Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

#PrayForDad: Learning the Semantics Behind Why Social Media Users Disclose Health Information

Zhijun Yin[⋆], You Chen[†], Daniel Fabbri[⋆†], Jimeng Sun[‡], Bradley Malin[⋆†]
{zhijun.yin, you.chen, daniel.fabbri, b.malin}@vanderbilt.edu, jsun@cc.gatech.edu
[⋆]Department of Electrical Engineering & Computer Science, School of Engineering, Vanderbilt University
[†]Department of Biomedical Informatics, School of Medicine, Vanderbilt University
[‡]School of Computational Science and Engineering, Georgia Institute of Technology

Abstract

User-generated content in social media is increasingly acknowledged as a rich resource for research into health problems. One particular area of interest is in the semantics individuals evoke because they can influence when health-related information is disclosed. While there have been multiple investigations into why self-disclose occurs, much less is known about when individuals choose to disclose information about other people (e.g., a relative), which is a significant privacy concern. In this paper, we introduce a novel framework to investigate how semantics influence disclosure routines for 34 health issues. This framework begins with a supervised classification model to distinguish tweets that communicate personal health issues from confounding concepts (e.g., metaphorical statements that include a health-related keyword). Next, we annotate tweets for each health issue with linguistic and psychological categories (e.g. social processes, affective processes and personal concerns). Then, we apply a non-negative matrix factorization over a health issue-by-language category space. Finally, the factorized basis space is leveraged to group health issues into natural aggregations based around how they are discussed. We evaluate this framework with four months of tweets (over 200 million) and show that certain semantics correspond with whom a health mention pertains to. Our findings show that health issues related with family members, high medical cost and social support (e.g., Alzheimer's Disease, cancer, and Down syndrome) lead to tweets that are more likely to disclose another individual's health status, while tweets with more benign health issues (e.g., allergy, arthritis, and bronchitis) with biological processes (e.g., health and ingestion) and negative emotions are more likely to contain self-disclosures.

Introduction

Social platforms have become popular environments for people to share, as well as seek, health-related information. For instance, it has been shown that certain people use Reddit to share information about their mental health, including symptoms, treatments received, and the influence such a problem has on their social life (De Choudhury and De 2014). Moreover, a non-trivial quantity of individuals want to integrate social platforms into the management of their care. As evidence, one survey indicated that 56% of participants wanted their providers to use social platforms to provide notifications of appointments, prescription availability, reporting of test results, and as a forum for asking general diagnostic- and procedure-related questions (Fisher and Clayton 2012).

Further evidence suggests that disclosing information about the self online can be intrinsically rewarding (Tamir and Mitchell 2012), while sharing one's health status can assist in organizing networks and obtaining social support (Claypoole 2014). Yet, health information is considered one of the most sensitive aspects about an individual (Pew Research Center 2014) and there is a perception that its disclosure has the potential to negatively impact personal privacy (van der Velden and El Emam 2013). This begs the question of why (and when) individuals choose to disclose such information.

Several recent studies have investigated this issue by inquiring individuals about which factors drive self-disclosure. In particular, one recent survey looked into the ways that youths disclose their personal health issues on social media (Lin et al. 2016), highlighting factors associated with trust and uncertainty. While it may be argued that concerns over personal privacy can be addressed by allowing an individual to choose when to disclose health information (Caine and Hanania 2013; Meslin et al. 2013), it must further be recognized that social media environments provide an opportunity for the disclosure of information about other individuals, often without consideration of their approval or consent. Specifically, it has been shown that individuals disclose information about a wide range of acquaintances, ranging from family members to friends to high profile persons in the media (Mao, Shuai, and Kapadia 2012; Yin et al. 2015).

An *ad hoc* review of the social media posts suggests that the decision to disclose information may be context-dependent. As such, we anticipate that the semantics (e.g., language categories) an individual evokes when discussing a health problem could influence when they choose to disclose. Moreover, such semantics may correlate with whom the disclosure pertains to (e.g., the author of a post or a related individual). The potential for a semantic analysis with respect to posts about health information in social media is justified through evidence from prior investigations. Notably, semantic analysis has been applied to compare how the severity and social stigma of health issues drive people to

seek via search engine or share in social media (De Choudhury, Morris, and White 2014). However, there has been little investigation into how collections of health issues, driven by communication semantics, relate to whom disclosures of health information on social media pertain to. Gaining an understanding of such factors could provide intuition into when an individual's privacy is put in jeopardy and if it is done so maliciously or simply to seek assistance or support. Moreover, by characterizing the semantics associated with such disclosure, it may be possible to develop programs to educate and, subsequently, mitigate the disclosure of other individual's information without their consent.

In this paper, we propose a novel framework that relies on semantic analysis and non-negative matrix factorization (NMF), to computationally uncover similar health issues communicated through social media, as well as their association with an individual's propensity to disclose. In this framework, we first apply a supervised classification model to distinguish tweets that communicate personal health issues from known confounding concepts (e.g., metaphorical statements that include a health-related keyword (Hayes et al. 2007; McNeil, Brna, and Gordon 2012)). Next, we annotate the tweets of each health issue with linguistic and psychological categories (e.g., social processes, affective processes and personal concerns). Then, we apply NMF over a space of health issue-by-language categories to obtain natural aggregations of the investigated health problems. Finally, we demonstrate that the semantics behind health issue mentions on Twitter are correlated with disclosure behavior.

There are several primary contributions of this paper:

- We show that language categories have a significant impact in health mention detection and the discovery of similar health issues communicated over social media.
- We introduce a health issue-by-language category model (as opposed to the traditional document-by-term model) to study groups of health issues. By applying NMF on this model, we show the existence of four groups of health issues and their semantics, which correspond to: 1) common semantics, such as feeling and cognitive processes (e.g., insight and tentative), 2) biological processes (e.g., health - medicine, clinic, ingestion - eat, and taste), 3) social processes (e.g., family, friends, humans - girl, and women), and 4) negative emotions.
- Using over 200,000 tweets from a four-month period, we find that disclosure behavior is associated with semantically similar groups of health issues. Specifically, we show that major life-altering health issues related with family members, high medical costs and searching for social support (e.g., Alzheimers Disease, cancer, and Down syndrome) are more likely to have tweets disclosing other individual's health status. By contrast, we show that more benign health issues related with simple chronic biological processes and negative emotions (e.g., allergy, arthritis, asthma, and bronchitis) tend to have tweets with self-disclosed health status.

