran, K. (2023).

AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces. Proceedings - International Conference on Software Engineering, 1958–1970. https://doi.org/10.1109/ICSE48619.2023.00166

how to use tf-idf with Naive Bayes? (n.d.). Https://Stackoverflow.Com/Questions/37405617/How-to-Use-Tf-Idf-with-Naive-Bayes.

Kirkman, D., Vaniea, K., & Woods, D. W. (2023). DarkDialogs: Automated detection of 10 dark patterns on cookie dialogs. Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023, 847–867. https://doi.org/10.1109/EuroSP57164.2023.00055

Lacey, C., & Caudwell, C. (2019). Cuteness as a “Dark Pattern” in Home Robots. ACM/IEEE International Conference on Human-Robot Interaction, 2019-March, 374–381. https://doi.org/10.1109/HRI.2019.8673274

Lingadeu. (n.d.). TF-IDF for Text Preprocessing in Machine Learning. Https://Medium.Com/@lingostat/Tf-Idf-for-Text-Preprocessing-in-Machine-Learning-A66b29774040.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep Learning Based Text Classification: A Comprehensive Review.

Nimkoompai, A. (2022). Risk Analysis of Encountering Dark Patterns of UX E-commerce Applications Affecting Personal Data. In 6th International Conference on Information Technology, InCIT 2022. https://doi.org/10.1109/InCIT56086.2022.10067640

Parrilli, D. M., & Hernandez-Ramirez, R. (2020). Re-Designing Dark Patterns to Improve Privacy. International Symposium on Technology and Society, Proceedings, 2020-November, 253–254. https://doi.org/10.1109/ISTAS50296.2020.9462197

Shen, Z., Zhang, Y., Wei, L., Zhao, H., & Yao, Q. (2018). Automated Machine Learning: From Principles to Practices.

Tiangpanich, P., & Nimkoompai, A. (2022). An Analysis of Differences between Dark Pattern and Anti-Pattern to Increase Efficiency Application Design. ICBIR 2022 - 2022 7th International Conference on Business and Industrial Research, Proceedings, 416–421. https://doi.org/10.1109/ICBIR54589.2022.9786470

Yada, Y., Feng, J., Matsumoto, T., Fukushima, N., Kido, F., & Yamana, H. (2022). Dark patterns in e-commerce: a dataset and its baseline evaluations. Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022, 3015–3022. https://doi.org/10.1109/BigData55660.2022.10020800