A Method for Analyzing Health Behavior in Online Forums

Rose Yesha
Department of Information Systems
University of Maryland Baltimore County
Baltimore, MD 21250
yrose@umbc.edu

Aryya Gangopadhyay
Department of Information Systems
University of Maryland Baltimore County
Baltimore, MD 21250
gangopad@umbc.edu

ABSTRACT

The prevalence of online social networks has enabled users to communicate, connect, and share content. Many of these networks serve as the de-facto Internet portal for millions of users. Due to the enormous popularity of these sites, the data about the users and their communications offer an enormous opportunity to analyze human behaviors on a large scale. It is important to analyze patterns within these records in order to more effectively treat individuals. In this paper, a method is presented for identifying these themes and patterns within forum data. This methodology includes automatic extraction of the main themes or patterns in the data, quantify the similarities and differences in the contents of different online forums, and finding similar documents based on user queries. We used data sets from four different forums. In this paper we describe a method that automatically differentiates between online discussion groups related to different behavioral health challenges and identifies the most appropriate discussion forum for a given input. Finally, we evaluated the efficacy of our method by using cross-validation.

Categories and Subject Descriptors

J.3 [**Health**]: Medical Information Systems—*Health IT, Behavioral Health*

1. INTRODUCTION

In the past decade, there has been a complete transformation in the type of data available on the Internet. Going forward from the year 2000, user-generated content has become increasingly popular on the web, and more and more users participate in content creation, instead of only participating in consumption [13]. Social media sites include web forums, photo and video sharing communities, and social networking platforms that offer combinations of all of these features. Particularly, there is an emphasis on the relationships among users within the community. Social media data has completely revolutionized the way in which human beings interact with one another. The prevalence of online social networks has enabled users to communicate, connect, and share content. Many of these networks serve as the de-facto Internet portal for millions of users [3]. Due to the enormous popularity of these sites, the data about the users and their communications offer an enormous opportunity to analyze human behavior on a large scale. When presented with unstructured data, such as that derived from raw forum data, the sheer volume of text makes the process seem impossible. This is where the role of text analytics takes the central stage. Text analytics seek to derive meaning from the text data. This task is very complex since human communication is so context dependent. This calls for well-designed text analytic techniques that are able to identify the major interactions between various terms on specific topics in big data platforms, and the ability to make sense of this data using prediction models and monitoring tools. This paper proposes the analysis of forum data for identifying behavioral health patterns, which aims to explain these techniques and their implications for behavioral health. The disorders explored in our methodology include posts from four various forums including OCD, Bluelight Marijuana, Substance Abuse, and suicide. This paper proposes the analysis of social media data for identifying behavioral health patterns, which aims to explain these techniques and their implications for behavioral health.

The rest of the paper is organized as follows: in Section 2 we present research that are related to this paper, in Section 3 we describe our methodology, and in Section 4 we present the experimental results followed by our conclusions in Section 5.

2. BACKGROUND AND RELATED WORK

Mental health conditions affect a large percentage of individuals each year. Traditional mental health studies have relied on information collected through contact with the mental health practitioner. There has been research on the utility of social media for depression, but there have been limited evaluations of other mental health conditions [13]. In the following part of this chapter, we will examine specific techniques that have previously been used to analyze forum data, define behavioral health and public health issues, and lastly, we will explore the implications that this research has for big data analytics.

2.1 Analysis of Social Media

In this part of the paper, we will explore the various techniques that have been previously used to analyze the data found in social media sites. The rise of social media sites, forums, blogs, and other communications tools has created an online community of individuals who are able to socialize and express their thoughts through various applications [22]. Microblogging has become a very popular tool for communication among users. The individuals who write these messages blog about their lives share opinions, and discuss current events. As more individuals participate in these micro blogging services, more information about their messages becomes available. The massive amount of data in user updates creates the need for accurate and efficient clustering of short messages on a larger scale [9]. Certain research areas have chosen to focus on the opinions and sentiments of these messages [25], community detection [19], politics [26], and user interests [15]. Techniques for clustering this data have included document clustering, topic modeling sentiment analysis, and text mining.