The remainder of the paper is organized as follows. We begin by reviewing related research in social media, health mentions, and inferential models for the analysis of such information. Next, we describe the methodology designed for amassing a dataset from a Twitter stream and define the health mention classification strategy. We then introduce an NMF-based approach for discovering semantically-similar health issues on Twitter and build a statistical model to learn the associations between the learned groups of health issues and disclosure behavior. Finally we discuss the notable findings of our experimental analysis and conclude the work.

Related Work

In this section, we summarize related research in several areas: 1) sharing and seeking of health information, 2) detection of health mentions in online postings, and 3) learning approaches to determine the factors driving health information disclosure in social media.

Seeking and Sharing Health Information. Various studies have provided intuition into what type of health information is communicated and/or sought over social media. For instance, it has been shown that, on Twitter, users with depression tend to publish content with a negative connotation and an expression of religious involvement (De Choudhury et al. 2013). On the other hand, as alluded to earlier, on Reddit (De Choudhury and De 2014), individuals with mental health problems provide information about challenges faced in daily life, as well as pose queries regarding certain treatments. Similarly, cancer survivors in online forums on Reddit often shared information with personal narratives, while other online participants tend to ask for assistance immediately after diagnosis (Eschler, Dehlawi, and Pratt 2015). One study on loneliness on Twitter (Kivran-Swaine et al. 2014) showed that female users tend to be more likely to express more severe, enduring loneliness, but receive less responses (and presumably support) than male users. Several studies have also shown parents often turn to Reddit or Facebook to seek social social support, but that their activities are often constrained by privacy concerns regarding the sharing of their childrens' health status (Ammari, Morris, and Schoenebeck 2014; Ammari and Schoenebeck 2015).

Detection of Personal Health Mentions. While the aforementioned studies discuss what behaviors people exhibit on social media, they do not necessarily address how to mine such information in an automated fashion. However, a growing number of approaches are being developed and applied to extract information from such settings. For instance, a character-based n-gram language model was shown to be effective for detecting tweets focused on post-traumatic stress disorder (PTSD) and depressio (Coppersmith, Harman, and Dredze 2014; Coppersmith et al. 2015a; 2015b). Additionally, language categories obtained from the Linguistic Inquiry Word Count (LIWC) framework, which we invoke in this study, have proven to be notable features for classifying specific health issues (De Choudhury et al. 2013; De Choudhury and De 2014; Tamersoy, De Choudhury, and Chau 2015). A word-based n-gram (Paul and Dredze 2014), as well as natural language processed outputs (Yin et al. 2015), have also been shown to be useful for building a universal health mention classifier, which cuts across a range of

health issues. We note that the classification model we introduce in this work differs from previous investigations in that 1) we build a model to classify tweets associated with a broad range of health issues and 2) we combine the character n-gram model (which focuses on the level of a single tweet) and language categories (which focus on the broader level of a health issue and incorporate multiple tweets).

Factors Driving Health Disclosure. Studies that investigate the driving factors behind information disclosure have relied upon direct inquiry through surveys and inference via computational methods. Notably, one recent survey delved into the ways that youths disclose their personal health issues on social media (Lin et al. 2016). The results suggested that these decisions were driven by 1) trust in social media platforms and 2) uncertainty about their physcian's advice. Severity and social stigma of health issues have also been shown to be factors that motivate people to seek health information via web searches (e.g., via the Bing search engine) or share information in social media (De Choudhury, Morris, and White 2014). It has also been shown that some individuals with serious mental problems comment or upload videos to YouTube to seek peer support (Naslund et al. 2014).

Data Preparation

For this study, we relied on the Twitter streaming API to collect tweets in English and published in the contiguous United States during a four-month window in 2014. The corpus of collected tweets (approximately 261 million) was filtered by a set of keywords associated with notable health issues. Specifically, we selected 34 health issues based on their high impact on healthcare as noted in the Medical Expenditure Panel Survey of the Agency for Healthcare Research and Quality (AHRQ) of the U.S. Department of Health and Human Services[1], as well as their popularity in Google Trends during the data collecting period. These health issues include chronic diseases (e.g., diabetes, hypertension, and arthritis), as well as more acute debilitating phenomena (e.g., stroke). Filtering the tweet stream resulted in a set of $281,357$ tweets (i.e., a reduction of 99.89%).

To obtain the ground truth, we implemented a survey on Amazon Mechanical Turk (AMT) to investigate whether a given tweet containing the keywords discloses personal health status. Specifically, we randomly selected 100 tweets for each of the 34 health issues. Due to the diversity of the content and to help participants better understand the task, we provided seven options, defined as follows:

1) The tweet discloses the health status of the author.

2) The tweet discloses the health status of the author's family members or friends.

3) The tweet discloses the health status of someone else, excluding the author, the author's family members and friends.

4) The tweet uses the health issue as a metaphor.

5) The tweet expresses a viewpoint on the health issue, or some kind of support to general patients with the health issue (excluding those specific persons mentioned in option 1, 2 and 3).

6) The tweet expresses a worry related with the health issue.

7) None of the above.

Each participant, who was a certificated AMT master that continuously demonstrated high accuracy in the AMT marketplace, was required to select one, and only one, of these options to best describe the given tweet.

The answers to the survey were designed hierarchically, such that the seven options were compressed into several types of information for each investigated health issue: i) the number of self-disclosed tweets (option 1) and the number of tweets disclosing others (option 2 and 3), and ii) the number of tweets on health disclosure, including their own, and other people' health status (which we refer to as the *positive* class: options 1, 2 or 3), and the number of tweets not on health disclosure (which we refer to as the the *negative* class: options 4, 5, 6, or 7). We apply the first type of information to assess, for a certain health issue, if individuals are more likely to disclose their own health status or that of another person. We leverage the second type of information as the gold standard when building a binary classifier to automatically detect tweets with health mentions.

