ANALYSIS OF SOCIAL MEDIA IMPACT ON EATING DISORDERS: SPECIAL FOCUS ON YOUTUBE

Kenza Battah

December 6, 2023

Abstract

This study investigates the impact of recommendation systems on information dissemination regarding eating disorders, with a specific emphasis on YouTube. The internet has revolutionized information accessibility, enabling widespread sharing of ideas. Indeed, alarming associations between social media use and disordered eating in young adolescents have been identified by Wilksch et al. (2020). Hence, by delving into the complex relationship between YouTube's recommendation systems and the spread of content related to eating disorders (EDs), we aim at highlighting the platform's dualistic role in shaping narratives. The research integrates various methodologies, including web scraping, analysis of recommendation algorithms, and survey experiments, to examine how YouTube may inadvertently foster environments conducive to pro-eating disorder (pro-ED) and anti-recovery content dissemination. A significant focus is placed on understanding the behavioral patterns and perceptions of users aged 19 to 32, a demographic particularly vulnerable to disordered eating and body image concerns. The study reveals that user engagement with YouTube, especially the frequency of use and interaction with the platform's recommendations, is closely linked to their perceptions and behaviors related to eating disorders and body image. Machine learning models, including Naive Bayes and LSTM, were employed to classify content into pro-ED and anti-eating disorder (con-ED) categories. The results indicate a nuanced complexity in text classification for such social issues, with simpler models outperforming more complex ones, underscoring the need for finer model tuning and larger datasets. The study also proposes the implementation of advanced text classification systems using the Snorkel framework to enhance context-aware content moderation. This research contributes to the understanding of social media's impact on mental health and underscores the importance of informed policy-making in the digital landscape.

Keywords— implicit bias, web scraping, recommendation system analysis, classification, YouTube, eating disorders, disordered eating, social media

Contents

1 Introduction

Eating Disorders (EDs), marked by the highest mortality rate among psychiatric illnesses, affect around 9% of the global population. The concerning rise of pro-Eating-Disorder (pro-ED) online communities, promoting unhealthy weight loss practices over the past decade, has gained even more significance during the COVID-19 pandemic. Indeed, this period witnessed an unprecedented increase in social media activity focusing on fitness and diet, paralleling a surge in eating disorder diagnoses (Toulany et al., 2023). While pandemic-related stress was the main factor, the greater exposure to content that encourages eating disorders through social media also contributed to this increase (Hensley, 2020). This has prompted platforms like YouTube to enforce policies that restrict content promoting unhealthy body images and eating habits, reflecting a heightened commitment to well-being (Arcelus and et al., 2011a). However, despite these efforts, users often find ways to circumvent safeguards, leading to the persistence of harmful content (Gerrard, 2018) — and the stark contrast between pro-ED communities and the counteracting pro-recovery or con-Eating-Disorder (con-ED) has hence created a complex digital battleground. These opposing communities actively promote vastly different messages, highlighting a pressing need to investigate the dynamics of this interaction. Thus, this study aims to delve into the role of YouTube's recommendation system, focusing on how it potentially navigates this polarized landscape. We seek to provide the grounds to understand better whether the system inadvertently leads users towards harmful content, thereby contributing to a siloing effect that separates and intensifies these opposing viewpoints. Firstly, we will examine the social dynamics and psychological aspects of EDs, as well as assess the YouTube recommendation system and the potential for algorithmic bias to influence the proliferation of pro-ED content. Then, we will integrate findings from our survey experiment and the application of classification models to critically understand how YouTube's algorithms may impact perceptions and behaviors related to eating disorders.

2 Contextual Overview

2.1 Social Media and EDs

The Social Psychology of EDs Eating disorders are fundamentally mental health disorders and are hence rooted in complex psychological frameworks involving perceptions of food and body image. They are the deadliest mental illnesses, with The National Association of Anorexia Nervosa and Associated Disorders (ANAD) reporting that "5% to 10% of anorexics die within ten years after contracting the disorder. Eighteen to 20% of anorexics will be dead after twenty years, and only 30% to 40% ever fully recover, while 20% bounce in and out of hospitals" (Arcelus and et al., 2011b). EDs frequently assume a crucial role in an individual's identity and coping mechanisms, necessitating a substantive reevaluation of lifestyle and self-concept during the recovery process. Indeed, these disorders can manifest as coping mechanisms in response to trauma, consequently often resulting in heightened proclivities towards social isolation. It is, however, important to note the difference between disordered a clinical ED which is a mental illness — and disordered eating habits — which do not manifest the full range of psychological traits usually associated (interpersonal distrust, isolation, perfectionism, frequent rituals around food, etc...). Hence it is thought that EDs exist along a spectrum — and many individuals, mostly women, engage in behaviors such as rigid exercise or calorie restriction influenced by cultural thinness mandates but do not display the full psychological profile of clinical EDs.

Therefore, what seems to make these illnesses so potent and recovery so complex is that EDs are rooted in societal norms, and symptomatic of a social problem. Gender dynamics and societal values such as beauty standards and diet culture, influence the

psychopathology of these disorders, by promoting food restraint and thinness as values of success. Feminist works study the influence of economic and social institutions that profit from the "cult of thinness" promoted by the mass media — such as diet culture, advertising, fitness, fashion, and cosmetic industries (Arcelus and et al., 2011b; Fallon et al., 1996; LaMarre et al., 2022). According to (Arcelus and et al., 2011b), the lucrative market of selling a beauty ideal classifies food and behaviors as moral values ("good" vs. "bad" or "healthy" vs. "unhealthy") which ultimately 1) makes individuals question their food choices, 2) perpetuates the belief that losing weight or achieving thinness will lead to happiness — hence capitalizing on insecurities to sell.

Moreover, the diagnosis and treatment of eating disorders face challenges due to historical biases in research and clinical paradigms, predominantly focusing on white, cisgender, heterosexual individuals (Root, 1990). For instance, the Diagnostic and Statistical Manual of Mental Disorders (DSM) has been subject to scrutiny for its historical lack of diversity in research samples, which potentially compromises diagnosis and treatment for sufferers who might not exhibit specific patterns/symptoms such as the ones highlighted in the DSM.

Furthermore, like the illness itself, the recovery from an eating disorder is a multifaceted and intricate process, characterized by a complex interplay of psychological, social, and physical factors. Recovery from eating disorders is non-linear, often marked by cycles of relapse and distress, shaped by sociocultural influences like wellness, diet, and fitness cultures. These cultural phenomena drive distorted body ideals and eating habits, highlighting the need for a nuanced understanding of recovery in the context of societal norms and social media's role in shaping body image perceptions.

All in all, examining the relationship between social media and eating disorders becomes crucial. Social media platforms play a pivotal role in shaping societal ideals and influencing individuals' perceptions of body image and/or dietary practices — and understanding this dynamic is essential in comprehending the multifaceted nature of eating disorder recovery in this digital age.

2.1.1 ED-centered online communities and echo chambers

YouTube, TikTok, Twitter, and Instagram have become fertile grounds for eating disorder communities, where echo chambers that amplify certain trends about food habits and body image can be found. Echo chambers are environments where users are only exposed to information and opinions that reflect their own, thus reinforcing them. They are fertile ground for misinformation and can distort users' perspectives so they have difficulty considering opposing viewpoints — this phenomenon is known as 'siloing'. We will provide an overview of firstly the pro-ED communities who promote EDs as a lifestyle choice more than a condition and then explore the dynamics of pro-recovery communities that seek to educate and support recovery. We contend that pro-ED communities work as echo chambers that scaffold pro-ED beliefs and practices and generate a trust barrier that can undermine therapeutic intervention.

Pro-ED echo chambers Pro-ED is a label used to define online communities aiming which actively promote negative ideals related to beauty standards, thinness, and misinformation about eating habits. Unlike the unintentional exposure to harmful content social media sustains, these communities deliberately perpetuate damaging notions about body image and eating habits (Giles, 2006). They take various forms, including static websites, blogs, forums, and communities on many social media platforms. Mainly, the controversy surrounding these websites stems from their "anti-recovery" stance, which, in its most extreme form, rejects the notion of anorexia (and sometimes bulimia) as illnesses, framing them as deliberate lifestyle choices (Strife and Rickard, 2011; Mulveen and Hepworth, 2006). This view reinforces the misinformed popular opinion which frequently associates EDs as lifestyles instead of illnesses.

