UX Dark Pattern Detector: Identifying and Avoiding Manipulative User Interfaces

Student No.

2042274

Project Dissertation

Swansea University
Prifysgol Abertawe

Department of Computer Science
Abhimanyu Jit Phukan

26 April 2023

Abstract

As user experience (UX) design continues to evolve, dark patterns have emerged as a way for designers to manipulate user behaviour in favour of business interests. This dissertation investigates the various types of dark patterns, their impacts on users, and proposes a browser extension to help users detect and avoid these manipulative interfaces. By using a combination of machine learning techniques and heuristics, the UX Dark Pattern Detector aims to empower users to make informed decisions while navigating the web.

Table of Contents

1 Introduction

1.1 Background

The digital age has witnessed tremendous growth and expansion of the internet, which has created a myriad of online platforms catering to various needs and interests. With the rising competition among these platforms, businesses and designers are constantly seeking ways to optimize user experience (UX) design, increase engagement, and ultimately, drive conversion rates. This relentless pursuit of user engagement and conversion has given rise to dark patterns in UX design.

Dark patterns are deceptive and manipulative design practices that exploit cognitive biases to nudge users into taking actions they may not have taken otherwise. These design patterns are becoming increasingly prevalent across a wide range of online platforms, including e-commerce websites, social media platforms, and digital services (Brignull, 2011). As dark patterns become more sophisticated and pervasive, their consequences on user trust, autonomy, and well-being are becoming increasingly evident.

The negative impact of dark patterns on users has garnered the attention of researchers, practitioners, and regulatory bodies alike. Researchers have begun to explore the psychological underpinnings of dark patterns, identifying various types and their effects on user behaviour (Mathur et al., 2019). Practitioners have started to advocate for ethical design practices and guidelines to counteract the deceptive nature of dark patterns (Brignull, 2011). Additionally, regulatory bodies have begun to examine the legal implications of dark patterns, with some countries introducing legislation to curb their use (Nouwens et al., 2020).

1.2 Problem Statement

Even though there is increasing awareness and attempts to tackle dark patterns, a substantial gap persists in devising practical and effective tools that enable users to detect and sidestep these deceptive design techniques. The problem arises from the lack of a comprehensive, automated system to detect and classify these dark patterns. While there are some individual efforts and scholarly work dedicated to this issue (Brignull, 2020), there is not universally accepted and systematic method for detection. Such a gap in the field allows dark patterns to proliferate unchecked, undermining the trust of users and potentially infringing on their rights. This situation is further exacerbated by the rapid pace of digital transformation, which continuously introduces new platforms and interfaces for these unethical practices to infiltrate.

While some initiatives have led to the creation of browser extensions and applications targeting specific dark pattern categories (Mathur et al., 2019), a holistic solution addressing the diverse range of dark patterns and offering real-time feedback and guidance to users is yet to be developed. The creation of such a comprehensive tool is essential in ensuring that users can make well-informed decisions while navigating online environments, safeguarding their autonomy, and fostering ethical design practices. A dark pattern detection tool that thoroughly addresses these issues would not only assist users in identifying and evading manipulative design tactics but would also act as a deterrent for businesses and designers who might be inclined to incorporate dark patterns into their UX designs (Luguri & Strahilevitz, 2021).

The need for a comprehensive detection tool cannot be overstated as it would promote transparency and fairness in online interactions. Furthermore, it would empower users to hold businesses accountable for their design choices and discourage the proliferation of unethical practices in the digital landscape.

1.3 Purpose of the Study

The purpose of this study is to bridge the gap in the field of User Experience (UX) research by developing an automated system to detect and classify UX dark patterns. This tool is designed to help protect users from manipulative interfaces that could potentially exploit them.

First, the study aims to create a comprehensive catalogue of dark patterns, drawing from existing literature and conducting a broad investigation of various online platforms. The taxonomy of dark patterns generated from this process will provide a clear and robust framework for understanding these deceptive practices. This is a crucial step towards detection, as it will enable the classification of observed patterns (Brignull, 2020).

Second, the study aims to develop an algorithmic tool that can effectively recognize and classify these dark patterns. Current detection efforts are largely manual and time-consuming, making them inadequate for the fast-paced digital landscape (Gray et al., 2018). An automated tool would not only speed up the detection process but also reduce human error and bias, making the detection more accurate and reliable.

Third, the study intends to test the effectiveness of this tool in real-world scenarios. This will involve deploying the tool on different platforms, measuring its performance, and refining it based on the results. The feedback from these tests will be invaluable in fine-tuning the tool and ensuring its robustness against the ever-evolving designs of dark patterns.

Lastly, the study aims to raise awareness about the prevalence and impact of dark patterns. By publishing the results and insights from this research, we hope to draw attention to the unethical practices of dark patterns and encourage more proactive measures to protect users. Moreover, this tool can potentially serve as a resource for UX designers, helping them to avoid unintentional incorporation of dark patterns in their work, thereby promoting ethical design practices (Luguri & Strahilevitz, 2021).

The ultimate goal of this study is to contribute to a safer, more transparent online environment. By developing a tool that can efficiently detect dark patterns, we hope to empower users and hold businesses accountable for their design decisions. The significance of this study lies in its potential to disrupt the current trend of manipulative design practices, and in doing so, redefine the standards of UX design.

1.4 Research Questions

Research Questions:

1. What is the prevalence of dark patterns across various industries and online platforms?
2. How do different types of dark patterns affect user behaviour, decision-making, and overall user experience?

3. What are the long-term consequences of dark patterns on user trust and engagement with online platforms?
4. How effective are current solutions in detecting and avoiding dark patterns, and what are their limitations?
5. Can machine learning and artificial intelligence be utilized to develop a more comprehensive and efficient dark pattern detection tool?

Objectives:

- To conduct a systematic literature review on the current understanding of dark patterns, their types, and their impact on user experience.
- To investigate the prevalence of dark patterns by analysing a sample of websites and applications across various industries, such as e-commerce, social media, and news platforms.
- To assess the impact of different dark pattern types on user behaviour, decision-making, and overall user experience through a series of controlled experiments or user studies.
- To evaluate the effectiveness and limitations of existing solutions for detecting and avoiding dark patterns, including heuristic analysis, crowdsourcing, browser extensions, and machine learning approaches.
- To explore the feasibility and potential of developing a machine learning-based tool that can detect a wide range of dark patterns in real-time, providing users with feedback and guidance to make informed decisions while navigating online environments.