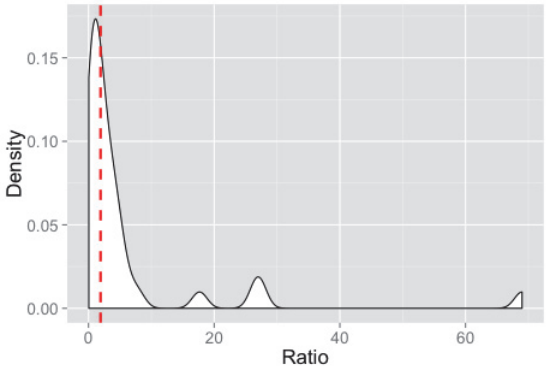

Figure 1: The density of the me vs. you (MvY) ratio. The red dashed line represents the median MvY ratio.

Each tweet was labeled by two masters, while a third master was employed to break the tie when there is a disagreement on whether the tweet disclose personal health status or not[2]. There were 65 AMT masters who participated in the labeling task and 21 AMT masters were invoked to assist in the breaking of conflicting labels. The kappa score of the agreement on the seven options level was 0.59 (an indicator of the complexity of the specific labeling task), but the kappa for the simpler positive- vs. negative-class level was 0.79. As such, the investigated tweets are categorized into two types: the positive class (41.5%) and the negative class

[1] http://meps.ahrq.gov/mepsweb/

[2] In other words, we only handle labeling conflicts at the positive- and negative-class level. If one tweet, for instance, received option 2 and option 3 from two masters, respectively, then it is labeled as positive.

(58.5%). Next, this dataset was applied to train a binary classifier to label additional personal health mention tweets in the larger corpus (i.e., the remaining 277,957 tweets) for further analysis. Additionally, in preparation for our analysis, we computed the ratio of the number of tweets disclosing the author's personal health status to the number of tweets disclosing another person's personal health status. We refer to this as the *Me vs. You*, or MvY ratio, for each health issue. The larger the MvY ratio for a health issue, the more likely it is that the corresponding tweets disclose their authors' personal health status. Figure 1 illustrates the density of the MvY ratio per health issue. It was observed that there was a strong positive skew suggesting that there are many health issues for which the author is more likely to self-disclose.

Health Mention Detection

Manual discovery and annotation of tweets that disclose personal health status is a timely, as well as costly, process. Thus, as alluded to in the previous section, we engineered a classification strategy, based on the labels provided by the AMT masters, to automatically detect tweets communicating the mentions of health status and augment the dataset for investigation[3].

Building Classification Models

We observed that the tweets can be naturally clustered by health issue keywords (e.g., all tweets associated with asthma), and each cluster may have different properties (e.g., one issue may be associated with a larger proportion of tweets that disclose health information than others). To incorporate these intercluster differences into classifiers, one natural candidate solution is to build a hierarchical model, where the parameters are governed by hyperparameters for each health issue. However, this would result in a very expensive computational model, due to the high-dimensionality of the features (e.g., word unigrams) involved in text classification.

Rather, in this paper, we introduce an approach based on language categories to reflect the differences between health issues. We construct features at both 1) the tweet-level and 2) the health-level. The tweet-level features consisted of 2000 character n-grams ($2 \leq n \leq 5$) from all of the labeled tweets, according to the ranking of their TF-IDF values. To obtain health-level features we extracted language categories using the Linguistic Inquery Word Count (LIWC) from all of the unlabeled tweets. LIWC has been invoked by many social data analysis studies with some successes (De Choudhury et al. 2013; De Choudhury and De 2014; De Choudhury, Morris, and White 2014; Coppersmith et al. 2015b). Basically, LIWC counts the number of words (e.g., from all the tweets related to a health issue) that match each of the language categories[4] and converts them as the percentages of total words. In total, we use 64 language categories as for health-level features. Note that all the tweets related to the same health issue are defined over the same language categories.

The baseline model trains classifiers with character n-gram features at the tweet-level. Our proposed model is an augmentation that includes the language categories as features at the health-level.

Predicting Health Mentions

We considered three common learners[5] for each model: i) a logistic regression, ii) a linear support vector machine (SVM), and iii) a random forest. All of the parameters were set to their defaults. The 3400 labeled tweets were applied as a gold standard. We applied 10-fold shuffled and stratified cross-validation and reported the mean and standard deviation of the area under the receiver operating characteristic curve (AUC) for each classifier in Table 1. A t-test was applied to assess if there is a statistically significant difference (at the 0.05 significance level) between the classifiers in their capability.

Model	Baseline	Proposed
Linear Regression	0.825 ± 0.007	0.837 ± 0.007
Linear SVM	0.811 ± 0.010	0.839 ± 0.007
Random Forest	0.823 ± 0.012	0.833 ± 0.012

Table 1: A comparison of the AUC for the baseline model and proposed model.

The t-test confirmed ($p < 0.05$) that introducing language categories as features (at health-level) can improve the performance of logistic regression and linear SVM. The results further indicate that the linear SVM with language categories as features (at health-level) significantly outperforms the classifiers that are devoid of such features ($p < 0.05$).

Rank	Feature	Rank	Feature
1	*health*	11	funct
2	nee	12	*negate*
3	*pronoun*	13	*conj*
4	*auxverb*	14	*we*
5	my	15	i'
6	*i*	16	*verb*
7	n	17	*bio*
8	has	18	*present*
9	*time*	19	i'm
10	ne	20	*humans*

Table 2: The top-20 most informative features selected by the random forest classifier. The group features (i.e., the aggregated language categories at the level of a health issue) are depicted in *blue italicized font*.

[3] It should be recognized that we aimed to build a classifier that is sufficient for detecting a large number of tweets with health mentions, so that we may investigate the extent to which language categories influence disclosure. It is impossible to engineer a perfectly accurate classifier by incorporating many other features (e.g., part of speech, grammar features and word2vec) and, thus, we acknowledge that there is a certain degree of error in the labels of the tweets we investigate.