2.1.1 Topic Modeling

Recent years have seen a surge in information that is both digitized and stored. As this trend continues, it has become increasingly difficult for users to find what they are looking for. Novel computational tools are needed to help organize, search, and comprehend these large amounts of data [8]. Currently, we are able to type keywords into a search and find documents that are related to them. However, there is a crucial element that is missing from this process. Specifically, it is important to utilize themes to explore specific topics. A thematic structure could serve as a portal through which users could explore and obtain knowledge about various topics. Topic modeling algorithms are statistical methods that analyze the words of the original documents and discover themes that occur. Furthermore, topic modeling analyzes how these themes relate to one another, and how they differ over time [6]. These algorithms do not need any previous annotations or labeling of the documents, these topics surface automatically form the analysis of the original texts. Blei [6] describes latent Dirichlet allocation (LDA), which is the simplest type of topic model. LDA is a statistical model of a collection of documents that tries to validate the intuition that documents exhibit multiple topics. The simple LDA model provides an effective and powerful way to discover and exploit the hidden thematic structures found in large amounts of text data.

2.1.2 Sentiment Analysis

Microblogging websites have developed into a source for varied types of information. Individuals post messages about their opinions, current events, complaints, and sentiments about products they use in their daily lives [16]. It is very often that companies study these user reactions on microblogging sites. The challenge then becomes how to build a technology that can detect and summarize an overall sentiment. A large amount of social media contains sentences that are sentiment-based. Sentiment is defined as a personal belief or judgment that is not founded on proof or certainty [11]. Sentiment involves the use of Natural Language Processing (NLP), statistics, or machine learning methods to extract, identify, or characterize the sentiment content of a text source [16]. The automated identification of sentiment types can be beneficial for many NLP systems.

2.1.3 Text Mining

Text mining is the discovery of new information by automatically extracting information from a large amount of various unstructured textual resources [2]. Text mining can help an organization gain valuable insights from text-based content such as word documents, email, and postings on social media sites like Facebook, Twitter and LinkedIn [24]. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging because natural language text is usually inconsistent. It contains ambiguities caused by inconsistent syntax and semantics. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. By using text analytics, an organization can successfully gain insight into content specific values such as emotion, sentiment, intensity and relevance. Text mining techniques include methods for corpus handling, data import, metadata management, preprocessing, and the creation of term-document matrices. The main structure for managing documents in is a corpus, representing a collection of text documents.

2.2 Behavioral Health

Behavioral health can be classified into several different categories, depending on the type and severity of the mental health disorder. Mental health care practitioners rely on specific evaluation criteria. This includes criteria found in Diagnostic and Statistical Manual of Mental Disorders (DSM) and data collected from self-reported instruments. Currently, over 61.5 million Americans experiences a mental illness in any given year. One in 17, about 13.6 million, have a serious mental illness such as major depression, schizophrenia, or bipolar disorder [18]. About 60 percent of adults and almost one half of youth ages 8 to 15 with a mental illness did not receive mental health services in 2013 [14].

2.2.1 Suicide

Many individuals at risk of suicide do not seek help prior to an attempt. Furthermore, they do not remain connected to any mental health services following the attempt [1]. E-12 health interventions are now being defined as a means to identify individuals who are at risk, offer self-help, or deliver interventions in response to user posts on the internet. Patterns found in users' social media usage can be especially indicative of suicide ideation. Research shows that there is some evidence to suggest that social media platforms can be used to identify individuals or geographical areas at particular risk for suicide. Specific language used in tweets can give practitioners and other Twitter users information about an individual's mental health status. Recent studies found specific tweets by users who both tweeted about suicidal ideations. One quote stated "people say "stop cutting! be happy with who you are." It's so much easier to say than do? i hate myself so much.." [7]. Another tweeter posted, "I'm so sick of being bullied. Everyone care about their problems and don't even bother to check on me. I'm going to kill myself!!" [7] It is evident from these tweets that intervention is possible. The few studies done in this area have shown that it is possible to use computerized sentiment analysis and data mining to identify users at risk for suicide.