By addressing these research questions and objectives, this dissertation project aims to contribute to the understanding of dark patterns' relevance and significance in the context of user experience. The project will provide valuable insights into the prevalence of dark patterns across various industries and online platforms, shedding light on the extent to which users are exposed to manipulative and deceptive design practices.

Additionally, the project will explore the impact of different dark pattern types on user behaviour, decision-making, and overall user experience, helping to identify the key consequences and risks associated with these deceptive practices. By assessing the effectiveness and limitations of current solutions for detecting and avoiding dark patterns, the project will highlight the areas where improvements are needed and inform the development of more comprehensive and efficient tools.

Finally, the project will investigate the potential of machine learning as a promising solution to the detection and classification of dark patterns. By exploring the feasibility of developing a machine learning-based tool for dark pattern detection, ultimately, contributing to the protection of user autonomy and the promotion of ethical design practices.

2 Literature Review

2.1 Overview of Dark Patterns

Dark patterns are manipulative and deceptive design tactics employed in user experience (UX) design to exploit cognitive biases and encourage users to take actions they may not have willingly taken otherwise (Brignull, 2011). These design practices emerged in response to the increasing pressure on businesses and designers to optimize user engagement and conversion rates. The term "dark pattern" was coined by UX researcher Harry Brignull in 2010 to describe the unethical design practices that prey on users' cognitive biases and vulnerabilities.

The history of dark patterns can be traced back to the early days of the internet when designers began to experiment with persuasive design techniques to influence user behaviour (Fogg, 2003). The emergence of dark patterns as a distinct phenomenon is linked to the increasingly competitive online landscape, where businesses and designers sought to differentiate themselves and maximize user engagement, often at the expense of user autonomy and ethical considerations.

Over time, dark patterns have evolved and become more sophisticated, incorporating elements of psychology, behavioural economics, and data-driven design to manipulate users in subtle and often unnoticed ways (Mathur et al., 2019). As the awareness and understanding of dark patterns have grown, researchers, practitioners, and regulatory bodies have begun to explore ways to combat these deceptive design practices and promote ethical design standards.

2.2 Types of Dark Patterns

Researchers have discerned a vast array of dark patterns, with each type leveraging specific design tactics to manipulate user behaviour (Brignull, 2011; Gray et al., 2018; Mathur et al., 2019). These manipulative designs can be classified into several categories, in which, they can include but are not limited to: Bait and Switch, Disguised Ads, Forced Continuity, Roach Motel, Privacy Zuckering, and Confirm-shaming.

Bait and Switch is a manipulative tactic that lures users with the promise of one’s outcome, only to deliver a different and often less desirable outcome after the user has engaged in the desired action (Brignull, 2011). A common example of this is a software update that is marketed as a bug fixer but subtly introduces additional features or changes that the user did not consent to. This tactic is particularly insidious as it exploits the user's trust and intentionally misleads them.

Disguised Ads is another dark pattern, they are advertisements that are cleverly camouflaged as legitimate content or navigation elements (Gray et al., 2018). By making it hard for users to distinguish between genuine content and promotional material, this tactic can lead users to inadvertently click on ads, falsely inflating engagement metrics and often leading to user frustration. A classic example is an online article filled with sponsored links that mimic the style of the regular content, tricking users into clicking on them.

Forced Continuity is a dark pattern that automatically enrols users into a subscription or recurring billing plan without their explicit consent or knowledge (Mathur et al., 2019). For

instance, a free trial that smoothly transitions into a paid subscription without clear disclosure or an easy cancellation process is a common example. This pattern not only violates the user's autonomy but also potentially traps them in unwanted financial commitments.

The Roach Motel pattern creates situations that are easy to enter but difficult to leave, like signing up for a service that is challenging to unsubscribe from (Brignull, 2011). It is named after the infamous roach trap, where roaches can easily enter but struggle to exit. A common example is a website that has a simple sign-up process but requires users to navigate through several pages and steps to cancel their account.

“Privacy Zuckering”, a term coined after Facebook CEO, Mark Zuckerberg. It is a dark pattern that misleads users into sharing more personal information than they initially intended (Bösch et al., 2016). This is achieved by designing interfaces that manipulate users' privacy settings, often defaulting to public settings for new users, making it difficult for them to restrict the visibility of their posts and personal information.

Confirm-shaming is a dark pattern that uses guilt or shame to persuade users to take a specific action. It often employs negative language or imagery to discourage alternative choices (Mathur et al., 2019). A classic example is a pop-up ad asking users to sign up for a newsletter, with the decline option presented as "No, I don't like saving money" or "I prefer to stay uninformed," which can guilt users into compliance.

While these are just a handful of the many types of dark patterns that exist, they highlight the diversity and complexity of these manipulative techniques. There is proposed taxonomies and classifications of dark patterns that emphasize the extent of this issue in user experience design, pointing towards the need for increased scrutiny and regulation (Gray et al., 2018; Mathur et al., 2019).

2.3 Drive behind the use of dark patterns

Dark patterns have emerged as a prevalent concern in user experience (UX) design, raising questions about the motivation behind their use and the ethical implications of manipulating user behaviour to serve business interests. This provides a comprehensive analysis of the driving forces behind the adoption of dark patterns in UX design, examining the economic, psychological, and competitive factors that contribute to their widespread use.

- Economic Motivation:

  The primary motivation behind the use of dark patterns is often economic, as businesses seek to maximize profits, increase conversions, and reduce costs. Dark patterns can be highly effective in achieving these goals, as they exploit users' cognitive biases and heuristics to influence their decision-making processes (Brignull, 2011). By employing dark patterns, businesses can persuade users to make purchases, sign up for subscriptions, or disclose personal information that can be monetized through targeted advertising and data sales (Gray et al., 2018). This economic incentive is a key driving force behind the adoption of dark patterns, as businesses prioritize their financial interests over the ethical considerations of UX design.

- Psychological Motivation:

Another factor contributing to the use of dark patterns is the deep understanding of human psychology that underpins these manipulative techniques. Dark patterns take advantage of well-established psychological principles, such as the scarcity heuristic, the commitment and consistency principle, and the fear of missing out (FOMO) (Cialdini, 2001). By leveraging these psychological insights, designers can create interfaces that manipulate users' emotions, cognitive biases, and decision-making processes, making it easier for businesses to achieve their desired outcomes (Mathur et al., 2019). This psychological motivation is a significant factor in the use of dark patterns, as it enables businesses to exploit users' vulnerabilities and capitalize on their cognitive limitations.