[4] http://www.kovcomp.co.uk/wordstat/LIWC.html

[5] As implemented in the Scikit-Learn package. http://scikit-learn.org/

To understand the importance of these features, we used the random forest classifier in the proposed model to select the top-20 most informative features for health mention detection, as shown in Table 2. It should be noted that 13 out of the 20 corresponding to features at health-level were obtained via the application of LIWC. Considering that only 64 out of the 2064 features are health-level features ($P(group\ feature) = 0.031$), a sign test implied there was a strong significant difference between these two types of features (where 13 successes out of was $= 20$ trials with $p < 0.001$). Table 2 also shows that biological processes, and the health language categories in particular, are critical for health mention detection. We further recognized notable language features pertained to time (notably the present time) and those associated with humans. Interestingly, pronouns are also important in both types of features. We suspect this stems from the fact that many tweets disclose health status about the authors' family members and friends. The following tweet is a clear example of this observation:

*Just found out that **my grandmother** has cancer. Thyroid cancer to be exact.*

Finally, we applied the logistic regression classifier and obtained $54,247$ health mention tweets (with an expected precision of 81.7%) to conduct the NMF analysis.

Discovery of Similar Health Issues

We aim to investigate if semantically similar health issues associate with the MvY disclosure rate. In this section, we show how the health issues were grouped according to their semantics.

Grouping Health Issues with NMF

We applied NMF to the set of tweets to learn similar health issues. We applied NMF, as opposed to another matrix factorization strategy like singular value decomposition, because it has been shown to have better interpretability when the original matrix values are all positive (Zhang 2012). However, applying the document-term model (as is traditional in matrix factorization) for the short texts encountered on Twitter will suffer from data sparsity. Many strategies have been proposed to overcome this problem, ranging from aggregation of documents (Quan et al. 2015) or words (e.g., the document by bi-terms model (Yan et al. 2013)) to a document-by-word embedding model (Sridhar 2015)). In this paper, we propose a health issue-by-language category model.

To build this model, we apply LIWC to extract language categories from the tweets with mentions for each health issue. This results in a matrix of 34 health issues by 64 language categories, which is subject to NMF[6]. We set the rank (i.e., the number of basis components in NMF) to 4 because this exhibited the best cophenetic correlation coefficient and dispersion coefficient. Note that this decision was based on the correlation of consensus matrix obtained from 100 NMF runs.

[6]https://cran.r-project.org/web/packages/NMF/

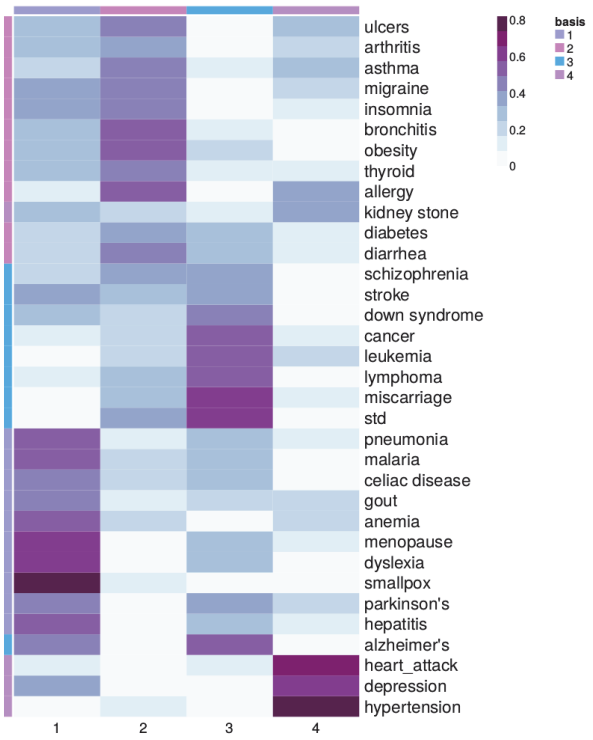

Figure 2: A heatmap of the basis components derived from NMF. Each cell is the probability that a health issue (along the row) belongs to the basis component (along the column).

Health Issue Groups and Semantics

Figure 2 depicts the heatmap of the four basis components (denoted by B_1, B_2, B_3 and B_4, respectively). Figure 3 illustrates the heatmap of the mixture coefficients for each basis component. The health issues are grouped by assigning them to their most associated basis component. The associated semantics are thus explained via the corresponding coefficients in the basis component. Note that certain health issues are affiliated with more than one group. Examples of such issues include *Alzheimer's Disease*, *Down Syndrome* and *Parkinson's Disease*, all of which can be characterized by the first (B_1) and the third (B_3) basis components.

Group I (corresponding to $\mathbf{B_1}$). This basis component, in comparison to the others (see Figure 3), exhibits a set of similar probabilities for a broad range of language categories. These include cognitive processes (e.g., think, know, guess, and stop), quantifiers (e.g., much and lot), non-fluencies and perceptual process, and feeling. This basis component covers common semantics shared by a wide range health issues, such as *dyslexia*, *gout*, *hepatitis*, *malaria*, *menopause*, *Parkinson's Disease*, *pneumonia*, and *smallpox*. The following tweets are clear examples:

*Not this year. She is **feeling** better from the pneumonia but still weak*

*I get picked on a **lot** for my dyslexia so I act like I **read** something faster than I really did.*