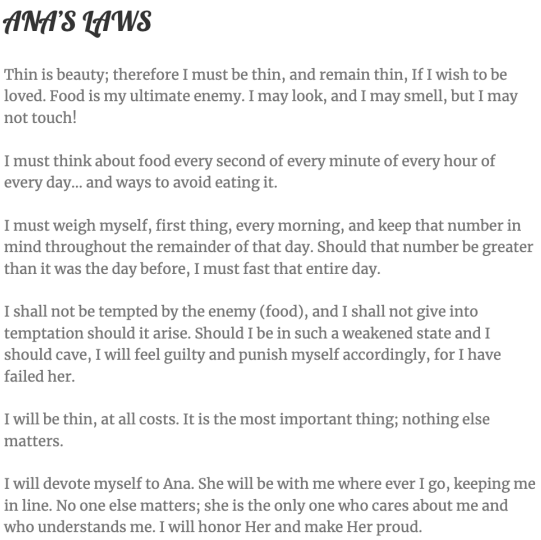

Figure 1: Illustration of "Ana's Laws": Common pro-Ana commandments shared within online communities.

Pro-ED communities are grounds for people to learn how to engage "successfully" in an ED. By sharing "tips" or "how to" content, which exchange technical weight-loss tips and strategies to conceal disordered eating, users learn to conceptualize their illness as a practice of self-discipline. Moreover, by disavowing entirely the associated health risks, and putting thinness as a worthy result, sufferers in these spaces enter a complete readjustment of seeing foods and bodies as moral values for themselves ("*If I do not stick to my restrictive diet, it is because I am lazy and "gave in*"). The illness becomes an identity, a certificate of authenticity, and proving this authenticity is an important part of being accepted as a member of many such communities. Terms such as "*wannarexic*" define members who raise doubts about being genuinely anorexic — which is often treated as an insult and can serve as a reason for banning individuals from the community (Boero and Pascoe, 2012). Media-rich content is used to provide "*thinspiration*" (or 'thinspo') images featuring extremely thin women that many users look up to for inspiration in pursuing their weight-related goals (Alberga et al., 2018). Some pro-Ana (pro-anorexia) sites/communities used to include a set core of belief, called an "Ana Creed" or "Ana Laws" — which served as "commandments" setting out the values of the group, usually the value of invoking self-control and discipline, keeping thin and losing weight, amongst other prescriptions (Stinson, 2019; Crowe and Watts, 2016).

More violent practices include the use of "*meanspiration*" (mostly on the blog-based social media Tumblr) consisting of requesting and giving mean messages to inspire users to lose weight.

Figure 2: Example of a X/Twitter user posting 'meanspiration' posts.

Hence, pro-ED communities, particularly exemplified by Pro-Ana groups, exhibit characteristics indicative of echo chambers. In this context, an echo chamber is identified as a social epistemic structure intentionally excluding dissenting voices, and demanding a general agreement with a core set of beliefs for membership. Osler and Krueger (2022) contends that all major properties of echo chambers are exemplified in Pro-ED groups. Firstly, by actively supporting and validating EDs as a lifestyle choice, and disavowing medical and scientific facts about the health risks of EDs, these groups successfully disseminate information on disordered eating behaviors. Moreover, pro-ED identities are created around a shared belief system, through the commandments, creeds, and prayers, as well as explicitly stating their exclusionary nature — this creates a need for an "all-or-nothing" commitment to the "ED lifestyle", and reinforces the isolation of members to opposing thoughts and opinions. This fosters an environment where dissenting opinions are not only unheard but actively suppressed — aligning entirely with the systematic mechanisms of an echo chamber. Indeed, being part of a Pro-ED group inoculates members not only from conceptualizing their need for medical treatment but also from friends and family members who could urge them to stop engaging in harmful and potentially fatal behaviors. Following the pro-ED rhetoric, outsiders just do not understand and are merely trying to hinder the individual's discipline and self-control. Resisting this outside influence is hence seen as a demonstration of strength and authenticity, which in turn reinforces commitment — an echo chamber phenomenon called disagreement reinforcement. This cyclic harmful reinforcement, as well as this shielding from any contradicting opinion, creates a bubble of thought where members are confirmed in their sustaining of harmful eating practices which become part of their own identity, deeply embedded in their cognitive frameworks.

This is, however, not to say all pro-ED communities can qualify as echo chambers, as they do not all employ these techniques, nor do they promote the same content. Indeed, there is a diversity of stances even in pro-ED groups. Firstly, about the type of illness that is encouraged — pro-bulimia (pro-mia) or pro-anorexia (pro-ana) (Falconberry, 2022). Additionally, perspectives on the illnesses themselves can vary. Some pro-ana groups view anorexia and bulimia as diseases and are critical of the lifestyle-choice approach, though the distinction between lifestyle and disease can sometimes blur — for instance, these groups frequently use warnings about dangerous practices such as the use of ipecac [1] and advice on how to be anorexic "safely" (Csipke and Horne, 2007). Many of these groups emphasize providing a non-judgmental space where individuals struggling with eating disorders —not yet ready for recovery or treatment — can connect and offer mutual support while promoting safer management (Mulveen and Hepworth, 2006). These groups primarily form around the need

[1] Ipecac syrup was once a common household remedy for situations where it was necessary to remove harmful substances from the stomach.

for connection and a sense of belonging. Members of these groups encourage healthier weight-loss and eating practices while discouraging dangerous behaviors (for instance, laxative abuse and purging) and even offer support when someone expresses a desire to seek recovery.

Hence, pro-ED groups can also offer non-judgmental social and emotional support that attracts many individuals. However, from a medical perspective, these groups are problematic as they often endorse the desire not to recover from eating disorders, which might normalize harmful behaviors and discourage sufferers from seeking professional help (Csipke and Horne, 2007).

Pro-Recovery community On the opposite side of the spectrum, Pro-recovery communities can be found — centered around sharing narratives of recovery, giving and finding support on the challenges of suffering and recovering from EDs. Greene et al. (2023) found the most common themes of pro-recovery content to be (1) the centrality of food to eating disorders and recovery, (2) what eating disorders look and feel like, (3) recovery as a process, (4) getting and giving help, and (5) negotiating diet culture in recovery. Another role endorsed by these communities by some members is to attempt to permeate pro-ED groups to educate them about the health risks of EDs, through the use of common pro-ED tags. The main role of these spaces is to provide social support in deconstructing the narrative of EDs as an identity.

There is significant proof that pro-recovery communities can be very beneficial for ED sufferers to rethink their illness in a less-identity coupled and more informed way, and to learn how to rethink their self-perception. Users usually engage in information exchange, seek recognition, and share their experiences as empowering processes. Aardoom et al. (2014) found that the most notable positive outcome users get from engaging in these communities was an enhanced sense of being well-informed. Additionally, some lesser degrees of reported outcomes included heightened tendencies to seek assistance, increased optimism and perceived control over the future, and enhanced confidence in both the treatment process and the therapeutic relationship. However, many sufferers engaging with pro-recovery content also revealed a major caveat of those spaces: the thin line between social support and social comparison. Indeed, it is a well-supported fact that social media act as significant arenas for comparisons, where users engage in upward social comparisons with inspirational bodies, potentially leading to body dissatisfaction — which can be crucial in promoting unhealthy body goals and other pro-ED content and hence be bias-confirming for users suffering from body ideation and negative body image. However, Au and Cosh (2022) about Instagram recovery community users also revealed the negative impact these recovery narratives can generate — namely a sense of competition and potential invalidation of one's progress — which can be crucially detrimental for ED recovering users.

