Exploring Online Anorexia (Pro-Ana) Communities: A Comprehensive Text Mining Exploration of Pro-Ana Discourses on Twitter

MSc. Thesis

Pia Kronenfeld

University of Twente
Faculty of Behavioural Management Sciences (BMS)
Positive Clinical Psychology and Technology

Supervision by
Dr. Peter M. ten Klooster
&
Dr. Jorge Piano Simoes

02.07.2024

Abstract

Pro-ana online communities on social media continue to celebrate anorexia despite its complex treatment and mortality rates. Previous research into pro-ana communities primarily relied on manual content analysis, limiting the quantity of analysed data. This study used automated text mining to identify discussed topics and sentiments and examined engagement metrics such as likes, retweets, replies, and quote count. A large corpus of almost 300,000 tweets spanning over ten years, containing at least one pro-ana hashtag per tweet, was used to perform topic modelling, sentiment analysis and correlation analyses. Common themes expressed in the tweets were explored utilising Latent Dirichlet Allocation (LDA) with ChatGPT 4.0 serving as a second labeller. Next, VADER sentiment analysis determined the emotional tone of the tweets. Finally, Spearman's correlations were used to identify the relationship between sentiment scores of the tweets and user engagement metrics such as likes, retweets, replies, and quote count.

The identified main themes were dieting, social engagement, and weight management, which aligned with prior research on anorectic preoccupations. Although LDA initially produced difficult-to-distinguish topics, focusing on disease-specific keywords allowed meaningful labelling. Afterwards, ChatGPT provided labels for the topics. Tweets exhibited a mainly neutral tone (94.7%), with less frequent overall positive (4.45%) and much less frequent overall negative sentiments (0.85%). Very weak but significant positive correlations (all p's $< .001$) were found between sentiment and likes ($\rho = .016$), retweets ($\rho = .042$), and replies ($\rho = .007$), and a negative correlation was found with quote count ($\rho = -.009$). Notably, only a limited number of tweets actually showed any engagement according to the engagement metrics.

This study confirms results from past manual content analysis of pro-ana communities, indicating a strong focus on content revolving around slim appearance, weight, and food management. It also expands knowledge of social media discourses supporting typically anorectic behaviour. The sentiment analysis suggests rather informative content-wise and less emotionally charged tweets. Emotionally charged tweets that, if present, evoke the most engagement. Due to the weak relationships with sentiment and engagement metrics, exploring other factors that promote interaction would be essential.

Introduction

Anorexia Nervosa (AN) is a severe eating disorder (ED) and mental illness associated with an increasingly high mortality rate (Van Hoeken & Hoek, 2020; Mehler et al., 2020). According to the DSM-5, AN involves restricted intake of calories intake leading to low body weight, coupled with an intense fear of weight gain despite being underweight, and a distorted perception of body weight or shape influencing self-evaluation and/or including a strong denial of the seriousness of low body weight (American Psychiatric Association, 2013). A very prevalent cause of death around AN is suicide, followed by the consequences of cardiovascular problems caused by prolonged restricted eating (Arnold et al., 2023; Kostro et al., 2014; Mehler et al., 2020; Mereu et al., 2022). Van Hoeken and Hoek (2020) further note that AN is associated with relatively high rates of other diagnosable mental illnesses. The prevalence of AN, which mainly affects young women, varies widely, with studies reporting prevalence rates as low as .16% to as high as 6.3% (Qian et al., 2021; Silén & Keski-Rahkonen, 2022). This variance in AN prevalence may be due to differences in study methods, populations, or locations, underscoring the complexity of diagnosing AN. There are effective psychological therapies for EDs, like Bulimia Nervosa and Binge Eating Disorder, particularly Cognitive Behaviour Therapy. A recent meta-analysis, however, could not prove that it was more effective than an active control condition for AN and that evidence-based treatments for severe and long-lasting instances of AN are scarce (Wonderlich et al., 2020).

Anorexia Nervosa and Social Media as a Refuge

Family, relatives, friends, and acquaintances of individuals with AN can react with anger, sadness, or disapproval, as they perceive the striving for extreme thinness as irrational, leading to stigma around the mental illness (Yeshua-Katz & Martins, 2012). Moreover, AN is often not perceived as a medical condition but as a deliberate choice with a societal belief that treatment solely revolves around "just eating" (Dimitropoulos et al., 2015). Consequently, adolescents and young adults with AN are likely to avoid discriminatory behaviour and search for support and coping possibilities outside regular treatment settings.

One method for individuals with stigmatised conditions is to escape to the internet to seek social support (Yeshua-Katz, 2015). In an attempt to find solace and empathy, many AN patients connect with like-minded peers online, browse through and consume media that present skinny body idols (Sukunesan et al., 2021). It has long been known that body dissatisfaction caused by media's portrayal, identity verification, and resistance to the public stigma of this thin ideal poses a severe risk for developing anorexia and serves as a maintenance factor (Ferguson et al., 2013; Yeshua-Katz, 2015). Social media has made it substantially easier to

find like-minded people who encourage or promote EDs, facilitating the sharing of experiences and views. This trend reaches vulnerable individuals and healthy people alike (Bond, 2012; Brown et al., 2023). Yeshua-Katz (2015) suggested that the large amount of online content of the so-called *pro-anorexia* (short: pro-ana) community, characterised as a safe space conducive to self-disclosure. This can offer a behind-the-scenes glimpse, making it of interest to discover for persons in (health)care professions (Brotsky & Giles, 2007; Cochran, 2010).

Pro-Anorexia Communities on Social Media

The pro-ana community is described as one in which AN is officially rejected as a mental illness while being anorexic is marketed as a lifestyle choice (Branley & Covey, 2017; Haas et al., 2010). This online community is driven by a desire for social support, a need perceived as lacking in real-life situations, with a central focus on content revolving around weight loss (Fettach & Benhiba, 2019; Rifai, 2020). Primarily consisting of individuals who identify or are diagnosed as anorectic, pro-ana groups share advice on achieving and maintaining low body weight (Rifai, 2020; Sheppard & Riccardelli, 2023), sometimes even providing strategies to conceal their attempts from family members (Arseniev-Koehler et al., 2016; Sheppard & Riccardelli, 2023). Regardless of whether individuals are diagnosed with AN or not, or are already thin, the entire community is united in promoting thinness as a central, common goal (Sukunesan et al., 2021).

Previous Research on Pro-Ana Communities on Social Media

A few studies have already explored the content and characteristics of the pro-ana communities on social media websites. A qualitative analysis of TikTok on pro-ED content found that the most prevalent theme was food and dietary restrictions (Greene et al., 2023). Utilising text mining on Reddit posts related to pro-ana, it was found that pro-ED communities predominantly discussed topics around weight loss and gain, shame, and mental and body dysfunction (Fettach & Benihiba, 2019). The content in pro-ana communications on Twitter and Tumblr has been described as “extreme” (Branley & Covey, 2017, p. 5). Especially the #thinspiration, as found by Branley and Covey (2017) and by Heinke (2023), appearing as the most prominent topic found in pro-ana discussions, was seen as a particularly dangerous trigger that could lead to harmful comparison for vulnerable users. A content analysis by Sheppard and Riccardelli (2023) and the interview study by Yeshua-Katz (2015) of the pro-ana movement on Twitter revealed that the main group effort seems to be to preserve and protect a specific group identity, rejecting the stigmas attached to EDs. Blocking, threatening, and labelling, identified as exclusion criteria, suggested potential group insecurity, and indicated intricate dynamics in online pro-ana environments (Branley & Covey, 2017). Arseniev-Koehler et al. (2016) delved

into pro-ana engagement between users. They found a link between the frequency of ED-related tweets and the presence of followers interested in such content, suggesting the emergence of social connections among individuals sharing similar interests within the network. A covert participant observation by Brotsky and Giles (2007) confirmed that pro-ana discussion content's primary function seems to be in its social nature, with users searching for connections. Exploring sentiments of pro-ana tweets, Heinke (2023) found that the tweets were rather more positively emotionally charged, less neutral and even less negative.

Twitter and Text Mining

Due to ethical concerns and user safety, social media websites like Tumblr, Pinterest, and Instagram have modified their service terms to prohibit posts supporting pro-ana. However, the content remains easily accessible (Sheppard & Ricciardelli, 2023). Twitter (now "X"), however, employs "Soft Moderation", where reported tweets marked as questionable or harmful are labelled as sensitive, remaining easily accessible (Sukunesan et al., 2021; Zannettou, 2021). Notably, Sanderson et al. (2021) found that tweets with warning labels even tend to spread more readily, diminishing the effectiveness of moderation, making Twitter a promising resource for finding unfiltered pro-ana content (Sukunesan et al., 2021). Moreover, the hashtag #proana was used by nearly 12,000 Twitter accounts (Sukunesan et al., 2021), and Twitter produced well over ten years' worth of AN communities' online exchanges, providing extensive opportunities for researchers and practitioners (Ghani et al., 2019; Hassani et al., 2020). Due to the vastness, complex and unstructured nature of their textual data, text mining can be essential for assessing social media material (Salloum et al., 2017).

Despite the vast amount of data available, most of the current research around online pro-ana content used manually carried-out analysis, limiting the amount of data that can be analysed (Au & Cosh, 2022; Brotsky & Giles, 2007; Greene et al, 2023; Haas et al., 2010; Sheppard & Riccardelli, 2023; Yeshua-Katz, 2015). Even with the apparent strengths of manual content analysis, the analysis also has disadvantages in terms of generalisation and scalability. Text mining can be a promising tool to overcome these limitations. It involves uncovering implicit, previously unknown, and potentially valuable information and patterns from extensive unstructured textual data. This is achieved using computer algorithms, often trained via machine learning, to operate semi or fully automatically (Hassani et al., 2020). Whereas conventional manual content analysis techniques usually create a codebook by examining a sample of data to identify themes within a collection of patterns. Manually, it is often impossible to recognise and cover all possible themes within a topic. Text mining usually comprises unsupervised machine learning, so it does not need such a codebook (Karami et al., 2020).