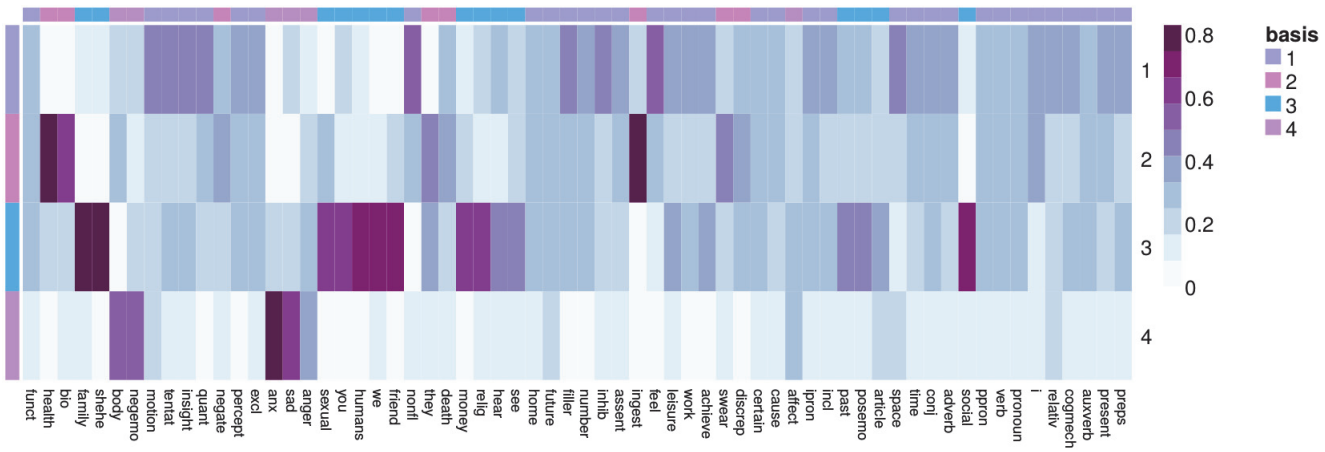

Figure 3: A heatmap of the mixture coefficients derived from NMF. Each cell is the probability that the language category (along by the column) is associated with the basis component (along the row).

*I **know** menopause only happens once, but my mother's an exception to that.*

*PLease help me to **move** back to Oklahoma to get on a clinical trial for my hepatitis c.*

Group II (corresponding to $\mathbf{B_2}$). This basis component exhibits a strong semantic of biological processes, especially health (e.g., clinic and pill) and ingestion (e.g., eat and taste). Note there are also less strongly semantic terms, such as swear words (e.g., damn) and the third person plural. Further note that swear words may be used to express negative experiences. The group of health issues corresponding to this basis component include more common disorders: *allergy*, *anemia*, *arthritis*, *asthma*, *bronchitis*, *celiac disease*, *diabetes*, *diarrhea*, *insomnia*, *migraine*, *obesity*, *thyroid* and *ulcers*. The following tweets serve as examples of this group:

*Seriously in need of some allergy **medicine**.*

*Well went to the **doc**. Gave me some shit for stomach Ulcers. Hopefully this works and I can **eat** in the next couple of days :)*

*I'm just glad my migraine went away, but I'm still **sick** to my stomach*

*I just need one of my friends to have insomnia like me so **they** can stay up and text me all night long*

Group III (corresponding to $\mathbf{B_3}$). This basis component has strong semantics associated with social processes (e.g., friends, family and humans - girl and woman), third person singular, first person plural, second person plural, money, religion (e.g., church and pray), and sexuality (e.g., love and incest). There are also less strong semantics associated with past tense and positive emotions. The group of health issues most associated with this basis component are more severe, and often debilitating, including: *Alzheimer's Disease*, *cancer*, *Down syndrome*, *miscarrage*, *leukemia*, *lymphonia*, *schizophrenia*, *sexually transmitted disease*, and *stroke*. The following tweets serve as examples of this group:

***#Ineedtoraisemoney** to help my **husband** who recently had a stroke need to raise $5000.00 any help out there!???*

*He has leukemia. His parents go to my **church**.*

*use **love** as means to solve it. My dad has awful violent schizophrenia he has tried to kill me and mom.*

*Dad is officially **cancer free**! They **caught** it in time and no chemo treatments are needed!!! **:)***

Group IV (corresponding to $\mathbf{B_4}$). This component is mainly about affective processes, such as negative emotions, anxiety, anger and sadness. The group of health issues most associated with this basis component are associated with chronic and painful problems, including: *depression*, *hypertension*, *heart attack* and *kidney stone*. Note that the semantic of *body* in this basis component may be due to the last two health issues. The following tweets serve as examples of this group:

*I'm so **afraid** that my depression is coming back. :(*

*I was so **nervous** ! I almost died of a heart attack*

*I **hate** medicine that's y I don't take it but right now my depression hittin hard af*

*reminds me of **grief** being diagnosed as depression and meds being prescribed*

Linking to MvY Disclosure

In this section, we investigate how the learned health issue groups associate with the rate at which information is disclosed about the author or other individuals. We first regress MvY ratio on the predictors extracted from the four NMF basis components, in order to examine how these basis components contribute to MvY disclosure when considered independently and when combined. Then, by connecting to the associated health issues in each basis component, we explore how the semantics (as factors) drive MvY disclosure.

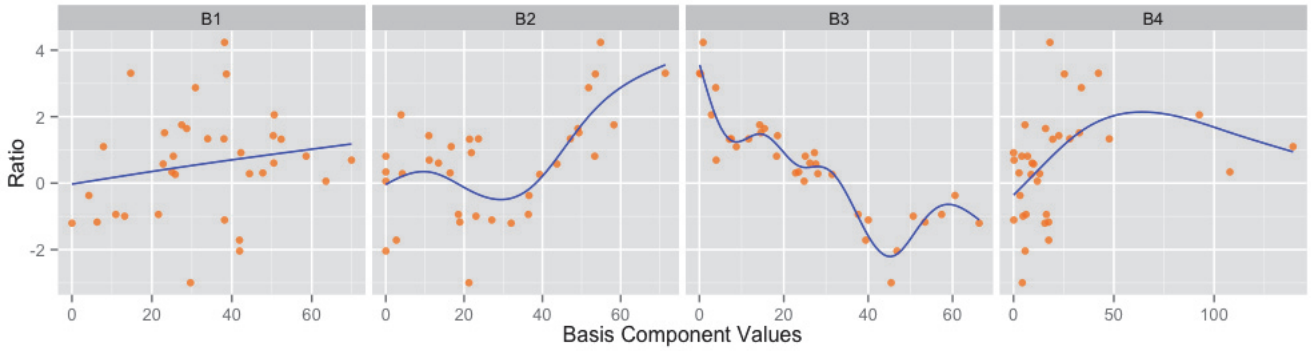

Figure 4: Correlation between the NMF basis components and the MvY ratio. The blue lines were smoothed via a thin plate regression spline. Note the positive effect of B_2 and the negative effect of B_3 with respect to the MvY ratio.