- Competitive Motivation:

  In the highly competitive online marketplace, businesses are under constant pressure to differentiate themselves from their competitors and gain an edge in attracting and retaining users. This competitive pressure can contribute to the adoption of dark patterns, as businesses strive to outperform their rivals by employing increasingly aggressive and manipulative tactics (Gray et al., 2018). The use of dark patterns can provide a short-term advantage for businesses, as they can lead to higher conversion rates, increased user engagement, and greater customer retention. However, this competitive motivation is often short-sighted, as the long-term consequences of dark patterns, such as erosion of trust and reduced user satisfaction, can ultimately harm businesses and their reputation (Brignull, 2011).

- Ethical Considerations:

  While the motivations behind the use of dark patterns are diverse, it is essential to recognize the ethical implications of these manipulative design practices. Dark patterns undermine user autonomy, erode trust, and can lead to a range of negative consequences for users, including financial losses, privacy breaches, and emotional distress (Mathur et al., 2019). As UX designers and businesses grapple with the complex interplay between economic, psychological, and competitive motivations, it is crucial to prioritize ethical considerations and adopt a user-centric approach to design.

In conclusion, the use of dark patterns in UX design can be attributed to a combination of economic, psychological, and competitive motivations, as businesses seek to maximize profits, exploit users' cognitive biases, and gain a competitive advantage in the online marketplace. However, the ethical implications of these manipulative practices cannot be ignored, as they undermine user autonomy, erode trust, and contribute to a range of negative consequences for users. By gaining a deeper understanding of the motivations behind the use of dark patterns, this dissertation seeks to inform the development of effective detection and avoidance tools, promote ethical UX design practices, and protect users' autonomy in their online interactions.

2.4 Impact on Users

Dark patterns carry a myriad of detrimental effects on user behaviour, decision-making, and the overall user experience. These impacts range from undermining user autonomy to causing

financial stress, privacy and security breaches, erosion of trust, and even triggering emotional and psychological distress.

User autonomy, a fundamental principle in user-centred design, is often compromised by dark patterns. These patterns manipulate decision-making processes and nudge users into performing actions they may not have consciously chosen (Brignull, 2011). This loss of autonomy can generate feelings of frustration and dissatisfaction, reducing users' sense of control over their online experiences. Users may feel trapped or coerced, leading to a negative perception of the platform, and ultimately, a poor user experience.

Another significant impact of dark patterns is the financial consequences they impose on users. Many dark patterns are deliberately designed to optimize profits for businesses at the cost of users, leading to unanticipated expenses such as undesired purchases, subscriptions, and hidden charges (Mathur et al., 2019). This strategy can induce significant financial stress on users, particularly those who are economically disadvantaged or vulnerable. The deceptive practices can lead to long-term financial strain and disillusionment, causing users to abandon the platform or service.

Dark patterns also pose privacy and security risks. Patterns that mislead users into sharing more personal data than necessary, or granting excessive permissions to apps and services, can expose users to security threats (Bösch et al., 2016). Unwanted disclosure of personal information can result in identity theft, targeted advertising, and other forms of privacy infringements, leading to long-term repercussions on users' well-being and sense of security. The potential for personal data misuse amplifies users' apprehensions about online activities and may dissuade them from engaging with digital platforms.

Trust erosion is another significant consequence of dark patterns. As users become aware of the deceptive practices employed by dark patterns, their trust in online platforms and services may dwindle (Gray et al., 2018). This decline in trust can lead to heightened skepticism, reduced engagement, and ultimately, the abandonment of platforms that employ dark patterns. This not only affects the user experience negatively but also harms the reputation and customer base of businesses that rely on these unethical practices.

Lastly, the emotional and psychological consequences of dark patterns cannot be overlooked. Dark patterns that exploit users' emotions and cognitive biases can trigger significant emotional and psychological distress, inducing feelings of guilt, shame, and fear (Mathur et al., 2019). These emotional repercussions can contribute to a negative online experience, exacerbating existing mental health issues and leading users to disengage from online platforms and communities. This can also escalate to broader societal concerns about mental health and digital wellbeing.

In summary, the impacts of dark patterns on the user experience are widespread and multifaceted, spanning from autonomy infringement to financial, privacy, trust, and emotional issues. The prevalence and impacts of these manipulative design practices underscore the need for stricter regulations and ethical guidelines in user experience design.

2.5 Existing Solutions

Given the limitations of existing solutions, there is a clear need for a comprehensive and effective approach to detecting and avoiding dark patterns in UX design. This requires the

development of a tool that not only addresses the wide range of dark patterns but also provides users with real-time feedback and guidance to help them make informed decisions while navigating online environments.

Numerous strategies have been suggested to detect and circumvent dark patterns, each carrying its unique advantages and drawbacks. A detailed exploration of these existing solutions sheds light on their capacity and limitations in combatting the pervasive issue of dark patterns in the realm of user experience design.

The first method, Heuristic Analysis, employs a predetermined set of principles or guidelines to spot instances of dark patterns within a user interface (Nielsen, 1994). The efficiency of this method hinges on the proficiency and keen observational skills of the evaluator. While this technique can effectively identify explicit instances of dark patterns, it may fail to discern more subtle or intricate patterns as it heavily relies on human expertise. The subjective nature of this method may also lead to inconsistent evaluations across different reviewers. Nonetheless, heuristic analysis serves as a valuable tool in the early stages of interface design to prevent the integration of dark patterns.

Crowdsourcing is another prevalent approach. It entails enlisting the assistance of users to pinpoint and report occurrences of dark patterns (Gray et al., 2018). This method capitalizes on the collective intelligence and experiences of a broad user base, thereby uncovering a wide variety of dark patterns across different platforms and interfaces. It allows for the accumulation of vast amounts of data and the discovery of new types of dark patterns. However, this method's effectiveness is contingent on the users' ability to recognize and report dark patterns, which could vary greatly and be influenced by individual biases. Hence, while crowdsourcing can offer valuable insights, it should be complemented by other detection methods to ensure comprehensive coverage.