Indeed, users who share their recovery practices on social media with a supportive intent often end up sharing experiences that only align with very linear and "traditional" ways of performing recovery, hence participating actively to the narrow ideas of how recovery can look like (Lamarre et al., 2017). Moreover, recovering users usually also post pictures of food and bodies (often white, thin, cisgender, female bodies) — even though media-rich content can be very sensitive for recovering ED sufferers. Moreover, trends in the pro-recovery communities such as "Before and After" where users are called to showcase their body transformations before and after recovery also join the dominant harmful narrative of what recovery literally ought to look like. These can, in turn, make users that might consider recovery feel like they 1) might not be at a point where recovery should be considered ("*I didn't look/feel as sick as they do, do I need to go back so that I can properly do recovery?*") 2) recovery should be a before/after linear progress instead of a full mental reconstruction of how to think about one's own body and food habits (Nikolova and LaMarre, 2023). Many findings

show that users posting this "*pseudo-recovery*" type of content can indicate that users still currently engage with eating disordered behaviors (Au and Cosh, 2022; Goh et al., 2022) — and that they often present eating specific foods as proof of recovery (usually foods that are relatively popular with affluent Western populations). Pro-recovery communities seem to be a significantly positive tool only when they foster a positive environment, with trauma-informed content sensitive to others' recovery timelines and supportive of the idea of recovery as being non-linear. However, it is clear that this social support has its limits, and many participants in Nikolova and LaMarre (2023) suggested that individuals who felt recovered often felt less compelled to perform and show their recoveries on social media, and usually exited ED-centered communities.

All in all, we have seen that ED-sensitive content and communities exist along a spectrum, and many sufferers will engage in different types of content and with different communities across social media platforms — depending on the broader context of the timeline of their illness, their set of beliefs, mental health resources available to them, amongst plenty of other external and internal factors. Hence, just like social factors are at the root of EDs as an illness, social media seems to be inherently part of the timeline of an individual with their illness. Social media identities around EDs are varied and many ED sufferers hence use social media to experience an important sense of belonging and ease in these online worlds — an ease that often stands in stark contrast to how they might feel in their offline worlds, considering the stigma around ED-related illnesses, or even the sense of isolation EDs push sufferers into. Therefore, one might inquire: to what degree should social media platforms have a responsibility to restrict users from accessing potentially harmful pro-eating Disorder spaces, all while avoiding stigmatization and refraining from depriving users of what might be one of their only available social support systems?

2.1.2 Content moderation

Content moderation on YouTube is a crucial aspect of the platform's operation and involves both algorithmic and human oversight. The former uses advanced machine learning models to scan videos for specific keywords and patterns signaling a violation of the platform's guidelines. These algorithms are supported by manual review from human moderators, who make judgments on content flagged by the system or reported by users.

However, moderating content related to EDs is particularly challenging. Chancellor et al. (2016) found the implementation of moderation by Instagram to have had adverse effects on the platform's strategy to eliminate pro-ED content in the long run. The first significant issue is that triggering content can be concealed within seemingly normal discussions, because of the prevalence of diet culture and beauty standards. This content often promotes unrealistic and harmful body goals, yet it's not explicitly against community guidelines, making it difficult to moderate without overreaching.

Moreover, users promoting EDs or encouraging unhealthy behaviors have developed sophisticated methods to bypass these moderation systems. Chancellor et al. (2016) noticed a rise in the use of varied language and the expression of increased toxic and vulnerable behavior over time as Instagram started moderating pro-ED content between 2011 and 2014. Indeed, users often employ unique hashtags and phrases like "*promia*" or "*thinspo*" — as well as variations like "*pr0mia*" and "*thynspo*", etc. — to bypass moderation filters.

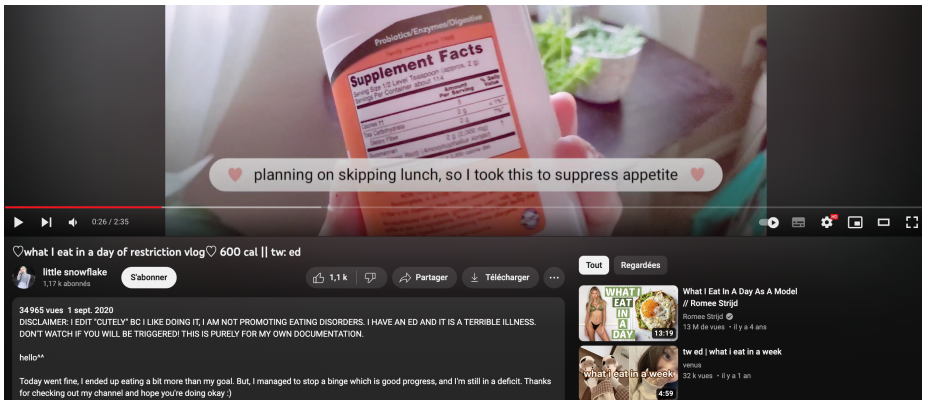

Figure 3: Example of pro-ED video on Youtube containing disclaimers to circumvent content moderation.

Disclaimers like "not pro-anything" in their profiles or trigger warnings are part of the language used by pro-recovery communities on Youtube — used to mask the true nature of their content, i.e. this user posts a "*high restriction*" diet, showing their calorie log and "*thinspiration*" and added a disclaimer and a link to recovery resources. This type of evasion complicates the task of content moderation, as it requires discerning the nuanced intent behind these postings, which becomes arduous without seeing the video entirely.

Furthermore, as studies and user experiences suggest, platforms like YouTube can inadvertently recommend pro-ED content, despite moderation efforts (Nikolova and LaMarre, 2023; Chancellor et al., 2016). This issue is exemplified by cases where for example the recommendation system — designed to tailor content based on user preferences — could fail to distinguish between a post about a diet promoting food neutrality or promoting a restrictive and moralizing approach. Such misinterpretations by algorithms, might equate posts promoting food neutrality with those advocating restrictive diets and hence present significant barriers to recovery journeys, underscoring the limitations of solely keyword-based moderation approaches.

Content moderation must also consider the balance between freedom of speech and the safety of users. Sufferers can find support and use in a variety of content types at different moments in the timeline of their illness.

Lastly, from a technical standpoint, content moderation involves constant updating of algorithms to keep pace with evolving language and user strategies — this requires a deep understanding of the subject matter and social dynamics within user communities. YouTube's approach includes consulting with third-party experts like the National Eating Disorders Association (NEDA), and aims to refine this balance but acknowledges the complexity and ongoing nature of the challenge.

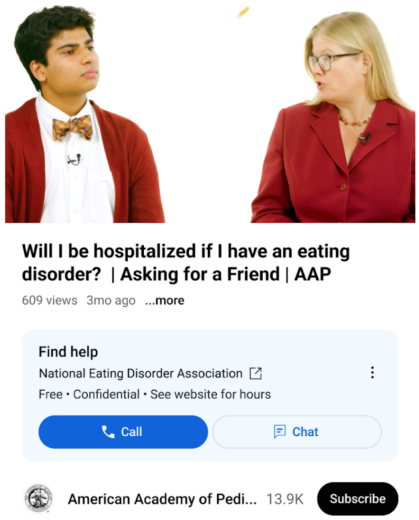

Figure 4: Youtube includes a link to NEDA's website and helpline on an ED-sensitive video.

2.2 The Youtube Recommendation System

Most recommendation systems use either content-based filtering — which selects items based on a user's past interactions, or collaborative-based filtering — which suggests content based on the preferences of similar users. YouTube, however, adopts a more complex approach due to its vast array of videos. Recommendations play a crucial role in how users decide which content to interact with videos — and often outweigh channel subscriptions and manual searches. To balance promoting new and popular content, YouTube employs a two-stage system.