Factors Driving MvY Disclosure

We use the four NMF basis components to predict MvY. Specifically, we adopt generalized additive models (GAMs) with a thin plate regression spline smoother (Wood 2011). In the process, we apply a log function to the response to account for the positive skewness of MvY. We apply an ANOVA to perform a Chi-square test on the deviance and compare the different GAMs at the 0.05 significance level.

Single Predictor Models. To examine the effect of each individual basis component, we begin by predicting MvY with a single predictor. This corresponds to models M_1, M_2, M_3 and M_4. (for each of the corresponding enumerated components). Figure 4 shows the relationship between the basis components and the MvY ratio. It can be seen that there is a direct correlation between the chance a health issue associates with Group II and the MvY ratio. By contrast, there is a negative correlation between a health issue associating with Group III and the MvY ratio. Neither Group I nor Group IV exhibits a strong association with the MvY ratio (and neither has a significant coefficient). The Chi-square tests indicate the best single predictor model is M_3, followed by M_2.

Multiple Predictors Models. Based on the results of M_2 and M_3, we investigated two additional GAMs: i) M_{2+3} by smoothing the marginal smooths of B_2 and B_3, and ii) M_{1+2+4} by applying a linear combination of the smoothed B_1, B_2 and B_4.

Table 3 summarizes the MvY predictive capability for the models. There are statistically significant effects for each of the predictors in each model. The effects of the predictors in M_{2+3} and M_{1+2+4} are shown in Figures 5 and 6, respectively. Notably, Figure 5 shows that a health issue tends to exhibit a higher MvY (i.e., self-disclosure rate) when it positively correlates with Group II and negatively correlates with Group III. Figure 6 shows that, when combined together, B_1, B_2 and B_4 enhance the prediction of a higher MvY.

Table 4 summarizes the results of the comparison on these models with a Chi-square test under an ANOVA function. Although M_{2+3} has a larger adjusted R^2 and deviance,

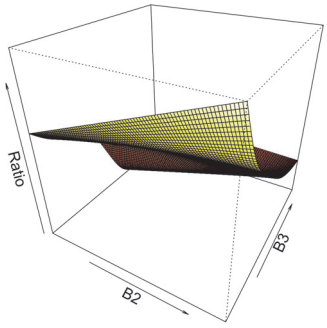

Figure 5: An illustration of the combined effect of B_2 and B_3 on predicting MvY. A larger MvY is positively correlated with Group II and negatively correlated with Group III.

there is not a statistically significant difference[7] with respect to M_3. It was, however, observed that M_3 outperforms M_{1+2+4} in a statistically significant manner. Furthermore, it was found that M_{1+2+4} outperforms M_2 in a significant manner as well. This suggests that Group III more strongly associates with an author disclosing another individuals' health status, whereas the combination of Group I, II and IV associate with self-disclosure.

Semantics Behind MvY Disclosure

To confirm these results, we ran a Spearman rank correlation test between the language categories and the MvY ratio. The results of the test, with correlation coefficient greater than or equal to 0.5, are shown in Table 5.

The results suggest that, as expected, there is a strong correlation between the use of first person singular and self-disclosure of health issues. In addition, tweets communicating self-disclosure tend to apply adverbs, time, quantifiers and present tense. For instance, it was observed that the top four health issues with the largest proportion of tweets using the words of "morning", "afternoon", "tonight", "tomor-

[7] Interestingly, the best model we obtained is M_{1+2+3}; however, there is no statistically significant effect on B_1 in M_{1+2+3}.

Model	Predictor	EDF	Ref.df	F	R-sq. (adj)	Dev.	GCV
M_2	s(B_2)	3.96 ***	4.83	5.85	0.45	51.5%	1.65
M_3	s(B_3)	7.91 ***	8.67	36.75	0.91	92.8%	0.33
M_{2+3}	s(B_2, B_3)	16.84 ***	21.40	17.73	0.92	96.1%	0.43
M_{1+2+4}	s(B_1)	3.65 ***	4.54	15.02	0.81	84.0%	0.62
	s(B_2)	1.00 ***	1.00	102.41			
	s(B_4)	1.21 ***	1.38	37.96			

Table 3: Smoothed terms for predicting the MvY ratio under different models. Note the linear effect of s(B_2) in M_{1+2+4}. *** $p < 0.001$.

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
M_2	29.042	40.958	-	-	-
M_{1+2+4}	27.143	13.509	1.899	27.449	***
M_3	25.088	6.083	2.055	7.426	***
M_{2+3}	16.162	3.330	8.926	2.753	0.143

Table 4: Comparison between models with Chi-square tests on deviance with ANOVA function. *** $p < 0.001$.

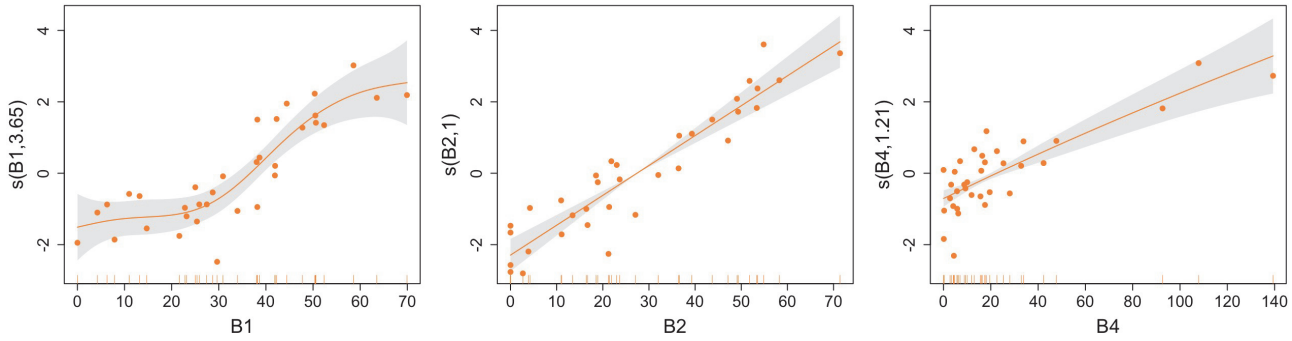

Figure 6: Effects of B_1, B_2 and B_4 when combined to predict the MvY ratio. Each of these basis components positively influence a higher MvY prediction.