Topic Modelling and *Sentiment Analysis* are the most frequently employed text mining applications. Topic modelling allows for the automatic extraction of a text's topic(s) (Albalawi et al., 2020), while sentiment analysis assigns a so-called sentiment score to a text to identify possible underlying attitudes, perspectives, opinions, or as the name suggests sentiments (Yadav & Vishwakarma, 2019).

Research Objective and Research Questions

To date, few studies have used text mining to explore the nature of pro-ana social media discourses. The current study uses text mining as a state-of-the-art analytical big data approach to provide further insights into the content and characteristics of a large corpus of pro-ana tweets. Using text mining, themes and sentiments expressed on pro-ana communities on Twitter were observed. Established sentiments were correlated with metadata, such as likes, retweets, replies, and quote count, to investigate links with the popularity of content and interaction dynamics. The following research questions (RQs) will be addressed:

RQ1: *What topics dominate pro-ana discussions on Twitter?*

RQ2: *What associations exist between sentiments expressed in pro-ana discussions and user engagement metrics (likes, retweets, replies, and quote count)?*

Method

This study aims to answer the RQs by exploring the presence of thematic and sentiment in pro-ana online communities through Twitter. Several steps were taken to collect, process, and analyse the tweets' content and perform the analysis using the available metadata. Data preparation, pre-processing, and subsequent analysis were partly done within the programming software *RStudio version R 3.3.0+* (*RStudio Desktop - Posit*, 2024) and partly via the text mining module of *Orange Data-Mining version 3.36.2* (*Bioinformatics Laboratory, University of Ljubljana*). Due to the size of the given data, the data mining platform Orange was used within the Jupyter Lab, a server provided by the University of Twente for bigger computational tasks (Research Support: Jupyter, University of Twente).

Data Collection

The data used in this study was previously scraped by Heinke (2023) and made available on GitHub (https://github.com/juliusheinke/Master-Thesis/tree/main/Datasets). This unprocessed dataset contains all tweets containing a selection of specific pro-ana-related hashtags posted from its launch in July 2006 (Broersma & Graham, 2013) until February 2023. Every tweet containing the primary hashtags #proana, including alternative writing styles or any other common AN-related hashtag was scraped with the help of the Python library *tweety* (for example, #proana, #pr0ana, #anasister, #anasisters, #anabuddy, #4nabuddy, #meanspo,

#meansp0, #bonespo, #bonesp0, #thinspiration, etc.) (Heinke, 2023). The data, in total 288,774 tweets containing at least one pro-ana hashtag per tweet, was stored in a comma-separated value file using Pythons *NumPy*. The steps from data preparation to analysis used in the current study are visualised in Figure 1.

Figure 1

Flowchart of Steps Done Within This Research

Data Preparation and Cleaning

First, the pro-ana dataset was processed and cleaned within RStudio. The R-Script can be found in Appendix A. The tweet content column was sorted by relevance, and unnecessary columns were removed. Next, unrecognised characters and placeholders, meaningless symbols, non-printing control characters, and emojis were removed. Even though Orange can handle some emojis, emojis from 2008 were not recognised anymore, leading to substantial noise within the data. Consequently, the dataset was more manageable in Orange after reorganisation, and its reduced size facilitated smoother processing within the software.

Preprocessing

Data pre-processing is crucial in preparing the data for in-depth text mining, especially given Twitter's linguistic diversity. It involves transforming the raw data into a format suitable for further analysis and enhancing its quality (Alexandropoulos et al., 2019; Chai, 2022; Raja & Thangavel, 2019). Data pre-processing within this thesis context consisted of tokenisation, transformation, and filtering.

Tokenisation

Tokenisation assigns the data into discrete parts, like words and phrases known as tokens, and it can also remove certain characters, such as accentuation marks. The tokenised corpus is then used for further processing (Naik et al., 2022). The text of the tweets was split into individual words and symbols (tokens) instead of whole related sentences. The pre-trained "Tweet" Tokeniser option in the "Preprocess Text" widget was used within Orange. This Tokeniser is optimised for the unique characterisations of Twitter data.

Transformation

Transformation means that all tokens were transformed to lowercase so that, for instance, "Twitter", "twitter" and "TWITTER" are seen as the same words, which reduces the

total number of unique words and simplifies subsequent steps like stop word removal to ensure uniformity (Raja & Thangavel, 2019).

Word Cloud

To visualise the pre-processing steps, a word cloud of the remaining tokens was created based on the bag-of-word (BoW) model to see which tokens are the most frequent within the pro-ana tweets. This word cloud created and adapted the final stop word list, especially regarding filtering below.

Filtering

Within filtering, multiple settings needed to be considered. Filtering is necessary to eliminate certain words and tokens to remove noise within the data. The options "Numbers" and "Regexp" (all punctuation marks) were removed from the dataset using Orange. Stop words are tokens frequently appearing with little context, such as pronouns and conjunctions, etc., which were a main part of the created stop word list ("for", "of", "be", "it", "in", "not", "just", etc.). In essence, words frequently appearing in text and containing insufficient information to provide relevant context while taking up space should be removed from the data (Naik et al., 2022). This was done iteratively with a text file (txt) uploaded into Orange (Appendix B). The first unit of stop words was the tweet scraping words. Due to the sheer quantity of these words, being present in all tweets, there was a risk of the tweets' content tokens being overshadowed. Some punctuation marks, "~" and "^", as well as some noisy tokens "https://t.co/u3ufes6wtj", "U0001faal", "tags", "abc", and "um" were added to the stop word list since they could not be removed by previous filtering and appeared as meaningless symbols in a word cloud, suggesting a high-frequency blocking space for more meaningful tokens.

Topic Modelling

Addressing the first RQ, Orange provides several options for BoW topic modelling algorithms to extract a defined number of topics, such as LDA and Latent Semantic Indexing. LDA is one of the first developed algorithms for topic modelling and is particularly well suited for analysing social media data (Krishna et al., 2019; Onan et al., 2016). LDA is a generative probabilistic technique for modelling collections of discrete data. It can handle lengthy documents and produces a complete generative model, fitting the unstructured data mass (Onan et al., 2016). Some studies have shown that LDA produces topics that can be quite general (Rizvi et al., 2019) or inconsistent (Egger & Yu, 2021). Nonetheless, the topic-term matrix leaves the detected topics available for further human interpretation. Applying it to short and long documents is also intuitive and simple (Egger & Yu, 2021). Due to Orange's computational limitations in handling the large text corpus, a random sample of 60% was

created with the “Data Sampler” widget, leaving 173,265 pro-ana tweets. However, the reduction of size was not necessary for the subsequent analysis.

The analysis involved identifying the main topics based on frequency and interpreting and labelling each topic by its most frequent terms. The optimal number of topics was determined by the topic coherence measure (one to minus one). A score near one indicates good coherence, a score near zero indicates unremarkable coherence between the topic words, and a score near minus one indicates poor coherence. Next, the topics were interpreted and manually labelled. This was done by seeking common ground, connection or meaning based on the topic keywords. ChatGPT 4.0 (*ChatGPT*) was used as a second labeller after the labels of the researcher were set to compare, broaden the perspective and explore the reliability and validity of human labelling (Rijcken et al., 2023). ChatGPT 4.0 was informed about the title of this thesis, RQ1, the topic modelling method section, the related topic words, plus their assigned numbers as depicted within Orange (and within Table 1). The prompt: “These were the generated topics with Orange. How would you label the topics based on the topic keywords?” was used to generate the labels.

Sentiment Analysis

Sentiment analysis is a natural language processing (NLP) technique determining a text's emotional undertone. Sentiment analysis of tweets falls within the categories of data mining and pattern classification. These two concepts can be defined as automatic (unsupervised) or semiautomatic (supervised) identifying valuable patterns in big data sets, which are intimately related and entwined (Choudhary, 2020). For this thesis, an unsupervised sentiment analysis via Orange was used, aiming to capture the emotional tone expressed in pro-ana discussions and rate them as positive, negative, or neutral based on a so-called (compound) sentiment score. Scores close to one represent a positive sentiment, scores close to zero a neutral sentiment, and scores close to minus one a negative sentiment. Furthermore, the sentiment scores are classified into two categories based on a threshold of .5 to distinguish between two different levels of sentiment intensity and to categorise clear and unclear trends in sentiment. Orange provides several sentiment analysis algorithms: Liu Hu, Vader, Multilingual sentiment, SentiArt, and Lilah sentiment. VADER (Valence Aware Dictionary and Sentiment Reasoner) includes a list of lexical features and labels texts according to their semantic orientation as either positive, negative, or neutral, as well as a set of grammatical and syntactical rules to help refine the sentiment scores based on the words’ context (rule-based adjustments). The advantage of VADER lies in its capability to handle slang, emojis, and abbreviations, and it is commonly used within social media data (Jain et al., 2023), making it a good fit for the current purpose

and was therefore chosen. Within Orange, the widget “Distributions” was connected to the widget “Sentiment Analysis” to produce and visualise the sentiment analysis results.

Correlation Analysis

Subsequently, overall compound sentiment scores for each tweet were correlated with four user engagement metrics: number of likes, retweets, replies, and quote count using Spearman’s ρ correlation coefficients in R-Studio for investigating the relationship between emotional tone and users’ interactions with the tweets. A Spearman’s correlation of one indicates a perfect positive linear correlation, minus one indicates a perfect negative linear correlation, and zero means no relationship between the variables.

