Received: 15 January 2018 | Revised: 3 April 2018 | Accepted: 13 April 2018

DOI: 10.1002/eat.22878

WILEY

International Journal of EATING DISORDERS

ORIGINAL ARTICLE

Analyzing big data in social media: Text and network analyses of an eating disorder forum

Markus Moessner PhD[1] | Johannes Feldhege MSC[1]* | Markus Wolf PhD[2] | Stephanie Bauer PhD[1]

[1]Center for Psychotherapy Research, University Hospital Heidelberg, Heidelberg, Germany

[2]Department of Psychology, University of Zurich, Zurich, Switzerland

Correspondence
Dr. Markus Moessner, Center for Psychotherapy Research, University Hospital Heidelberg, Bergheimer Str. 54, 69115 Heidelberg, Germany.
Email: markus.moessner@med.uni-heidelberg.de

Abstract

Objective: Social media plays an important role in everyday life of young people. Numerous studies claim negative effects of social media and media in general on eating disorder risk factors. Despite the availability of big data, only few studies have exploited the possibilities so far in the field of eating disorders.

Method: Methods for data extraction, computerized content analysis, and network analysis will be introduced. Strategies and methods will be exemplified for an ad-hoc dataset of 4,247 posts and 34,118 comments by 3,029 users of the *proed* forum on Reddit.

Results: Text analysis with latent Dirichlet allocation identified nine topics related to social support and eating disorder specific content. Social network analysis describes the overall communication patterns, and could identify community structures and most influential users. A linear network autocorrelation model was applied to estimate associations in language among network neighbors. The supplement contains R code for data extraction and analyses.

Discussion: This paper provides an introduction to investigating social media data, and will hopefully stimulate big data social media research in eating disorders. When applied in real-time, the methods presented in this manuscript could contribute to improving the safety of ED-related online communication.

KEYWORDS

big data, eating disorders, social media, social network analysis, text analysis

1 | OBJECTIVE

Social media are an important part in young people's lives. All kinds of topics including health and mental health related issues are discussed online among peers, producing huge amounts of content and communication related data (see e.g., www.internetlivestats.com). A number of experimental studies investigated the effects of social media exposure on eating disorders (ED) or ED risk factors mostly in controlled laboratory settings (e.g., Cohen & Blaszczynski, 2015), several cross-sectional questionnaire studies explored the relationships between the use of social network sites and different kinds of (ED) pathology (Becker et al., 2011; Eckler, Kalyango, & Paasch, 2017; Murray, Maras, & Goldfield, 2016; Saffran et al., 2016; Sidani, Shensa, Hoffman, Hanmer, & Primack, 2016; Valkenburg, Koutamanis, & Vossen, 2017; Walker et al., 2015). Meta analyses provide further support for the relevance of (social) media for ED (Holland & Tiggemann, 2016; Mingoia, Hutchinson, Wilson, & Gleaves, 2017; Rodgers, Lowy, Halperin, & Franko, 2016). Yet, naturalistic studies of the processes and effects of social media in ED are still sparse.

Every day users write or upload billions of words via personal webpages, blogs, social media websites, or other digital communication channels for self-presentation, self-expression, and interpersonal exchange. Data of social media websites can be accessed through application programming interfaces (API). Digital communication provides researchers with a rich and steadily growing resource for

Markus Moessner and Johannes Feldhege should be considered joint first author

naturalistic analyses of human behavior, affect, and attitudes allowing unobtrusive, nonreactive observations in various nonclinical and clinical contexts, for example, in ED-related research (e.g., Wolf et al., 2007, 2013). Social media communication is text-based, ready transcribed and easily accessible. Powerful big data approaches for the retrieval, processing and quick and reliable automated analysis of language and communication are increasingly applied on social media data in the social sciences and medical and health-related research to learn about psychological structures and behaviors under real-world conditions (e.g., Golder & Macy, 2011; O'Dea et al., 2017; Schwartz et al., 2013).

Big data research differs from more traditional psychological research in a number of ways (Qiu, Chan, & Chan, 2018). First, big data research tends to be more data-driven. Most often, researchers extract and describe big data applying exploratory, bottom-up approaches. Yet, there are also theory-driven approaches to the extraction of big data from online sources (Landers, Brusso, Cavanaugh, & Collmus, 2016). Secondly, the extracted data are a lot noisier than the data collected in traditional research. Instead of carefully planning laboratory experiments or developing questionnaires in order to capture the phenomena relevant for the research question as precisely as possible, large amounts of data are extracted that often require extensive cleaning before any analysis can be conducted. Another aspect is that the storage of big datasets requires careful planning, because storage can be challenging (Chen & Wojcik, 2016). In addition, depending on the kind of analyses conducted, the computational power of regular computers might not be sufficient.

Various studies demonstrate the potential of big data. Researchers have employed big data methods in the study of depression support groups on the internet: they studied the language used in the groups (Carron-Arthur, Reynolds, Bennett, Bennett, & Griffiths, 2016; Gkotsis et al., 2017; Xu & Zhang, 2016), predicted users leaving the groups (Sadeque et al., 2016) or showed long-term effects of participation (Park & Conway, 2017). Other health-related issues that have been explored using big data methods include smoking cessation (Zhao et al., 2016), suicidal ideation (De Choudhury & Kiciman, 2017), cancer (Wang, Kraut, & Levine, 2015), HIV/AIDS (Wang, Shi, Chen, & Peng, 2016), and obesity (Chou, Prestin, & Kunath, 2014). A number of systematic reviews have been conducted on all kinds of topics related to research using social media, online communities, and support groups (Carron-Arthur, Ali, Cunningham, & Griffiths, 2015; Eysenbach, Powell, Englesakis, Rizo, & Stern, 2004; Moorhead et al., 2013; Seabrook, Kern, & Rickard, 2016; Sinnenberg et al., 2017; Wongkoblap, Vadillo, & Curcin, 2017).

Yet, there are only few studies that investigated ED-related research questions using a big data approach. Only recently, a sophisticated approach to detecting and characterizing ED communities on Twitter was published (Wang, Brede, Ianni, & Mentzakis, 2017). The authors found that ED communities on Twitter are tightly linked: ED users mostly communicate with and follow other ED users. By comparing the ED sample to two other samples, they showed that ED users have fewer interactions, are active for a shorter time, and express more negative emotions than the comparison samples (Wang et al., 2017). Other studies have analyzed the usage of weight loss apps by users with underweight BMI goals (Eikey et al., 2017), tags on Instagram posts to infer mental illness severity (Chancellor, Lin, Goodman, Zerwas, & De Choudhury, 2016) or modeled networks of Pro-ED websites (Casilli, Pailler, & Tubaro, 2013).

