HHS Public Access

Author manuscript

Int J Eat Disord. Author manuscript; available in PMC 2019 October 12.

Published in final edited form as:

Int J Eat Disord. 2019 October ; 52(10): 1150–1156. doi:10.1002/eat.23148.

Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention

Hao Yan, PhD[1], **Ellen E. Fitzsimmons-Craft, PhD**[2], **Micah Goodman, BS**[2], **Melissa Krauss, MPH**[2], **Sanmay Das, PhD**[1], **Patricia Cavazos-Rehg, PhD**[2]

[1]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri

[2]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri

Abstract

Objective: Online forums allow people to semi-anonymously discuss their struggles, often leading to greater honesty. This characteristic makes forums valuable for identifying users in need of immediate help from mental health professionals. Because it would be impractical to manually review every post on a forum to identify users in need of urgent help, there may be value to developing algorithms for automatically detecting posts reflecting a heightened risk of imminent plans to engage in disordered behaviors.

Method: Five natural language processing techniques (tools to perform computational text analysis) were used on a dataset of 4,812 posts obtained from six eating disorder-related subreddits. Two licensed clinical psychologists labeled 53 of these posts, deciding whether or not the content of the post indicated that its author needed immediate professional help. The remaining 4,759 posts were unlabeled.

Results: Each of the five techniques ranked the 50 posts most likely to be intervention-worthy (the "top-50"). The two most accurate detection techniques had an error rate of 4% for their respective top-50.

Discussion: This article demonstrates the feasibility of automatically detecting—with only a few dozen labeled examples—the posts of individuals in need of immediate mental health support for an eating disorder.

Keywords

eating disorders; machine learning; mass screening; natural language processing; social media

Correspondence: Patricia A. Cavazos-Rehg, Department of Psychiatry, Washington University School of Medicine, 660 South Euclid Avenue, Box 8134, St. Louis, MO 63110. pcavazos@wustl.edu.

CONFLICTOFINTEREST

The authors have no conflicts of interest to declare.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

1 | INTRODUCTION

Eating disorders are serious mental illnesses, and anorexia nervosa is associated with the highest mortality rate of any psychiatric illness (Klump, Bulik, Kaye, Treasure, & Tyson, 2009). Websites and online forums that encourage eating disorder behaviors have become a.growing concern in recent years and are thought to promote “pro-anorexic” and “pro-bulimic” lifestyles. Online communities such as these may endorse “thinspiration” (e.g., pictures or quotes that serve as motivation for a person trying to achieve or maintain a very low weight) as well as “tipsandtricks” for promoting thinness and maintaining disordered eating behaviors (Norris, Boydell, Pinhas, & Katzman, 2006). Such interactive forums may not only provide a venue for individuals to focus on ways to achieve “success” with their eating disorder but may also provide an outlet to disclose struggles without feeling stigmatized and to have a supportive network that is not achievable offline (Ransom, LaGuardia, Woody, & Boyd, 2010)

The ego-syntonic nature of many eating disorders may result in the individual’s resistance to or ambivalence about change (Williams & Reid, 2010), which is in contrast to the ego-dystonic nature of most other mental disorders and the desire to want to be rid of one’s problem. For these reasons, there may be particular value in studying online forums related to eating disorders, which previous research has begun to explore (Branley & Covey, 2017; Moessner et al., 2018). For example, one study found that the writings of pro-eating disorder bloggers have amore closed-minded style, are less emotionally expressive, and refer to social interactions less often relative to blogs about recovering from eating disorders and blogs unrelated to eating disorders (Wolf, Theis, & Kordy, 2013). These findings indicate that writing can be analyzed with natural language processing techniques to predict whether or not authors may be opposed to overcoming their eating disorder.