Results

The following analyses were done within a total sample of 288,774 and pro-ana tweets. Due to Orange's computational limitations, 173,265 (60% of the total sample) tweets were used solely for the topic modelling.

Table 2 provides descriptive statistics for the engagement metrics, including likes, retweets, replies, and quote count. The mean, median, SD, and range of values show that all metrics are skewed to the right. Most tweets received no or little engagement, while a few received exceptionally high numbers of likes and retweets.

Table 1

Descriptive Statistics (with Outlier)

	Pct. of 0s	Mean	Median	SD	Min	Max
Likes	52.93 %	5.26	0	1.53	0	79,806
Retweets	74.89 %	1.14	0	0.52	0	34,344
Replies	83.69 %	0.32	0	0.09	0	4714
Quote Count	97.64 %	0.06	0	0.06	0	2919

Word cloud

Figure 2 gives an overview of the final BoW model’s most frequent words within the pre-processed data of the pro-ana tweets. The bigger the words’ appearance, the more frequently they are used. The most used tokens were “#skinny,” “fat,” “want,” “#ed,” “#ana,” and “need,” all suggesting a strong focus on body appearance and aspiration.

Figure 2

Word Cloud

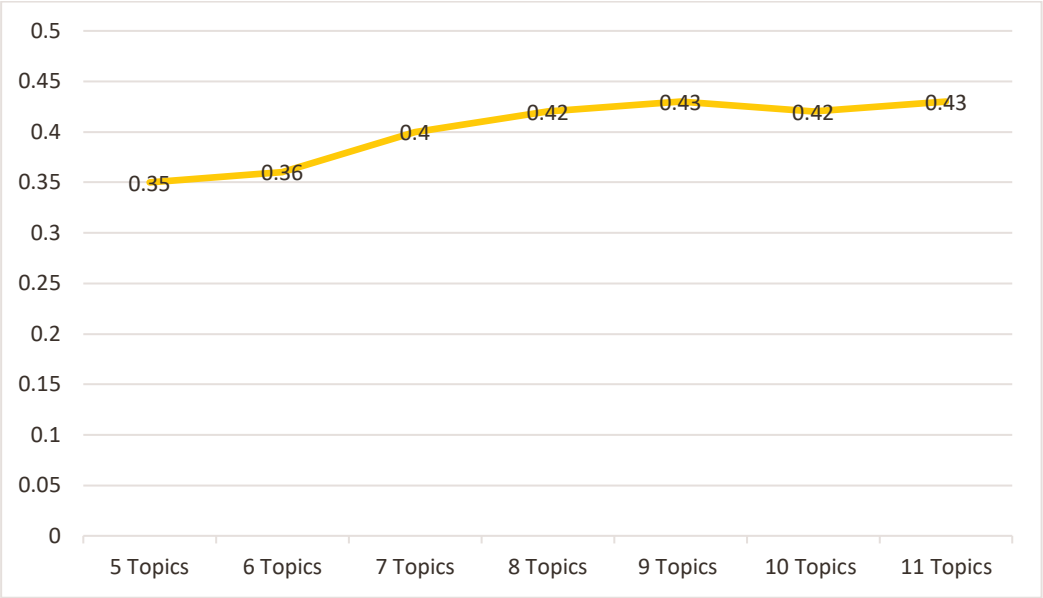
Note. The colours are for visual distinction and are randomly assigned. "#mia" is the hashtag used for the disorder Bulimia Nervosa. "[…]twt" as found behind "edtwt" and "proanatwt" is the short term for Tweet. "#fatspo" is the opposite of "#thinspo", as explained in the introduction, using an image of a person perceived as fat to motivate oneself or others not to gain too much weight. Something similar applies to "#fitspo", the description for a post used to motivate oneself or others to stay fit.

Topic Modelling

Figure 3 presents the LDA-generated topics for the pro-ana Twitter data and the accompanying topic coherence scores for creating five to 11 topics. Topic coherence scores slightly increased until nine topics and then appeared to plateau around .43. Therefore, nine topics were considered an optimal solution, giving the highest observed coherence score and the lowest number of created topics.

Figure 3

Topic Coherence (y-axis) of the Generated Pro-Ana Topics (Number of Topics = x-axis)



The nine themes and the ten most frequent words within each theme are listed in Table 1. Each topic was labelled by the researcher based on a central theme connecting the topic keywords. Next, the labels provided by ChatGPT were quite similar to those of the researcher, enhancing the likelihood of an appropriate label for the generated topics. The researcher- or ChatGPT-based labels considered to best reflect the keywords according to the researcher were kept. ChatGPT suggested naming one topic "body positivity", but this label was considered inappropriate in the theme of pro-ana. Most tweets revolved around "dieting" (18%) and "social engagement" (12.9%). This suggests a strong focus on eating behaviour and finding like-minded individuals because these topics are frequently followed by tweets on "weight management" (10.5%) and ideal body pursuit (10.25%). One example Tweet per Topic can be found in Appendix C. All in all, each topic appeared to have some unique content, although the emphasis on being and remaining skinny appeared in some form in every topic.

Table 2

Top 9 Most Relevant Pro-Ana Topics

Topic	Topic keywords	Approx. Number (%) of Tweets	Researcher Label	ChatGPT Label
1	want, thinspo, people, #proanatwt, go, see, ed, thread, follow, friends	22,384 (12.9%)	Social Connections	Social Engagement*
2	fat, #thinspo, thin, going, morning, thighs, girl, someone, keep, skinny	16,869 (9.7%)	Body Image	Physical Self-Image*
3	#ricecaketwt, please, fast, think, help, edtwt, anyone, ugh, hunger, guys	16,367 (9.5%)	Group Identity*	Community Identity
4	skinny, need, day, one, food, beautiful, water, time, fasting, ice	17,622 (10.17%)	Lifestyle*	Dietary Routine
5	new, perfect, today, calories, know, wanna, would, fucking, keep, #diet	31,225 (18%)	Dieting*	Diet and Aspirations
6	#ana, #ed, eating, legs, bones, #fatspo, #edtwitter, #anatwt, #mia, can’t	16,056 (9.3%)	Eating Disorders*	Eating Disorders Awareness
7	weight, diet, good, #anorexia, feel, lose, ached, great, #eatingdisorder, #diet	18,208 (10.5%)	Food	Weight Management*
8	body, love, eat, #weightloss, #promia, much, really, gonna, #motivation, omg	12,704 (7.3%)	Self-perception*	Body Positivity
9	#skinny, look, #thin, #thighgap, thigh, gap, looking, wish, ana, #ana	17,756 (10.25%)	Body Image	Ideal Body Pursuit*

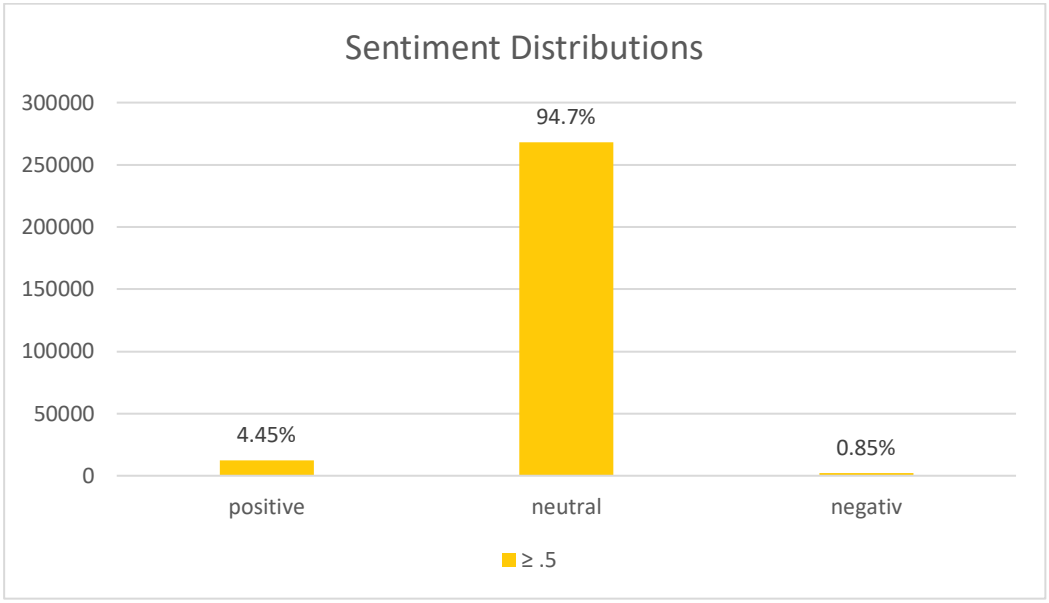
Note. *Topic label kept for this study

Sentiment Analysis

The assigned sentiment proportions across the total corpus of Tweets are shown in Figure 4. The mean sentiment compound score was .13 (*SD* = .40), suggesting an overall neutral trend regarding sentiment. The VADER sentiment analysis also revealed that neutral sentiment was present in most tweets (94.7%; *n* = 267,904). Tweets with a clear positive (≥ .5) sentiment (4.45%; *n* = 12,598) and a clear negative (≥ .5) sentiment (0.85%; *n* = 2,414) were much less frequently present. On the one hand, the neutral sentiment might suggest that most tweets would rather report with a more neutral tone. On the other hand, it could also mean that many tweets mix positive and negative sentiments, complicating the algorithm's ability to detect a clear trend and making interpreting pro-ana sentiment more complex.