Pro-ED, also known as Pro-Ana for anorexia and Pro-Mia for bulimia, refers to material and content on the internet that promotes the development and maintenance of ED. Pro-ED content can be found on websites, forums, personal blogs (Borzekowski, Schenk, Wilson, & Peebles, 2010), and in communities on social networks (Juarascio, Shoaib, & Timko, 2010). Common features are motivational or "thinspiration" content, and "thin commandments", that is, instructions and techniques for maintaining an ED and rapid weight loss (Borzekowski et al., 2010; Lapinski, 2006; Norris, Boydell, Pinhas, & Katzman, 2006; Steakley-Freeman, Jarvis-Creasey, & Wesselmann, 2015). Experimentally induced exposure to Pro-ED websites has been found to be associated with increases in dieting, body dissatisfaction, and negative affect (for a review see Rodgers, Lowy, Halperin, & Franko, 2016). Similarly, surveys found higher body dissatisfaction, drive for thinness, perfectionism, and disordered eating behaviors in Pro-ED website users compared to non-users (Custers & Van den Bulck, 2009; Harper, Sperry, & Thompson, 2008; Peebles et al., 2012). Analyses of user generated content in Pro-ED communities show that it corresponds with markers of disordered eating behaviors (Sowles et al., 2018), and that mental illness severity increases over time (Chancellor, Lin, Goodman, et al., 2016). Users, however, also report various benefits of participation in Pro-ED communities: receiving social support, a sense of belonging, improved self-esteem, and feeling less alone (Csipke & Horne, 2007; Haas, Irr, Jennings, & Wagner, 2011; Ransom, La Guardia, Woody, & Boyd, 2010; Rodgers, Skowron, & Chabrol, 2012).

Politicians as well as the operators of social media websites and webhosts have recognized the harmful nature of Pro-ED and have taken action. In France, a law was passed that penalizes websites for the encouragement of excessive thinness and food intake restriction. Social media websites have begun to censor or remove Pro-ED content (Chancellor, Lin, & De Choudhury, 2016), and webhosts have used warning texts or advertisements in Pro-ED search results or websites to prevent people from visiting these sites (Martijn, Smeets, Jansen, Hoeymans, & Schoemaker, 2009; Yom-Tov, Brunstein-Klomek, Mandel, Hadas, & Fennig, 2018). However, users find ways to circumvent these efforts (Chancellor, Pater, Clear, Gilbert, & De Choudhury, 2016). More research is needed in order to identify and possibly change harmful online communities. Analyzing big data often requires specific methods. Two commonly applied approaches are computer-assisted analyses of language, and social network analysis.

1.1 | Computer-assisted analyses of natural language and large text corpora

Computerized quantitative text-analyses have a comparably long tradition in psychology and psychotherapy research with the first — usually top-down driven — word-count approaches being around since the appearance of personal computers in the late 60s (e.g., Gottschalk & Bechtel, 1993; Stone et al., 1966). Most recently, with the growing

computational power and the possibility to connect with large text corpora, bottom-up approaches applying machine learning or text-mining allow for "deeper" and more complex analyses of communication structure in social media (e.g., Boyd, 2017).

Roughly, methods of computerized text analysis aim to operationalize, measure and quantify human communication. Thus, they allow researchers to extract various linguistic, stylistic, or content features of interest from (digital) natural language samples, such as social media transcripts, amongst others:

- Linguistic and lexical features and categories on the word and sentence level, such as part-of-speech and certain word classes (e.g., nouns, verbs, pronouns, adverbs, prepositions, numbers, etc.), punctuation (e.g., question marks, exclamation marks, etc.), emoticons, or other corpus-linguistic features such n-grams.
- Lexical language markers for specific contents, semantic or meaning-related concepts, or psychological constructs such as cognitive processes or emotional tone, for example, the frequency of positive emotion words written in a text (such as happy, friend, love, etc.) as a proxy for positive affectivity.
- Common themes, topics, or semantic clusters, based on co-occurrences of words, or tokens, in a given text.

The first two groups of criteria refer to top-down text analyses that usually draw on a software that allows to categorize certain words or parts of the text according to an inbuilt algorithm or rule, or based on a dictionary that assigns single words, word stems or other text units to predefined language categories. A widespread used example for such a dictionary-based text-analysis program is the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015). Roughly speaking, the LIWC consists of a word count software and a dictionary that assigns >6,000 words or word stems (English version) to approximately 90 different linguistic and language categories, such as positive and negative emotion words, cognitive processes, or social processes (Pennebaker et al., 2015). While processing a text, the program compares each word of a given text with the words of the dictionary; if a word is a match it is automatically counted in the respective category. The results of the text analyses are then displayed as the relative frequency of a certain category relative to the total word count of that text. While LIWC is one of the most comprehensive dictionary-based programs, there are other programs and dictionaries around that might have a different conceptual focus, but their general approach is usually the same. By combining single extractable language features a researcher can increase the precision of predictive models. For instance, similar to the genetic fingerprint, "linguistic fingerprints" or footprints refer to a broad set of relevant linguistic features that allow to reliably identify a certain individual, text document, or genre, based on the distribution of its features, or profile (see Boyd, 2017). Such profiles can then be used to assess, for instance, a certain text's narrative coherence, its cognitive complexity, or determine its synchronicity, similarity or "match" with other texts (Ireland & Pennebaker, 2010).

While these approaches are mostly top-down, or theory driven (i.e., the categories are conceptually predefined proxy for a certain psychological or linguistic concept or construct), bottom-up or data driven approaches are focusing on word patterns or co-occurrences in the data structure. Primary aim is to derive models that explain or uncover a text's underlying, or latent, structure or content, for instance its most prevalent topics or themes. Important examples in this rapidly growing field are the latent semantic analysis (Landauer & Dumais, 1997), the meaning extraction method (Chung & Pennebaker, 2008), or latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003).

Latent semantic analysis and LDA (Blei, Ng, & Jordan, 2003) are unsupervised, bottom-up methods that do not require a-priori classifications of documents before conducting the analysis. They can produce short and descriptive summaries of each document and at the same time help discover underlying statistical relations between documents that can be used in further analyses. Topic modeling with LDA, has found applications in research on online communities and forums relating to mental disorders. Topic models have been used to differentiate between regular users and "super users" in an online support group for depression (Carron-Arthur, Reynolds, Bennett, Bennett, & Griffiths, 2016), to investigated the topics users in a weight loss forum on reddit.com talked about, and how the use of these topics related to actual weight loss (Pappa et al., 2017). Furthermore, a number of studies have used topics as features in machine learning processes to differentiate social media content from other online communities (Nguyen et al., 2017; Nguyen, Phung, Dao, Venkatesh, & Berk, 2014).

In LDA, a specified number of latent topics are modeled in a corpus of text documents. The combination of words in a document can be explained by underlying, unobserved topics. The method assumes that a document is generated as a mixture of topics according to a Dirichlet distribution and that each word in a document is drawn from a topic's multinomial distribution (Blei et al., 2003). Each word and each document has a probability of being generated by each topic. Using the observed documents and words, LDA estimates the latent topics that generated them. Frequency and co-occurrence of words in a document are relevant parameters in topic modeling whereas their order and syntax in the document are ignored (Roberts, Stewart, & Airoldi, 2016). For example, a document could contain multiple mentions of words like "binge" and "junk food" and one mention each of "therapist" and "inpatient". In that case the model would determine that the document is largely about a binge-eating topic with "binge" and "junk food" as words with high probability for the topic and to a smaller part about an eating disorder treatment topic with "therapist" and "inpatient" as words with high probability. The probability with which each document and each word can be assigned to each latent topic are the main parameters estimated in LDA. By reviewing the most characteristic words for each topic, determined by highest probability or other related measures, the content of the topics can be determined.