2.2.2 *Depression*

Many have begun turning towards online communities for help in understanding and dealing with symptoms. Nimrod [20] examined the content online forum discussion of depression in order to explore the potential benefits they could offer people with depression. Quantitative content analysis of one year of data from 25 top online communities was performed, using the Forum Monitoring System. Content analysis revealed nine main subjects discussed in the communities, including (in descending order) “symptoms”, “relationships”, “coping”, “life”, “formal care”, “medications”, “causes”, “suicide”, and “work”. The results indicated that online depression communities serve as a place for sharing experiences and receiving techniques for coping [20]. Searching for online health information and searching within social media sites are both on-going difficulties users face [27]. There are many reasons that these social media platforms are a valuable source of health information. For example, social media provides an important tool for people with health concerns to talk to one another. Also, these sites are well known as a source of tacit information, that is less common online. Wilson et al. [29] focused their study on a prevalent mental health issue, depression. Depression has increased substantially in developed and developing countries [5], and it is estimated to affect over 350 million people [28]. Depression affects more than 27 million Americans and is believed to be responsible for more than 30,000 suicides every year [8, 17]. Although discussing issues related to depression with others is seen to be an important facet of coping, personal factors discourage people from doing so in real life [4]. Therefore, social media sites provide an outlet for people to communicate with potentially millions of people, while reducing the consequences of real life disclosure [12, 21]. More users are choosing to share their thoughts and emotions that encompass their daily lives. The language and emotion used in social media posts may include feelings of worthlessness, helplessness, guilt, and self-hatred, which are all characteristic of depression. The characterization of social media activity can provide a measurement of depression symptoms in a manner that could help detect depression in populations. Choudhury et al. [10] examined the use of social media as a behavioral assessment tool. In contrast to behavioral health surveys, social media measurement of behavior captures social activity and language expression in a naturalistic setting [10].

3. METHODOLOGY

3.1 Research Objectives

The overall objective of our research is to develop an automated method to extract the main topics of discussions in online forums related to behavioral health. A practical application of such a methodology in the context of online forums is to be able to identify the appropriate discussion forum for a given user interest. In order to achieve that goal, our method needs to be able to distinguish between discussion forums on different behavioral health issues. In this paper we describe a method that automatically differentiates between online discussion groups related to different behavioral health challenges and identifies the most appropriate discussion forum for a given input (query). Our main research objectives are as follows:

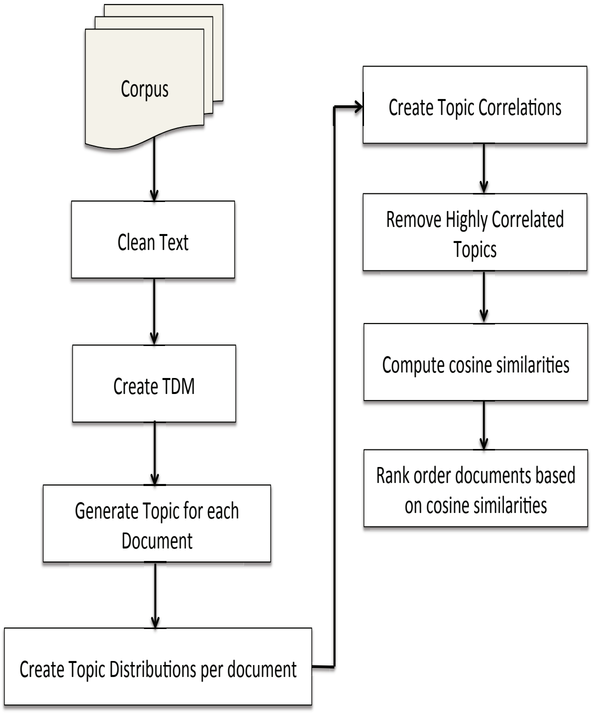
Corpus
Clean Text
Create TDM
Generate Topic for each Document
Create Topic Distributions per document
Create Topic Correlations
Remove Highly Correlated Topics
Compute cosine similarities
Rank order documents based on cosine similarities

Figure 1: Overall Architecture

1. How to we automatically discover the main topics in online discussion forums?
2. How do we differentiate between discussion forums on different behavioral health issues?
3. How do we determine the similarities between discussion forums?
4. How do we retrieve the most appropriate discussion forum given a user query?

In this paper we have explored *topic modeling* [6] as the underlying methodology for automatically discovering the topics in a discussion forum. However, there are two challenges with topic modeling: determining the appropriate number of topics, and identifying the set of topics that will enable us to differentiate between discussion forums for different behavioral health challenges. We will discuss how we have dealt with both issues in the following sections. Furthermore we present a novel approach of scoring discussion forums in terms of their similarities and differences. High similarity score is equivalent to a low difference score between two discussion forums.

