The Influence of Dark Patterns on User Behavior: Evaluating Social Media Design Choices

CS4991 Capstone Report, 2025

Jennifer Vo
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
phb8pt@virginia.edu

ABSTRACT

Users turn to social media for connection and to access information, but these platforms employ dark patterns, prioritizing business interests, highlighting the urgent need for greater awareness in platform design. To address this, I evaluated the ten most popular social media platforms in the US for features that influence users' actions and analyzed the comments on YouTube videos to find a correlation between emotion and engagement. Specifically, I examined HCI elements that undermine user agency through emotion or misdirection. Additionally, I applied sentiment analysis to the comments on popular Youtube videos and compared the sentiment score to engagement metrics. The results revealed that X (formerly Twitter) employed the most dark patterns. Additionally, Youtube videos that sparked negative emotion had more overall engagement, whereas neutral videos had less engagement. I encourage other computer scientists to perform sentiment analysis across various platforms, particularly on polarizing content as breaking news unfolds in real time.

1. INTRODUCTION

Internet usage has become nearly universal in the U.S., rising from just over half of adults in 2000 to 95% today (Pew Research Center, 2024). As digital connectivity increases, so do concerns about how online platforms shape behavior. Research has linked excessive screen time to lower physical activity, poor sleep quality, and higher stress (Vizcaino et al., 2020). Social media platforms provide community and information but are also engineered to maximize engagement—often at the expense of user autonomy. Exploitative design, or "dark patterns," subtly manipulate attention and decision-making that ultimately undermine user autonomy for commercial gain. As Kozyreva et al. (2020) highlight, these techniques include misleading information, behavioral nudges, and the commodification of attention. Because interactions vary by individual psychology, regulation remains difficult. The central problem, however, is that most users remain largely unaware of these underlying techniques and their potential effects, often engaging with platforms from a place of trust rather than skepticism, leaving users vulnerable to exploitation.

2. RELATED WORKS

Numerous studies have examined dark patterns and their impact on user behavior in online environments. Kozyreva et al. (2020) categorize these challenges into four major types: persuasive and manipulative choice, AI-assisted information, false and misleading information, and distracting environments. These manipulative design strategies shape user decision-making, often reinforcing digital dependency. In response to these challenges, they explore behavioral and cognitive interventions as a means to enhance user

autonomy. In a related analysis, Schaffner et al. investigated account deletion processes across the top 20 social media platforms in the United States (2022). Their findings highlight significant variations in account deletion options and reveal that the language used to describe these options is often unclear, further complicating user agency in managing their online presence.

Beyond dark patterns, research has explored user engagement and motivation in online spaces. Khan (2017) examined user interactions with YouTube and identified five key motivations for content consumption and engagement: information seeking, giving information, self-status seeking, social interaction, and relaxing entertainment. These motivations extend beyond YouTube, encapsulating broader reasons users gravitate toward digital platforms. By analyzing how specific UI elements leverage these motivations, we can better understand their role in shaping user behavior and fostering engagement.

3. PROJECT DESIGN

I selected the ten most widely used social media platforms in the United States based on Pew Research Center's latest report. YouTube ranked highest, with 80% of respondents—ages 18 and older—reporting that they have used the platform. From this list, I analyzed platform features related to urgency, misdirection, social proof and obstruction, building on the Kozyreva, et al. (2020) categorization of dark patterns.

Urgency examines platform notifications and language designed to create a sense of immediacy, such as countdowns or time-sensitive prompts. Misdirection focuses on how platforms steer users away from certain choices through emotional appeals, visual design, or ambiguous language. Social proof explores how user behavior is influenced by cues highlighting the experiences and actions of others. Lastly, obstruction assesses barriers such as mandatory account creation and deliberately complex account cancellation processes.

For each of the ten social media platforms analyzed, I assessed the presence of each feature, using a scale from 0 to 3, where 0 indicates the feature is not implemented, 1 means it is minimally implemented, and 2 signifies that it is frequently used or fully integrated. These social media platforms were used in the desktop and mobile version.

In addition to evaluating the presence of dark patterns, I extracted textual content from various social media platforms to analyze language patterns that may influence user behavior. I compiled this text into a structured dataset, pre-processing it by tokenizing words. To capture manipulative language, I extracted key linguistic features, focusing on directive verbs that influence user behavior. These include urgency verbs (hurry, act now); commitment-driven verbs (subscribe, guarantee); social influence cues (join, discover); emotionally charged words (shocking, unbelievable); and authority-based phrases (proven, experts say).