2.2.1 Mechanics

Indeed, the YouTube recommendation system is a multi-stage process using deep learning models. It is designed in two key stages: candidate generation and ranking. Each stage uses a complex neural network in a different manner: the former focuses on classifying videos based on user engagement from a vast pool hence optimizing recall [2], while the latter aims at predicting the engagement likelihood of the user on a specific video using weighted logistic regression, which optimizes precision [3]. Indeed, the first candidate stage is essentially a learning problem formulated as an extreme multi-class classification problem: out of all existing videos, predict the ones that the user would engage with. The input is a large corpus (around 800 million videos) generated via collaborative filtering — which matches each of YouTube's user's *implicitly* rated items to similar items and then compiles those similar items into a list. From this, the system needs to pick out a few hundred mainly using user engagement data. Engagement data is composed of demographic data — but also of *implicit* and *explicit* feedback from user history. Explicit feedback such as likes, subscriptions, and clicks are direct interactions the user has with a video or a creator with a specific intent. However, implicit feedback like watch times, shares, browsing history, and rewatches are significant because they rely on a user's unconscious, automatic way of engaging

[2]High recall means less false positives.
[3]High precision means less false negatives.

with content — and the order of magnitude of implicit feedback is also much greater than the explicit feedback data. For instance, the sequence formed by a specific user's watch history contains particular patterns with asymmetric co-watch probabilities. Indeed, if the user has viewed two videos by the same content creator, it is probable they will choose to watch another from the same source. A predictive model that doesn't take into account this pattern of viewing behavior, focusing only on whether a user might watch a suggested video without considering their recent viewing history, is less effective. Hence, it uses the user's latest watch (and search) history — by feeding the user's latest 50 watched videos and 50 search queries, at the time of the training example, as features into the model. Hence this is why implicit user history is mostly used to train the candidate generation model [4]. For newer videos, to avoid the "cold start problem" [5] using NLP to group similar words and concepts in video descriptions to recommend fresh videos based on content similarities, circumventing the need for prior user interactions.

From these hundreds of videos, the system then ranks them by engagement — this is the second stage. The ranking also uses implicit feedback from the user: using weighted logistic regression, the clicked impressions (positive training examples) are weighted according to the duration they were watched, while unclicked ones (videos shown but not watched: negative examples) receive standard weights. However, the system also uses other sources: it assigns videos with more views, likes, shares, and comments with a higher score. These techniques approximate the odds learned by the model to expected watch time — hence the priority of the model is not to use click-through rate but user engagement /footnoteThis method de-emphasizes "clickbait" content, which typically attracts clicks but receives short watch times — and instead promotes videos that are likely to engage viewers for longer periods.

Instead, YouTube's model learns to predict the next watch, given the user's latest watch (and search) history. Technically, it does this by feeding the user's latest 50 watched videos and 50 search queries, at the time of the training example, as features into the model.

The ranking model is continuously optimized through A/B testing, which primarily focuses on maximizing this expected watch time per impression to enhance user engagement. Additionally, some randomness is built-in for novelty purposes (so you avoid getting bored). All in all the system is ever-changing, constantly evolving based on a vast array of data and user interactions.

By dividing the process into these two distinct stages, YouTube efficiently manages the vast scale of its content library. The candidate generation stage effectively reduces the massive pool of videos into a manageable subset, while the ranking stage then meticulously orders these videos to best match the user's interests, optimizing both computational resources and recommendation quality.

2.2.2 Algorithmic bias

While YouTube's recommendation system is a powerful tool to make users engage with diverse content, it is not immune to algorithmic bias. Algorithmic bias refers to the presence of unfair or discriminatory outcomes resulting from the use of algorithms. In the context of YouTube, bias can manifest in the system's promotion or demotion of certain types of content based on factors such as user demographics, implicit feedback, or content characteristics.

[4] This also explains why some platforms like TikTok provide top-quality recommendations with a much lesser need for explicit feedback — where users do not need to select videos, and merely swipe

[5] New pieces of content are not served up as recommendations until they are sufficiently clicked on and watched.

Firstly, one aspect of bias in YouTube recommendations is related to content characteristics. We have seen that YouTube's system relies on user engagement data, including implicit feedback such as watch times, likes, and shares — hence, it becomes evident that the candidate generation network is a component susceptible to introducing confirmation bias into recommended videos even before the ranking process begins. Since the candidate generation system takes into account various user-specific inputs, including the user's watch vector, search vector, geographic information, age, gender, etc. — this initial filtration process significantly narrows down the pool of potential recommended videos to a subset aligned with the user's preferences, potentially reinforcing their existing views. The primary objective of these recommendations is to generate clicks and maximize watch time [6]. However, this incentive places recommendation systems in an ethically questionable territory — as they may inadvertently manipulate confirmation bias in users. Indeed, this reliance on historical data may perpetuate biases present in the content itself. For instance, if certain types of content are historically favored/disfavored by users, the algorithm may inadvertently reinforce these preferences, leading to a lack of novelty in the content recommended to them — potentially encouraging the formation of echo chambers.

Moreover, YouTube's recommendation system takes into account demographic data and user engagement history. However, if the system disproportionately favors or disfavors certain demographics, it could potentially result in biased recommendations. For instance, if the algorithm consistently suggests content preferred by a specific age group or cultural background, it may inadvertently exclude or underrepresent other demographics. Moreover, while there is yet little evidence to support that users' age, gender, and geolocation play any significant role in amplifying misinformation in search results or recommended videos for brand new accounts (Ribeiro et al., 2020) — one could imagine that YouTube's ever-changing recommendation system might still take in that data to fine-tune its model based on implicit data.

Since biased algorithms can contribute to the formation of echo chambers, it is crucial to audit recommendation systems — however, auditing involves addressing various challenges to ensure meaningful investigations. The difficulty in auditing YouTube's personalized recommendation system stems from a combination of factors. Firstly, as highlighted by Burrell (2016), the algorithm's intricate and opaque nature presents a substantial barrier. Cristos Goodrow, the VP of Engineering at YouTube, emphasizes this complexity — stating, "It's constantly evolving, learning every day from over 80 billion pieces of information we call signals. That's why providing more transparency isn't as simple as listing a formula for recommendations, but involves understanding all the data that feeds into our system" (Cristos Goodrow, 2021). The complexity of the algorithm makes it challenging for external researchers to fully comprehend its workings and mechanisms. Secondly, the restricted access to user activity data is a significant hurdle. This data is exclusively available to social scientists employed by YouTube, limiting external researchers' ability to access comprehensive information about user interactions. The lack of transparency and data accessibility hampers the creation of models that accurately reflect the actual processes within the YouTube recommender system. As a result, external researchers can only conduct audits on YouTube's recommendation system in a 'black-box' manner — which may not capture the full scope of the system's intricacies, potentially leading to an incomplete understanding of the pathways to problematic content. Hence, researchers have to control their experimental setup carefully — requiring decisions on data collection frameworks, components to audit, and confounding factors. Indeed, to effectively analyze the output label, such as misinformative content on YouTube, clear criteria are essential. In our case, labeling videos based on their promotion, debunking, or neutral stance on eating habits and body perceptions provides a structured approach to

[6] This ultimately serves more advertisements

evaluating content.

All in all, YouTube's recommendation system, while a powerful tool for content engagement, is susceptible to algorithmic bias — of which is essential to understand the consequences.

3 Survey Experiment

3.1 Survey design

3.1.1 Purpose

In recent years, there has been growing interest in understanding the relationship between social content consumption, particularly on platforms like YouTube, and the potential impacts on individuals' attitudes toward food, body image, and the development of eating disorders. Many survey-based papers investigate gender and age demographics, as well as time spent on the platform to show the causation between social media use and deteriorating mental health — a fact that is now well-supported by decades of studies and meta-analyses. However, while existing research has explored the effect of demographics and time spent on social media platforms on mental health, there is a notable gap in investigating users' 1) beliefs about recommendation systems and 2) the impact of their behavioral patterns on algorithmic processes.

Hence, this survey aims to address this gap by examining the associations between YouTube use and the development of eating disorders, focusing on users aged 19 to 32, a demographic statistically more susceptible to body image concerns and disordered eating thoughts. We seek to understand how users' self-reported behaviors with their recommendations, the type of content recommended to them, and their tendency to enter "rabbit hole" sessions on YouTube may be linked to negative body image and disordered eating behaviors.

