Mining Twitter Data to Improve Detection of Schizophrenia

Kimberly McManus, BS[*1], Emily K. Mallory, BS[*1], Rachel L. Goldfeder, MS[*1], Winston A. Haynes, BA[*1], Jonathan D. Tatum, BS[*1]
[1]Stanford University, Stanford, CA

Abstract

Individuals who suffer from schizophrenia comprise 1 percent of the United States population and are four times more likely to die of suicide than the general US population. Identification of at-risk individuals with schizophrenia is challenging when they do not seek treatment. Microblogging platforms allow users to share their thoughts and emotions with the world in short snippets of text. In this work, we leveraged the large corpus of Twitter posts and machine-learning methodologies to detect individuals with schizophrenia. Using features from tweets such as emoticon use, posting time of day, and dictionary terms, we trained, built, and validated several machine learning models. Our support vector machine model achieved the best performance with 92% precision and 71% recall on the held-out test set. Additionally, we built a web application that dynamically displays summary statistics between cohorts. This enables outreach to undiagnosed individuals, improved physician diagnoses, and destigmatization of schizophrenia.

Background

Individuals who suffer from schizophrenia comprise 1 percent of the United States population and are four times more likely to die of suicide than the general US population.[1] Schizophrenia presents as a combination of positive and negative symptoms. Positive symptoms, including hallucinations, delusions, and thought disorders, generally develop between the ages of 16 and 30, with earlier onset in men than women.[2] Negative symptoms, those that "disrupt normal emotions and behaviors", include an inability to finish tasks and a lack of emotions and/or speech and may be misdiagnosed as depression or other disorders. Schizophrenia affects men and women equally and is found in several ethnicities at similar rates. Despite the high prevalence of schizophrenia, there is a lack of broad discussion about the disease, especially when compared to other, less prevalent disorders.[3]

Schizophrenia can have numerous debilitating consequences (e.g., individually, socially, economically) if left untreated. Individuals with schizophrenia are at an increased risk for suicide, homelessness, and incarceration, with 6% of individuals with schizophrenia are either homeless or live in a shelter and another 6% are incarcerated.[3] This negatively impacts the afflicted individual and their family, and leads to an increased economic burden. In the US, the majority of individuals with schizophrenia in prison were found guilty of non-serious charges, most often caused by lack of treatment.[4] Treatment for schizophrenia effectively reduces symptoms and improves quality of life; however, nearly 50% of afflicted individuals remain untreated, mostly due to a lack of awareness for their illness.[5] This lack of awareness hinders individuals from seeking treatment, which in turn makes the identification of individuals with schizophrenia challenging.

Microblogging is a general term for a form of social media where users share abbreviated messages. Prominent examples of microblogging include Twitter, Facebook, Sina Weibo, Google+, and Tumblr. An estimated 73% of adults use some form of social media.[6] In our analysis, we focus on Twitter, which, as of January 2014, has 645 million active users generating 58 million tweets (Twitter posts) every day.[7] Twitter posts are a maximum of 140 characters in length and can include photographs and/or links. Twitter use is particularly prevalent in individuals between the ages of 18 and 30 (31% usage), which overlaps with the standard onset age for schizophrenia.[8]

Sentiment analysis is the identification of mood characteristics from textual word usage.[9–11] In particular, paranoia shows specific language patterns which can be detected using sentiment analysis.[12] Several groups have developed straightforward and effective adaptations of existing sentiment analysis approaches for analyzing sentiment using tweets.[13–15] Features of Twitter text, including the use of hashtags and emoticons, can also be mined for fine-tuned sentiment analysis.[16]

Table 1 summarizes existing approaches for identifying depression based on social media presence. Across these studies, the manual curation required for identifying a depressed cohort was a major challenge. Additionally, feature identification for input into the machine learning models required careful consideration. In particular, social interactions and micro-blog usage patterns were useful traits that are not included in traditional sentiment analysis. BlueFriends draws on the methodologies detailed above to visualize the percentage of a user's Facebook friends that show signs of depression, with the goal of reducing the stigma of depression.[17]

---

*Co-first authors

Table 1. Approaches for classifying depression using social media.