The LDA algorithm tries to find values for the model parameters with maximum likelihood using Bayesian inference methods, it is sensitive to the starting values (Roberts, Stewart, & Tingley, 2016). One method of dealing with this problem is to generate a number of candidate models with different starting values and discarding models based on semantic coherence and exclusivity. The final decision of how many topics to select usually needs to involve human judgment based on

model parameters for coherence, exclusivity, and interpretability of the model topics.

1.2 | Social network analysis

In addition to the content of communication, social media data allow to investigate social structures and communication patterns in online communities. Most influential users (e.g., influencers) can be identified and the kind of content they are contributing can be investigated. In addition, these network structures allow to analyze how content is disseminated in an online community. Finally, social network analysis (SNA) can be applied to investigate the effects and consequences of participating in such an online community.

Network analysis models relational patterns between a predefined set of elements (network nodes). In this case, the elements are social media users. Relations between nodes are called edges. These edges can be directed (from-to relations) or undirected (connected yes vs. no), and weighted vs. unweighted. On the basis of the structure of communication, SNA can help to identify important players in the communication network (centrality measures), and individuals who are crucial for the dissemination of potentially harmful and/or helpful content. Communities within the network can be identified and analyzed. Communities are groups of nodes that are densely connected with each other and less connected to nodes outside the community (see Girvan & Newman, 2002).

In addition to describing the communication patterns, other properties can be attached to network nodes. These could be sociodemographic variables (e.g., age, gender), ED-related impairment, the language of users (e.g., the results of topic modelling), or any other available property. On the basis of the network structure and the node properties, regression methods can be applied to research questions referring to clustering, social contagion, or causal effects within a network. Instead of just assuming that writing and/or reading thinspiration content negatively influences others, one could actually test whether this is the case. Investigating social contagion and interpersonal influences in social networks is closely related to the work of Christakis and Fowler, who developed methods to identify causal effects in networks, and conducted a number of studies in different health-related areas (see Christakis & Fowler, 2013), including obesity (Christakis & Fowler, 2007). On the basis of the structure of the connections within the network, network influence models can be applied to model interdependencies and even causal effects within the network (O'Malley, 2013). Effects take place via network edges, and the general idea is to explain or predict a dependent variable by a number of predictors/explanatory variables that execute their effects via network edges. A weight matrix quantifies the total influence acting on a person (node). The weight matrix can be directly derived from the network's adjacency matrix (O'Malley, 2013). It might be the case that not all members of a network are assumed to be equally influential. These assumptions can be modelled by including different weight matrices for different kinds of nodes (celebrities, etc.) or by running separate models for each type of peer (O'Malley, 2013). Indirect influences can be included in the models, too. A detailed description of regression models in SNA would be beyond the scope of this article. Interested readers will find detailed information on different model types elsewhere (Christakis & Fowler, 2013; O'Malley, 2013; O'Malley & Marsden, 2008 provide R-code for fitting different kinds of models).

This manuscript outlines a big data approach to social media. We will provide an introduction to (a) data extraction, (b) the analysis of contents, (c) the analysis of communication patterns, and (d) social contagion. We will exemplify the approach by analyzing an ad-hoc dataset from the pro-ED forum of a social media website (Reddit).

2 | METHOD

The procedures and analyses reported in this article were conducted in R (R Development Core Team, 2017). R benefits from having user-created packages that provide functionality such as authenticating with APIs, extracting data from social media websites, and analyzing the extracted data (for a list of R packages that are useful for these tasks see appendix A). A number of R packages have been developed for specific social media websites (such as Facebook, Instagram, Twitter) that include functions to extract data based on user-set criteria. Appendix B contains the R code for data extraction and the analyses conducted in this article.

2.1 | Data extraction

For this article, we extracted data from the *proed* forum on www.reddit.com. With 542 million monthly visitors (234 million unique visitors), Reddit is ranked as the fourth most visited website in the US and ninth worldwide in 2017 (Wikipedia; 09.11.2017, https://en.wikipedia.org/wiki/Reddit). It contains over 11,400 active user-created forums (subreddits), that are formed around specific common interests such as sports, movies, or mental disorder. Within the ED-related subreddits, *proed* was the most active with 18,305 subscribers (www.reddit.com/r/proed 12.01.2018). The official description of the *proed* subreddit is: "This is a support subreddit for those who are suffering with an ED or disordered eating behaviors but are not ready for recovery".

A linux server with an instance of an RStudio Server and a cron implementation (i.e., a time-based job scheduler) were set up to execute the data extraction scripts to access Reddit's API. New posts and comments in the *proed* subreddit were downloaded every five minutes for two months. Data were written to a MySQL database using the DBI and RMySQL packages. The extracted dataset included N = 40,048 comments and N = 4,395 posts by N = 3,193 users. Comments that did not have a corresponding post in the dataset, and comments and posts by bot accounts were deleted. Data from users that had incomplete data in any relevant variables were removed from the dataset. The final dataset contains N = 4,247 posts and N = 34,118 comments by N = 3,029 users.

2.2 | Data analysis

2.2.1 | Latent Dirichlet allocation

The analysis of the *proed* subreddit was conducted on the user level in order to estimate the prevalence of each topic for each user. To

prepare the data for the analysis, all comments and posts of each user were concatenated in a single document. Posts and comments in a language other than English were removed. Several text transformation steps were necessary to bring the documents into a suitable form for topic modelling using the tm package. All text was converted to lower case. HTML code, hyperlinks, punctuation marks, and numbers were removed from the documents. A stopword list from the Snowball stemmer project included in the tm package (Feinerer & Hornik, 2017) was used to remove common words such as "the" or "a", as removing stopwords has been shown to increase the coherence of topics (Schofield, Magnusson, & Mimno, 2017). Furthermore, words consisting of just one letter or words that appeared in only one document were removed. Several users were removed because their texts were entirely made up of text elements that were deleted in the text transformation process. The dataset for topic modelling contains texts by $N = 3{,}009$ users. The latent Dirichlet allocation implementation in the stm package (Roberts, Stewart, & Tingley, 2017) was applied as it allows for correlated topics and the optional addition of external covariates.

2.3 | Model selection

A main question in topic model analyses is the number of topics to model. There are different methods of determining the optimal number of topics. We followed recommendations by Roberts et al. (2014) in focusing on semantic coherence and exclusivity to narrow down the number of candidate models. Words can have high probabilities in different topics. Semantic coherence measures the degree to which words that are characteristic for a model's topic are jointly present in documents. Exclusivity is low in cases in which words are characteristic for multiple topics. In general, an increasing number of topics in a model yields decreases in semantic coherence and increases in exclusivity.

Initially, models with number of topics $K = 5, 10 \ldots 50$ were estimated and compared based on the average exclusivity and semantic coherence scores. For models with K larger than 30 very small increases in exclusivity came with large decreases in semantic coherence. Therefore, in a second run models with $K = 3, 6 \ldots 30$ topics were estimated and evaluated. Here, models with K between 6 and 15 topics showed the highest values in both parameters. The next run with $K = 6$–15 topics offered two models with the best combination of the two parameters, $K = 9$ and 11 topics. For $K = 9$ and 11, 50 models with different initializations were estimated. Out of the ten models with the highest likelihood, the models with the best scores on average exclusivity and semantic coherence for $K = 9$ and 11 were selected.