Figure 4

Sentiment Distributions in Bar Charts of Pro-Ana Data in Clear Sentiment Threshold of ≥ .5



Correlation Analysis

One extreme outlier in the number of likes and retweets, also visualised in Figures D1 to D4, was removed from the corpus before computing the Spearman's correlations between the sentiment compound scores and the engagement metrics to minimise disruption in the compound. This outlier concerned a tweet posted by Australian musician Calum Hood, who appeared to have nothing to do with the pro-ana community as no other tweet ever appeared from him in the pro-ana context besides this (https://twitter.com/Calum5SOS/status/759269069606576129).

All engagement metrics were very weak, yet statistically significant ($p < .001$), correlated with the compound sentiment scores, as depicted in Table 2. The number of likes, retweets and replies were slightly positively related to more positive sentiments of the tweets, while quote count was slightly negatively correlated with negative sentiments.

Table 3

Pearson Correlation between Sentiment Compound and Engagement Metrics

	Likes	**Retweets**	**Replies**	**Quote Count**
Pearson's ρ	.016	.042	.007	- .009
P-value	p < .001*	p < .001*	p < .001*	p < .001*

*Indicates significance

For visualisation purposes, the scatterplots below display the correlations (Figures 5 to 8). In Appendix D (Figures D1 to D4) there are scatterplots visible with the outlier. Based on the statistical significance and the observation that most tweets received little to no engagement, it can be recognised that when tweets are interacted with, they tend to be rather positive in sentiment and quoted when a more negatively connotation is present.

Figure 5

Scatterplot of the Correlation between Sentiment Compound and Likes

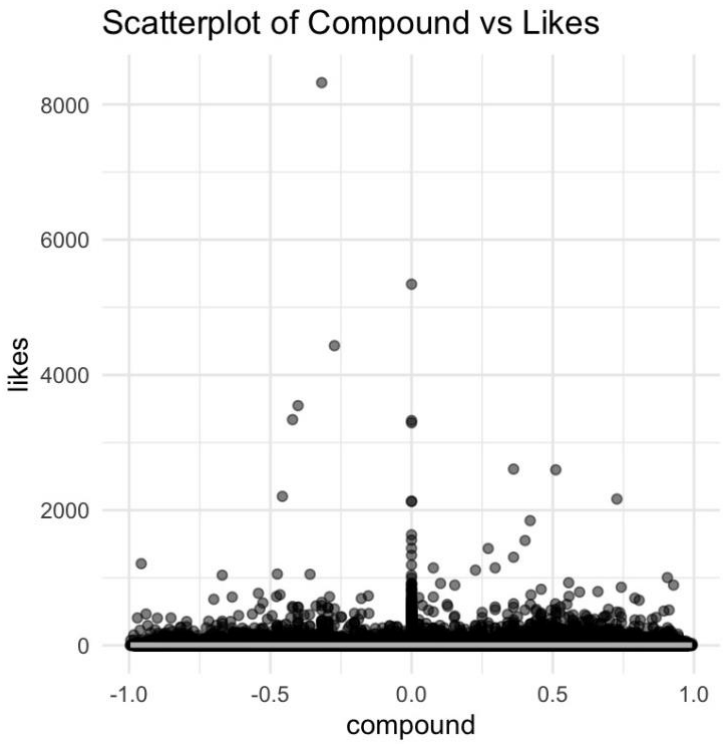


Figure 6

Scatterplot of the Correlation between Sentiment Compound and Retweets

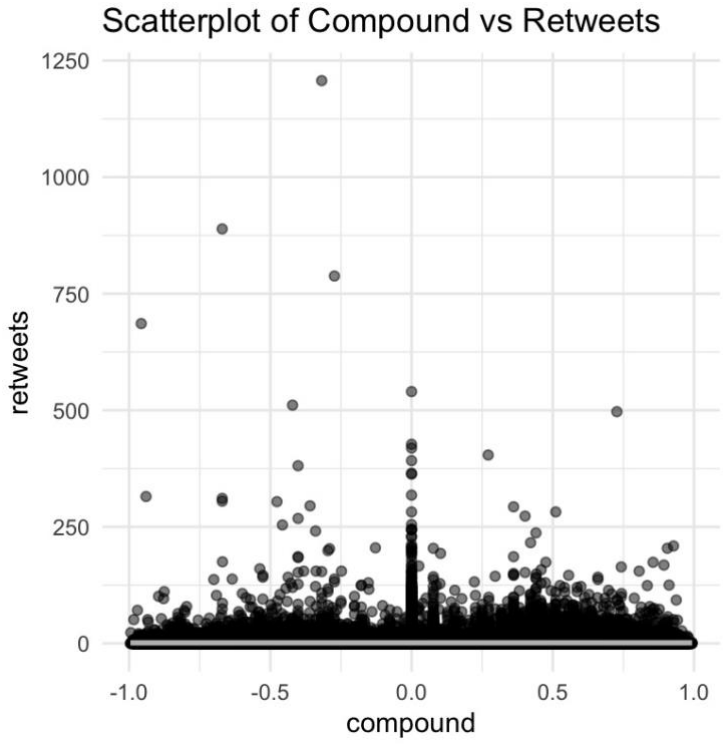


Figure 7

Scatterplot of the Correlation between Sentiment Compound and Replies

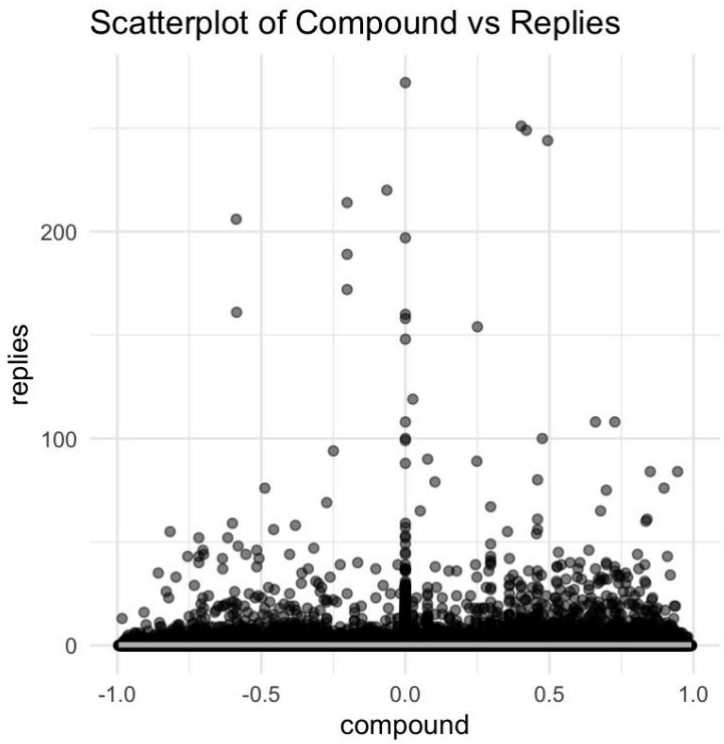
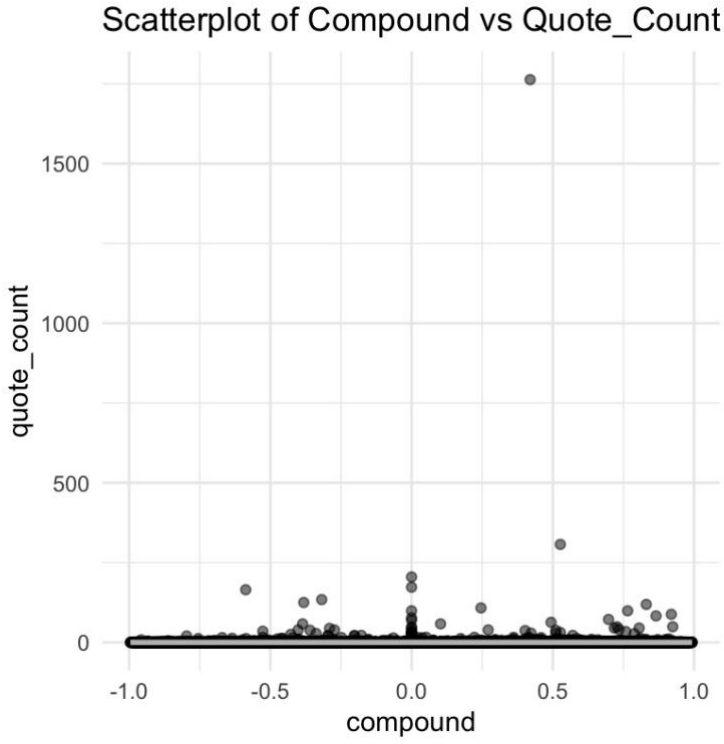


Figure 8

Scatterplot of the Correlation between Sentiment Compound and Quote Count



Discussion

This research employed text mining to analyse a large corpus of pro-ana discourse, collecting over ten years' worth of Tweets to investigate the pro-ana community on Twitter regarding its content, sentiment, and interactive nature. According to the results, users exclusively tweet about their illness and are fully occupied with their low body weights. The predominance of mainly neutral sentiments found within pro-ana discussions provides valuable insights into the general interest levels of the pro-ana group on Twitter. It reveals a rather reporting and informative tone without focusing on emotional talk. The very weak yet statistically significant relationships between the engagement metrics and sentiment support the low engagement in general. Other factors might reveal higher reasons for interaction within the community.