Ref	Media	Cohort Acquisition	Features	Approach	Results
[18,19]	Twitter	Clinical depression surveys	Interactions, emoticons, vocabulary, drugs, linguistic style, behaviors	Support vector machine	0.74 precision 0.63 recall
[20]	Bulletin boards	Prozac post, doctor curation	Vocabulary	2-step support vector machine	0.82 accuracy AUC 0.88
[21]	Sina	Psychologist curation	Pronoun use, emoticons, interactions, behaviors	Weka, BayesNet	0.91 accuracy AUC 0.90

While applications of machine learning algorithms to microblog and other online posts have been successful for identifying depression, further work is needed to show applicability of microblog posts to other mental disorders. Schizophrenia provides a distinct challenge from depression due to its heterogeneous presentation. In this work, we built a framework to distinguish individuals with schizophrenia from control individuals using Twitter data. Furthermore, we provide a web application for interrogation of our results.

Methods

We identified a cohort of Twitter users who self-identify, as part of their user profile or Twitter posts, as having schizophrenia (cases) and another group of Twitter users who do not self-identify as having any mental disorder (controls). All data is from English-speaking users and was extracted in April 2014. We defined a user as self-identifying with schizophrenia if two or more of the following held: the user self-identifies in user description; the user self-identifies in status updates; the user follows @schizotribe, a known Twitter community of users with schizophrenia. We utilized the Twitter API to extract relevant English-speaking users and status updates.

We utilized the Twitter API's 1% random stream, a random 1% sample of statuses as they are posted, and manually curated Twitter posts to define the control set of Twitter users. To control for potential biases of users posting at particular times of day, we extracted users every two hours (from 9am-5pm PST) throughout one day. All potential control Twitter accounts were also manually curated. Spam accounts and accounts where a user self-identified with any mental disorder were excluded from the control group

From this initial set of control users, we developed a second, smaller control group [Table S2] by matching the age distribution of our cohort of users with schizophrenia. Twitter does not explicitly save the ages and genders of users, so user accounts were manually curated for this information.

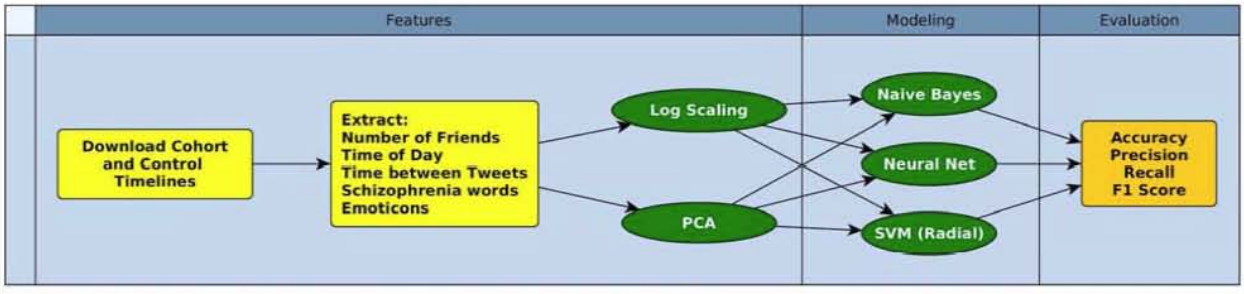

Figure 1. Analysis workflow for feature extraction, model building, and evaluation.

With this set of 96 cases and 200 controls, we used Twitter data, data processing techniques, and machine learning techniques to discover patterns of Twitter usage. We first extracted features from the Twitter text and usage patterns, selected a reduced set of these features, trained several machine learning models, and evaluated the performance of these models [Figure 1].