The two final models were evaluated manually by looking at the most characteristic words and documents for each topic in order to determine whether they were intelligible. Finally, based on interpretability the model with $K = 9$ topics was selected. Topic labels were determined based on the FREX metric. The FREX metric is the harmonic mean of a word's ranks probability and exclusivity. As a result, it yields more distinct topic descriptions (Roberts et al., 2017).

2.4 | Social network analysis

A weighted, undirected network was modelled with users as network nodes. Two nodes were considered connected, when they were part of the same conversation (i.e., thread). The edge weight between two nodes was defined as the number of common conversations. Network density was calculated as the sum of absolute edge weights divided by the number of possible edges, that is, density describes the mean edge weight across the network. The higher a network's density the stronger are the interconnections between its nodes. Network centrality measures allow the identification of key players within the reddit *proed* forum. There are a variety of measures to assess the centrality of a user within a network. Depending of the specific process under investigation, different measures might be more appropriate. Common centrality measures are degree (for undirected networks the sum of a node's edge weights), closeness centrality (reflects the invert of average length of the shortest paths from this node to all other nodes), and betweenness centrality (the number of shortest paths between two other nodes in which this node is included). For this study, we analyzed degree and betweenness centrality, and provide plots for the centrality distributions, and scatter plots of centralities × LDA topics. A network community is a subset of nodes which are strongly interconnected. Different algorithms to analyze communities are available. For this example we applied the algorithm implemented in Gephi (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). On the basis of the overall structure of the network, and the availability of node properties (LDA results), the clustering of specific properties within the network can be explored by plotting the graph and color-coding the nodes according to the property of interest.

2.5 | Linear network autocorrelation model

A linear network autocorrelation model was estimated (see the following equation):

$$y_i = W * y_i + X * \beta + \varepsilon \quad ([1])$$

The LDA topic 6 "Social Support" was explained by (a) the other eight LDA topics as covariates (X) and the adjacency weighted "Social Support" ($W^* y_i$). The weight matrix W was derived through devision of the adjacency matrix by the row sums (in this case the row sums are equal to the degree centralities). The adjacency weighted "Social Support" ($W^* y_i$) models the effect of the "Social Support" usage of direct neighbors in the network on a user's "Social Support" usage. Several software solutions for SNA are available. Analyses for this paragraph were performed with the R package igraph and sna (see appendix B for R code), and the open-source software Gephi (https://gephi.org/).

3 | RESULTS

3.1 | Latent Dirichlet allocation

The final model contains $K = 9$ topics, which were manually annotated with a label by studying the most characteristic words and the most representative documents for the topic. The most characteristic words,

TABLE 1 Main result of topic modelling: Topic labels, FREX, and high probability words

Annotated labels	FREX words	High probability words
Thinspiration and Appearance	Hip im dont ribcage kms (= kill myself) your smol don waist isnt	Like look lol just body fuck weight im people fat
Treatment and Recovery	Anorexia suffer behavior therapist valid treatment disorder recovery harm sub	People ED disorder can person body go think self see
Prescription Drugs and their abuse	mg stimulant Adderall stack Ephedrine ec Bronkaid Wellbutrin med blood	Take help can go use time will also run try
Weight Gain/Loss	kg lbs pound gain lost scale lose weight goal cycle	Weight go binge lose day now start gain week lbs
Binge eating and high calorie foods	Sunday lunch Cheeto pizza dinner poop fri bag tonight chip	Day go week binge today one food last fuck ate
Feedback and social support	Sorry know feel really thank think hope better honest okay	Like feel just know really think get much make try
Family	Brother son friend mom dad phone sister met text husband	Friend get go time want year love eat look one
Low Calorie Foods	Flavor spice syrup tuna onion stevia unsweetened Splenda mustard roast	Like cal calorie can use one tea make also milk
Nutrition	Intake fast hungry eat food burn avoid meal Keto junk	Eat food calorie fast day get water drink restrict work

as calculated by the highest probability and FREX (Airoldi & Bischof, 2016), and the annotated label of each topic can be seen in Table 1.

The proportion of topics in the corpus is displayed in Figure 1. The most common topic of discussion is giving and receiving social support and expressing one's feelings. The second most common topic is centered on weight, weight loss, or gain. The next four topics are roughly equal in prevalence and are concerned with family, treatment and recovery, low calorie foods, and nutrition. Less common topics are about appearance and thinspiration, binge eating and high calorie foods, and prescription drugs and their abuse. The prescription drugs topic contains mentions of prescription drugs (e.g., Wellbutrin, Adderall, Bronkaid) that have appetite suppression as a side effect. The topic "Thinspiration and Appearance" was labelled this way because it contains references to body parts like hip, ribcage, or waist that are focused on in an effort to look thinner. The topic "Low Calorie Foods" is characterized by mentions of two sweeteners, "stevia" and "Splenda", and of "spice" and "syrup" that can give flavor to low calorie foods. Table 2 shows intercorrelations between the topics.

Correlations show which topics are combined, for example, treatment topics are not commonly discussed in the context of eating and nutrition. Topic 2 "Treatment and Recovery" is negatively correlated with topic 5 "Binge eating and High Calorie Foods", topic 8 "Low Calorie Foods", and topic 9 "Nutrition", and positively correlated with topic 6 "Feedback and Social Support".

3.2 | Communication patterns

The *proed* subreddit network has a size of 3,029, that is, it contains 3,029 nodes. The network's density is $d = 0.06$. Figure 2 displays the degree distribution of the network (in this case, degree is the number of threads, in which a user posted, see below).

The degree distribution shows that the majority of users contribute to only a few threads, but that there is a small subgroup of very active users that contributed to >1,000 threads. This subgroup is responsible for the majority of content, and the overall activity in the forum. Within this network three communities could be detected, consisting of 1,263, 1,146, and 360 users. All other communities consisted of <150 users (see Supporting Information Figure S1 in the supplement).

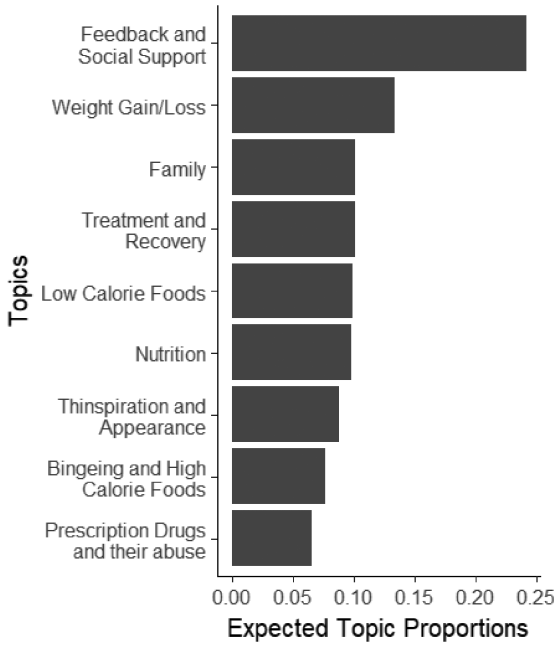

FIGURE 1 Expected topic proportions of a model with $K = 9$ topics in a corpus of posts and comments ($N = 3,009$)