3.2 Overall Architecture

There are eight steps in our proposed methodology, as shown in Figure 1. The input to the system is a corpus of documents that consist of discussion threads in online forums. Since most of the discussions are fairly informal the documents need to be cleaned before further processing. Examples of cleaning includes removing non-alphanumeric characters, redundant spaces, headers, URLs., punctuations, etc. All characters in the input documents are converted to lowercase characters, and common stop words such as articles, prepositions, and conjunctions are removed. We did not do any stemming in this work but it would be part of the cleaning process in future work. Next we create a Term-Document Matrix (TDM) that consists terms in the rows and the documents in the matrix. The entries in the TDM corresponds to the frequency of occurrence of each term in each document. Instead of term frequencies one can use other measures such as TF/IDF (term frequency/ inverse document frequency). Next, the topics are created for each document. From the topic distributions across all documents we create correlations among the topics and remove the topic(s) that are highly correlated with other topics. We identified the most appropriate forum for a user query by calculating the cosine similarity between the topic distribution of the user query against that of all documents and ordered from high to low in the cosine similarities.

3.3 Topic Generation and Similarity Scoring

In this research we used LDA for topic modeling. The assumption is that each discussion forum is a random mixture of latent topics where each topic is a discrete distribution over the vocabulary in the corpus of discussion forums. The number of latent topics is assumed to be known and can be varied. In our experiments we assumed that the number of latent topics is 10. Each topic is a discrete probability distribution over a fixed vocabulary (set of all terms found in the corpus) represented as ϕ, and each document is a discrete probability distribution over the available topics, denoted by θ. Both distributions are assumed to have been drawn from the symmetric Dirichlet distribution with hyperparameters β and α respectively. Furthermore, each word w_i has a topic index z_i.

The goal is to determine the unobserved (latent variables) z, ϕ, and θ, given the observed data (words in the forums). This is often referred to as *posterior inference*, which is an intractable problem. However, approximate inference techniques are available including variational methods and Gibbs sampling. In this paper we have used collapsed Gibbs sampling [23] that reduces the problem to the estimation of the probability z from which θ and ϕ can be derived. In this method we start with random topic assignment of each word in each document. Next we calculate the probability of topic assignment for word w_i given the topic assignments of all other words: $p(z_i|z_{-i}, \alpha, \beta)$, where z_{-i} indicates all topic assignments except z_i using a Markov Chain Monte Carlo (MCMC) method that constructs a Markov chain of posterior distributions of the latent variables given the observed data. The method eventually converges to a stationary distribution.

4. RESULTS

4.1 Data

The data for Suicide, Substance Abuse, and OCD (obsessive compulsive disorder) was all taken from forums found within the website takethislife.com. These forums do not serve as a substitute for professional help, however, many users find that sharing experiences with others and connecting with others helps them alleviate their current mental state. Members from many different countries, vary in age, and all have different circumstances. These posts may include inspirational pictures, art and other creative works, and other helpful information on mental health conditions.

The data for Marijuana use was taken from a forum thread found on bluelight.org, Bluelight is an international, online harm-reduction community, committed to reducing the risk associated with drug use. Bluelight also hosts a recovery community for those seeking an end to their drug abuse and addiction. These forums invite visitors to discuss addiction and sobriety, share recovery resources and encourage members to seek help.

4.2 Topic Generation

As mentioned in a previous section we used Gibbs sampling to generate the topics from the online discussion forums. In each iteration we calculate the log-likelihoods of the probabilities of the word assignments to topics and topic assignments to the documents. The log-likelihoods represent how well the models capture the documents which are the observed data. The actual values of the log-likelihoods are not relevant, but of relevance are the relative values after each iteration. After a number of iterations the log-likelihood values tend to reach a steady state and remains flat afterwards, When this point is reached we stop. The log-likelihoods are shown in Figure 2 where the number of iterations are shown on the x-axis and the log-likelihood values are shown in the y-axis for one of the datasets. As can be seen from Figure 2, the log-likelihoods approached convergence after 100 iterations. This was consistent across all of the data sets in our experiments. We generated 10 topics for each document and 20 terms with the highest probabilities for each topic.