In order to distinguish Twitter users with schizophrenia from controls, we extracted a set of features from each user's profile and posting history. The feature set was largely derived from one shown to work well in the

classification of users with depression [Figure 1, Left].[18,19] Additionally, we added more refined features for describing emoticon use and the use of schizophrenia-related words [Supplementary Methods]. We used the Natural Language Toolkit (NLTK) for Python to perform tokenization and lemmatization, before extracting textual features and NumPy for generating the final numeric feature vectors. [22,23]

The final 28 numerical features included: number of Twitter followers and number of followed users, proportion of tweets using schizophrenia-related words, emoticon usage, posting time of day, and posting rate [Tables S1-S4]. To include information about posting time and posting rate, we chose to use quantiles of the users' distributions of tweet time of day and the delay between tweets [Table S2]. We observed that these distributions were not Gaussian at the population level, so we suspected that using quantiles would provide a more stable signal. We chose the median, upper and lower quartiles, and the 10th and 90th percentile of each distribution to provide estimates of both central events and events in the tails of the distribution. In addition to the raw feature vectors, we tested two transformations of the feature vectors for each of the models: log scaling of the delay between tweets and Principal Components Analysis (PCA).[24]

We trained several classification models with complementary strengths and weaknesses [Figure 1, Middle, Supplementary Methods]. These models included Naive Bayes (NB), artificial neural networks (ANNs), and support vector machines (SVMs).[25,26] The specific feature set used by the model and the hyperparameters of that model were tuned using 5-fold cross validation on the training data.

We separated the data into an 80% training set and 20% held-out test set for model tuning and evaluation. To provide a less biased estimate, tuning was performed exclusively on the training set and final performance is reported on the test set. Performance of tuned models was evaluated by calculating accuracy, precision, recall, and the F1 score [Figure 1, Right] on the held-out set. Precision and recall represent the goals of maintaining a low false positive rate and a low false negative rate, respectively, whereas the F1 score gives a single metric for precision and recall that is high when both of these values are high.

Results

We identified 96 users who met the criteria to be included in the case cohort. We developed two cohorts of controls. Our initial control cohort had about five times the number of users as our case cohort. During the curation process, we noticed a large difference in the age distribution of our cases and initial control group. Therefore, we created a control group subset that more closely matched the age distribution of the cases. This age matched control group still lacks users older than 30 due to the low quantity of older users in the initial control group. Interestingly, this age distribution correction inadvertently also corrected for the increased proportion of females in the initial cohort.

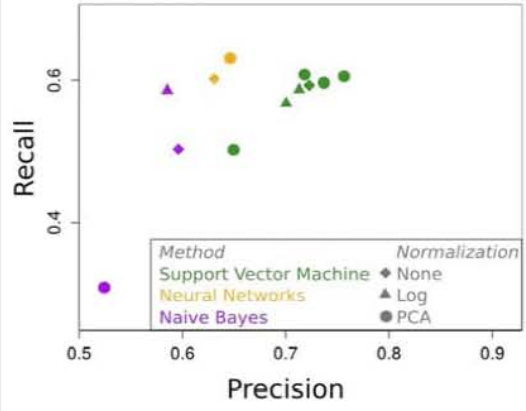

Figure 2. 5-Fold cross validation results.

SVM using the PCA transformed data performed the best in terms of F1 score and precision, and ANN using the PCA transformed features performing similarly, scoring slightly higher in recall [Figure 2]. It is worth noting that reducing the number of principal components used did not improve performance of the models. Given the number of features relative to the size of our data set, the models were able to fit the training data adequately with 28 features.

NB using the log transformed data had the third highest F1 score and provides intuition about relative feature importance. Under the assumptions of the NB model, the most important features of the model are: frequency of tweets with schizophrenia related words, time between tweets, happy emoticon usage, tweet time of day, and tweet frequency in the morning.

Table 2 presents the performance of our trained models on unseen data. Since these data are unseen during the model building, these values are lower than observed on the cross-validation data in Figure 2. The precision of SVM is high compared to both other models and related works using similar methodologies. As seen in cross validation, ANN has better recall, but worse precision. NB again performs worse, but is still reasonably high performing.

We developed an intuitive web application, available at https://hayneswa.shinyapps.io/twitterCohort/, which enables public interactions with the data and analysis results [Figure S2, Supplementary Results]. The web application generates dynamic output that is immediately responsive to user input changes. The interface renders an

assortment of plots, which illustrate properties such as vocabulary use, time of day usage patterns, summary statistics (tweet length, number of users followed, etc), and word clouds that show commonly used terms. Notably, our interface is generalizable to other Twitter cohort analyses.