Topic Modelling

To answer RQ1, the primary topics observed within pro-ana Tweet discourse included dieting, social engagement, and weight management. Despite the challenge of distinguishing topics due to similar keywords and apparently overlapping contexts, naming was possible by carefully considering tokens that reflect specific traits of AN. Key characteristics defined by the DSM-5 for AN include restricted energy intake and the occupation of achieving and maintaining a low body weight (American Psychiatric Association, 2013). This served as guiding factor. Identifying these differences in topics relied on the researcher's understanding of the mental illness and sporadic scanning of example tweets allocated to the topic. While necessary in topic modelling analysis, this qualitative approach has an inherent bias like all qualitative approaches. However, it was shown that ChatGPT named the topics with a same focus, while also supporting and improving the choice of the researcher's labels. In line with Rahman et al. (2023) using ChatGPT like this as an enhancer in the research process can be very valuable. For the eighth topic, the AI chose the label "body positivity", which is discussed in social terms but based on the acceptance of a broad range of body sizes, tended to be used for normal to higher weight classes (Cohen et al., 2019; Lazuka et al., 2020; Sastre, 2014; Stewart & Ogden, 2019). One definition by Cohen et al. (2019) is, for instance, promoting respecting the body's physical attributes, capacities, and well-being. In the clinical context of AN, body positivity towards AN was not considered appropriate. As expected by previous studies and the clinical definition from the DSM-5, a strong focus could be found on remaining and being skinny, covering all topics typically associated with AN behaviours and mentalities.

The most dominating topic, "dieting", was in line with the mixed-methods studies of Fettach and Benhiba (2019), Rifai's (2020) and Sheppard and Riccardellis's (2023) qualitative

approach. The concept of “thinspiration”, also mentioned by Branley and Covey (2017) and found by Heinke (2023), was also very frequently observed in the current study, most commonly reflected in the shorten version of the term “thinspo”. Thinspiration or Thinspo demonstrates the intention to inspire oneself and others to become or remain thin, for instance, with a picture of a thin body or by mentioning a food restriction. "thinspo" was one of the most frequently expressed tokens in the most prominent topic around “dieting” and the second most prevalent topic of “social engagement”. This often-observed theme across tweets corroborates findings from earlier research (Brotsky & Giles, 2007; Fettach & Benhiba, 2026; Rifai, 2020; Sukunesan et al., 2021; Yeshua-Katz, 2015) that highlighted the importance of social interaction within pro-ana groups as central.

Although Fettach and Benhiba (2019) and Rifai (2020) found that the pro-ana groups were formed out of a search for social belonging and social support, often lacking in their real-life environment, the very low observed engagement metrics in the current study suggest that most pro-ana Twitter users may actually fail to achieve this supposed goal. Interestingly, Arseniv-Koehler et al.’s. (2016) quantitative content analysis on Twitter found a strong relationship between the number of ED-related tweets and ED content followers, potentially indicating the formation of a social network of like-minded people, while the current study found a potential scarcity of interaction with pro-ana tweets. This could imply that individuals with AN-related interests may follow similar profiles, forming a "bubble" of exposure to ED content but without much active participation. The primary function of pro-ana groups on Twitter may not solely be social but rather a potentially failed attempt to fulfil social functions.

The third-most prevalent topic was “weight management”, closely followed by “ideal body pursuit” and “lifestyle”, which showed similar content. As Greene et al.’s (2023) mixed-methods exploratory study already pointed out, users in pro-ana communities frequently share content on daily food and dietary restrictions. Notable, the most frequent terms “water” and “ice”, as found in the topic “lifestyle”, could suggest a dietary “hack” of individuals with AN. In this dataset, several tweets reported that the corresponding users often drink ice cold water as an “appetite suppressant” or to try to become colder to shiver attempting to burn calories Similarly, Wood and Knight (2015) reported that some young people with AN restrict their fluid intake due to beliefs that water has calories or to reduce weight through dehydration. Conversely, overconsumption of water is also common among them to artificially increase perceived weight, suppress hunger, purify their system, or due to compulsive behaviour.

Another interesting term used in the topic group related to identity is “#ricecaketwt”, as also found by Lernan et al. (2023). This hashtag appears to have evolved as a codeword to

distinguish genuine members from those who might infiltrate the “#edtwt” (= ED Tweet) hashtag with less “genuine” intentions. The term "rice cake" seems to have been chosen to symbolise a highly restrictive diet, as rice cakes are very low in calories, reflecting the community's extreme dietary practices (Alderton, 2022; thintokyo, 2020). This fits to previous research (Sheppard & Ricardelli, 2023; Yeshhua-Katz, 2015) that found Twitter’s pro-ana community primarily aims to maintain a distinct group identity and combats other opinions.

Sentiments of Pro-Ana Tweets and Correlations with Engagement Metrics

Regarding RQ2, predominantly 95% of neutral sentiment results were found, and 4.5% of tweets could be categorised as positive and an even smaller 0.85% as negative. They can be attributed to a genuinely neutral reporting tone in pro-ana Tweets. On the other hand, it could be due to the instinctive complicity of emotional expression, away from positive, neutral, or negative tones in the content. As concluded by Geethanjali and Valarmathi (2022) and Jindal and Aron (2021), who support this view, people do not often express their emotions straightforwardly, which commonly results in neutral categorisations of automated sentiment analysis algorithms on social media. If the latter is the case, these insights emphasise the challenges sentiment analysis models face in capturing the full spectrum of human emotions communicated online (Kanavos et al., 2017). Nevertheless, an interesting perspective on the utility of the domination of neutral sentiments in social media analysis is that neutrally balanced tweets may also provide valuable insights into the general level of interest in pro-ana discussions. Namely, that the tone of voice is reporting and informative without focusing on emotional talk.

A very weak yet statistically significant positive relationship was observed between positive sentiment and the simpler engagement metrics on Twitter, such as likes and retweets. This correlation showed a slightly stronger ρ than the weaker positive relationship between replies and sentiment requiring more than just a button press. Conversely, a significant negative correlation was identified between quote count and sentiment. Two observational studies from Al-Rawi (2019) and Rathje et al. (2021) found that content that evokes emotions in any direction, positive or negative, leads to more interaction in social media content, as this is also the case in this study. This pattern was, however, evidenced in news and political posts, complicating the comparison with and direct applications to pro-ana. The general (healthy) public trends to view AN and especially pro-ana content negatively, as reflected in the stigma (Yeshua-Katz & Martins, 2012). However, the dangerous physiological picture of pro-ana (Arnold et al., 2023; Kostro et al., 2014; Mehler et al., 2020; Mereu et al., 2022), such as the content that celebrates extreme thinness (Branley & Covey, 2017; Haas et al., 2010), is

considered positively within these groups themselves. Consequently, such content may evoke engagement from members who share this "positive" perspective. This underscores the importance of considering community-specific values. Sheppard and Riccardelli (2023) and Yeshua-Katz (2015) have already researched and analysed a defended internal group identity, as they can potentially significantly influence engagement patterns. The observed weak correlations between sentiment and engagement metrics fit the results from the overall low levels of interaction within the pro-ana discourses and the low amount of explicit positive and negative sentiment content. Potentially, other factors, such as content type, user network dynamics, the timing of posts, algorithmic influences, etc., might play more critical roles in shaping engagement patterns, showing that even minimal participation in social media can reveal significant underlying patterns, demonstrating the complexity of the dynamics.

Strengths, Limitations and Future Implications

Utilising text mining provided a more replicable and, at least partly, more objective analysis of the topics and sentiments expressed in a vast corpus of pro-ana Tweets. The actual topics found suggest a valid corpus of Tweets was used. The substantial dataset of nearly 300,000 Tweets provided considerable statistical power, enabling the detection of even subtle correlations within the data. Theoretically, this study enriches our knowledge by confirming how anorexia seems to be discussed in online communities. As shown in this study, there is only minor engagement, even if wished. Also, if further confirmed, the observation that most users do not find social support, despite they are searching for it may be an exciting finding supporting the isolated nature that individuals with AN perceive (Fettach & Benhiba, 2019; Rifai, 2020). However, several limitations must be acknowledged.

Reliance on social media data may also introduce bias, affecting generalisability due to selection bias of individuals with AN using Twitter, self-representation bias, and platform-specific bias not represented in the broader AN patient population. In addition, some of the Tweets and accounts could have been generated by bots, which has neither been ruled in nor out (Alothali et al., 2018). Furthermore, LDA provided topics that needed to be classified with caution, as Rizvi et al. (2019) and Egger and Yu (2021) noted. Moreover, the overall interaction levels with the Tweets were notably low, as expressed by the low means of likes, retweets, replies, replies, and quote count. These skewed engagement metrics may have also suppressed correlations with the sentiment scores of the Tweets. Future studies are warranted to isolate and analyse other influences of additional variables, providing valuable insights into the complex mechanisms that drive interaction within and across various online platforms to better understand the activities and participation.

References

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modelling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence, 3.* https://doi.org/10.3389/frai.2020.00042

Alderton, Z. (2022). *Preventing harmful behaviour in online communities: Censorship and interventions*. Routledge.

Alexandropoulos, S. N., Kotsiantis, S., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *Knowledge Engineering Review, 34.* https://doi.org/10.1017/s026988891800036x

Alothali, E., Zaki, N., Mohamed, E. A., & Alashwal, H. (2018). Detecting Social Bots on Twitter: A Literature Review. *IEEE Xplore*. https://doi.org/10.1109/innovations.2018.8605995

Al-Rawi, A. (2019). Networked emotional news on social media. *Journalism Practice, 14*(9), 1125–1141. https://doi.org/10.1080/17512786.2019.1685902

Arnold, S., Correll, C.U., & Jaite, C. (2023). Frequency and correlates of lifetime suicidal ideation and suicide attempts among consecutively hospitalized youth with anorexia nervosa and bulimia nervosa: Results from a retrospective chart review. *Borderline Personality Disorder and Emotion Dysregulation, 10*(1). https://doi.org/10.1186/ s40479-023-00216-1

Arseniev-Koehler, A., Lee, H., McCormick, T. H., & Moreno, M. A. (2016). #Proana: Pro-eating disorder socialization on Twitter. *Journal of Adolescent Health, 58*(6), 659–664. https://doi.org/10.1016/j.jadohealth.2016.02.012

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

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

Bioinformatics Laboratory, University of Ljubljana. (n.d.). *Orange data mining.* Retrieved April 25, 2024, from https://orangedatamining.com/

Bond, E. (2012). Virtually anorexic – Where's the harm? A research study on the risks of Pro-anorexia websites. *Children's Media Foundation*.