TABLE 2 Spearman-correlations of topics in $N = 3009$ documents in the topic model with $K = 9$ topics

Topics		1	2	3	4	5	6	7	8	9
1	Thinspiration and Appearance	–	−0.02	−0.07***	0.02	−0.16***	0.05†	0.22***	−0.06***	−0.16***
2	Treatment and recovery	−0.02	–	0.09***	0.03	−0.61***	0.35***	0.17***	−0.43***	−0.30***
3	Prescription drugs and their abuse	−0.07***	0.09***	–	0.07***	−0.03	0.01	−0.12***	0.04	0.33***
4	Weight gain/loss	0.02	0.03	0.07***	–	−0.12***	0.28***	−0.11***	−0.45***	0.31***
5	Binge eating and high calorie foods	−0.16***	−0.61***	−0.03	−0.12***	–	−0.04	0.09***	0.51***	0.40***
6	Feedback and social support	0.05†	0.35***	0.01	0.28***	−0.04	–	0.21***	−0.33***	0.03
7	Family	0.22***	0.17***	−0.12***	−0.11***	0.09***	0.21***	–	−0.11***	−0.44***
8	Low calorie foods	−0.06***	−0.43***	0.04	−0.45***	0.51***	−0.33***	−0.11***	–	0.23***
9	Nutrition	−0.16***	−0.30***	0.33***	0.31***	0.40***	0.03	−0.44***	0.23***	–

Note. † $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

3.3 | Combined LDA and SNA results

Figure 3 displays scatter plots for normalized degree centrality and the relative frequency of the nine LDA topics (see Supporting Information Figure S4 in the supplement for betweenness centrality scatter plots).

Based on the scatter plots, users that are crucial for the dissemination of specific topics can be identified. These users are in the upper right quadrant of the bivariate distributions. Supporting Information Figures S2 and S3 explore the clustering of the two LDA topics "Thinspiration and Appearance" and "Feedback and Social Support" within one of the communities.

Table 3 displays the results of the network autocorrelation model. In addition to the other eight LDA topics (covariates), the proportion of the topic "Social Support" is associated with the proportion of this topic among the user's direct neighbors in the network.

4 | DISCUSSION

The manuscript outlines a big data social media research strategy. It exemplifies how to extract data from social media websites, and demonstrates methods to automatically analyze content and communication patterns for the *proed* forum on Reddit, one of the most active social media websites. The extraction of data via APIs is convenient, feasible, and does not require elaborate programming skills. Open source software packages are available for the majority of social media platforms. There are other approaches to analyze big data, the methods illustrated in this article do not cover all of them. Both computerized language analysis and SNA of complex networks are elaborate methods, the manuscript only gives a brief introduction. Yet, the manuscript provides a good starting point for readers who plan to analyze social media data. The software packages in appendix A and the R code in appendix B will be valuable resources to get started.

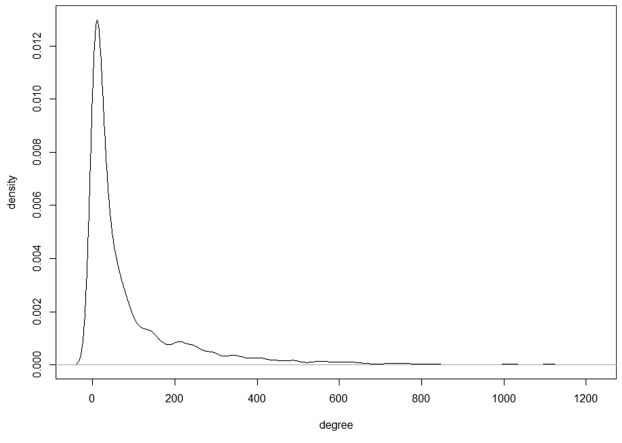

FIGURE 2 Degree centrality distribution of the communication network ($N = 3,029$)

The results presented in this manuscript are preliminary results within an ad-hoc dataset. The aim was not to report results but to demonstrate strategies and the potential of big data approaches in social media. Yet, the topics extracted in the LDA approach are plausible, and meaningful. The most frequent topic "Feedback and Social Support" is not ED specific, but highlights that interacting with peer and peer support are important reasons to engage in online exchange (Juarascio, Shoaib, & Timko, 2010). Additionally, a number of topics more characteristic of ED were extracted from the data, which highlights the heuristic value that these automated, unsupervised approaches can generate in big datasets. Correlations between the within person frequencies of topics can provide insights about the contexts in which these topics are mentioned. Results like these can provide valuable information on how to approach users on social media and facilitate help-seeking behavior.

SNA provided descriptive information on the communication structures in the *proed* forum. The degree distribution illustrated the huge variance in the way users contribute contents. While some users only write few posts, the most active ones contributed >1,000.

On a node level, these network metrics (centralities) can help to identify key players. Merging node properties to the network opens plenty of opportunities for analysis. We demonstrated how to identify key players, and outlined a strategy to investigate clustering graphically. Yet, the potential of combining SNA with node properties is most striking when regression models are applied. Based on the network structure modelled within the SNA, individual outcome regression models