4.3 Topic Selection

The next task in our method is to select the topics that would allow us to best discriminate the documents. This is crucial to achieve high predictive accuracy in document (forum) identification for user queries. In order to accomplish this we first created the topic distributions across documents. From the topic distributions we calculated the correlations among the topics which are shown in Figure 3, where high positive correlations are shown in blue with hashing from lower left to upper right, high negative correlations are shown in red with hashing from upper left to lower right. The darker colors show stronger correlations. Weaker correlations are shown in lighter colors. Thus, it can be seen that topic 8 has strong positive correlations with topics 3, 7, 6, 5, 1 and 4 in the order of high to low.

We used the topic correlations to remove topics that are potentially less useful in document discrimination and hence document prediction give user queries. This is illustrated in the stacked bar plots shown in Figure 4, where each stacked

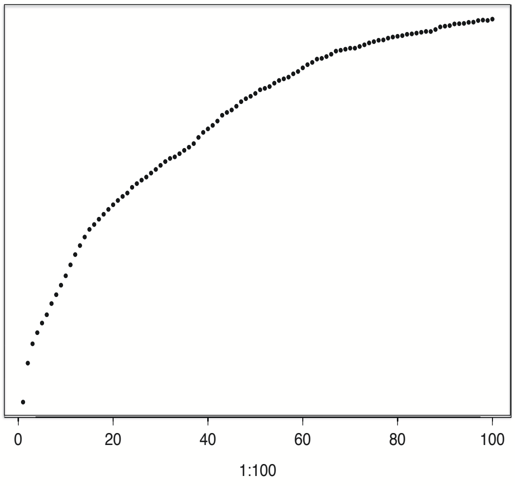

Figure 2: Convergence of log-likelihoods using collapsed Gibbs sampling

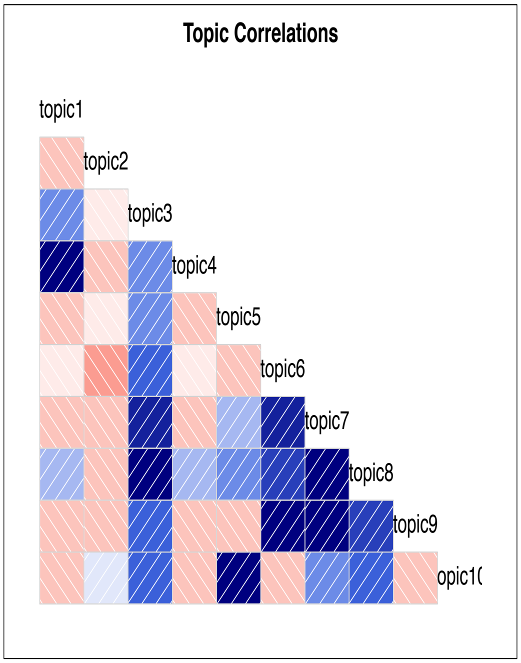

Figure 3: Correlations among Topics

bar shows a different color coded topic. The bars on the left of the figure shows the topic distributions for the four documents with all 10 topics. As can be noted the bar corresponding to topic 8 is the largest for all four documents indicating that topic 8 has the maximum number of occurrence of words in all four documents. Hence it would be difficult to differentiate between the documents using topic 8. We found out that the terms in topic 8 are mostly irrelevant to any of the behavioral health issues. The stack bars on the right show the topic distributions in the same four documents after removing topic 8. It can be seen that removing topic 8 resulted in a visibly different pattern in each of the four documents.