Branley, D. B., & Covey, J. (2017). Pro-ana versus Pro-recovery: A content analytic comparison of social sedia users' communication about eating disorders on Twitter and Tumblr. *Frontiers in Psychology, 8.* https://doi.org/10.3389/fpsyg.2017.01356

Broersma, M., & Graham, T. (2013). Twitter as a new source. *Journalism Practice, 7*(4), 446–464. https://doi.org/10.1080/17512786.2013.802481

Brotsky, S. R., & Giles, D. (2007). Inside the "Pro-ana" community: a covert online participant observation. *Eating Disorders*, *15*(2), 93–109. https://doi.org/10.1080/10640260701190600

Brown, R. D., Sillence, E., Coventry, L., Branley-Bell, D., Murphy-Morgan, C., & Durrant, A. (2023). Health stigma on Twitter: Investigating the prevalence and type of stigma communication in tweets about different conditions and disorders. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1264373

Chai, C. P. (2022). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/s1351324922000213

ChatGPT. (n.d.). https://chat.openai.com/

Choudhary, S. (2020). Opinion mining and sentiment analysis on big data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 410–415. https://doi.org/10.32628/cseit2063100

Cochran, A. (2010). Blogging the recovery from anorexia: A new platform for the voice of ED. *Young Scholars In Writing*, *7*, 122-128.

Cohen, R., Irwin, L., Newton-John, T., & Slater, A. (2019). #bodypositivity: A content analysis of body positive accounts on Instagram. *Body Image*, *29*, 47–57. https://doi.org/10.1016/j.bodyim.2019.02.007

Dimitropoulos, G., Freeman, V., Muskat, S., Domingo, A., & McCallum, L. (2015). "You don't have anorexia, you just want to look like a celebrity": perceived stigma in individuals with anorexia nervosa. *Journal of Mental Health*, *25*(1), 47–54. https://doi.org/10.3109/09638237.2015.1101422

Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: A topic modelling comparison. *Tourism Review*. https://doi.org/10.1108/tr-05-2021-0244

Ferguson, C. J., Muñoz, M. E., Garza, A., & Galindo, M. (2013). Concurrent and prospective analyses of peer, television and social media influences on body dissatisfaction, eating disorder symptoms and life satisfaction in adolescent girls. *Journal of Youth and Adolescence, 43*(1), 1–14. https://doi.org/10.1007/s10964-012-9898-9

Fettach, Y., & Benhiba, L. (2019). Pro-eating disorders and Pro-recovery communities on Reddit: Text and network comparative analyses. In the 21st international conference on information integration and Web-based applications & services (iiWAS2019) (pp. 1–10). ACM. https://doi.org/10.1145/3366030.3366058

Geethanjali, R., & Valarmathi, A. (2022). Issues and future challenges of sentiment analysis for social networks - a survey. *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS).* https://doi.org/10.1109/icacrs55517.2022.10029070

Ghani, N. A., Hamid, S. B. B. O. A., Hashem, M., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior, 101*, 417–428. https://doi.org/10.1016/j.chb.2018.08.039

Greene, A. K., Norling, H. N., Brownstone, L. M., Maloul, E. K., Roe, C., & Moody, S. A. (2023). Visions of recovery: a cross-diagnostic examination of eating disorder pro-recovery communities on TikTok. *Journal of Eating Disorder, 11(1).* https://doi.org/10.1186/s40337-023-00827-7

Haas, S. M., Irr, M. E., Jennings, N. A., & Wagner, L. M. (2010). Communicating thin: A grounded model of online negative enabling support groups in the pro-anorexia movement. *New Media & Society, 13*(1), 40–57. https://doi.org/10.1177/1461444810363910

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing, 4*(1), 1. https://doi.org/10.3390/bdcc4010001

Heinke, J. (2023). "You are what you tweet!": Analysing Pro-anorexia and Pro-recovery messages on Twitter using transformer-based text mining applications *(Master's thesis, University of Twente).* https://essay.utwente.nl/96993/

Hwang, Y. (2023). When makers meet the metaverse: Effects of creating NFT metaverse exhibition in maker education. *Computers and Education/Computers & Education, 194*, 104693. https://doi.org/10.1016/j.compedu.2022.104693

Jacobs, A. M., & Kinder, A. (2019). Computing the affective-aesthetic potential of literary texts. *AI, 1*(1), 11–27. https://doi.org/10.3390/ai1010002

Jain, R., Kumar, A., Nayyar, A., Dewan, K., Garg, R., Raman, S., & Ganguly, S. (2023). Explaining sentiment analysis results on social media texts through visualization. *Multimedia Tools and Applications, 82*(15), 22613–22629. https://doi.org/10.1007/s11042-023-14432-y

Jindal, K., & Aron, R. (2021). Withdrawn: A systematic study of sentiment analysis for social media data. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.048

Kanavos, A., Nodarakis, N., Sioutas, S., Tsakalidis, A., Tsolis, D., & Tzimas, G. (2017). Large scale implementations for Twitter sentiment classification. *Algorithms*, *10*(1), 33. https://doi.org/10.3390/a10010033

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. *IEEE Access*, *8*, 67698–67717. https://doi.org/10.1109/access.2020.2983656

Kostro, K., Lerman, J. B., & Attia, E. (2014). The current status of suicide and self-injury in eating disorders: A narrative review. *Journal of Eating Disorders*, 2(1). https://doi.org/10.1186/s40337-014-0019-x

Krishna, V. K., Kumar Pandey, A. K. P., & Kumar, S. (2019). Efficient topic level opinion mining and sentiment analysis algorithm using Latent Dirichlet Allocation model. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(5), 2568–2572. https://doi.org/10.30534/ijatcse/2019/105852019

Lavis, A. (2018). Not eating or tasting other ways to live: A qualitative analysis of 'living through' and desiring to maintain Anorexia. *Transcultural Psychiatry*, *55*(4), 454–474. https://doi.org/10.1177/1363461518785796

Lazuka, R. F., Wick, M. R., Keel, P. K., & Harriger, J. A. (2020). Are we there yet? Progress in depicting diverse images of beauty in Instagram's body positivity movement. *Body Image*, *34*, 85–93. https://doi.org/10.1016/j.bodyim.2020.05.001

Lei, A., Willems, R. M., & Eekhof, L. S. (2023). Emotions, fast and slow: Processing of emotion words is affected by individual differences in need for affect and narrative absorption. *Cognition and Emotion*, *37*(5), 997–1005. https://doi.org/10.1080/02699931.2023.2216445

Malecki, J., Rhodes, P., & Ussher, J. (2018). Childhood trauma and anorexia nervosa: From body image to embodiment. *Health Care for Women International*, *39*(8), 936–951. https://doi.org/10.1080/07399332.2018.1492268

Marucci, S., Ragione, L. D., De Iaco, G., Mococci, T., Vicini, M., Guastamacchia, E., & Triggiani, V. (2018). Anorexia nervosa and comorbid psychopathology. *Endocrine, Metabolic & Immune Disorders. Drug Targets*, *18*(4), 316–324. https://doi.org/10.2174/1871530318666180213111637

Mehler, P. S., Watters, A., Joiner, T. E., & Krantz, M. J. (2022). What accounts for the high mortality of anorexia nervosa? *International Journal of Eating Disorders*, *55*(5), 633–636. https://doi.org/10.1002/eat.23664

Mereu, A., Fantoni, T., Caini, S., Monzali, F., Roselli, E., Taddei, S., Lucarelli, S., & Pisano, T. (2022). Suicidality in adolescents with onset of anorexia nervosa. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, *27*(7), 2447–2457. https://doi.org/10.1007/s40519-022-01384-9

Naik, D. A., Mythreyan, S., & Seema, S. (2022). Relevance feature discovery in text mining using NLP. *2022 3rd International Conference for Emerging Technology (INCET).* https://doi.org/10.1109/incet54531.2022.9824807

Nemes, L., & Kiss, A. (2020). Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, *5*(1), 1–15. https://doi.org/10.1080/24751839.2020.1790793

Qian, J., Wu, Y., Liu, F., Zhu, Y., Jin, H., Zhang, H., Wan, Y., Li, C., & Yu, D. (2021). An update on the prevalence of eating disorders in the general population: A systematic review and meta-analysis. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, *27*(2), 415–428. https://doi.org/10.1007/s40519-021-01162-z

Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., & Rahaman, S. (2023). ChatGPT and Academic Research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*, *3*(1), 1–12. https://doi.org/10.52631/jemds.v3i1.175

Raja, P. S., & Thangavel, K. (2019). Missing value imputation using unsupervised machine learning techniques. *Soft Computing*, *24*(6), 4361–4392. https://doi.org/10.1007/s00500-019-04199-6

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(26). https://doi.org/10.1073/pnas.2024292118

Research Support: Jupyter, JupyterLab, Jupiter, Cloud Computing, Service Portal, University of Twente. University of Twente. https://www.utwente.nl/en/service-portal/research-support/it-facilities-for-research/jupyterlab#about-jupyterlab

Rifai, E. (2020). Digital waistlands: Pro-ana communities, religion, and embodiment. *Journal of Religion, Media and Digital Culture*, *9*(2), 207-227.