Because many people turn to online forums to disclose mental health struggles (Balani & De Choudhury, 2015), forums could be used for identifying and reaching out to people who could potentially benefit from early or immediate intervention. However, manually reading every online post and comment to clinically appraise which individuals may benefit from such interventions would be impractical. Therefore, developing methods for automatically detecting posts with content indicating potential for imminent harm (e.g., within 24hr) could have value and would make it feasible to offer help to those who need it. While early research focused on manual evaluation of social media posts (e.g., Holleran, 2010), in the last few years researchers have developed methods for automatically detecting particular sentiments in posts, for example, depression and suicidality on Twitter (e.g., De Choudhury, Gamon, Counts, & Horvitz, 2013; Nambisan, Luo, Kapoor, Patrick, & Cisler, 2015; O’Deaetal., 2015). Furthermore, De Choudhury, Kiciman, Dredze, Coppersmith, and Kumar (2016) developed a method to identify which Reddit users are more likely to develop suicidal ideation. In the area of eating disorders, researchers have used the tags associated with photos shared on Instagram to automatically estimate the mental illness severity of users posting eating disorder-related content (Chancellor, Lin, Goodman, Zerwas, & De Choudhury, 2016). The evidence suggests that language analytic tools can detect mental illness symptoms that are posted on social media by someone who could potentially benefit from timely mental health outreach.

It is of note that past research in this area has relied on a large amount (i.e., hundreds or even thousands) of manually coded posts that are then used to train machine-learning algorithms. The goal of the current study was to measure the effectiveness of different algorithms trained on only a few dozen labeled examples to identify markers of eating disorder psychopathology that signal need for a mental health intervention. While most posts on such forums reveal disordered behaviors, this investigation focused on posts that indicate a need for immediate intervention—those posts that could be interpreted by a clinician as describing a heightened risk of near-term engagement in eating disorder behaviors. We hypothesized that a machine-learning classification model would be able to correctly identify the vast majority of posts as either indicating a need for immediate intervention, or not indicating such a need, even when only trained on a few dozen labeled examples. A model would achieve this by leveraging the power of unlabeled examples, using semi-supervised learning techniques (Das, Saier, & Elkan, 2007; Elkan & Noto, 2008; Zhu & Goldberg, 2009). Semi-supervised learning is a machine-learning method in which a small fraction of the dataset is labeled and is used to categorize the unlabeled data, yielding a more accurate model than unsupervised learning (which uses no labeled data) with-out the extensive human time and/or cost that may be required to acquire labels for fully supervised learning (in which all data are labeled). It represents a much lower burden than needing to manually code hundreds or even thousands of posts in order to establish such an automatic detection model.

2 | METHOD

Forums utilized for the present study were on the website reddit.com. Reddit, one of the most-visited websites in the world (Alexa Internet, 2018; https://www.alexa.com/siteinfo/reddit.com), is a collection of forums (called “subreddits”) about specific topics. Users can belong to as many subreddits as they choose, and within the subreddits, they can create posts (with text, links, photos, and other media), comment on posts, and vote for or against posts and comments. Reddit is particularly well suited for this investigation because it is public, widely used, and designed for discussion. Additionally, posts are semi-anonymous (associated with the poster’s username) and can be as long or as short as the poster wishes.

The overarching approach was to train a machine-learning classifier using only a few labeled examples and a significant portion of unlabeled data. Two licensed clinical psychologists categorized 53 posts as either in need of immediate intervention (positive) or not in need of immediate intervention (negative). These licensed psychologists had experience in the research of eating disorders (Patricia Cavazos-Rehg: 2.5 years, Ellen Fitzsimmons-Craft: 12 years). The data were first coded independently by each of these two coders. A subse-quent meeting occurred to review coding and reach consensus on discrepant codes (n=5). Reddit posts do not have a character limit and it can be time-intensive to human code these data; therefore, the two coders set aside 3 hr to label the posts. At the end of that time, they had labeled 53 posts. It is worth noting that experiments on active learning in the natural language processing literature typically assume that human labor is only available for a few minutes (e.g., Settles, 2011), in contrast with the several hours needed to label 53 posts in this study. With the labeled data and the remaining universe of unlabeled examples (see below for information on how those were collected), we used several different techniques to

train a classifier that predicted the probability of each unlabeled post being intervention-worthy or not intervention-worthy.