row", "now" and "soon" (more than 8%) were *insomnia*, *migraine*, *kidney stone* and *asthma*. Additionally, the top four health issues with the largest proportion of tweets containing words of "pill" and "med(icine)" were *malaria*, *thyroid*, *anemia* and *hypertension* (more than 8%).

Next, we turn our attention to tweets in which the author discloses the health status of another person. As expected, we find a strong correlation with the 3rd person singular. We also find that authors tend to disclose information about family members. Moreover, it appears that the religion category indicates social support. For instance, we observed that *Alzheimer's Disease*, *leukemia*, *Parkinson's Diseases* and *cancer* are the top four health issues with the largest proportion of tweets containing the words "mother (mom)" and "father (dad)" (more than 20%). Also, it should be noted that more than 15% of *Alzheimer's Disease* tweets contain words relating to "grandmother (grandmom)" and "grandfarther (grandpa)". Note that *cancer*, *lymphoma*, *leukemia* and *pneumonia* are the top four health issues with the largest proportion of tweets containing words of "bless", "pray" and "support" (more than 15%).

Looking back at the NMF results in Figures 2 and 3, it can be seen that, for tweets disclosing another person's health status, the related health issues[8] and language categories are

Category	CC.	Statistic
1st person singular	0.82 ***	1170.45
Adverbs	0.69 ***	2051.63
Time	0.58 ***	2768.42
Quantifiers	0.56 ***	2847.81
Present tense	0.55 ***	2977.65
Body	0.52 **	3169.48
Relativity	0.51 **	3196.49
2nd person	−0.50 **	9833.52
Religion	−0.52 **	9934.07
1st person plural	−0.59 ***	10412.72
Humans	−0.67 ***	10934.02
Family	−0.79 ***	11746.19
Social processes	−0.92 ***	12561.19
3rd person singular	−0.93 ***	12642.93

Table 5: Language categories with a Spearman correlation ≥ 0.50 for the MvY ratio. Note that "CC" represents the correlation coefficient. **$p < 0.01$, ***$p < 0.001$.

[8]Note that pneumonia has a weak signal in Group III.

consistent with Group III. For tweets with a self-disclosed health mention, the related health issues[9] and language categories distribute in Group I, II and IV. This is consistent with M_{1+2+4}, which indicates that all three basis components, in combination, have a positive effect on predicting a higher MvY ratio.

Discussion

This work presented a framework, relying on language categories, to demonstrate that groups of health issues and their semantics are associated with the rate of disclosure for one's self vs. another individual on Twitter. While it is not necessarily the case that disclosure of another individual's health information has transpired without their consent, it is likely that many such disclosures have not been approved. As such, we believe this investigation shows there are opportunities to develop support programs for individuals to utilize (e.g., via private discussion or counseling) before unveiling the health status of their relative or friends. At the same time, we believe that our ability to automatically detect such revelations, suggesting that interventions can be invoked after an initial disclosure to mitigate further revelations.

Effective Language Categories

It is important to recognize that the language categories extracted by LIWC are essential to our framework in several ways. First, the language categories play a more significant role for health mention detection than traditional features based on character n-grams. Second, the language categories enable an avoidance of data sparsity when applying NMF. Third, the groups of health issues, driven by language categories and their semantics (as expressed by the associated language categories), act as factors for learning the motivation behind MvY disclosure.

Groups of Health issues

By applying NMF on the health issue-by-language category model, our investigation suggests there are (at least) four groups of health issues. It is interesting to note that, although B_3 (Group III) is associated with the high cost of medicine and social support, there is a relatively strong signal for the semantic of a positive emotion. This may be due to the fact that some tweets celebrate reversals in diagnosis (e.g., a family member is now cancer free) or an expression of support, such as the following tweet:

> *Love 2 my wife, my hero, whose done w radiation treatment today ...*

Basis component B_4 (Group IV) is also notable because it mainly focuses on negative emotions. These emotions appear to cut across various health issues, including depression, heart attack, hypertension, and kidney stones. Such health issues appear to align with literature on these topics, particularly (De Choudhury et al. 2013; De Choudhury and De 2014) where users with mental health problems (on Twitter and Rediit) have been shown to express negative emotions.

[9]Note that half of the health issues belong to Group II: *asthma*, *insomnia*, *migraine* and *thyroid*.

MvY Disclosure

Our findings show that basis component B_3 has a stronger impact on predicting the MvY ratio than the combination of the other three components (i.e., B_1, B_2 and B_4). At the same time, B_3 has a negative effect, while the other three groups tend to have positive effects. This suggests that the health issues that occur for family members, are associated with the high cost of medicine, and require social support, tend to have a lower MvY ratio. By contrast, for other health issues, where the semantics are associated with biological processes (e.g., health and ingestion) and negative emotions, the authors tend to disclose their own health status.

Limitations and Future Work

There are several limitations in this paper we wish to highlight, which we believe can serve as the basis for further investigation in the topic of health information disclosure in social media. The first limitation is in the fidelity of the health mention prediction model. We tuned the precision of the logistic regression classifier to 81.7%, thus mixing tweets without health mentions into the NMF analysis. As this research evolves, it will be useful to build more robust health mention classifiers, which may be possible by incorporating more variety in the training space (e.g., via additional health issues) and features (e.g. extracted from rich social context). The second limitation is in the factorization and resulting groups. Specifically, it should be noted that the number of basis components is determined by optimization for several coefficients associated with NMF decomposition. Regularizations, constrained by external factors (e.g., the severity and social stigma of health issues), are worth considering to derive a more interpretable matrix factorization. In particular, it would be worthwhile to investigate if there exist certain clinical factors that drive the formation of groups of health issues and MvY disclosure.