Table 2. Testing data results.

Model	Precision	Recall	Accuracy	F1
SVM + PCA	*0.923*	0.706	*0.893*	*0.800*
ANN + PCA	0.813	*0.765*	0.875	0.788
NB + Log	0.688	0.647	0.803	0.667

Discussion

One of the most important features, according to the NB model, is the frequency of tweets with schizophrenia related words. We note this word set does not include any variant of "schizophrenia" and thus should not be confounded by our self-identification method of selecting schizophrenia cases. Interestingly, happy emoticon usage also appears more common in our schizophrenia cases. This is converse to what we originally expected, as individuals with schizophrenia are known to present decreased outward emotion. However, studies have shown that, though afflicted individuals lack outward emotion, they report feeling emotions at least as strongly as control individuals. Thus, social media may provide a unique outlet of emotion for individuals with schizophrenia and would be an interesting complement to current research.[27,28] Other important features are temporal: time between tweets, tweet time of day, and tweet frequency in the morning. We find that the frequency of tweets in the early morning hours is higher, indicating that Twitter users with schizophrenia tend to tweet earlier than the general Twitter population. Lastly, the 90th and 75th percentile of time between tweets is larger in our cases, indicating afflicted users generally exhibit larger gaps between tweets.

Our best performing model was a SVM with PCA transformed features. Interestingly, this model also performed best in a similar study on depression and Twitter data.[18,19] The two best performing models, based on the F1 score, both involve PCA transformed features. This makes sense as we have a feature set including highly correlated features (e.g. time percentiles), and PCA produces a representation with linearly uncorrelated features.[29]

This study has a few limitations. First of all, our case cohort contains individuals who identify as having schizophrenia on Twitter. By including users with self-identified schizophrenia, the study will be biased toward detecting diagnosed individuals. However, this is still an at-risk demographic as many people with schizophrenia do not continue medication use. Future work will include criteria to find undiagnosed individuals with a Twitter account that can be used for the undiagnosed classification task. Additionally, we eliminated variables that were explicitly used to identify individuals with schizophrenia on Twitter. There may be some bias in using only public profiles. If there is a meaningful difference between public and private users with schizophrenia, we will be unable to incorporate it into our models. As control users were identified via the Twitter API's 1% random stream, they may also be biased toward users who are currently active. While our sample size of 296 Twitter users limits the applicability of our results to a broader population, we intend this as a pilot study for the application of social media platforms to the identification of at-risk individuals. In future work, this sample size should be increased in order to draw stronger conclusions. In future studies we will explore additional machine learning and topic modeling methods. We acknowledge the possibility that users in the control group self-identify with mental disorders that they do not discuss on Twitter. In the future and with IRB approval, we could increase the accuracy of these groups by connecting Twitter accounts with external validation of psychological status, such as through surveys or medical records.

By discovering microblogging tendencies that distinguish individuals with schizophrenia from the general population, we were able to mine Twitter data to identify individuals with schizophrenia. This identification has the potential to enable outreach to untreated or undiagnosed individuals. Quick, automatic identification is a huge improvement on the current process for diagnoses where individuals are diagnosed on an individual basis during a clinic visit or led to treatment by a friend, colleague or loved one. Identifying users who have schizophrenic microblogging tendencies enables other groups to develop communities and provide outreach to affected individuals. Additionally, increasing awareness of the prevalence of schizophrenia can help destigmatize the disease. Finally, this work enables clinicians to incorporate Twitter posts into a diagnostic tool for diagnosing schizophrenia on an individual level. Together, this will lead to an increase in the number of individuals with schizophrenia receiving treatment, which will in turn improve their quality of life.