Posts were coded as “positive” if its author seemed to be in immediate distress and/or at risk of immediate engagement in disordered eating behaviors based on the clinical interpretation of their respective online posts. A binary categorization of either “positive” or “nega-tive” was chosen to distinguish whether or not a post’s author should immediately be offered mental health support. Investigators intention-ally had the coders label only 53 posts initially, 38 of which were “pos-itive” and 15 “negative,” to test the accuracy of classifiers that only have access to a small amount of training data. Table 1 shows paraphrased (to protect the original poster’s privacy) excerpts from “positive” and “negative” posts.

Only the text of posts was considered in this study: images, links, videos, and other media were disregarded. Five different machine-learning methods were used to classify new posts as either “positive” or “negative.” First, the text of each post was preprocessed (explained below) to prepare it for conversion into a mathematical format suited for the classification techniques. The text was converted into either a term frequency-inverse document frequency (TF-IDF) reweighting of a bag-of-bigram vectorizer or a word embedding space vector, depending on the learning method used. The learning methods were some form of either logistic regression with TF-IDF vectors, positive and unlabeled learning, or distance-based methods using the word mover’s distance. These methods, described in more detail below, all return a score with their classifications that can be interpreted as a confidence score.

3 | DATA

The sample started with 6,000 posts: the 1,000 “hottest” submissions (i.e., recent, highly supported posts, as measured in the difference between “up votes”—votes supporting the post—and “down votes”—votes disapproving of the post) from the subreddits EatingDisorders, BingeEatingDisorder, eating_disorders, bulimia, proED, and fuckeatingdisorders as of April 2018. These sub-reddits were chosen because they were the most popular eating disorder sub-reddits at the time (as measured by the number of people subscribed) and covered a variety of topics.

The first step of preprocessing involved removing URLs, applying a word tokenizer (breaking the text into individual words and removing punctuation when appropriate), making all words lowercase, removing stop words (i.e., common words that provide little information about content, such as “and” and “the”), and removing words containing nonalphabetic characters. These steps standardized the text and allowed the classifiers to ignore unimportant words (Vijayarani, Ilamathi, & Nithya, 2015). Numbers were removed because they increase the feature dimensions (the number of characteristics of the passages the machine-learning algorithms have to account for) and lead to overfitting.

In the second step of preprocessing, each post was converted into one of two versions. The first stems the words (reducing them to their linguistic roots, e.g., “fishing” and “fisher” are both stemmed to “fish”). This version was used for logistic regression (Genkin, Lewis, & Madigan, 2007; Nigam, Lafferty, & McCallum, 1999) and positive-and-unlabeled learning

(Li & Liu, 2003), because they are bag-of-*n*gram-based language processing methods (in which the algorithm goes through a post in groups of *n* words; see Table 2 for examples of bigrams, where $n = 2$). The Porter stemmer in the NLTK package was used. The second version of the post removes words that do not exist in the Google Word2Vec word embedding dictionary (the "W2V version"). Google Word2Vec is a dictionary that maps words to a vector in the word embedding space for computational analysis, as explained further below. This version was used for the Word Mover's Distance method (Kusner, Sun, Kolkin, & Weinberger, 2015).

Any post whose stemmed version or W2V version was less than 10 words long was removed, leaving $N = 4{,}759$ submissions in the unlabeled data set. Posts shorter than 10 words were removed because they were often not about eating disorders and typically did not contain enough information to be categorized. The training data set is a random sample of $N = 53$ posts labeled by domain experts, with $N = 38$ "positive" and $N = 15$ "negative" posts. The preprocessing steps described above were also applied to these labeled posts.

4 | WORDS AS FEATURE VECTORS

Computational language models rely on converting words into feature vectors. The following models were used.