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., & Kaymak, U. (2023). Towards Interpreting Topic Models with ChatGPT. Paper presented at the 20th World Congress of the International Fuzzy Systems Association, Daegu, Korea, Republic of.

Rizvi, R., Wang, Y., Nguyen, T. H., Vasilakes, J., Bian, J., He, Z., & Zhang, R. (2019). Analyzing social media data to understand consumer information needs on dietary supplements. *PubMed*, *264*, 323–327. https://doi.org/10.3233/shti190236

RStudio Desktop - Posit. (2024, January 11). Posit. https://posit.co/download/rstudio-desktop/

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: Facebook and Twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal*, *2*(1), 127–133. https://doi.org/10.25046/aj020115

Sanderson, Z., Brown, M., Bonneau, R., Nagler, J., & Tucker, J. (2021). Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-77.

Sastre, A. (2014). Towards a radical body positive. *Feminist Media Studies, 14*(6), 929–943. https://doi.org/10.1080/14680777.2014.883420

Sheppard, A., & Ricciardelli, R. (2023). Bio-citizens online: A content analysis of pro-ana and weight loss blogs. *Canadian Review of Sociology/Revue Canadienne De Sociologie*, *60*(2), 259–275. https://doi.org/10.1111/cars.12426

Silén, Y., & Keski-Rahkonen, A. (2022). Worldwide prevalence of DSM-5 eating disorders among young people. *Current Opinion in Psychiatry*, *35*(6), 362–371. https://doi.org/10.1097/yco.0000000000000818

Stewart, S. F., & Ogden, J. (2019). The role of BMI group on the impact of weight bias versus body positivity terminology on behavioral intentions and beliefs: An experimental study. *Frontiers in Psychology, 10.* https://doi.org/10.3389/fpsyg.2019.00634

Sukunesan, S., Huynh, M. Q., & Sharp, G. (2021). Examining the Pro-eating disorders community on Twitter via the hashtag #proana: Statistical modeling approach. *JMIR Mental Health*, *8*(7), e24340. https://doi.org/10.2196/24340

thintokyo. (2020, October 20). [Tweet]. X. https://x.com/thintokyo/status/1318887739157323778?lang=de

Van Hoeken, D., & Hoek, H. W. (2020). Review of the burden of eating disorders: Mortality, disability, costs, quality of life, and family burden. *Current Opinion in Psychiatry*, *33*(6), 521–527. https://doi.org/10.1097/yco.0000000000000641

Wonderlich, S. A., Bulik, C. M., Schmidt, U., Steiger, H., & Hoek, H. W. (2020). Severe and enduring anorexia nervosa: Update and observations about the current clinical reality. *International Journal of Eating Disorders*, *53*(8), 1303–1312. https://doi.org/10.1002/eat.23283

Yadav, A., & Vishwakarma, D. K. (2019). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, *53*(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yeshua-Katz, D. (2015). Online stigma resistance in the Pro-ana community. *Qualitative Health Research*, *25*(10), 1347–1358. https://doi.org/10.1177/1049732315570123

Yeshua-Katz, D., & Martins, N. (2012). Communicating stigma: the Pro-ana Paradox. *Health Communication*, *28*(5), 499–508. https://doi.org/10.1080/10410236.2012.699889

Zannettou, S. (2021). "I won the election!": An empirical analysis of soft moderation interventions on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 865–876. https://doi.org/10.1609/icwsm.v15i1.18110

Appendix A

R-script

```
###install standard packages
install.packages("readxl")
install.packages("lavaan")
install.packages("lavaanPlot")
install.packages("tidyr")
install.packages("dplyr")
install.packages("haven")
install.packages("ggpubr")
install.packages("ggplot2")
install.packages("semPlot")
install.packages("MVN")
install.packages("tidyverse")
install.packages("WriteXLS")
install.packages("ltm")
install.packages("outliers")
install.packages("EnvStats")
install.packages("lme4")
install.packages("lmtest")
install.packages("openxlsx")
install.packages("psych")
install.packages("olsrr")
install.packages("jtools")
install.packages("moments")
install.packages("lmtest")
install.packages("prettyR")
install.packages("stringi")
install.packages("stringr")
install.packages("tm")

###load standard packages
library(readxl)
library(lavaan)
```

```
library(lavaanPlot)
library(tidyr)
library(dplyr)
library(haven)
library(ggpubr)
library(ggplot2)
library(semPlot)
library(MVN)
library(tidyverse)
library(WriteXLS)
library(ltm)
library(outliers)
library(EnvStats)
library(lme4)
library(lmtest)
library(openxlsx)
library(psych)
library(olsrr)
library(jtools)
library(moments)
library(lmtest)
library(prettyR)
library(stringi)
library(stringr)
library(tm)

setwd("/Users/piakronenfeld/Library/Mobile
      Documents/com~apple~CloudDocs/Masters/Second Quarter & following/Master
      Thesis/Original Datasets")

Pro_Ana_Data <- read.csv("ProAnaSentiment.csv", header = TRUE)

#first few rows of the data
head(Pro_Ana_Data)
```

```
#removing columns
names(Pro_Ana_Data)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -X)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -Unnamed..0.2)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -Unnamed..0.1)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -Unnamed..0)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -author_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -user_created_at)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -author_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -user_withheld)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -conversation_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -tweet_created_at)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -tweet_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -pinned_tweet_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -in_reply_to_user_id)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -reply_settings)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -source)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -withheld)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -verified)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -name)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -username)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -description)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -location)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -walabel)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -url)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -user_withheld)
Pro_Ana_Data <- dplyr::select(Pro_Ana_Data, -conversation_id)

View(Pro_Ana_Data)

#sorting columns, moving column (text) to first position
Pro_Ana_Data <- Pro_Ana_Data[c(names(Pro_Ana_Data)[9], names(Pro_Ana_Data)[-9])]

View(Pro_Ana_Data)
```

```
#remove random things - general cleanup for common unrecognized characters placeholders
Pro_Ana_Data$text <- gsub("[�•◦]", "", Pro_Ana_Data$text)
Pro_Ana_Data$text <- gsub("\u2665", "", Pro_Ana_Data$text)

#removing all emojis, symbols, and non-printing control characters
Pro_Ana_Data$text <- stringi::stri_replace_all_regex(Pro_Ana_Data$text, "[\\p{So}\\p{C}]",
        "")

#save the modified dataframe to a new CSV file with a new name
write.csv(Pro_Ana_Data, "Pro_Ana_Data.csv", row.names = FALSE)
write.csv(Pro_Ana_Data, "~/Desktop/Pro_Ana_Data_new.csv", row.names = FALSE)

#_____
        _____

#counting zeros
counting_zero_percentage <- function(column) {
  return(sum(column == 0) / length(column) * 100)
}

zero_percentages <- Pro_Ana_Data %>%
  summarise(
    likes_zero_pct = calculate_zero_percentage(likes),
    retweets_zero_pct = calculate_zero_percentage(retweets),
    replies_zero_pct = calculate_zero_percentage(replies),
    quote_count_zero_pct = calculate_zero_percentage(quote_count)
  )

print(zero_percentages)

#_____
        _____

###Correlation Analysis
```

```
setwd("/Users/piakronenfeld/Library/Mobile
      Documents/com~apple~CloudDocs/Masters/Second Quarter & following/Master
      Thesis/New sorted Datasets")

Cor <- read.csv("Compound_Engagement.csv", header = TRUE)
View(Cor)

#outlier things
outlier <- Cor[Cor$likes == 79806, ]

outlier <- Cor %>%
  filter(likes == 79806)

print(outlier)

#deteling the first rows and checking data
Cor <- Cor[-c(1, 2), ]
sum(is.na(Cor$compound))
sum(is.na(Cor$likes))
sum(is.na(Cor$retweets))
sum(is.na(Cor$replies))
sum(is.na(Cor$quote_count))

Cor$compound <- as.numeric(gsub("[^0-9.-]+", "", Cor$compound))
Cor$likes <- as.numeric(gsub("[^0-9.-]+", "", Cor$likes))
Cor$retweets <- as.numeric(gsub("[^0-9.-]+", "", Cor$retweets))
Cor$replies <- as.numeric(gsub("[^0-9.-]+", "", Cor$replies))
Cor$quote_count <- as.numeric(gsub("[^0-9.-]+", "", Cor$quote_count))

#remove outlier
outlier_row <- Cor[Cor$likes == 79806, ]
print(outlier_row)
Cor <- Cor[Cor$likes != 79806, ]
```

```
#Pearson
cor_l <- cor(Cor$compound, Cor$likes, method = "spearman", use = "complete.obs")
print(cor_l)

cor_ret <- cor(Cor$compound, Cor$retweets, method = "spearman", use = "complete.obs")
print(cor_ret)

cor_rep <- cor(Cor$compound, Cor$replies, method = "spearman", use = "complete.obs")
print(cor_rep)

cor_q <- cor(Cor$compound, Cor$quote_count, method = "spearman", use = "complete.obs")
print(cor_q)

test_likes <- cor.test(Cor$compound, Cor$likes, method = "spearman", use = "complete.obs")
print(paste("Correlation (likes):", test_likes$estimate))
print(paste("P-value (likes):", test_likes$p.value))

test_retweets <- cor.test(Cor$compound, Cor$retweets, method = "spearman", use =
        "complete.obs")
print(paste("Correlation (retweets):", test_retweets$estimate))
print(paste("P-value (retweets):", test_retweets$p.value))

test_replies <- cor.test(Cor$compound, Cor$replies, method = "spearman", use =
        "complete.obs")
print(paste("Correlation (replies):", test_replies$estimate))
print(paste("P-value (replies):", test_replies$p.value))

test_quote_counts <- cor.test(Cor$compound, Cor$quote_count, method = "spearman", use =
        "complete.obs")
print(paste("Correlation (quote counts):", test_quote_counts$estimate))
print(paste("P-value (quote counts):", test_quote_counts$p.value))

#scatterplots
p1 <- ggplot(Cor, aes(x = compound, y = likes)) +
```

```
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "grey") +
  ggtitle("Scatterplot of Compound vs Likes") +
  theme_minimal()
print(p1)

p2 <- ggplot(Cor, aes(x = compound, y = retweets)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "grey") +
  ggtitle("Scatterplot of Compound vs Retweets") +
  theme_minimal()
print(p2)

p3 <- ggplot(Cor, aes(x = compound, y = replies)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "grey") +
  ggtitle("Scatterplot of Compound vs Replies") +
  theme_minimal()
print(p3)

p4 <- ggplot(Cor, aes(x = compound, y = quote_count)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "grey") +
  ggtitle("Scatterplot of Compound vs Quote_Count") +
  theme_minimal()
print(p4)

#save
install.packages("openxlsx")
library(openxlsx)
write.xlsx(Cor, "Cor1.xlsx", rowNames = FALSE)
```