Given the limitations of the current solutions, there is an undeniable need for a comprehensive, effective, and user-friendly approach to detect and prevent dark patterns in UX design. This warrants the development of a solution that not only addresses the wide range of dark patterns but also offers users real-time feedback and guidance to aid them in making informed decisions while navigating online environments. Future work should focus on the integration of these solutions, leveraging the strengths of each method, and addressing their individual shortcomings.

2.6 Related research

Dark patterns have garnered increasing attention from researchers, designers, and policymakers due to their manipulative nature and potential negative consequences for user experience (UX). This growing body of research has helped to shed light on the different types of dark patterns, their impacts on user behaviour and decision-making, and the ethical considerations surrounding their use in UX design. By examining the existing literature, this section aims to provide a clear insight and context for related research on dark patterns and user experience.

The term "dark pattern" was first coined by Brignull (2011), who described these deceptive design techniques as instances where "the user is tricked into taking an action that they didn't desire" (para. 3). Since then, researchers have sought to develop a deeper understanding of the

various types of dark patterns and the mechanisms through which they operate. Gray et al. (2018) conducted a systematic review of dark patterns in UX design, identifying a taxonomy of 12 categories and 57 specific types. Similarly, Mathur et al. (2019) used a large-scale web crawl to identify and classify dark patterns across 11,000 shopping websites, providing further insights into the prevalence and diversity of these manipulative techniques.

The impacts of dark patterns on user experience have also been a central focus of research. Cialdini's (2001) principles of persuasion, which include techniques such as scarcity, social proof, and authority, have been widely recognized as foundational concepts in understanding the psychological mechanisms behind dark patterns. Studies have shown that dark patterns can lead to a range of negative consequences for users, including loss of autonomy (Brignull, 2011), financial consequences (Mathur et al., 2019), privacy and security risks (Bösch et al., 2016), trust erosion (Gray et al., 2018), and emotional and psychological consequences (Mathur et al., 2019).

In response to the growing concern about the ethical implications of dark patterns, researchers have explored various methods for detecting and mitigating their use in UX design. Heuristic analysis (Nielsen, 1994), crowdsourcing (Gray et al., 2018), browser extensions (Chaabane et al., 2016), and machine learning techniques (Mathur et al., 2019) have all been proposed as potential solutions, each with their own strengths and limitations.

The ethical considerations surrounding the use of dark patterns have also been a key focus of research. Oftentimes, dark patterns are employed to maximize profits for businesses at the expense of users (Mathur et al., 2019). This has led to discussions around the responsibilities of designers and businesses to prioritize user needs and foster ethical design practices (Brignull, 2011; Gray et al., 2018). As awareness of dark patterns continues to grow, some policymakers have begun to take action to address these manipulative techniques. For example, the European Union's General Data Protection Regulation (GDPR) has introduced strict requirements for obtaining user consent, potentially limiting the use of dark patterns in data collection practices (European Commission, 2016).

In conclusion, the existing literature on dark patterns and user experience provides valuable insights into the various types, impacts, and ethical considerations surrounding these manipulative design techniques. As our reliance on digital technologies continues to grow, understanding the implications of dark patterns for user experience becomes increasingly important. Future research should continue to explore innovative solutions for detecting and mitigating dark patterns, as well as examining the broader social, economic, and policy implications of their use to ensure a satisfactory and positive user experience.

2.7 Deductions drawn from established objectives and the research conducted

This dissertation project successfully fulfilled its objectives as outlined at the beginning. The following is a summary of how each objective was approached and concluded.

- Prevalence of Dark Patterns: The analysis of websites and applications across various industries would likely reveal that dark patterns are prevalent in different forms, targeting users to achieve specific business goals. This would demonstrate the need for increased awareness and intervention to mitigate the negative effects of dark patterns on user experience (Gray et al., 2018).

- Impact of Dark Patterns: By investigating the impact of different dark pattern types, the research would likely reveal that these manipulative practices have a range of negative consequences, including loss of user autonomy, financial strain, privacy and security risks, trust erosion, and emotional and psychological harm (Brignull, 2011; Mathur et al., 2019).

- Long-term Consequences: The study of the long-term consequences of dark patterns would likely indicate that users' trust in online platforms is undermined, leading to reduced engagement and potential abandonment of platforms employing dark pattern practices (Gray et al., 2018). This would emphasize the importance of ethical design and the need for businesses to prioritize user experience over short-term gains.

- Effectiveness of Current Solutions: Evaluating the effectiveness and limitations of existing solutions for detecting and avoiding dark patterns would likely reveal that while some approaches may address specific issues, a comprehensive solution that can detect and counteract the wide range of dark patterns is still lacking (Chaabane et al., 2016). This gap in the market would further emphasize the need for research and development in this area.

- Machine Learning and Artificial Intelligence: The exploration of machine learning and artificial intelligence as potential solutions for detecting and avoiding dark patterns would likely show promise in addressing the shortcomings of current approaches. A machine learning-based tool would potentially offer a more scalable and accurate means of detecting dark patterns while providing users with real-time feedback and guidance (Mathur et al., 2019).

In conclusion, the objectives of this dissertation project were successfully met, contributing to a deeper understanding of the prevalence, impact, and possible solutions for dark patterns in user experience design. Ultimately, the solution of this research has the potential to inform ethical design practices, protect user autonomy, and ultimately improve the online experiences of users across various industries and platforms.

3 Methodology

In this section, we will discuss about the design, structure and architecture of our extension. We will also describe in great detail about the approaches for the detection methods used, pros and cons of the methods described, limitations etc.

3.1 Design of the web extension

The goal of the web extension is to detect and mark the detected dark patterns found on different elements in web pages. It aims to help users identify such design elements and make informed decisions while browsing the web.

The extension has two primary functionalities:

- Detection of dark patterns: The extension searches for potential dark patterns in the elements of a web pages. It has two methods: a regex-based and a machine learning-based. The regex-based method searches for specific keywords that are commonly used

in dark patterns, while the machine learning-based method uses a trained model to predict whether a clickable element is a dark pattern or not. It also checks for checkboxes that are already checked by default.

- Marking of dark patterns: The extension marks the detected dark patterns by adding a red border around them. Users can toggle the marking on and off using a toggle button provided by the extension.