4.1 | Bag-of-bigram vectorizer

A bigram is two consecutive words. Each post was separated into all possible bigrams, and each bigram corresponds to a certain position within a vector. If a bigram is in a given sentence, the vector position that corresponds to the bigram is "1," and if it is not in the sentence, the position is "0." Table 2 shows an example of this vectorization technique below.

4.2 | TF-IDF reweighting

Certain bigrams hold more information than others. In the sentences "Tomorrow will be sunny" and "Tomorrow will be cloudy," the most important bigrams are "be sunny" and "be cloudy." As a result, the model should pay more attention to these more valuable bigrams. To achieve this, bigrams that appear less often will have a greater weight. Just as "tomorrow will" and "will be" appear in both sentences and do not provide much information, other bigrams that frequently reappear are less likely to hold key information.

4.3 | Word embeddings

With enough passages, it is possible to map the spatial relationships between words. This method uses vectors that represent words to mathematically map out words' position in the word embedding space. Finding the distance between the words' vectors can reveal meaningful relationships (Mikolov, Yih, & Zweig, 2013). Figure 1 shows an illustration of word embeddings below.

5 | TEXT CLASSIFICATION MODELS

After preprocessing the posts and converting their contents into feature vectors, the following machine-learning models were used to classify the text as either "positive" or "negative." Here, we briefly summarize the main ideas behind the models. See Appendix S1 for the formal mathematical definitions of these models.

5.1 | Logistic regression

The model is trained on prelabeled data to find a relationship between the content of Reddit posts and the category they fall into (i.e., "positive" or "negative").

5.2 | Positive and unlabeled learning (PU learning)

As mentioned, logistic regression uses training data. However, the relative shortage of "negative" training posts reduces the accuracy of standard logistic regression. Therefore, a modified version of logistic regression could yield better results. Instead of "positive" and "negative" categories, the data are divided into "positive" and "unlabeled." The "negative" labeled posts are ignored and replaced with all posts that were not human-labeled. To find the probability that a post is "positive," take the probability that the label of that post is known, and divide it by the probability that, for any post in the data set, the label for a post given that it is "positive" is known.

5.3 | Word mover's distance

Instead of finding the distance between words with the word embedding space, the distance between entire passages of text is found. This model finds the minimum distance that the embedded words in one passage must travel to reach the embedded words of the other passage, with the goal of measuring how similar two passages are.

6 | TRIALS

6.1 | TF-IDF with logistic regression

This method is used for two different trials. In the first (pos_neg), a classifier is trained on the 53 (35 "positive" and 18 "negative") labeled posts. The classifier is then tested on the 4,759 test posts. This approach follows the steps of standard logistic regression. In the second (pos_all), a classifier is trained with the 35 "positive" labeled posts and the 4,759 unlabeled posts that are treated as "negative" posts. This classifier is also tested on the 4,759 unlabeled posts.

6.2 | PU learning (pu_learning)

Because the PU method makes a prediction by dividing the probability of a given post being "positive" by the probability of *any* post being "positive" given that it is revealed, and because TF-IDF with logistic regression is used to make the prediction model, the pu_learning is mathematically equivalent in terms of the ranking to pos_all. However, because 20% of the "positive" labeled examples must be removed from the training sample (to allow the PU algorithm to compute the constant), the pu_learning results will not actually

be the same as the pos_all results. Therefore, pu_learning is repeated 20 times and the 50 posts with the highest median output are used as the top 50 results.

6.3 | Word mover's distance

The word mover's distance from each of the 4,759 unlabeled posts to all of the 38 "positive"l abeled posts is calculated. In one trial, "wmd_1," each unlabeled post's score was its distance to the closest "positive" post. In another trial, "wmd_5," each unlabeled post's score was the average of its distances to the closest five "positive" posts.

7 | RESULTS

A total of 6,000 posts were analyzed; after removing stop words and selecting for posts > 10words, $n = 4{,}759$ posts remained for analysis. A separate initial set of 53 earlier Reddit posts from the same forums was coded by two clinical psychologists, flagging each post that would indicate a need for intervention. For example, if a person expresses strong urges to immediately engage in disordered eating, this would indicate the need for intervention.