Last, I selected the top ten trending YouTube videos and extracted the top 100 comments for each using YouTube's API. The top comments refer to the first comment in a thread, rather than replies. To analyze engagement patterns, I pre-processed the data by tokenizing the text, removing stop words, and normalizing the words through lemmatization. I then applied sentiment analysis using a pre-trained model, VADER (Valence Aware Dictionary and sentiment Reasoner), to classify comments as positive, negative or neutral. Finally, I examined the relationship between sentiment polarity and engagement metrics, such as the number of likes, replies and overall comment

visibility, to understand how sentiment influences audience interaction.

4. RESULTS

The analysis revealed notable differences in the prevalence of dark patterns across the ten social media platforms, with X having the highest dark pattern score at 2.6, while Pinterest had the lowest at 1.3. Across platforms, the most commonly employed strategy involved using style and visual presentation to manipulate user decisions, often through tactics such as infinite scrolling, the use of small or hidden text in terms and conditions and the strategic contrast of large buttons. These techniques effectively guided users toward specific actions or prevented them from noticing alternatives, reinforcing manipulative design choices across multiple platforms.

Notifications, particularly those related to urgency and commitment to using the platform, were prevalent across various social media platforms. Urgency-driven verbs, such as hurry and see now, were frequently employed in notifications and prompts, especially in time-sensitive contexts, like a friend posting content. Interestingly, emotionally charged words, such as sorry or sad, were predominantly utilized in features related to account deletion or irreversible actions, reflecting the emotional appeal often used to deter users from completing these tasks.

Sentiment analysis of YouTube comments revealed that negative comments received the highest engagement, with an average of 30% more replies compared to neutral or positive comments. Additionally, negative comments, while less frequent, were more likely to generate discussion, often forming long reply chains. The data suggest that emotionally charged or opinion-driven comments tend to attract higher interaction, aligning with platform design strategies that amplify engagement. This supports the idea that both platform features and language patterns work in unison to shape user behavior and sustain digital engagement.

5. CONCLUSION

This study reveals how social media platforms employ dark patterns to subtly manipulate user behavior, with X exhibiting the highest prevalence of these tactics. Features related to urgency, misdirection, social proof, and obstruction were found across all ten platforms, reinforcing engagement-driven design choices that prioritize platform retention over user autonomy. Additionally, sentiment analysis of YouTube comments showed that negative sentiment correlates with higher engagement, suggesting that emotionally charged content fuels discussion and interaction. These findings support the argument that both platform design and persuasive language influence user behavior in ways that often go unnoticed.

A major challenge in studying manipulative design is the lack of structured datasets on manipulative language. While sentiment analysis detects emotional tone, it does not capture the persuasive techniques embedded in social media interactions. This study addresses that gap by identifying key linguistic patterns—such as urgency-driven verbs, social influence cues, and emotionally charged phrasing—offering a more systematic approach to analyzing manipulative language. Unlike traditional sentiment analysis, which broadly categorizes emotions, this project focuses on persuasive strategies that shape user decision-making. By contributing to the development of structured datasets for manipulative text, this research lays the groundwork for more advanced detection methods and a deeper understanding of how language influences digital engagement.

6. FUTURE WORK

Future research could extend sentiment analysis across multiple platforms to determine whether engagement patterns observed on YouTube hold true elsewhere. Analyzing polarizing content in real-time events, such as breaking news, could further reveal how social media amplifies emotional responses. Additionally, refining the linguistic dataset and leveraging machine learning to detect manipulative language and dark patterns at scale could support regulatory efforts and ethical design standards. By deepening our understanding of the intersection between interface design, persuasive language, and engagement metrics, this work contributes to a more transparent and user-conscious digital landscape.

REFERENCES

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Pew Research Center (2024, January 31). *Americans' use of mobile technology and home broadband* [Report]. https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/

Schaffner, B., Lingareddy, N., & Chetty, M. (2022). Understanding account deletion and relevant dark patterns on social media. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1-43.

Vizcaino, M., Buman, M., DesRoches, T., & Wharton, C. (2020). From TVs to tablets: The relation between device-specific screen time and health-related behaviors and characteristics. *BMC Public Health*, *20*(1), 1295. https://doi.org/10.1186/s12889-020-09410-0