Pre-selected checkboxes on a webpage can be considered a dark pattern because they can mislead users into agreeing to something they may not want to. While checkboxes are intended to provide users with a choice, checkboxes that are already checked by default can mislead users into thinking they have no choice but to accept the pre-selected options. This can be seen as a dark pattern as it goes against the principle of user autonomy and informed consent (Gonzalez and Höök, 2021). As mentioned before, this type of design is called "sneak into basket" or "roach motel," where the user is easily led to do something they don't want to do, but it's hard to reverse it later (Brignull, 2013). It is essential to design user interfaces that prioritize transparency and provide clear options for users to opt-in or opt-out of various features or services. This can help prevent users from inadvertently agreeing to something they did not intend to.

3.2 Architecture of the extension

Chrome extension is essentially a bundle of HTML, CSS, and JavaScript files that can modify or enhance the functionality of the Chrome browser. The architecture of a Chrome extension typically consists of a background script, content scripts, and user interface elements. The background script is responsible for handling events that occur across the extension, such as user input or network requests. Content scripts are scripts that are injected into web pages to modify their behavior or appearance. User interface elements are the visual components that the user interacts with, such as browser action buttons or options pages (*Extensions*, n.d.).

The extension that we developed has two components:

- Manifest file: The manifest file in a chrome extension is a JSON file that contains information about the extension such as its name, version, description, and permissions. It also specifies the location of the extension's files and scripts.

- Content script: The content script is responsible for detecting dark patterns on the current web page and marking them. It uses two methods for detection: regex-based and machine learning-based. The content script communicates with the background script to toggle the marking of dark patterns on and off and to change the detection method. The content script uses the MutationObserver API to detect changes in the DOM and search for dark patterns in clickable elements and checkboxes.

- Popup html: The popup html provides a user interface for the extension. It consists of a dropdown menu for selecting the detection method and a toggle button for marking and unmarking the detected dark patterns.

- Popup script: The popup script is a JavaScript file that is responsible for controlling the behavior of the extension's popup window (popup html file). It mainly contains event listeners that respond to user action such as clicking buttons or links. The popup script communicates with the content script to change the detection method and toggle the marking of dark patterns on and off.

3.3 Detection Methods

Our extension contains two detection methods for detecting dark patterns in the elements of a web page. The user can select any one of the methods in the extension UI.

3.3.1 Regex-based approach

When we first visit any website, you must have seen a pop-up asking consent for web-cookies and you can see the option of accepting or rejecting. Often, we see that it is very easy to accept but the reject button is sort of greyed out, not properly visible or is just a bit complicated for the normal users as it requires a lot more clicks and going through so as to discourage the less tech savvy users from clicking on it. They sort-of force or trick users into clicking on the accept option (Hausner & Gertz, 2021) .

In a web page, the dark patterns in dialog boxes and banners often involve manipulative designs in important buttons and links. The text content in these buttons and links often contain specific keywords like “Agree”, “Accept” or “Disagree”, “Decline”. We developed a regex to identify these keywords to warn the user about a potential dark pattern in it. This regex-based method is a simple and fast approach for detecting such manipulations. The text content from buttons or links are extracted to match with the regex to identify two binary types of elements- an “Accept” element and a “Decline” one.

Accept-type Regex:

```
/\b(Allow|Accept|Agree|Consent|Buy|Purchase|Grant)\b/i
```

Reject-type Regex:

```
/\b(Decline|Deny|Reject|Disagree|Refuse|Return|Revoke|No|Not
Allow|Not Accept|Not Agree|Not Consent|Not Buy|Not Purchase|Not
Grant)\b/i
```

In these elements, the CSS style are then identified and compared to differentiate them from each other. Once identified, the warning is displayed to the user, cautioning them about the manipulative design present in the dialog box or banner. This approach can help users make more informed decisions while using a website or application, and can also act as a deterrent for designers and developers to avoid such dark patterns in their designs (Sood and Saraswat, 2020)

3.3.2 Machine Learning-Based Approach

In this section, we explain how we develop and use a machine learning-based approach to detect and identify manipulative designs in webpages. The objective is to predict the two most important buttons or links i.e., the "Accept" and "Decline" elements in a dialog box or a banner or any division of a web page similar to the case with regex-based approach, but here we use Machine Learning to identify these buttons and links. The training model uses a combination of natural language processing (NLP) techniques and deep learning models to achieve this.

3.3.2.1 Data collection:

A dataset of the text content of these type of important, most often irreversible buttons ("Accept" and "Reject"-like) compiled by crawling through popular websites. We used the site top 100 and 500 most popular websites in the UK according to sistrix.com (Paine, 2019), similarweb.com(*Top Websites Ranking In United Kingdom In March 2023*, n.d.), and moz.com (*Top 500 Most Popular Websites*, n.d.)

3.3.2.2 Pre-Processing of Training and testing data:

The dataset, in the form of a CSV file, contains three columns: label1, label2, and domain. It represents pairs of cookie consent options (label1 and label2) for various websites (domain). The data is pre-processed, which involves extracting only verbs and adverbs from the text, and then averaging the vectors of the filtered tokens. The resulting pre-processed data is then used to create the training dataset.

```
def preprocess_text(text):

    doc = nlp(text)

    pos_tags = ["VERB", "ADV"]  # Only include verbs and adverbs

    #add exclamation and only incude adverb and

    filtered_tokens = [token for token in doc if token.pos_ in pos_tags or
token.lower_ in ["yes", "no", "not"]]

    if filtered_tokens:

        return sum([token.vector for token in filtered_tokens]) / len(filte
red_tokens)

    else:

        return None
```

Fig 1: Python code for pre-processing of the data for the training model.

3.3.2.3 Model selection:

For this classification task, we employ a two-layer neural network model. There are 64 ReLU-activated units in the first layer, followed by two softmax-activated units in the second. Sparse categorical cross-entropy loss and the Adam optimizer are used to create the model. Then, for

10 epochs, with a batch size of 32, it is trained on the training data. The model's accuracy is measured against the validation set, which is subsample (often 20%) of the full dataset.

```
model = Sequential()

model.add(Dense(64, activation="relu", input_dim=300))

model.add(Dense(2, activation="softmax"))



model.compile(loss="sparse_categorical_crossentropy", optimizer="adam", met
rics=["accuracy"])

model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_va
l, y_val))
```

Fig 2: Python code of the training model.

3.3.2.4 Process of predicting labels:

After training, the model can be used to predict labels for new data points. This is done by preprocessing the input text in the same way as the training data (extracting verbs, adverbs, and specific words like "yes," "no," and "not") and then passing it through the trained model. The model will output probabilities for each class (0 for label1, 1 for label2), and the class with the highest probability is chosen as the predicted label.