Each classifier outputs the 50 posts it ranked as most likely to be "positive." One hundred fourteen different posts were included in the "top 50" posts across all five classifiers, revealing significant overlap in the posts the classifiers identified for their "top-50." To determine the error rates, the same two domain experts who categorized the original 53 training posts labeled these 114 posts as either "positive" or "negative." Although these posts were categorized by the classifiers, they were labeled by humans as well to assess the accuracy of the classifiers. The error rate was the number of samples classified incorrectly in a method's top-50 list, divided by 50. The most successful methods were pos_neg and wmd_5, with a 4% error rate. The results are shown in Table 3, with the trial names corresponding to the descriptions above.

Because the learning techniques of pos_all and pu_learning are so similar, they have similar error rates. On the other hand, although wmd_1 and wmd_5 both use word mover's distance, they have very different error rates because of the small number of labeled training posts, leading to the "positive" labeled posts being scattered across the word embedding space. While all posts categorized incorrectly by the best performing models were false positives, there was only one other point of similarity between the posts the models labeled incorrectly. One post was in the top 50 of all five trials. This likely happened because the post contained multiple words often used in "positive" posts, revealing a shortcoming of these classification methods.

8 | DISCUSSION

Concerning posts about mental health struggles have been found on social media (Balani & De Choudhury, 2015). Accordingly, the mental health field has a high interest in understanding ways to harness patients' electronic footprints to facilitate timely treatment initiation and/or to inform care fo rthose already engaged (Fisher & Appelbaum, 2017). Thus, our findings indicate that it is possible and practical (i.e., it does not require a prohibitive amount of labeled training data) to build a machine-learning classifier that

identifies social media posts from authors in need of immediate intervention. These findings align with previous related work, demonstrating that machine learning is a powerful tool for categorizing the content of posts on social media (Chancellor et al., 2016). Leveraging social media for early intervention may be especially important for individuals with an eating disorder because untreated symptoms tend to become more frequent, severe, and persistent over time, and shorter latency between eating disorder onset and start of treatment is associated with better out comes (Lewinsohn, Striegel-Moore, & Seeley, 2000; Loeb, Craigen, & Goldstein, 2011). Making use of these algorithms would require part-nerships with social media platforms. The platforms could, for example, implement a feature that uses algorithms to detect posts indicating a need for immediate intervention, and then automatically send the post's author a message containing helpful resources. Additionally, our findings have applications outside of organic social media. For example, if a private online social network was created to connect participants in an eating disorder treatment program in order to support their recovery, these algorithms could be used to automatically monitor posts from users to rapidly and efficiently identify patterns of risky social networking behaviors within the forum and intervene as needed.

It would be valuable to expand upon this work by generating more training data (labeling more "positive" and "negative" posts) and using that data to train the five classification methods discussed here. Although the classifiers in this study achieved high rates of accuracy, that accuracy was only for the 50 posts most likely to be labeled correctly; more training data would enable the classifiers to go beyond 50 posts. In addition, it would be worthwhile to evaluate the error rate for a greater number of posts to understand how these classification methods perform, for example, for the social network's top 500 posts. In this study, only a single post was used to determine whether a user may be in need of immediate intervention for their eating disorder. In the future, it would be worth investigating if taking all of a user's posts and comments into account allows for more accurate predictions.

While these results demonstrate the promise of training a classifier to detect "positive" posts, it is important to note that the low error rates were only for the 50 posts most likely to be labeled correctly. Therefore, in thinking of the practical implementations of this automatic detection, it may be of minimal consequence for the submitter of a "negative" post to receive an intervention (e.g., be given the opportunity to receive mental health); however, if the intervention consists of using limited resources, such as a therapist's time, then it is important to offer the intervention to those who need it most. As past research suggests, individuals may be more receptive to help if they know they were specifically targeted for the intervention (Noar, Benac, & Harris, 2007). As research in this area progresses and findings are applied in nonclinical settings, it will be critical to implement these algorithms cautiously, carefully considering their impacts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study was funded by the National Institute of Mental Health (grant numbers R21 MH112331 and R34 MH119170).