Conclusion

This novel synthesis of sentiment analysis techniques with large-scale Twitter data allowed us to identify previously undescribed schizophrenic microblogging tendencies and accurately classify Twitter users with schizophrenia and control users. Our analysis included a cohort of 96 twitter users with schizophrenia and 200 age-matched controls. We used an SVM to separate the cohorts based on their Twitter usage patterns with 92% precision. Finally, we created an interactive data visualization tool, which is generalizable to other projects, to look at the results. This work will have great impact on the identification of individuals with schizophrenia.

Funding. WAH and RLG: NSF GRFP DGE-114747. RLG and EKM: NIH NLM T15-LM007033. KM: NIH 2T32GM007276-39 and CEHG. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or NIH.

References

1. Hor K, Taylor M. Suicide and schizophrenia. *J Psychopharmacol*. 2010;24(4 Suppl):81-90.
2. NIMH. NIMH · Schizophrenia. Available at: http://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml.
3. Torrey EF. *Surviving Schizophrenia: A Manual for Families, Patients, and Providers*. HarperCollins; 2006:576.
4. Keefe R, Harvey PD. *Understanding Schizophrenia*. Simon and Schuster; 2010:283.
5. Kessler RC, Berglund PA, Bruce ML, et al. The prevalence and correlates of untreated serious mental illness. *Health Serv Res*. 2001;36(6 Pt 1):987-1007.
6. Pew Research. Social Networking Fact Sheet | Pew Research Center's Internet & American Life Project. 2014.
7. Webstralia. Visualistan: Twitter Facts And Figures 2014 [Infographic]. 2014.
8. Duggan M, Smith A. Demographics of key social networking platforms | Pew Research Center's Internet & American Life Project. 2013.
9. Pang B, Lee L. A sentimental education. In: *Proc 42nd Ann Mtg Association Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics; 2004:271-es.
10. Pang B, Lee L. Opinion Mining and Sentiment Analysis. *Found Trends® Inf Retr*. 2008;2(1–2):1-135.
11. Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proc Conf Human Lang Tech Empirical Methods Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics; 2005:347-354.
12. Oxman TE, Rosenberg SD, Tucker GJ. The language of paranoia. *Am J Psychiatry*. 1982;139(3):275-82.
13. Bollen J, Pepe A, Mao H. Modeling public mood and emotion. 2009:17-21.
14. Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis. *ICWSM*. 2011:538-541.
15. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*. 2010:1320-1326.
16. Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using Twitter hashtags and smileys. *Proc COLING*. 2010:241-249.
17. Haimson O, Ringland K, Simpson S, Wolf C. Using Depression Analytics to Reduce Stigma via Social Media: BlueFriends. *Proc iConference*. 2014:1-5.
18. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: *Proc 5th Annual ACM Web Science Conference*. New York, New York, USA: ACM Press; 2013:47-56.
19. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression via Social Media. *ICWSM*. 2013;2.
20. Shen Y, Kuo T, Yeh I, Chen T, Lin S. Exploiting Temporal Information in a Two-Stage Classification Framework for Content-Based Depression Detection. *Adv Knowl* 2013:276-288.
21. Wang X, Zhang C, Ji Y, Sun L, Wu L. A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. *Proc PAKDD Work*. 2013.
22. Bird S, Klein E, Loper E. *Natural Language Processing with Python*. "O'Reilly Media, Inc."; 2009:504.
23. Van der Walt S, Colbert SC, Varoquaux G. The NumPy Array. *Comput Sci Eng*. 2011;13(2):22-30.
24. R Development Core Team R. R: A Language and Environment for Statistical Computing. Team RDC, ed. *R Found Stat Comput*. 2011;1(2.11.1):409.
25. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. Misc functions of the Department of Statistics (e1071), TU Wien. *R Packag version 16-2*. 2014:http://cran.r-project.org/package=e1071.
26. Günther F, Fritsch S. neuralnet : Training of Neural Networks. *R J*. 2010;2:30-38.
27. Kring AM, Moran EK. Emotional response deficits in schizophrenia *Schizophr Bull*. 2008;34(5):819-34.
28. Kring AM, Caponigro JM. Emotion in Schizophrenia. *Curr Dir Psychol Sci*. 2010;19(4):255-259.
29. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933.