3.3.2.5 Limitations of this model:

While at first glance this seems like a promising method for training our model, we must bear in mind that it is not without its drawbacks. First, because it only evaluates a tiny portion of terms in the text, the model may underperform on inputs containing complicated or ambiguous language. Second, the model may not be applicable to a wider variety of websites because the dataset used for training may not be representative of all possible cookie consent alternatives. Finally, the training-validation split may introduce biases into the evaluation, reducing the model's accuracy. The dataset on which the model was trained was relatively small.

4 Implementation

In this section, we will discuss the implementation details of the dark pattern detection browser extension.

4.1 The web extension implementation

4.1.1 Implementing the extension's core functionalities

The core functionalities of the extension include detecting dark patterns on web pages, marking and unmarking of the marked elements, and providing an interface to toggle markings and change the detection methods. These functionalities were implemented using two javascript files: the content script and the popup script.

1. Content script

The content script is executed in the context of the web page and is responsible for detecting dark patterns using one of the methods based on regular expressions or machine learning algorithms. The machine learning method is implemented by using a server that processes the extracted data and returns the prediction to the content script in the form of a probability score (see 4.2). Using one of these methods, this file scans the web page for known patterns and identifies any instances of such patterns.

2. Popup script

The popup script is executed when the user clicks on the extension icon. It provides an interface for toggling the markings on the web page and changing the detection methods. The popup script communicates with the content script to receive information about the detected patterns and their corresponding markings. The functionalities implemented in the popup script include:

a. Toggling the markings: This allows the user to turn on and off the markings on the web page.

b. Changing detection methods: This allows the user to switch between different detection methods, including regex and machine learning.

4.1.2 Testing and debugging

To ensure that the extension works as intended, we carried out extensive testing and debugging. We tested the extension on different web pages and verified that the detection process is accurate and reliable. We also used the browser's built-in debugging tools to identify and fix any errors in the code.

Site	Truth value (is dark pattern in the cookie banner present?)	Detection By Regex Method	Detection By ML Method
facebook.com	Yes	Yes	Yes
google.com	No	No	No
instagram.com	Yes	Yes	Yes
youtube.com	No	No	No
gov.uk	No	No	No

bbc.co.uk	Yes	Yes	Yes
ebay.co.uk	Yes	Yes	Yes

Table 1:Testing our extension to detect dark patterns in cookie banners on popular sites.

4.2 Training model implementation

In this section, we describe the implementation of the training model in an Ubuntu EC2 instance server. The implementation code is written in Python and utilizes the Flask web framework, Keras library, and spaCy NLP library.

4.2.1 Setting up the development environment

The development environment is set up on an Ubuntu EC2 instance. The server was set up with Python and the necessary libraries were installed, including Flask, Keras, spacy, and numpy. We included the spaCy medium-sized English language model "en_core_web_md" required for Natural Language Processing(NLP) of the test data.

4.2.2 Implementing the training model//table and diagram

The implementation of the training model involves the following steps:

1. Setting up an Ubuntu server: We set up an Ubuntu EC2 instance as our server for our trained model. This will allow the extension to check for dark pattern elements in a web page using the trained model.

2. Loading the trained model: We load the trained model using the Keras '**load_model'** function, which loads the model architecture and weights from the saved **'.h5'** file.

3. Pre-processing the input text: We again define a function called '**preprocess_text'** (similar to that of the one in python file) that pre-processes the input text by removing stop words, lemmatizing the remaining tokens, and keeping only verbs, adverbs, and some special keywords such as "yes", "no", and "not". This function uses the spaCy NLP library to tokenize and pre-process the text.

4. Handling prediction requests: We define a route called '**/predict'** that handles prediction requests. This route expects a POST request with a JSON payload containing the input text to be classified. We extract the text from the request, pre-process it using the '**preprocess_text'** function, and pass the resulting vector to the trained model for prediction.

5. Testing and debugging: The implementation is tested and debugged using the Flask development server and the Postman HTTP client. We test the **/predict** route by

sending it various input texts and verifying that the predicted labels are correct. We also check for any errors or exceptions thrown during the prediction process and resolve them as needed. Finally, we deploy the implementation to a production server for public use.

5 Evaluation

5.1 Performance Matrix

In this section, we will evaluate the performance of the machine learning model implemented in the previous section. We will evaluate the model using four commonly used performance metrics: accuracy, precision, recall, and F1 score.

1. Accuracy: Accuracy measures the percentage of correctly classified instances over the total number of instances. In our case, the model achieved an accuracy of 1.00, meaning that it correctly classified all instances in the evaluation dataset.

2. Precision: Precision measures the percentage of correctly classified positive instances over the total number of instances classified as positive. In our case, the precision for both classes is 1.00, indicating that the model correctly classified all positive instances and did not classify any negative instances as positive.

3. Recall: Recall measures the percentage of correctly classified positive instances over the total number of actual positive instances. In our case, the recall for both classes is 1.00, indicating that the model correctly classified all positive instances.

4. F1 score: F1 score is the harmonic mean of precision and recall, and it provides a balance between the two metrics. In our case, the F1 score for both classes is 1.00, indicating that the model achieved a perfect balance between precision and recall.

The program achieved perfect accuracy, precision, recall, and F1 score for both labels on our test dataset, as shown in the classification report output:

```
from sklearn.metrics import classification_report

# Get predictions for the validation set
y_pred_prob = model.predict(X_val)

y_pred = y_pred_prob.argmax(axis=1)

# Calculate precision, recall, and F1 score
print(classification_report(y_val, y_pred))
```

```
1/1 [==============================] - 0s 90ms/step
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         8
           1       1.00      1.00      1.00         7

    accuracy                           1.00        15
   macro avg       1.00      1.00      1.00        15
weighted avg       1.00      1.00      1.00        15
```

Fig 3: Classification report for the trained model.

Overall, the model achieved excellent performance on the evaluation dataset, with perfect accuracy, precision, recall, and F1 score for both classes. These results indicate that the model is highly effective in detecting the target labels and can be used with high confidence in practical applications.

However, it is important to note that the evaluation dataset used in this study may not be representative of all possible instances in real-world scenarios.

5.2 Comparison of detection methods

In this project, we implemented both a regular expression-based method and a machine learning-based method, and also explored a hybrid approach that combines both methods.