Our survey comprises three sections. The first section includes demographic questions alongside inquiries about the type of content consumed, daily time spent on YouTube, and patterns of interaction with recommendations. These data enable us to establish categorical predictors such as type of content watched, age, gender, time spent on YouTube, and a continuous predictor: the tendency to follow YouTube recommendations. The second section focuses on self-reported perceptions about eating habits, body image, and disordered eating behaviors. These sensitive yet crucial questions aim to generate profiles related to engagement in disordered eating behaviors and negative feelings about food and body image. Additionally, this section explores whether participants experience negative feelings about their bodies after consuming content on YouTube. The third section delves into participants' beliefs about the YouTube algorithm's impact on their experience.

3.1.2 Distribution

We surveyed to gather data for this study, and respondents were kept anonymous to encourage honesty and accuracy, in hopes of reducing bias. There were 21 questions in the survey, and they covered our predictors of body comparison and disordered eating. Only accessible through a link, the survey was designed on Google Forms — which provides a cost-effective user-friendly interface, all the while allowing us to restrict IP tracking, hence safeguarding participant privacy.

To address the sensitive nature of our study, we included a prominent warning in the preface of the survey, alerting participants to the sensitive content and providing a link to the Canadian National Eating Disorders Information Centre (NEDIC) for comprehensive informational resources.

We advertised our study among McGill University students. Additionally, we leveraged SurveyCircle, a platform designed to connect researchers with potential study participants. This platform offers a broader reach beyond personal circles and provides access to a more diverse pool of respondents.

3.2 Survey results and insights

3.2.1 Descriptive Analysis

Content Preferences and Viewing Behaviors Regarding content preferences, a majority of 68.7% primarily watched entertainment videos, 52.2% favored educational content, and 43.3% were inclined towards watching vlogs. The data also revealed a people rely significantly on the recommendation system during their watch sessions: a combined total of 69.6% of respondents frequently watched videos consecutively as recommended in a single session, and the most common clickthrough rate was watching 1-2 videos back to back (49.3%) — followed by 3-4 videos (29.9%).

Interactions with Recommendations The majority of the participants (82%), reported starting their viewing sessions from their homepage recommendations. The survey indicated that a majority believe YouTube accurately understands their preferences and is precise in its recommendations: patterns in recommendations were noticed by 70.2% of users. Moreover, a significant number of respondents, 56.7%, did not express concerns regarding the recommendations and their accuracy. However, a majority of 66.6% of users did report experiencing a 'rabbit hole' effect' on YouTube — where they find themselves continuously watching videos suggested by their algorithm.

Food Habits and Body Image Perceptions 1 out of 4 respondents has reported their tendency to over-exercise (despite weather, fatigue, illness, or injury), and 46.2% admit food and caloric intake occupy their thoughts and behaviors daily. 20.9% revealed that they are currently or have in the past engaged with disordered eating behaviors such as purging, restricting, or using diuretics. This is consistent with the results from many national surveys of college students with about 20-28% of total respondents reporting they suspected that they had suffered from an ED at some point in their lives (Daly and Costigan, 2022).

While this descriptive analysis provides an overview of user behaviors and perceptions regarding YouTube, it lacks depth in understanding the complexities of these interactions. To gain more nuanced insights, especially in uncovering underlying relationships and causal effects, a further statistical analysis such as regression and ANOVA is necessary. These methods allow us to gain a more comprehensive understanding of how user behavior and the effectiveness of recommendation algorithms affect the way people feel about food and their bodies.

3.2.2 Linear regressions, MANOVA and ANOVA results

To prepare for analysis, the data comprising 65 responses was processed using the Python libraries Pandas and SciPy. Key pre-processing steps included tokenizing, mapping responses to numerical values for quantification, and creating custom variables averaging similar questions to create ED-score and body-perception scores — which will be used mainly as our predictors.

We used 3 methods. The first one is Multivariate Analysis of Variance (MANOVA): this was used to assess the impact of multiple independent variables (like gender, and YouTube usage patterns) on dependent variables (eating disorder score and body comparison score). Then we leveraged linear regressions: two models were fitted – one predicting eating disorder scores and another predicting body comparison scores, using variables such as frequency of YouTube use and content watched. Finally, we used Analysis of Variance (ANOVA) to explore the effects of various factors (such as

following recommendations and perception of YouTube's algorithm) on eating disorders and body image scores. Using all these techniques, we aimed to uncover insights into the potential impact of social media use on perceptions about food habits and body image. The point of using different techniques is to try to cover two different hypotheses about the interactions between our ED and body perception scores.

Indeed, with MANOVAs, we evaluate the impact of independent variables (like gender, and YouTube usage patterns) on our related dependent variables (eating disorder score and body comparison score) simultaneously — thus assuming that these dependent variables are expected to be correlated or influence each other. The assumption here is that our independent variables might jointly affect multiple outcomes. However, ANOVA is used to assess the effects of different independent variables on a single dependent variable at a time — which allows us to isolate and understanding the impact of each dependent variable, like the influence of following YouTube recommendations on either eating disorder scores or body image scores separately. Indeed, the assumption in ANOVA is that any observed differences in means across groups for a single dependent variable are due to the independent variable — rather than chance. Hence, by employing both methods, the analysis gains both depth (through MANOVA) and specificity (through ANOVA).

MANCOVA Results The MANCOVA analysis indicated significant effects for some independent variables. Specifically, time spent on YouTube (Wilks' lambda = 0.7725 and p = 0.0006) and tendency to follow recommendations (Wilks' lambda = 0.8871, p = 0.0329) showed statistically significant impacts on the dependent variables, namely eating disorder score and body comparison score. Other variables like gender and perception of algorithm accuracy did not demonstrate significant effects.

Linear Regression Results

- **Eating Disorder Score Regression:** The model's R-squared value was 0.577, indicating that approximately 57.7% of the variance in eating disorder scores is explained by the independent variables. Notably, the frequency of YouTube use (coef = 0.4881, p = 0.005) and body comparison score (coef = 0.6985, p ¡ 0.001) were significant positive predictors.
- **Body Comparison Score Regression:** This model showed an R-squared value of 0.576. The tendency to follow recommendations (coef = 0.2440, p = 0.033) and the eating disorder score itself (coef = 0.5952, p ¡ 0.001) emerged as significant predictors.

Pearson's Correlation Coefficients The Pearson correlation analysis revealed a moderate correlation between the tendency to follow recommendations and disordered eating scores (correlation = 0.1658). However, the rest of the correlations were relatively weak.

ANOVA Results

- **Tendency to follow recommendations vs. Eating Disorder/body image scores:** There was no significant effect of the tendency to follow recommendations on eating disorder/body image scores (F = 0.749 and p = 0.390).
- **Time and frequency of YouTube use:** The frequency of YouTube use had a significant impact on eating disorder/body image scores (F = 13.498 and p = 0.0005), whereas time spent on YouTube did not.
- **Perception of YouTube algorithm:** Neither awareness nor distrust of the YouTube algorithm showed a significant impact on eating disorder/body image scores.

3.2.3 Conclusion and Discussion

All in all, the results from this survey experiment provide valuable insights into the complex relationship between social media use, particularly YouTube, and the perceptions students have about eating disorders and body image.

The MANOVA analysis indicated significant effects from variables such as the time spent on YouTube and the tendency to follow recommendations, suggesting that these factors play a considerable role in influencing users' eating disorder scores and body comparison perceptions. Interestingly, gender and perception of the algorithm's accuracy did not show significant effects, which might suggest that these aspects are not as influential as the actual interaction patterns with the platform. Linear regression results further substantiate this, revealing that the frequency of YouTube use and body comparison scores are significant predictors of eating disorder scores. This indicates a direct link between how often users engage with YouTube content and their perceptions related to body image and eating disorders. A moderate correlation between following recommendations and disordered eating scores, as indicated by Pearson's analysis, points towards the nuanced influence of YouTube's algorithm on users' perceptions and behaviors. While the dynamics of algorithm awareness and distrust did not yield conclusive results — we think there is need for further research to explore whether algorithm literacy, especially for younger generations, might help mitigate the potentially harmful consequences of social media use on self-perceptions.