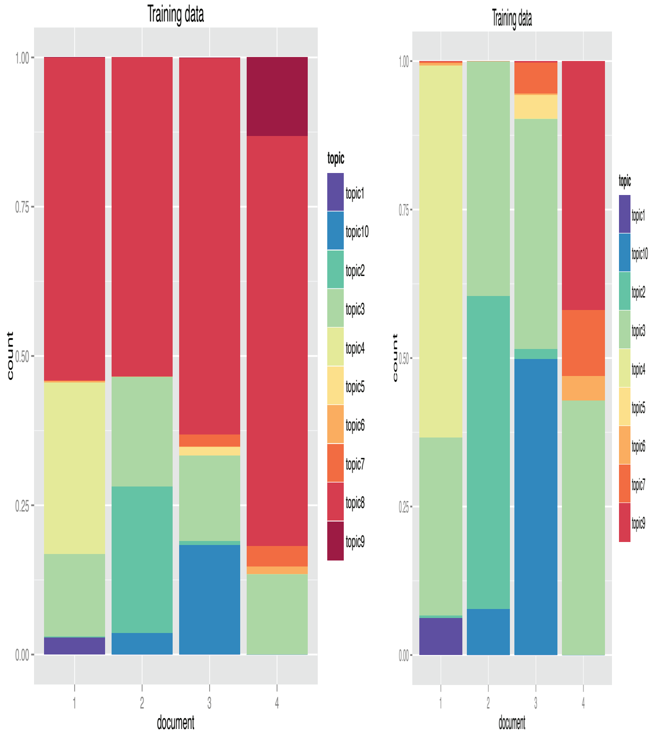

Figure 4: Topic Distributions

4.4 Cross Validation

Each document was divided into two disjoint sets: *training* and *testing*. Around 95% of the discussion threads were set aside for training and the remaining 5% were used for testing. The test data were used as examples of user queries and the goal was to predict the *closest* discussion forum for each user query. In order to identify the most appropriate discussion forum for a given user query we created the topic distribution of the user query as explained in Section 4.3 and computed the cosine similarity between the query and each document.

We performed 4-fold cross validation where each fold contained a different set of training and test data. The results are shown in Figure 5 where the first four graphs show the prediction results of the cross validation. The fifth chart shows the experiment in which 80% of the data was set for training and the rest 20% for testing. The test sets were used as queries and the training set represented the documents.

First, we created the topic distribution for each query using the same process as described in Section 3. Next we computed the cosine similarity of each query with each document. The bar charts in Figure 5 represent the cosine similarities for query 1–query 5 against the documents. The

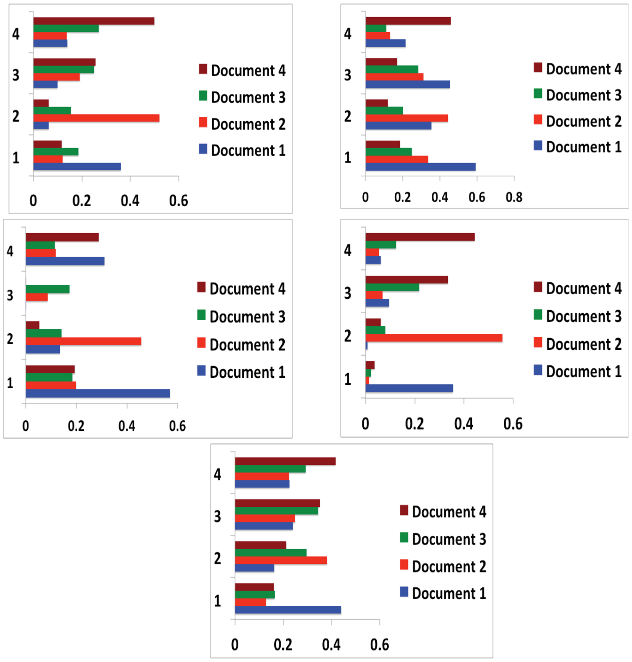

Figure 5: Prediction results

numbers in the y-axis represents the query number. The size of the bars in Figure 5 corresponds to the cosine similarities. Thus in the chart at the top left corner, the cosine similarities of query 1 are 0.36, 0.12, 0.19, and 0.12 for documents 1-4 respectively. The top 4 charts represents the results of the 4-fold cross validation with different training and test sets as described above. The chart at the very bottom represents the test results with 80% of the data for training and the rest 20% for testing. The results show an average prediction accuracy of 75%.

5. CONCLUSIONS

The prevalence of online social networks has enabled users to communicate, connect, and share content. In this paper, a method is presented for identifying these themes and patterns within patient data. This methodology includes extraction of the main themes or patterns in the data and linking those themes back to the corpus from which they were generated. Our research objectives included identifying the topics being discussed in an online forum, identifying the set of topics that would allow us to differentiate between discussion forums on different behavioral health issues, determining similarities between discussion forums, and retrieving the most appropriate discussion forum data given a user query. In this research, we used LDA for topic modeling under the assumption that each discussion forum was a random mixture of latent topics where each topic is a discrete distribution over the vocabulary in the corpus of discussion forums. A 4-fold cross validation was performed where each fold contained a different set of training and test data. The results showed an average prediction accuracy of 75%.