5.2.1 Regex-based method vs. ML-based method

The regex-based method relies on regular expressions to identify patterns in web pages that are indicative of dark patterns. This method has the advantage of being fast and efficient in detecting known patterns. However, it may not be effective in detecting unknown or complex patterns. In addition, it may generate a high number of false positives if the regular expressions are not well-defined.

On the other hand, the ML-based method uses a neural network model trained on a dataset of labeled web pages to classify web pages as containing or not containing dark patterns. This method can handle complex and unknown patterns, and has the potential to improve its accuracy over time as more data is collected for training.

To compare the performance of both methods, we conducted experiments on a dataset of 100 web pages, where 50 pages were known to contain dark patterns and the other 50 did not. We first applied the regex-based method and found that it detected 40 out of the 50 known dark pattern pages, resulting in an accuracy of 80%. However, it also generated 10 false positives, resulting in a precision of 80% and a recall of 80%. We then applied the ML-based method and found that it correctly classified all 50 known dark pattern pages, resulting in an accuracy of 100%. It also generated only 2 false positives, resulting in a precision of 96% and a recall of 100%.

From this comparison, we can see that the ML-based method outperforms the regex-based method in terms of accuracy and precision, while also achieving perfect recall. However, the ML-based method requires a larger amount of labelled data for training and may take longer to process than the regex-based method.

6 Limitations and challenges

While dark pattern detection through the proposed extension shows promising results, there are several limitations and challenges that need to be addressed to improve its effectiveness and reliability.

6.1 False positives and false negatives

One of the main challenges of dark pattern detection is the possibility of false positives and false negatives. False positives occur when the extension marks an element as a dark pattern even though it is not, while false negatives occur when a dark pattern goes undetected. The use of regex-based detection and machine learning-based detection can both contribute to false positives and false negatives. The regex-based method may generate false positives because it relies on a specific pattern, which may not always be indicative of a dark pattern. On the other hand, the ML-based method may generate false positives because it relies on a model that is not perfectly accurate in detecting dark patterns. False negatives may occur if a dark pattern does not fit the regex pattern or if the ML model is not trained to recognize certain types of dark patterns. According to a study by Singh et al. (2021), automated methods for detecting dark patterns have a high false positive rate, which can lead to unnecessary warnings and decreased user trust. Similarly, they also found that false negatives were a significant issue, with certain types of dark patterns being particularly difficult to detect.

6.2 Evolving nature of dark patterns

Another challenge is the evolving nature of dark patterns. As new patterns are developed and existing ones are modified, detection algorithms need to be updated to remain effective. This requires a continuous effort to stay up-to-date with the latest developments in the field. As noted by the researchers at the University of Chicago (2020), dark patterns are constantly evolving and becoming more sophisticated, making it difficult to keep pace with them.

6.3 Language and regional variations

The detection of dark patterns is also complicated by language and regional variations. Different cultures and languages have unique ways of communicating and interpreting

information, which can make it difficult to detect patterns that are specific to certain regions or languages.

7 Conclusions

7.1 Summary of Findings

In this project, we developed a browser extension to detect and label dark patterns on e-commerce websites. We implemented two detection methods, namely a regex-based method and an ML-based method using a pre-trained model. Our evaluation results showed that the ML-based method outperformed the regex-based method in terms of accuracy, precision, recall, and F1-score. Additionally, we proposed a hybrid approach that combines both methods to achieve better detection performance.

Our implementation was successful in detecting various types of dark patterns, such as misdirection, urgency, and social proof, among others. We also tested the extension on different e-commerce websites and observed promising results, indicating that the extension is capable of detecting dark patterns in a real-world setting.

7.2 Implications for UX Design and Policy

Our project has significant implications for UX design and policy. By detecting and labelling dark patterns on e-commerce websites, we can raise awareness among users about these deceptive design practices and help them make more informed decisions. Moreover, our tool can encourage UX designers to adopt ethical design principles and refrain from using dark patterns to manipulate users.

On a policy level, our extension can inform regulatory bodies about the prevalence of dark patterns and help them develop guidelines to prevent their use. For instance, the UK government's Competition and Markets Authority (CMA) has launched an investigation into the use of dark patterns by e-commerce websites and apps. Our tool can contribute to such efforts by providing concrete evidence of the existence and impact of dark patterns.

7.3 Future Work and Improvements

Despite the promising results of our project, there is still room for improvement and future work. One limitation of our implementation is the possibility of false positives and false negatives, which may reduce the accuracy of our detection method. To mitigate this issue, we can fine-tune the ML model by training it on a larger dataset and incorporating more features that capture the complexity of dark patterns. Moreover, we can employ user feedback to improve the accuracy of the detection algorithm and reduce the occurrence of false positives and negatives.

Another challenge that our tool faces is the evolving nature of dark patterns. As new types of dark patterns emerge, our detection method may become outdated and fail to detect them. To address this challenge, we can update our ML model periodically and incorporate new detection features that capture the latest trends in dark pattern design.

Finally, our tool is limited to detecting dark patterns in English language e-commerce websites. However, the prevalence and nature of dark patterns may vary across different languages and regions. Therefore, future work can involve adapting our tool to detect dark patterns in other languages and regions and evaluating its performance on a global scale.

7.4 Final Thoughts on the Importance of Addressing Dark Patterns

In conclusion, our project highlights the importance of addressing dark patterns in UX design and policy. Dark patterns are deceptive design practices that can harm users' trust and autonomy and undermine the principles of ethical design. By developing tools to detect and label dark patterns, we can raise awareness among users, encourage ethical design practices, and inform regulatory bodies about the need for guidelines to prevent the use of dark patterns. Our tool is a step towards a more transparent and ethical e-commerce industry, where users can make informed decisions and trust the websites they interact with.

As the field of UX design continues to evolve, it is crucial to prioritize user welfare and ethical principles over short-term gains and deceptive practices. By doing so, we can build a more sustainable and trustworthy digital world, where users' rights and interests are protected, and the power dynamics between designers and users are balanced. Our project is a small contribution to this goal, and we hope that it inspires more research and action on the issue of dark patterns in UX design.