4 Classification Model

4.1 Methods

Using the YouTube Data v3 API, we collected our dataset for both pro-ED and con-ED sides. To ensure a broad and unbiased collection of data, we utilized eight different YouTube API keys, generated from newly created Google accounts — this approach was intended to reduce potential biases from implicit bias. The initial phase of our study involved a manual examination of YouTube videos to understand the common themes and subthemes within pro-ED and con-ED content. Guided by (Lookingbill et al., 2023), we identified various types of videos, such as motivational, vlog-type, fashion and lifestyle content, as well as fitness videos promoting unrealistic ideals. This manual examination was complemented by a snowball sampling method, where we added specific harmful and pro-ED terms like "*thinspiration*", "*skinny*", "*size0*", and their spelling variations to enhance the depth of our search. We also utilized *www.hashtagify.me*, a tool primarily used for identifying popular hashtags on Twitter, to gather relevant search terms related to eating disorders. This initial list of terms, further refined through manual searches on Twitter and Tumblr, led to the compilation of 32 search queries representing both sides of the ED spectrum. For each query, we scraped the first 20 videos using the YouTube Data v3 API, collecting titles, full descriptions, and the first page of comments. The preprocessing of this data involved several steps to ensure its suitability for machine learning models and NLP.

With our textual data cleaned, the next phase involved transforming this text into a format suitable for our models. We used Term Frequency-Inverse Document Frequency (TF-IDF). Two separate TF-IDF vectorizers were employed for the 'Title' and 'Description' columns. Class labels, pro-ED content labeled as 1 and pro-recovery labeled as 0, were encoded using LabelEncoder.

We employed a variety of machine learning models to classify YouTube videos into pro-eating disorder (pro-ED) and anti-eating disorder (con-ED) categories. The models tested include Naive Bayes, Support Vector Machine (SVM), AdaBoost, and Long Short-Term Memory (LSTM) networks. Each of these models offers unique strengths and is relevant for our classification task due to their different approaches

to pattern recognition in textual data. Firstly, Naive Bayes (NB) is a probabilistic classifier known for its simplicity and efficiency in handling large datasets — it is particularly effective for text classification due to its assumption of independence among features. Support Vector Machine (SVM) is a robust classifier that works well on high-dimensional spaces like text data. It is effective in cases where there is a clear margin of separation between classes. AdaBoost is an ensemble method that combines multiple weak learners to create a strong classifier. It is used to improve the accuracy of decision trees — which are typically the base learners in this approach. Finally, LSTM, a type of recurrent neural network, is capable of learning order dependence in sequence prediction problems. This is particularly useful in text data where the context provided by the sequence of words is crucial.

Given the complex nature of text data in our dataset, our initial hypothesis would be that LSTM will perform the best. Indeed, the ability of LSTM to understand the context and sequence of words in text data is often important in distinguishing subtle nuances between pro-ED and con-ED content.

4.2 Results

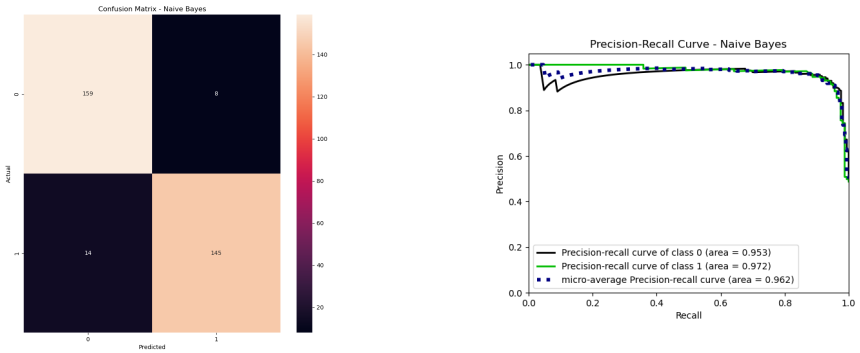

Figure 5: Naive Bayes results

Naive Bayes: This model demonstrated exceptional performance, achieving an overall accuracy of 0.93. The precision was marginally higher for pro-ED data, indicating a slightly better model performance in identifying this category.

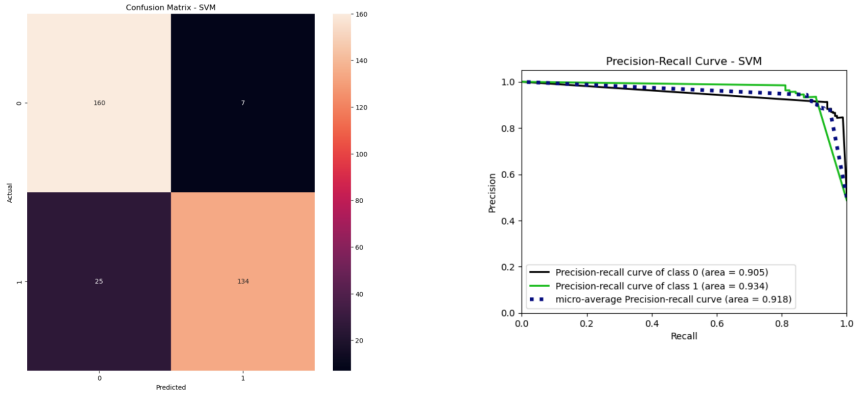

Figure 6: SVM results

Support Vector Machine The SVM model also performed well, with an overall accuracy of 0.90. While slightly lower than Naive Bayes, it maintained high precision and recall scores, showcasing its robustness in classifying complex textual data.

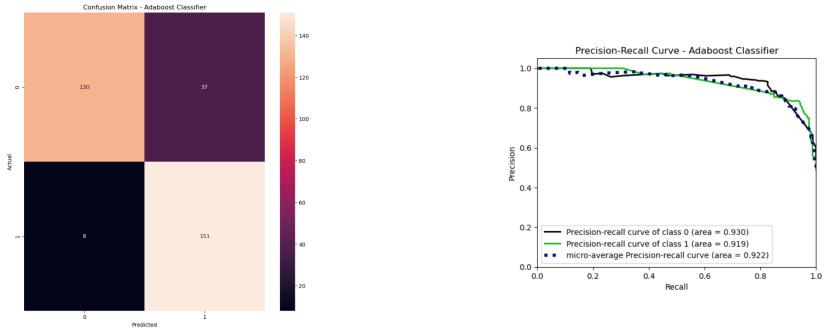

Figure 7: AdaBoost results

AdaBoost The AdaBoost model achieved an accuracy of 0.86.

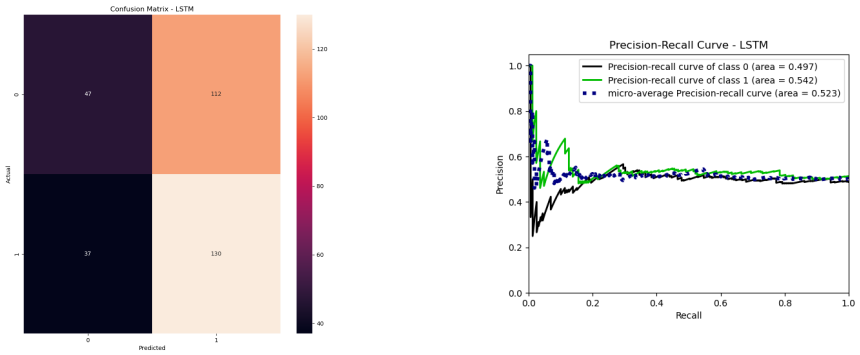

Figure 8: LSTM results

LSTM The LSTM model faced challenges, as evidenced by fluctuating validation accuracy and high validation loss (accuracy of approximately 54%, with a precision of 0.55%, recall of 0.54%, and an F1-score of 0.51.) — which suggests potential overfitting or that the model's complexity might not align well with the characteristics of our dataset.

Hypothesis refutation Contrary to our initial hypothesis, the LSTM model did not perform as well as expected. With only one layer, this cannot be attributed to the complexity of the model being too high for the relatively simpler task of classifying the given dataset. Hence, this hints at the fact that our dataset needs more careful tuning of parameters. They also need a larger amount of data to train effectively compared to simpler models, 1303 videos might not be enough. On the other hand, simpler models like Naive Bayes and AdaBoost showed strong performance, with Naive Bayes achieving an overall accuracy of 0.93. This suggests that the task of classifying pro-ED versus con-ED content might not require the capturing of long-term dependencies or the understanding of the context to the extent that an LSTM does — this is counter-intuitive as pro-recovery and pro-ED language classification is a complex task (as highlighted in section 2.1.2), hence this calls for more insights. The effectiveness of Naive Bayes and AdaBoost can be attributed to their ability to efficiently handle high-dimensional text data and focus on the most relevant features for classification.