Conclusion

This investigation illustrates that users of Twitter disclose health information about themselves as well as others. We introduced a novel framework to detect tweets associated with 34 health issues and associate their underlying semantics with the rate at which (self vs. another person) disclosure transpires. This framework consists of 1) data collection (extracting tweets from Twitter stream), 2) health mention annotation (AMT labeling), 3) health mention classification (constructing more tweets with health mentions), 4) similar health issue discovery (a non-negative matrix factorization (NMF) approach), and a me vs. you (MvY) analysis, which is based on the learned groups and semantics discovered through factorization. Our findings highlight that the authors of tweets tend to disclose information about another persons health status when talking about the high cost of medicine or treatment and when searching for social support, but disclose information about their own health status when talking more benign health issues related with simple chronic biological processes and negative emotions. We anticipate extending this work to include more robust health

mention classifiers and regularizing NMF to obtain more interpretable basis components from the factorization process.

Acknowledgements

This research was supported, in part, by grants from the Smart and Connected Health Program of the U.S. National Science Foundation (IIS-1418504 and IIS-1418511), the Patient Centered Outcomes Research Institute (1501-26498, the Mid-South Clinical Data Research Network) and National Institutes of Health (K99LM011933).

References

Ammari, T., and Schoenebeck, S. 2015. Understanding and supporting fathers and fatherhood on social media sites. In *Proc. CHI*, 1905–1914.

Ammari, T.; Morris, M. R.; and Schoenebeck, S. Y. 2014. Accessing social support and overcoming judgment on social media among parents of children with special needs. In *Proc. ICWSM*, 22–31.

Caine, K., and Hanania, R. 2013. Patients want granular privacy controls over health information in electronic medical records. *Journal of the American Medical Informatics Association* 20(1):7–15.

Claypoole, T. F. 2014. Privacy and social media. *Business Law Today*.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015a. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proc. of the Workshop on CLPsych*.

Coppersmith, G.; Dredze, M.; Harman, C.; and Hollingshead, K. 2015b. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proc. of the Workshop on CLPsych*.

Coppersmith, D.; Harman, C.; and Dredze, M. 2014. Measuring post traumatic stress disorder in Twitter. In *Proc. ICWSM*, 579–582.

De Choudhury, M., and De, S. 2014. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proc. ICWSM*, 71–80.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Proc. ICWSM*, 128–137.

De Choudhury, M.; Morris, M. R.; and White, R. W. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proc. CHI*, 1365–1376.

Eschler, J.; Dehlawi, Z.; and Pratt, W. 2015. Self-characterized illness phase and information needs of participants in an online cancer forum. In *Proc. ICWSM*, 101–109.

Fisher, J., and Clayton, M. 2012. Who gives a tweet: assessing patients interest in the use of social media for health care. *Worldviews on Evidence-Based Nursing* 9(2):100–108.

Hayes, M.; Ross, I.; Gasher, M.; Gutstein, D.; Dunn, J.; and Hackett, R. 2007. Telling stories: News media, health literacy and public policy in Canada. *Social Science and Medicine* 64(9):1842–1852.

Kivran-Swaine, F.; Ting, J.; Brubaker, J. R.; Teodoro, R.; and Naaman, M. 2014. Understanding loneliness in social awareness streams: Expressions and responses. In *Proc. ICWSM*, 256–265.

Lin, W.-Y.; Zhang, X.; Song, H.; and Omori, K. 2016. Health information seeking in the web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. *Computers in Human Behavior* 56:289–294.

Mao, H.; Shuai, X.; and Kapadia, A. 2012. Loose tweets: an analysis of privacy leaks on Twitter. In *Proc. WPES*, 1–12.

McNeil, K.; Brna, P.; and Gordon, K. 2012. Epilepsy in the Twitter era: a need to re-tweet the way we think about seizures. *Epilepsy and Behavior* 23(2):127–130.

Meslin, E.; Alpert, S.; Carroll, A.; Odell, J.; Tierney, W.; and Schwartz, P. 2013. Giving patients granular control of personal health information: using an ethics "Points to Consider" to inform informatics system designers. *International Journal of Medical Informatics* 82(12):1136–1143.

Naslund, J. A.; Grande, S. W.; Aschbrenner, K. A.; and Elwyn, G. 2014. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. *PLoS One* 10:e110171.

Paul, M. J., and Dredze, M. 2014. Discovering health topics in social media using topic models. *PLoS One* 9(8):e103408.

Pew Research Center. 2014. Public perceptions of privacy and security in the post-Snowden era.

Quan, X.; Kit, C.; Ge, Y.; and Pan, S. J. 2015. Short and sparse text topic modeling via self-aggregation. In *Proc. ICWSM*, 2270–2276.

Sridhar, V. K. R. 2015. Unsupervised topic modeling for short texts using distributed representations of words. In *Proc. NAACL-HLT*, 192–200.

Tamersoy, A.; De Choudhury, M.; and Chau, D. H. 2015. Characterizing smoking and drinking abstinence from social media. In *Proc. HT*, 139–148.

Tamir, D., and Mitchell, J. 2012. Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences USA* 109(21):8038–8043.

van der Velden, M., and El Emam, K. 2013. 'Not all my friends need to know': a qualitative study of teenage patients, privacy, and social media. *Journal of the American Medical Informatics Association* 20(1):16–24.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(1):3–36.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proc. WWW*, 1445–1456.

Yin, Z.; Fabbri, D.; Rosenbloom, S. T.; and Malin, B. 2015. A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research* 17(6):e138.

Zhang, Z.-Y. 2012. Nonnegative matrix factorization: Models, algorithms and applications. In *Data Mining: Foundations and Intelligent Paradigms*. Springer. 99–134.