REFERENCES

Alexa Internet. (2018). Reddit Competitive Analysis, Marketing Mixand Traffic. Retrieved from https://www.alexa.com/siteinfo/reddit.com

Balani S, & De Choudhury M (2015). Detecting and characterizing mental health related self-disclosure in social media. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, South Korea.

Branley DB, & Covey J (2017). Pro-ana versus pro-recovery: A content analytic comparison of social media users' communication about eating disorders on Twitter and Tumblr. Frontiers in Psychology, 8, 1356. [PubMed: 28848472]

Chancellor S, Lin Z, Goodman EL, Zerwas S, & De Choudhury M (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, CA.

Das S, Saier MH, & Elkan C (2007). Finding transport proteins in a general protein data base. Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery, Warsaw, Poland.

De Choudhury M, Gamon M, Counts S, & Horvitz E (2013). Predicting depression via social media. ICWSM, 13, 1–10.

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, & Kumar M (2016). Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Elkan C, & Noto K (2008). Learning classifiers from only positive and unlabeled data. In Proceedings of the 14 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV.

Fisher CE, & Appelbaum PS (2017). Beyond Googling: The ethics of using patients' electronic footprints in psychiatric practice. Harvard Review of Psychiatry, 25(4), 170–179. [PubMed: 28504978]

Genkin A, Lewis DD, & Madigan D (2007). Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3),291–304.

Holleran S (2010). The early detection of depression from social networking sites [Doctoral dissertation]. University of Arizona, Tucson, AZ.

Klump KL,Bulik CM,Kaye WH,Treasure J,&Tyson E(2009).Academy for eating disorders position paper: Eating disorders are serious mental illnesses. International Journal of Eating Disorders, 42(2), 97–103. [PubMed: 18951455]

Kusner M, Sun Y, Kolkin N, & Weinberger K (2015). From word embed-dings to document distances. Paper presented at the International Conference on Machine Learning, Lille, France.

Lewinsohn PM, Striegel-Moore RH, & Seeley JR (2000).Epidemiology and natural course of eating disorders in young women from adolescence to young adulthood. Journal of the American Academy of Child and Adolescent Psychiatry, 39(10), 1284–1292. [PubMed: 11026183]

Li X, & Liu B (2003).Learning to classify texts using positive and unlabeled data. Paper presented at the IJCAI (Vol.3, pp.587–592).

Loeb KL, Craigen KE, & Goldstein MM (2011).Early treatment for eating disorders In Lock J & Le Grange D (Eds.), Eating disorders in children and adolescents: A clinical handbook (pp.337–361). New York, NY: Guilford Press.

Mikolov T, Yih WT, & Zweig G (2013).Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA.

Moessner M, Feldhege J, Wolf M, & Bauer S(2018).Analyzing big data in social media: Text and network analyses of an eating disorder forum. International Journal of Eating Disorders, 51, 656–667. [PubMed: 29746710]

Nambisan P, Luo Z, Kapoor A, Patrick TB, & Cisler RA (2015).Social media, big data, and public health informatics: Ruminating behavior of depression revealed through Twitter. Paper presented at the 2015 48th Hawaii International Conference on System Sciences (HICSS), Kauai, HI.

Nigam K, Lafferty J, & McCallum A (1999). Using maximum entropy for text classification. Paper presented at the IJCAI-99 Workshop on Machine Learning for Information Filtering, Stockholm, Sweden.

Noar SM, Benac CN, & Harris MS (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin, 133(4), 673–693. [PubMed: 17592961]

Norris ML, Boydell KM, Pinhas L, & Katzman DK (2006).Ana and the internet: A review of pro-anorexia websites. International Journal of Eating Disorders, 39 (6), 443–447. [PubMed: 16721839]

O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, & Christensen H (2015).Detecting suicidality on Twitter. Internet Interventions, 2(2), 183–188.