Hence, contrary to initial expectations, the simplicity of the model cannot explain its performance issues, suggesting the need for more fine-tuning and a potentially larger dataset. While simpler models like Naive Bayes and AdaBoost demonstrated strong performance, this prompts further investigation into the specific requirements of the classification task and the potential for future improvements in model design and data collection strategies.

5 Discussion

5.1 Lessons Learned (Negative)

The underlying social issue underscores the complexity of this research. Human expression, especially concerning mental illness within sensitive echo chambers, deviates from straightforward patterns. Indeed, individuals often employ positive language to

convey deeply disturbing and harmful ideas. Moreover, attempts to circumvent content moderation pose an additional challenge. Beyond using positive terms related to food and body image, individuals resort to deceptive tactics, including the use of trigger warnings, to convey harmful messages — and such nuanced expressions can confound models that are not finely tuned to recognize these subtleties.

Moreover, I learned that data preprocessing entails navigating a landscape filled with assumptions. One notable challenge was the decision of whether to exclude non-English comments from the dataset. However, this approach yielded suboptimal results, as words associated with eating disorder (ED) language, such as 'ana' or 'mia,' ceased to appear. Furthermore, tokenization and linguistic processes like lemmatization and stemming can potentially erode the intricate relationships between words. However, in the context of addressing a deeply rooted social problem like EDs, preserving these linguistic dependencies is imperative.

My initial intentions to create automated bots for traversing recommendations were met with the realization that distinguishing between pro-ED and con-ED content is a far more intricate task than anticipated. This underscores the inherent complexity of text classification, particularly within the realm of social issues. Indeed, unlike more clear-cut categories, classifying trends in EDs and echo chambers is highly subjective. It seems like the subjective and sensitive nature of these classifications poses a significant challenge, as it involves deciphering implicit meanings, intentions, as well as the context within which the content is shared.

5.2 Lessons Learned (Positive)

Positively, this research journey has provided me with profound insights into the intricate dynamics of eating disorder related communities. The extensive background research I undertook in this study stems from the unique nature of EDs — which occupy a distinctive position within the realm of mental illnesses. These disorders intersect with pressing contemporary issues, including the pervasive influence of social media and technology. The prevalence of hyper-advertising, the burgeoning *influencer* industry, and the widespread consumption of context-limited short-form videos on platforms such as TikTok and YouTube have collectively heightened the risk of an upsurge in eating disorder-related challenges in today's digital landscape.

A particularly encouraging finding from the survey was the awareness among respondents regarding the limitations of algorithmic recommendations, notably the lack of novelty. This discovery underscores the need to delve into algorithm literacy among younger generations. Exploring to what extent early exposure to social media and tech proficiency could make younger individuals more resilient to algorithmic echo chambers.

Furthermore, the study shed light on the isolation and harm prevalent within the YouTube eating disorder community. It challenged my previously held assumptions about the decline of pro-ED content, revealing that such content, exemplified by terms like "ana creed" and "meanspo" still remain alive and well within these online spaces.

5.3 Future Work

- Conducting sentiment analysis on comment sections to ascertain sentiment ratios, offering an additional feature for distinguishing between pro and con eating disorder content.
- Manual labeling of data to leverage labeled samples and facilitate additional fine-tuning of models. This step is essential for enabling the model to discern the subtle nuances in language, particularly within the context of text classification surrounding social issues like pro-ED and con-ED communities.

- Exploring the application of BERT models, particularly those tailored to mental health-related projects like the MentalBERT model, which could yield valuable insights and improvements.

While some models performed well in my analysis, I need to test these results more thoroughly. In parallel, I plan on implementing a more refined text classification system using the Snorkel framework — leveraging multiple labeling functions (LFs) applying heuristics, such as markup terms detection, sentiment analysis, and emotional tone assessment, to generate labeled data. These LFs include checks for specific pro-ED and con-ED keywords, sentiment polarity using TextBlob, and emotional analysis. The combination of these varied LFs enables the creation of a rich, probabilistically labeled dataset. This dataset could then be used to train a more accurate machine learning classifier, enhancing the system's context awareness. I plan I used the snorkel API's .fit() function to run SGD optimizer to train the label model to ultimately produce a single set of noise-aware training labels.

```
@labeling_function()
def contains_proEDkeywords(x):
    string = str(x.text).lower()
    for keyword in pro_ED_keywords:
        if keyword in string:
            return PRO
    return ABSTAIN

@labeling_function()
def contains_conEDkeywords(x):
    string = str(x.text).lower()
    for keyword in con_ED_keywords:
        if keyword in string:
            return CON
    return ABSTAIN

@labeling_function()
def contains_tw_keywords(x):
    string = str(x.text).lower()
    for keyword in tw_keywords:
        if keyword in string:
            return CON
    return ABSTAIN

#allows us to still label videos with pro-ed themes but that circumvent using pro-recovery terms
@labeling_function()
def markup(x):
    string = str(x.text).lower()
    for keyword in tw_keywords:
        if x.labels == 1 and keyword in string:
            return PRO
        elif keyword in string:
            return CON
    return ABSTAIN
```

Figure 9: New labeling functions that will be used to finetune a BERT based model.

6 Conclusion

This investigation into the impact of YouTube's recommendation system on the dissemination of information regarding EDs has yielded crucial insights about the interplay between social media, mental health, and algorithmic content curation. Through a comprehensive approach encompassing web scraping, survey experiments, and machine learning classification, we have gained insights into the nuanced yet significant influence of social media, and particularly YouTube's recommendation algorithm in perpetuating pro-ED content. The survey experiment conducted revealed a notable relationship between YouTube usage patterns, such as frequency of use and reliance on recommendations — and the perceptions of users related to eating disorders and body image. This suggests that the manner in which users interact with YouTube content can significantly impact their mental health. Furthermore, the machine learning models deployed to classify YouTube videos into pro-ED and con-ED categories underscored the complexity of text classification in the context of social issues. Despite initial expectations, simpler models like Naive Bayes demonstrated stronger performance than more complex models like LSTM — highlighting the need for more nuanced model tuning and a larger dataset for effective classification. For future work,

this study suggests the need for more in-depth sentiment analysis of user-generated content, manual labeling of data for improved model accuracy, and exploration of specialized models like MentalBERT. Deepening this work would contribute greatly to the ongoing discourse on the intersection of technology, mental health, and social media — providing valuable insights for future policy-making.

References

J. J. Aardoom, A. E. Dingemans, L. H. Boogaard, and E. F. Van Furth. Internet and patient empowerment in individuals with symptoms of an eating disorder: A cross-sectional investigation of a pro-recovery focused e-community. *Eating Behaviors*, 15(3):350–356, 2014. ISSN 1471-0153. doi: 10.1016/j.eatbeh.2014.04.003. URL https://www.sciencedirect.com/science/article/pii/S1471015314000452.

A. S. Alberga, S. J. Withnell, and K. M. von Ranson. Fitspiration and thinspiration: A comparison across three social networking sites. *Journal of Eating Disorders*, 6(1):39, 2018. doi: 10.1186/s40337-018-0227-x. URL https://doi.org/10.1186/s40337-018-0227-x.

J. Arcelus and et al. Mortality rates in patients with anorexia nervosa and other eating disorders. a meta-analysis of 36 studies. *Archives of General Psychiatry*, 68(7):724–731, 2011a. doi: 10.1001/archgenpsychiatry.2011.74. URL https://doi.org/10.1001/archgenpsychiatry.2011.74.

J. Arcelus and et al. Mortality rates in patients with anorexia nervosa and other eating disorders: A meta-analysis of 36 studies. *Archives of General Psychiatry*, 68(7):724–731, 2011b. doi: 10.1001/archgenpsychiatry.2011.74. URL https://doi.org/10.1001/archgenpsychiatry.2011.74.