Appendix B

Stop Word List

#proana
#pr0ana
#anasister
#anasisters
#anabuddy
#4nabuddy
#meanspo
#meansp0
#sweetspo
#sweetsp0
#bonespo
#bonesp0
#thinspiration
#thinspotwt
#thinspo
#thinsp0
#th1nspo
#th1nsp0
#thynspo
~
^
https://t.co/u3ufes6wtj
U0001faal
back
even
tags
instead
abc
um
actually
you
for

of
be
it
in
not
just
that
are
have
has
before
then
than
got
any
get
make
still
like
hi
im
rt
i'd
i'm
could
u

Appendix C

Example Tweet per Topic

Topic 1: Looking for positive thinspo people to follow #fitspo #thinspo #wieghtloss #running #yoga

Topic 2: have a skinny morning#thinspo #thinspiration

#ed #thin #skinny #edtwt #edtwitter #프로아나_트친소 #프로아나 #よかったらダイエットのモチベーション上がる写真載せてくれませんか #ダイエット刺激画像 #ダイエット刺激画像

Topic 4: Day 1 of my fast im so sick of my ugly disgusting body no wonder no boys like me , they all probably laugh and whisper about me when i walk past thats why i need to get skinny ill keep you updated x#proana #edtwt #thinspo #meanspo

Topic 3: People eating so you don’t have to thread #edtwt #ricecaketwt #foodspo #grossfood #fatspo,

Topic 4: Water fasting isn't thigh gap so why beautiful #proana

Topic 5: Today’s food log:Fucking spooned spelt pasta and ground beef out of the pot like a savage animal + a plumTotal Cals: Around 960? Idk. I can only estimate. I didn’t eat enough to have surpassed that. I don’t think.#edtwt #shtwt #198twt #54twt #254twt #ricecaketwt

Topic 6: a this look disordered" pics from people on edtwt ~ a thread #ana #ed #edtwt #edtwtthread #thinspo #meanspo #sweetspo #fatspo #edtwtfood #mia

Topic 7: Day 1 of fastCurrently weighing : 122 poundsCurernt goal weight : 102#thinsp0 #anorexia #starve

Topic 8: I love it when people see me an I’m known as the thin or skinny girl, especially when they say I need to eat more when I am eating, proud to know I can eat and it’s not noticable on my body.. #thinspo #th1nspo #ed #edtwt #bonespo #skinspo #edtwtdiet #weightloss

Topic 9: Wish I could climb out of bed rn to the perfect thigh gab #Thinspo #thighgap #skinny

Appendix D

Scatterplots from Correlation Analysis with Outlier

Figure D1

Scatterplot of the Correlation between Sentiment Compound and Likes with Outlier

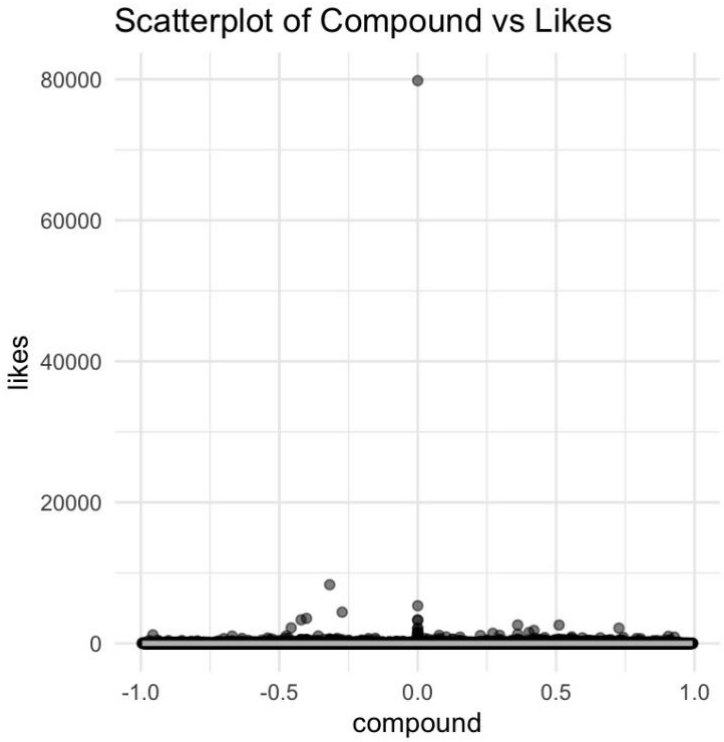


Figure D2

Scatterplot of the Correlation between Sentiment Compound and Retweets with Outlier

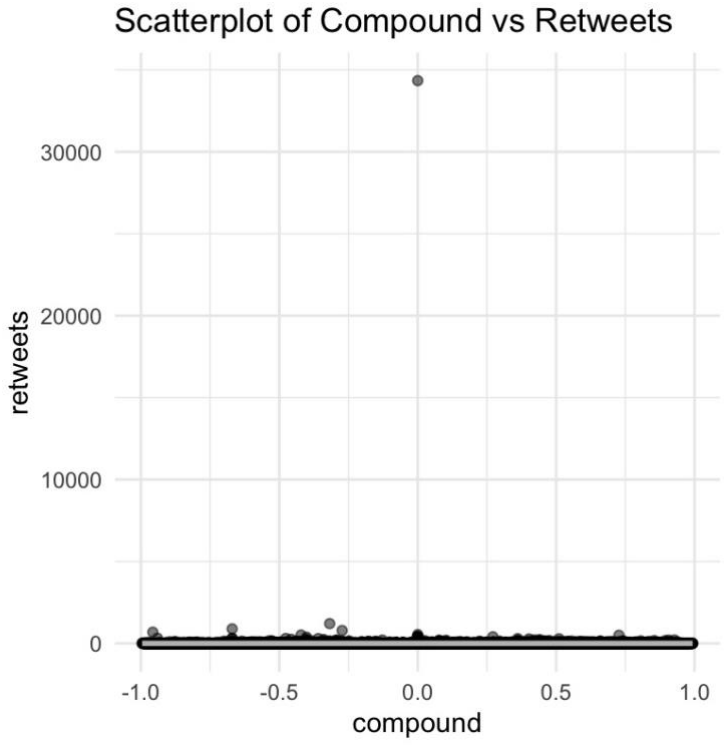


Figure D3

Scatterplot of the Correlation between Sentiment Compound and Replies with Outlier

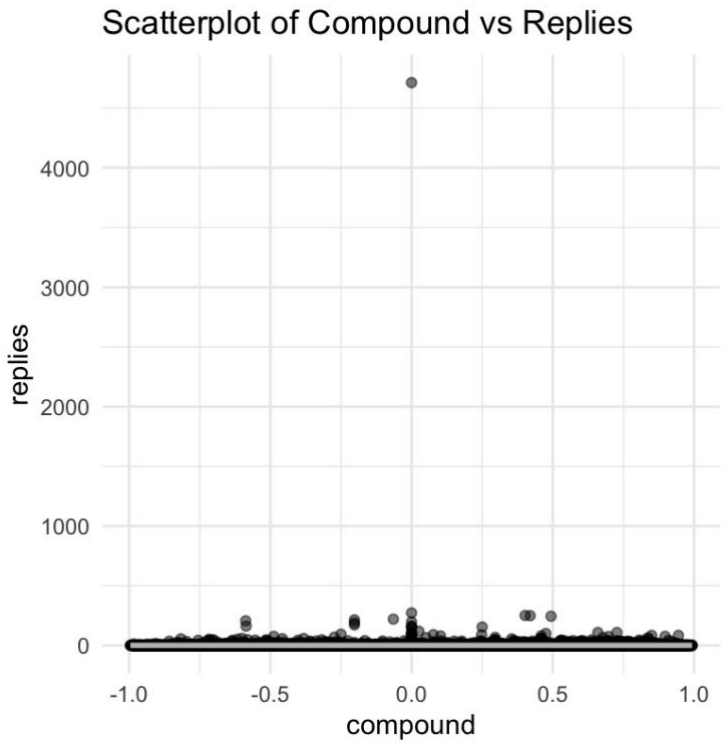


Figure D4

Scatterplot of the Correlation between Sentiment Compound and Quote Count with Outlier