Ransom DC, LaGuardia JG, Woody EZ, & Boyd JL (2010). Inter-personal interactions on online forums addressing eating concerns. International Journal of Eating Disorders, 43(2), 161–170. [PubMed: 19308991]

Settles B (2011).Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp.1467–1478), Edinburgh, Scotland.

Vijayarani S, Ilamathi MJ, & Nithya M (2015).Preprocessing techniques for text mining—Anoverview. International Journal of Computer Science & Communication Networks, 5(1), 7–16.

Williams S, & Reid M (2010).Understandingtheexperienceofambiva-lenceinanorexianervosa: The maintainer's perspective. Psychology and Health, 25(5), 551–567. [PubMed: 20204933]

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.

Zhu X, & Goldberg AB (2009).Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1), 1–130.

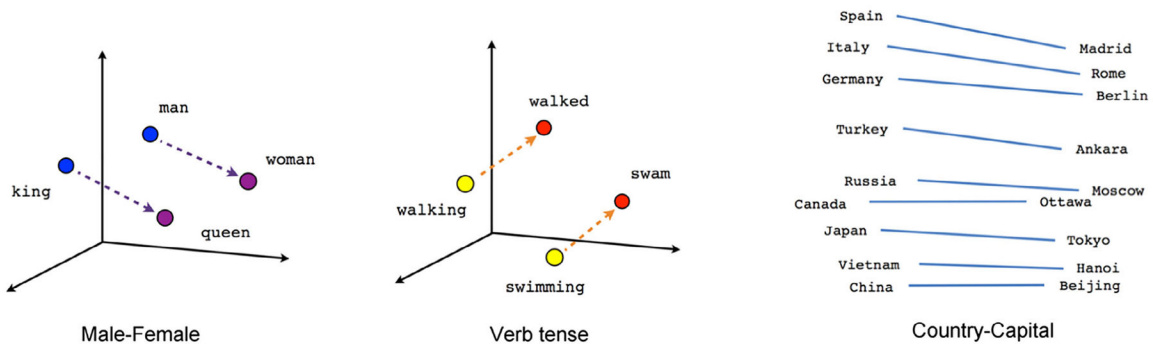

FIGURE 1.

These diagrams show how word embeddings allow the discovery of relationships between words.

TABLE 1

Paraphrased passages from posts that do (positive) and do not (negative) indicate an immediate need for intervention

Passage	Label
This experience has taught me so much. I now know that my body needs food to work properly, and that starving my body only harms it.	Negative
My urge to binge went away after I ate the dessert, so I did not binge for the whole day.	Negative
I used to binge for a whole week once I slipped up just once, but this time will be different.	Negative
My eating disorder controls everything I do.	Positive
I do not know how much longer I can go on like this before it feels like death is a better option.	Positive
I binged again last night and my body and brain keep telling me that I need to fast today to counteract the binging.	Positive

TABLE 2

The bag-of-bigram representation for a document with the phrases “tomorrow will be sunny” and “tomorrow will be cloudy”

Bigram	tomorrow will	will be	be sunny	be cloudy
Sentence a	1	1	1	0
Sentence b	1	1	0	1

TABLE 3

The error rates for the 50 posts most likely to be labeled correctly for each of the learning methods

Trial:	pos_neg (classifier trained on all 53 labeled posts)	pos_all (classifier trained on the 35 "positive" labeled posts and the 4,759 unlabeled posts, treating the unlabeled posts as if they were labeled "negative")	pu_learning (classifier trained on 80% of the 35 "positive" labeled posts, and the 4,759 unlabeled posts, treating the unlabeled posts as if they were labeled "negative")	wmd_1 (classifier using the word mover's distance to the single closest "positive" post)	wmd_5 (classifier using the average of the word mover's distance to the five closest "positive" posts)
Top-50 error rate:	4%	12%	8%	12%	4%

Int J Eat Disord. Author manuscript; available in PMC 2019 October 12.