Our future research plan includes investigating the efficacy of our methods in larger number of discussion forums, improving search results to retrieve relevant forums based on user queries, and visualizing the topics discovered.

6. REFERENCES

[1] Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining twitter for suicide prevention. In *Natural Language Processing and Information Systems - 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*, pages 250–253, 2014.

[2] Charu C. Aggarwal and ChengXiang Zhai, editors. *Mining Text Data*. Springer, 2012.

[3] R. Albert and A.-L. Barbassi. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[4] M. Back, J.Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3):372–374, 2010.

[5] BBC, 2013. Available online at http://www.bbc.com/news/health-23192252.

[6] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April 2012.

[7] Scott Burton, Christophe Giraud-Carrier, and Carl Hanson. Tracking suicide risk factors through twitter in the us. *Crisis*, 35(1), 2014.

[8] CDC. Available online at http://nccd.cdc.gov/s_broker/ WEATSQL.exe/weat/freq_year.hsql.

[9] Zhiyuan Chen and Bing Liu. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 703–711, 2014.

[10] Munmun De Choudhury, Michael Gamon, Aaron Hoff, and Asta Roseway. ”moon phrases”: A social media faciliated tool for emotional reflection and wellness. In *7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, PervasiveHealth 2013, Venice, Italy, May 5-8, 2013*, pages 41–44, 2013.

[11] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 241–249, 2010.

[12] Miles Efron and Megan Winget. Questions are content: A taxonomy of questions in a microblogging environment. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 27:1–27:10, Silver Springs, MD, USA, 2010. American Society for Information Science.

[13] Jan-Are K. Johnsen, Jan H. Rosenvinge, and Deede Gammon. Online group interaction and mental health: An analysis of three online discussion forums. *Scandinavian Journal of Psychology*, 43(5):445–449, 2002.

[14] Byron W. Keating, John A. Campbell, and Peter Radoll. Evaluating a new pattern development process

for interface design: Application to mental health services. In *Proceedings of the International Conference on Information Systems, ICIS 2013, Milano, Italy, December 15-18, 2013*, 2013.

[15] Huayi Li, Arjun Mukherjee, Bing Liu, Rachel Kornfield, and Sherry Emery. Detecting campaign promoters on twitter using markov random fields. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 290–299, 2014.

[16] Bing Liu. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[17] J. Luoma, C. Martin, and J.Pearson. Contact with mental health and primary care providers before suicide: a review of the evidence. *American Journal of Psychiatry*, 159(6):909–916, 2002.

[18] Mark Matthews, Saeed Abdullah, Geri Gay, and Tanzeem Choudhury. Tracking mental well-being: Balancing rich sensing and patient needs. *IEEE Computer*, 47(4):36–43, 2014.

[19] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review*, 2004.

[20] Galit Nimrod. From knowledge to hope: online depression communities. *International Journal on Disability and Human Development*, 11(1):23–30, 2012.

[21] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010.

[22] Georgios Paltoglou and Mike Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Trans. Intell. Syst. Technol.*, 3(4):66:1–66:19, September 2012.

[23] Ian Porteous, David Newman, Alexander T. Ihler, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 569–577, 2008.

[24] Maria-Evgenia G. Rossi, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. Spread it good, spread it fast: Identification of influential nodes in social networks. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 101–102, 2015.

[25] Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. Exploiting social relations and sentiment for stock prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1139–1145, 2014.

[26] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[27] Ryen W. White and Eric Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst.*, 27(4), 2009.

[28] WHO. Available online at http://www.who.int/mental_health/prevention/suicide/wspd/en/.

[29] R. Wilson, A. Capuano, P. Boyle, G. Hoganson, L. Hizel, R. Shah, S. Nag, J. Schneider, S. Arnold, and D. Bennett. Clinical-pathologic study of depressive symptoms and cognitive decline in old age. *Neurology*, 83(8):702–709, 2014.