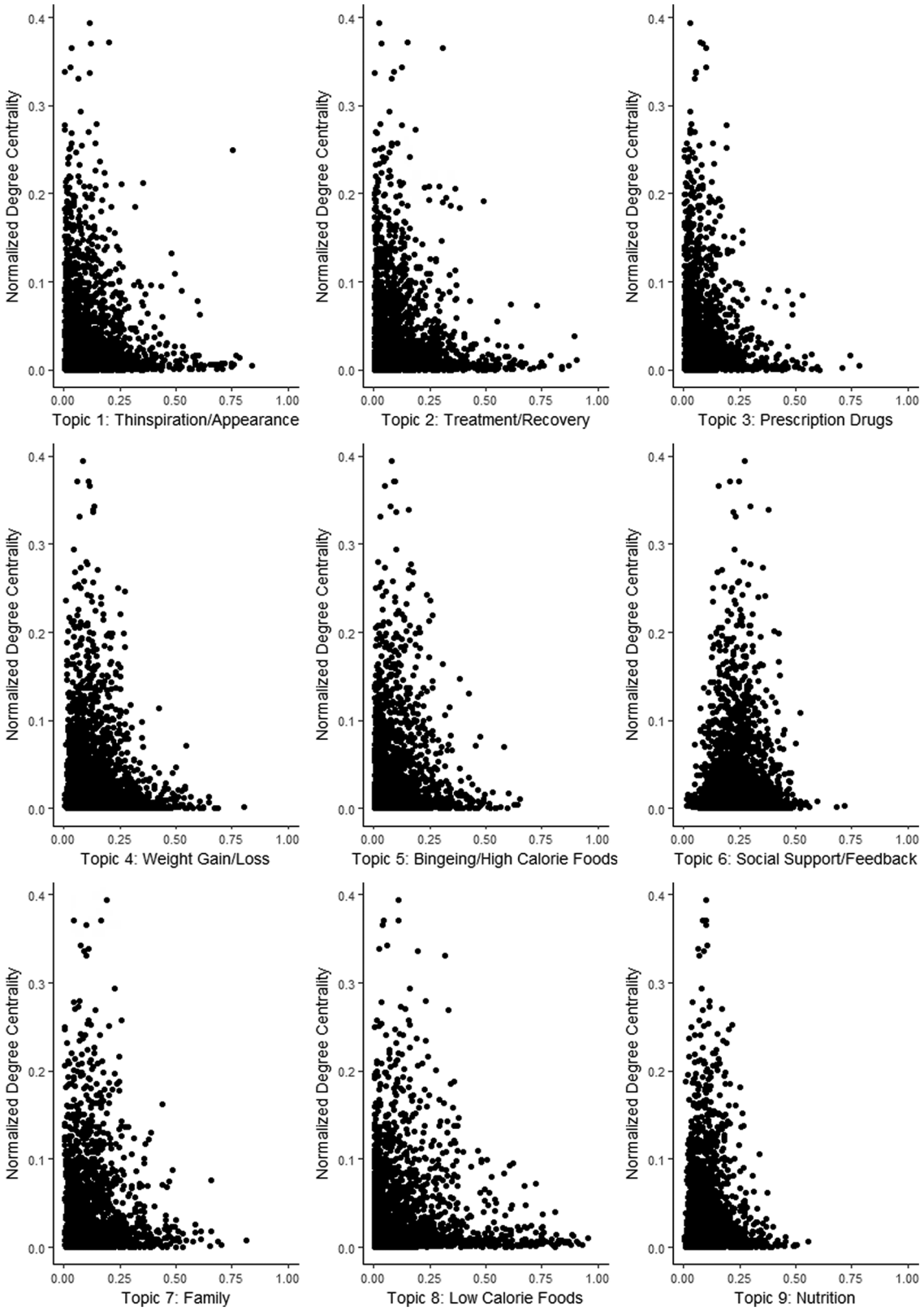

FIGURE 3 Scatter plots of normalized degree centrality and topic proportions ($N = 3{,}009$)

can be used to research social contagion and causal effects within the network. Without SNA these kinds of research questions cannot be addressed convincingly, because SNA models the communication pathways on which these effects take place. These analyses go far beyond traditional studies that compare social media users to a sample of not-users, or studies that investigate changes in a user's properties (e.g., ED-related impairment) over the course of participation in an online community. Because these methods actually allow to model the

TABLE 3 Autoregressive outcomes for LDA topic 6 „*Social Support*"

Term		Estimate	Std. Error	Z–value	p
1	Thinspiration and Appearance	−0.045	0.012	−3.831	<.001
2	Treatment and Recovery	0.087	0.011	7.955	<.001
3	Prescription Drugs and their Abuse	−0.087	0.016	−5.565	<.001
4	Weight Gain/Loss	0.088	0.012	7.146	<.001
5	Binge eating and high calorie foods	−0.001	0.016	−0.087	.930
7	Family	0.116	0.013	8.720	<.001
8	Low calorie foods	−0.149	0.009	−17.273	<.001
9	Nutrition	0.147	0.020	7.468	<.001
	Adjacency weighted "*Social Support*"	0.886	0.013	67.977	<.001

Note. Adjacency weight matrix, adjusted R-squared = 0.2551.

processes that cause these changes. Although their potential is obvious, they require data on node properties, which might not be extractable from social media data alone. We demonstrated the general approach by applying a linear autoregressive outcome model. The results show that controlled for a person's own language, its language is correlated with its direct neighbors' in the network. ED-related measures like ED pathology, ED-related behaviors (e.g., binge eating, vomiting, etc.), or ED risk factors (e.g., body dissatisfaction, etc.) would be more interesting than properties that can be extracted from social media data alone (e.g., the LDA results). Additionally, longitudinal data would allow to analyze changes and time-lagged effects. Yet, these data are difficult to collect and require additional assessment strategies. Given that the article aims at introducing methods to analyze big data, the example provides a good starting point although LDA topics might not be the most interesting node properties in this context.

In every model building process, a series of decisions have to be made. Usually, there is no clear right or wrong, and modelling decisions are based on the study's objectives and weighing advantages and disadvantages. The same is true for the analyses we illustrated in this manuscript. The model selection approach we presented is methodologically sound, intelligible, and reproducible. Yet, other approaches exist, and although we are convinced that we applied the best available strategy, there is no clear consensus within the scientific community. Within the SNA approach, we modelled undirected network edges. Users that wrote a post or comment in the same thread were considered connected. The weight of that connection was defined by the number of shared threads. Yet, these network edges do not have to be undirected. In most cases the comments section in online forums allows direct replies to other comments and often even contains a hierarchical structure. One could define comments to posts as "directed", this would change the model structure significantly. There are good arguments for directed networks. We decided to model undirected edges, because we think that reading a harmful post can be harmful or influential even if the post was not written as a reply to an own post. Within a directed network, regression models would not take these into account. Additionally, there are numerous centrality measures, algorithms to identify communities in networks, and ways to define a weight matrix. Which algorithms and measures are the best depends in the end on the research question and the properties of the data. In that sense, modelling decisions should be guided by the research question.

Finally, the approaches outlined in this manuscript are predominantly exploratory, and the results are mainly descriptive. Some of the descriptive results (e.g., network density) would profit from the availability of comparative data, because their interpretation is challenging. Nevertheless, given the general lack of big data approaches to social media in ED, descriptive, explorative approaches are an appropriate first step.

Social media are of great importance in the daily life of young people. Especially for the field of ED, their potentially harming effects on ED-related risk factors are often highlighted in the literature. Yet, big data approaches to social media are still sparse. We hope that this introductory illustration might motivate others and initiate more studies on this important topic. From a public health point of view, applications focusing on identifying and counteracting influencial and potentially harmful users on social media sites are promising. Additionally, big data approaches of social media sites might help to identify nontreatment seeking individuals, and facilitate referral to routine care.

5 | CLINICAL AND POLICY APPLICATIONS

Social media websites play an important role in everyday life, and can have both beneficial and harmful effects in adolescents and young adults (Campaioli, Sale, Simonelli, & Pomini, 2017). They can introduce users to and reinforce potentially harmful behaviors such as abusing prescription drugs or viewing and identifying with "thinspiration" materials. On the other hand, they can also be a space for receiving social support from peers and encouragement to seek treatment.

Form a public health point of view, beneficial content and forums should be promoted. Potentially harmful online communication should be prevented or reduced. The approaches illustrated in this article can help to identify harmful communities, and key users that have a central role in the dissemination of harmful content.

Text analyses can be helpful in identifying potentially harmful posts and comments. Network analyses can help to identify starting points to disrupt harmful online communities. When applied in real-time, they could become useful tools for forum moderators to improve the safety of ED-related online communication.