E. S. Au and S. M. Cosh. Social media and eating disorder recovery: An exploration of instagram recovery community users and their reasons for engagement. *Eating Behaviors*, 46:101651, 2022. doi: 10.1016/j.eatbeh.2022.101651. URL https://doi.org/10.1016/j.eatbeh.2022.101651.

N. Boero and C. Pascoe. Pro-anorexia communities and online interaction: Bringing the pro-ana body online. *Body & Society*, 18(2):27–57, 2012. doi: 10.1177/1357034X12440827. URL https://doi.org/10.1177/1357034X12440827.

J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2016. doi: 10.1177/2053951715622512.

S. Chancellor, J. Pater, T. Clear, E. Gilbert, and M. Choudhury. thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. pages 1199–1211, 02 2016. doi: 10.1145/2818048.2819963.

V. o. E. a. Y. Cristos Goodrow. On youtube's recommendation system. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/, Sept. 2021. Accessed on 23 November 2023.

N. Crowe and M. Watts. 'we're just like gok, but in reverse': Ana girls – empowerment and resistance in digital communities. *International Journal of Adolescence and Youth*, 21(3):379–390, 2016. doi: 10.1080/02673843.2013.856802.

E. Csipke and O. Horne. Pro-eating disorder websites: users' opinions. *European Eating Disorders Review*, 15:196–206, 2007. doi: 10.1002/erv.789. URL https://doi.org/10.1002/erv.789.

M. Daly and E. Costigan. Trends in eating disorder risk among u.s. college students, 2013–2021. *Psychiatry Research*, 317:114882, 2022. ISSN 0165-1781. doi: 10.1016/j.psychres.2022.114882. URL https://www.sciencedirect.com/science/article/pii/S0165178122004747.

K. Falconberry. Pro-ana and pro-mia sites explained: Why they're dangerous, 2022. URL https://toledocenter.com/adolescents/. Toledo Center — Eating Disorders Treatment Center.

P. Fallon, M. A. Katzman, and S. C. Wooley. *Feminist perspectives on eating disorders*. Guilford Press, 1996.

Y. Gerrard. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511, 2018. doi: 10.1177/1461444818776611.

D. Giles. Constructing identities in cyberspace: The case of eating disorders. *British Journal of Social Psychology*, 45:463–477, 2006. doi: 10.1348/014466605X53596. URL https://doi.org/10.1348/014466605X53596.

A. Q. Goh, N. Y. Lo, C. Davis, et al. eatingdisorderrecovery: A qualitative content analysis of eating disorder recovery-related posts on instagram. *Eat Weight Disord*, 27:1535–1545, 2022. doi: 10.1007/s40519-021-01279-1. URL https://doi.org/10.1007/s40519-021-01279-1.

A. Greene, H. Norling, L. Brownstone, et al. Visions of recovery: A cross-diagnostic examination of eating disorder pro-recovery communities on tiktok. *Journal of Eating Disorders*, 11(109), 2023. doi: 10.1186/s40337-023-00827-7. URL https://doi.org/10.1186/s40337-023-00827-7.

L. Hensley. Why the coronavirus pandemic is triggering those with eating disorders, 2020. URL https://globalnews.ca/news/6735525/eating-disorder-coronavirus/. Accessed on 25 October 2023.

A. Lamarre, C. Rice, and G. Jankowski. Eating disorder prevention as biopedagogy. *Fat Studies*, 6, 03 2017. doi: 10.1080/21604851.2017.1286906.

A. LaMarre, M. P. Levine, S. Holmes, and et al. An open invitation to productive conversations about feminism and the spectrum of eating disorders (part 1): basic principles of feminist approaches. *Journal of Eating Disorders*, 10(54), 2022. doi: 10.1186/s40337-022-00532-x. URL https://doi.org/10.1186/s40337-022-00532-x.

V. Lookingbill, E. Mohammadi, and Y. Cai. Assessment of accuracy, user engagement, and themes of eating disorder content in social media short videos. *JAMA Network Open*, 6(4):e238897, 2023. doi: 10.1001/jamanetworkopen.2023.8897.

R. Mulveen and J. Hepworth. An interpretative phenomenological analysis of participation in a pro-anorexia internet site and its relationship with disordered eating. *Journal of Health Psychology*, 11(2):283–296, 2006.

I. Nikolova and A. LaMarre. "if i unfollow them, it's not a dig at them": A narrative analysis of instagram use in eating disorder recovery. *Psychology of Women Quarterly*, 47(3):pp. TBD, 2023. doi: 10.1177/03616843231166378. URL https://doi.org/10.1177/03616843231166378. Published online: April 11, 2023; Published in print: September 2023.

L. Osler and J. Krueger. Proana worlds: Affectivity and echo chambers online. *Topoi*, 41:883–893, 2022. doi: 10.1007/s11245-021-09785-8. URL https://doi.org/10.1007/s11245-021-09785-8.

M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr. Auditing radicalization pathways on youtube. pages 131–141, 2020.

M. P. P. Root. Disordered eating in women of color. *Sex Roles*, 22:525–536, 1990. doi: 10.1007/BF00288168. URL https://doi.org/10.1007/BF00288168.

C. Stinson. The absent body in psychiatric diagnosis, treatment, and research. *Synthese*, 196(6):2153–2176, 2019. URL http://www.jstor.org/stable/45147717.

S. R. Strife and K. Rickard. The conceptualization of anorexia: The pro-ana perspective. *Affilia: Journal of Women & Social Work*, 26(2):213–217, 2011. doi: 10.1177/0886109911405592. URL https://doi.org/10.1177/0886109911405592.

A. Toulany, N. R. Saunders, P. Kurdyak, R. Strauss, L. Fu, N. Joh-Carnella, S. Chen, A. Guttmann, and T. A. Stukel. Acute presentations of eating disorders among adolescents and adults before and during the covid-19 pandemic in ontario, canada. *CMAJ*, 195(38):E1291–E1299, October 2023. doi: 10.1503/cmaj.221318.

S. M. Wilksch, A. O'Shea, P. Ho, S. Byrne, and T. D. Wade. The relationship between social media use and disordered eating in young adolescents. *The International journal of eating disorders*, 53(1):96–106, 2020. doi: 10.1002/eat.23198.

Table 1: Themes and Definitions of Pro–Eating Disorder, Anti–Eating Disorder, and Prorecovery Content from Lookingbill et al. (2023)

1. Encouraging the Development or Sustainment of Eating Disorders: *Creators promote eating disorders as a lifestyle.*
    (a) Thinspiration or meanspo: *Creator posts images or text to promote thinness, such as images of appearance-related ideals or insulting language to encourage other users to lose weight.*
    (b) Eating with an eating disorder
        i. Counting calories
        ii. Dieting
        iii. Food guilt
        iv. What I eat in a day (WIEIAD)
    (c) Losing weight to achieve goal weight: *Creator discusses or makes reference to losing weight to achieve their goal weight, including losing weight to achieve an unhealthy body weight.*
        i. Exercising to achieve goal weight
2. Sharing Physical and Emotional Experiences with Eating Disorders: *Creators share their experiences with living with an eating disorder, including how they developed an eating disorder and some of the physical and/or emotional symptoms.*
    (a) Best practices for talking to someone with an eating disorder
    (b) Challenging misconceptions of eating disorders
        i. Revealing the "unglamorous" side of eating disorders
    (c) Combatting pro-eating disorder content
    (d) Personifying ED
    (e) Sharing onset of eating disorder
    (f) Using humor
3. Sharing Narratives of Recovery: *Creators express their personal experiences with recovering from an eating disorder by sharing advice and words of encouragement, as well as the challenge(s) they face during recovery.*
    (a) Interacting with healthcare professionals
    (b) Recovery
        i. Celebrating recovery
        ii. Eating in recovery
        iii. Recovery struggles
4. Social Support: *Creators provide social support to other users, seek social support from other users, or share their experiences with receiving social support from individuals on- and offline.*
    (a) Providing social support
    (b) Receiving social support
    (c) Seeking social support