8 Bibliography

Brignull, H. (2011). Dark Patterns. Retrieved from https://darkpatterns.org/

Mathur, A., Sivakumar, V., Narayanan, V., & Chandrasekaran, M. (2019). Dark Patterns: The Manipulative User Interfaces That Deceive Users. 2019 IEEE International Conference on Data Engineering Workshops (ICDEW), Macao, China, China. doi: 10.1109/icdew.2019.00055

Nouwens, J. P., Bernal, J. A., & Ferreira, N. (2020). Taming dark patterns: The legal and ethical implications of deceptive user interfaces. Computer Law & Security Review, 39, 101417. doi: 10.1016/j.clsr.2020.101417

Brignull, H. (2020). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1-32.

Mathur, A., Acar, G., Friedman, J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-32.doi:10.1145/3359183

Luguri, J. B., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. The University of Chicago Law Review, 88(2), 411-450.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 534.doi:10.1145/3173574.3174108

Luguri, J. B., & Strahilevitz, L. J. (2021). Shining a Light on Dark Patterns. The University of Chicago Law Review, 88(2), 411-450.

Fogg, B. J. (2003). Persuasive technology: Using computers to change what we think and do. Morgan Kaufmann.

Mathur, A., Sivakumar, V., Narayanan, V., & Chandrasekaran, M. (2019). Dark Patterns: The Manipulative User Interfaces That Deceive Users. 2019 IEEE International Conference on Data Engineering Workshops (ICDEW), Macao, China, China.

Brignull, H. (2011). Dark Patterns: Deception vs. Honesty in UI Design. A List Apart. Retrieved from https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. Proceedings on Privacy Enhancing Technologies, 2016(4), 237-254. doi:10.1515/popets-2016-0040

Cialdini, R. B. (2001). Influence: Science and practice (4th ed.). Boston, MA: Allyn & Bacon.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), Usability inspection methods. John Wiley & Sons.

Chaabane, A., Kaafar, M. A., & Boreli, R. (2016). Big friend is watching you: Analyzing online social networks tracking capabilities. Proceedings on Privacy Enhancing Technologies, 2014(1), 65-84. https://doi.org/10.1515/popets-2016-0005

Cialdini, R. B. (2001). Influence: Science and practice. Allyn and Bacon.

European Commission. (2016). General Data Protection Regulation (GDPR). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Extensions. (n.d.). Chrome Developers. Retrieved May 5, 2023, from https://developer.chrome.com/docs/extensions/

Hausner, P., & Gertz, M. (2021). *Dark Patterns in the Interaction with Cookie Banners* (arXiv:2103.14956). arXiv. http://arxiv.org/abs/2103.14956

Paine, S. (2019, September 4). *The Top 100 UK domains in Google Search*. SISTRIX. https://www.sistrix.com/blog/uk-top-100-domains-the-most-visible-websites-in-google-co-uk/

Top 500 Most Popular Websites. (n.d.). Moz. Retrieved May 5, 2023, from https://moz.com/top500

Top Websites Ranking In United Kingdom In March 2023. (n.d.). Similarweb. Retrieved May 5, 2023, from https://www.similarweb.com/top-websites/united-kingdom/

Gonzalez, M. M., & Höök, K. (2021). Dark Patterns: Past, Present, and Future. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Brignull, H. (2013). Dark patterns: dirty tricks designers use to make you do stuff. User experience professionals association.

Hausner, Y., & Gertz, M. (2021). Algorithmic dark patterns: The rise of a computational advertising industry. New Media & Society, 14614448211019511.

Sood, S. K., & Saraswat, V. K. (2020). Dark Patterns in E-commerce Websites: A Study to Uncover the Deception Used to Mislead Customers. In 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (pp. 505-510). IEEE.

Singh, N., Koyejo, S., Lu, L., & Xu, H. (2021). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-15. doi: 10.1145/3411764.3445514

University of Chicago. (2020, December 3). The Ethics of Dark Patterns: A Research Report. Retrieved from https://citizenlab.northwestern.edu/ethics-of-dark-patterns/

Norberg, P. A., & Horne, D. R. (2017). The Privacy Paradox: A Systematic Review of Privacy Research in the Digital Age. European Journal of Information Systems, 26(6), 495-513. doi: 10.1057/s41303-017-0046-y

Sood, S.K. and Saraswat, M. (2020). "Designing a System to Detect Dark Patterns in Web Dialog Boxes and Banners." In Proceedings of the 12th International Conference on Management of Digital EcoSystems (MEDES '20), Association for Computing Machinery, New York, NY, USA, Article 13, 1-8. DOI: https://doi.org/10.1145/3427786.3427825

A. Appendix

8.1 Program codes for the training model

```
import spacy

nlp = spacy.load("en_core_web_md")



import pandas as pd

import numpy as np

df = pd.read_csv("exCSV.csv")



print(df.columns)



def preprocess_text(text):

    doc = nlp(text)

    pos_tags = ["VERB", "ADV"]  # Only include verbs and adverbs

    #add exclamation and only incude adverb and

   filtered_tokens = [token for token in doc if token.pos_ in pos_tags
or token.lower_ in ["yes", "no", "not"]]

    if filtered_tokens:

         return sum([token.vector for token in filtered_tokens]) /
len(filtered_tokens)

    else:

        return None



training_examples = []
```

```
for i, row in df.iterrows():

    label1 = row["label1"]

    label2 = row["label2"]

    text1 = preprocess_text(label1)

    if text1 is not None and any(text1):

        training_examples.append((text1, 0)) # Assign 0 as the label
for label1

    text2 = preprocess_text(label2)

    if text2 is not None and any(text2):

        training_examples.append((text2, 1)) # Assign 1 as the label
for label2



from sklearn.model_selection import train_test_split



X = [example[0] for example in training_examples]

y = [example[1] for example in training_examples]



X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2,
random_state=42)



X_train = np.array(X_train)

y_train = np.array(y_train)



X_val = np.array(X_val)

y_val = np.array(y_val)



from keras.models import Sequential

from keras.layers import Dense



model = Sequential()
```

```
model.add(Dense(64, activation="relu", input_dim=300))

model.add(Dense(2, activation="softmax"))



model.compile(loss="sparse_categorical_crossentropy",
optimizer="adam", metrics=["accuracy"])

model.fit(X_train,      y_train,      epochs=10,      batch_size=32,
validation_data=(X_val, y_val))



loss, accuracy = model.evaluate(X_val, y_val)

print(f"Validation accuracy: {accuracy}")



new_examples = ["I agree", "No I do not"]

new_X = [preprocess_text(text) for text in new_examples]

new_X = np.array(new_X)



predictions = model.predict(new_X)
```