ORCID

Markus Moessner http://orcid.org/0000-0002-1215-1055

REFERENCES

Airoldi, E. M., & Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, *111*(516), 1381–1403. https://doi.org/10.1080/01621459.2015.1051182

Becker, A. E., Fay, K. E., Agnew-Blais, J., Khan, A. N., Striegel-Moore, R. H., & Gilman, S. E. (2011). Social network media exposure and adolescent eating pathology in Fiji. *British Journal of Psychiatry*, *198*(01), 43–50. https://doi.org/10.1192/bjp.bp.110.078675

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Blondel, V. D., Guillaume, J. -L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borzekowski, D. L., Schenk, S., Wilson, J. L., & Peebles, R. (2010). e-Ana and e-Mia: A content analysis of pro-eating disorder web sites. *American Journal of Public Health*, *100*(8), 1526–1534. https://doi.org/10.2105/ajph.2009.172700

Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In S. Hai-Jew, (Ed.), *Data analytics in digital humanities. Multimedia systems and applications*. Cham, Switzerland: Springer. pp. 161–189. https://doi.org/10.1007/978-3-319-54499-1_7

Campaioli, G., Sale, E., Simonelli, A., & Pomini, V. (2017). The dual value of the web: Risks and benefits of the use of the internet in disorders with a self-destructive component in adolescents and young adults. *Contemporary Family Therapy*, *39*(4), 301–313. https://doi.org/10.1007/s10591-017-9443-9

Carron-Arthur, B., Ali, K., Cunningham, J. A., & Griffiths, K. M. (2015). From help-seekers to influential users: A systematic review of participation styles in online health communities. *Journal of Medical Internet Research*, *17*(12), e271. https://doi.org/10.2196/jmir.4705

Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., & Griffiths, K. M. (2016). What's all the talk about? Topic modelling in a mental health internet support group. *BMC Psychiatry*, *16*(1), 367. https://doi.org/10.1186/s12888-016-1073-5

Casilli, A. A., Pailler, F., & Tubaro, P. (2013). Online networks of eating-disorder websites: Why censoring pro-ana might be a bad idea. *Perspectives in Public Health*, *133*(2), 94–95. https://doi.org/10.1177/1757913913475756

Chancellor, S., Lin, Z., & De Choudhury, M. (2016). "This Post Will Just Get Taken Down": characterizing removed pro-eating disorder social media content. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp 1157–1162). https://doi.org/10.1145/2858036.2858248

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp 1171–1184). https://doi.org/10.1145/2818048.2819973

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp 1201–1213). https://doi.org/10.1145/2818048.2819963

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Chou, W-y. S., Prestin, A., & Kunath, S. (2014). Obesity in social media: A mixed methods analysis. *Translational Behavioral Medicine*, *4*(3), 314–323. https://doi.org/10.1007/s13142-014-0256-1

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, *357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*(4), 556–577. https://doi.org/10.1002/sim.5408

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, *42*(1), 96–132. https://doi.org/10.1016/j.jrp.2007.04.006

Cohen, R., & Blaszczynski, A. (2015). Comparative effects of Facebook and conventional media on body image dissatisfaction. *Journal of Eating Disorders*, *3*(1), 23. https://doi.org/10.1186/s40337-015-0061-3

Csipke, E., & Horne, O. (2007). Pro-eating disorder websites: Users' opinions. *European Eating Disorders Review*, *15*(3), 196–206. https://doi.org/10.1002/erv.789

Custers, K., & Van den Bulck, J. (2009). Viewership of pro-anorexia websites in seventh, ninth and eleventh graders. *European Eating Disorders Review*, *17*(3), 214–219. https://doi.org/10.1002/erv.910

De Choudhury, M., & Kıcıman, E. (2017). The language of social support in social media and its effect on suicidal ideation risk. *Proceedings of the. International Aaai Conference on Weblogs and Social Media. International AAAi Conference on Weblogs and Social Media*, *2017*, 32–41.

Eckler, P., Kalyango, Y., & Paasch, E. (2017). Facebook use and negative body image among U.S. college women. *Women & Health*, *57*(2), 249–267. https://doi.org/10.1080/03630242.2016.1159268

Eikey, E. V., Reddy, M. C., Booth, K. M., Kvasny, L., Blair, J. L., Li, V., & Poole, E. S. (2017). Desire to be underweight: exploratory study on a weight loss app community and user perceptions of the impact on disordered eating behaviors. *JMIR Mhealth and Uhealth*, *5*(10), e150. https://doi.org/10.2196/mhealth.6683

Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., & Stern, A. (2004). Health related virtual communities and electronic support groups: Systematic review of the effects of online peer to peer interactions. *BMJ*, *328*(7449), 1166. https://doi.org/10.1136/bmj.328.7449.1166

Feinerer, I., & Hornik, K. (2017). *tm: Text Mining Package* (Version 0.7-3). Retrieved from https://CRAN.R-project.org/package=tm

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, *7*, 45141. https://doi.org/10.1038/srep45141

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881. https://doi.org/10.1126/science.1202775.

Gottschalk, L. A., & Bechtel, R. J. (1993). *Psychologic and neuropsychiatric assessment survey: Computerized content analysis of natural language or verbal texts*. Palo Alto, CA: Mind Garden.

Haas, S. M., Irr, M. E., Jennings, N. A., & Wagner, L. M. (2011). 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

Harper, K., Sperry, S., & Thompson, J. K. (2008). Viewership of pro-eating disorder websites: Association with body image and eating disturbances. *International Journal of Eating Disorders*, *41*(1), 92–95. https://doi.org/10.1002/eat.20408

Holland, G., & Tiggemann, M. (2016). A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. *Body Image*, *17*, 100–110. https://doi.org/10.1016/j.bodyim.2016.02.008

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, *99*(3), 549–571. https://doi.org/10.1037/a0020386.

Juarascio, A. S., Shoaib, A., & Timko, C. A. (2010). Pro-eating disorder communities on social networking sites: A content analysis. *Eating Disorders*, *18*(5), 393–407. https://doi.org/10.1080/10640266.2010.511918

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/met0000081

Lapinski, M. K. (2006). StarvingforPerfect.com: A theoretically based content analysis of pro-eating disorder web sites. *Health Communication*, *20*(3), 243–253. https://doi.org/10.1207/s15327027hc2003_4

Martijn, C., Smeets, E., Jansen, A., Hoeymans, N., & Schoemaker, C. (2009). Don't get the message: The effect of a warning text before visiting a proanorexia website. *International Journal of Eating Disorders*, *42*(2), 139–145. https://doi.org/10.1002/eat.20598

Mingoia, J., Hutchinson, A. D., Wilson, C., & Gleaves, D. H. (2017). The relationship between social networking site use and the internalization of a thin ideal in females: A meta-analytic review. *Frontiers in Psychology*, *8*, 1351. https://doi.org/10.3389/fpsyg.2017.01351

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. (2013). A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research*, *15*(4), e85. https://doi.org/10.2196/jmir.1933

Murray, M., Maras, D., & Goldfield, G. S. (2016). Excessive time on social networking sites and disordered eating behaviors among undergraduate students: Appearance and weight esteem as mediating pathways. *Cyberpsychology, Behavior, and Social Networking*, *19*(12), 709–715. https://doi.org/10.1089/cyber.2016.0384

Nguyen, T., O'Dea, B., Larsen, M., Phung, D., Venkatesh, S., & Christensen, H. (2017). Using linguistic and topic analysis to classify subgroups of online depression communities. *Multimedia Tools and Applications*, *76*(8), 10653–10676. https://doi.org/10.1007/s11042-015-3128-x

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., & Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, *5*(3), 217–226. https://doi.org/10.1109/TAFFC.2014.2315623

Norris, M. L., Boydell, K. M., Pinhas, L., & Katzman, D. K. (2006). Ana and the internet: A review of pro-anorexia websites. *International Journal of Eating Disorders*, *39*(6), 443–447. https://doi.org/10.1002/eat.20305

O'Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A linguistic analysis of suicide-related twitter posts. *Crisis*, *38*(5), 319–329. https://doi.org/10.1027/0227-5910/a000443.

O'Malley, A. J. (2013). The analysis of social network data: An exciting frontier for statisticians. *Statistics in Medicine*, *32*(4), 539–555. https://doi.org/10.1002/sim.5630

O'Malley, A. J., & Marsden, P. V. (2008). The analysis of social networks. *Health Services and Outcomes Research Methodology*, *8*(4), 222–269. https://doi.org/10.1007/s10742-008-0041-z

Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Couto Silva, A. P., Meira, W., Jr., & Beleigoli, A. M. R. (2017). Factors associated with weight change in online weight management communities: A case study in the LoseIt Reddit Community. *Journal of Medical Internet Research*, *19*(1), e17. https://doi.org/10.2196/jmir.5816

Park, A., & Conway, M. (2017). Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online depression community. *Journal of Medical Internet Research*, *19*(3), e71. https://doi.org/10.2196/jmir.6826

Peebles, R., Wilson, J. L., Litt, I. F., Hardy, K. K., Lock, J. D., Mann, J. R., & Borzekowski, D. L. G. (2012). Disordered eating in a digital age: eating behaviors, health, and quality of life in users of websites with pro-eating disorder content. *Journal of Medical Internet Research*, *14*(5), e148. https://doi.org/10.2196/jmir.2023

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science*, *1*(1), 59–66. https://doi.org/10.1007/s42001-017-0013-6

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ransom, D. C., La Guardia, J. G., Woody, E. Z., & Boyd, J. L. (2010). Interpersonal interactions on online forums addressing eating concerns. *International Journal of Eating Disorders*, *43*(2), 161–170. https://doi.org/10.1002/eat.20629

Roberts, M. E., Stewart, B., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. 51–97). New York: Cambridge University Press.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2017). *stm: R Package for Structural Topic Models*. Retrieved from http://www.structuraltopicmodel.com

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rodgers, R. F., Lowy, A. S., Halperin, D. M., & Franko, D. L. (2016). A meta-analysis examining the influence of pro-eating disorder websites on body image and eating pathology. *European Eating Disorders Review*, *24*(1), 3–8. https://doi.org/10.1002/erv.2390

Rodgers, R. F., Skowron, S., & Chabrol, H. (2012). Disordered eating and group membership among members of a pro-anorexic online community. *European Eating Disorders Review*, *20*(1), 9–12. https://doi.org/10.1002/erv.1096

Sadeque, F., Pedersen, T., Solorio, T., Shrestha, P., Rey-Villamizar, N., & Bethard, S. (2016). Why do they leave: Modeling participation in online depression forums. *Proceedings of the 4th Workshop on Natural Language Processing and Social Media* (pp 14–19).

Saffran, K., Fitzsimmons-Craft, E. E., Kass, A. E., Wilfley, D. E., Taylor, C. B., & Trockel, M. (2016). Facebook usage among those who have received treatment for an eating disorder in a group setting. *International Journal of Eating Disorders*, *49*(8), 764–777. https://doi.org/10.1002/eat.22567

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (432–436). https://doi.org/10.18653/v1/E17-2069

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, *25*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

Seabrook, E. M., Kern, M. L., & Rickard, N. S. (2016). Social networking sites, depression, and anxiety: A systematic review. *JMIR Mental Health*, *3*(4), e50. https://doi.org/10.2196/mental.5842

Sidani, J. E., Shensa, A., Hoffman, B., Hanmer, J., & Primack, B. A. (2016). The association between social media use and eating concerns among US young adults. *Journal of the Academy of Nutrition and Dietetics*, *116*(9), 1465–1472. https://doi.org/10.1016/j.jand.2016.03.021

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, *107*(1), e1–e8. https://doi.org/10.2105/AJPH.2016.303512

Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., … Cavazos-Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image*, *24*, 137–144. https://doi.org/10.1016/j.bodyim.2018.01.001

Steakley-Freeman, D. M., Jarvis-Creasey, Z. L., & Wesselmann, E. D. (2015). What's eating the internet? Content and perceived harm of pro-eating disorder websites. *Eating Behaviors*, *19*, 139–143. https://doi.org/10.1016/j.eatbeh.2015.08.003

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Valkenburg, P. M., Koutamanis, M., & Vossen, H. G. M. (2017). The concurrent and longitudinal relationships between adolescents' use of social network sites and their social self-esteem. *Computers in Human Behavior*, *76*, 35–41. https://doi.org/10.1016/j.chb.2017.07.008

Walker, M., Thornton, L., De Choudhury, M., Teevan, J., Bulik, C. M., Levinson, C. A., & Zerwas, S. (2015). Facebook use and disordered eating in college-aged women. *Journal of Adolescent Health*, *57*(2), 157–163. https://doi.org/10.1016/j.jadohealth.2015.04.026

Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2017). Detecting and characterizing eating-disorder communities on social media. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp 91–100). https://doi.org/10.1145/3018661.3018706

Wang, X., Shi, J., Chen, L., & Peng, T. -Q. (2016). An examination of users' influence in online HIV/AIDS communities. *Cyberpsychology, Behavior & Social Networking*, *19*(5), 314–320. https://doi.org/10.1089/cyber.2015.0539

Wang, Y.-C., Kraut, R. E., & Levine, J. M. (2015). Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of Medical Internet Research*, *17*(4), e99. https://doi.org/10.2196/jmir.3558

Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders*, *40*(8), 711–717. https://doi.org/10.1002/eat.20445

Wolf, M., Theis, F., & Kordy, H. (2013). Language use in eating disorder blogs: Psychological implications of social online activity. *Journal of Language and Social Psychology*, *32*(2), 212–226. https://doi.org/10.1177/0261927x12474278

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research*, *19*(6), e228. https://doi.org/10.2196/jmir.7215

Xu, R., & Zhang, Q. (2016). Understanding online health groups for depression: Social network and linguistic perspectives. *Journal of Medical Internet Research*, *18*(3), e63. https://doi.org/10.2196/jmir.5042

Yom-Tov, E., Brunstein-Klomek, A., Mandel, O., Hadas, A., & Fennig, S. (2018). Inducing behavioral change in seekers of pro-anorexia content using internet advertisements: Randomized controlled trial. *JMIR Mental Health*, *5*(1), e6. https://doi.org/10.2196/mental.8212

Zhao, K., Wang, X., Cha, S., Cohn, A. M., Papandonatos, G. D., Amato, M. S., … Graham, A. L. (2016). A multirelational social network analysis of an online health community for smoking cessation. *Journal of Medical Internet Research*, *18*(8), e233. https://doi.org/10.2196/jmir.5985

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Moessner M, Feldhege J, Wolf M, Bauer S. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *Int J Eat Disord.* 2018;51:656–667. https://doi.org/10.1002/eat.22878