Linguistic Analysis of Schizophrenia in Reddit Posts

Jonathan Zomick
Psychology Department
Hofstra University
Hempstead, NY 11549
jzomick1@pride.hofstra.edu

Sarah Ita Levitan
Computer Science Department
Columbia University
New York, NY 10027
sarahita@cs.columbia.edu

Mark Serper
Psychology Department
Hofstra University
Mount Sinai School of Medicine
mark.r.serper@hofstra.edu

Abstract

We explore linguistic indicators of schizophrenia in Reddit discussion forums. Schizophrenia (SZ) is a chronic mental disorder that affects a person's thoughts and behaviors. Identifying and detecting signs of SZ is difficult given that SZ is relatively uncommon, affecting approximately 1% of the US population, and people suffering with SZ often believe that they do not have the disorder. Linguistic abnormalities are a hallmark of SZ and many of the illness's symptoms are manifested through language. In this paper we leverage the vast amount of data available from social media and use statistical and machine learning approaches to study linguistic characteristics of SZ. We collected and analyzed a large corpus of Reddit posts from users claiming to have received a formal diagnosis of SZ and identified several linguistic features that differentiated these users from a control (CTL) group. We compared these results to other findings on social media linguistic analysis and SZ. We also developed a machine learning classifier to automatically identify self-identified users with SZ on Reddit.

1 Introduction

Schizophrenia is a serious mental illness that affects roughly 1% of the US population (NIMH, 2019) and is reportedly one of the 25 top causes of disability around the world (Vos et al., 2015). Symptoms of the disorder are categorized as positive symptoms (e.g., delusions, hallucinations, disorganized thinking) or negative symptoms (e.g., diminished emotional expression, anhedonia, asociality) (APA, 2013). Individuals with SZ are at an elevated risk for suicide; an estimated 4-5% of people diagnosed with SZ die from suicide (Hor and Taylor, 2010; Carlborg et al., 2010). Early detection and diagnosis of the disorder has been speculated to improve long-term outcomes for people suffering with SZ (Birchwood et al., 1997). However, early detection and diagnosis of SZ is challenging given that it is a relatively uncommon disease and diagnostic measures are reliant on self-report measures. Additionally, many people suffering from the disorder genuinely do not believe they have SZ (Rickelman, 2004).

Linguistic abnormalities are prominent symptoms of SZ (APA, 2013). Some of the linguistic markers associated with people with the illness include diminished emotional expression, incoherence, derailment, tangentiality, co-reference failure and lexical and syntactical errors (Rochester and Martin, 1979; Harvey and Serper, 1990; Hoekert et al., 2007; Covington et al., 2005; Kuperberg, 2010). Much of the research on language and SZ has focused on analyzing transcriptions of spoken language and handwritten samples, which tend to be small, manually collected datasets.

Some recent research has focused on analyzing language from social media posts (Birnbaum et al., 2017; Lyons et al., 2018; Coppersmith et al., 2015; Mitchell et al., 2015). With the advent of social media, many people who suffer from various forms of mental illness have found a sense of community and support, and these platforms offer a mode of expression for discussing their experiences openly online. Additionally, many online platforms allow users to post anonymously, giving users a sense of security and anonymity to discuss their experiences and struggles without the fear of being stigmatized or discriminated against (Balani and De Choudhury, 2015; Berry et al., 2017; Highton-Williamson et al., 2015).

There are many advantages to leveraging social media data for analyzing the linguistic characteristics of SZ. This open discussion enables the collection and annotation of social media posts of relatively uncommon disorders such as SZ. These corpora can be collected using automated or

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 74–83
Minneapolis, Minnesota, June 6, 2019.

semi-automated methods, and enable analysis on a much larger scale. Regular social media use has risen above two billion users worldwide (Kemp, 2014), and youth comprise the largest and fastest growing demographic of social media users – over 90% of youth in the US reportedly engage in social media on a daily basis (Lenhart et al., 2015). Studying SZ among social media users can be useful for identifying early stages of the disorder, which is critical for early intervention.

Most of the research on social media posts and SZ has focused on Twitter data. In this paper we explore another popular social media platform: Reddit. Reddit is one of the fastest growing and widely used social media platforms, averaging over 330 million active monthly users, and as of 2018 was the fourth most visited website in the US (Hutchinson, 2018). Unlike Twitter, Reddit imposes no limits on the length of posts, enabling an analysis of longer language samples. In addition, Reddit is composed of subreddits, which are forums dedicated to specific topics. We leverage subreddits that are communities for individuals with SZ for identifying potential Reddit users with SZ, in order to collect a corpus of posts from these users (as described in Section 3).

These online posts provide a rich source of language data which we use to identify linguistic markers of SZ. We also use this data to train a machine learning classifier to automatically identify individuals with SZ using linguistic cues. Hopefully, an improved understanding of linguistic patterns unique to this population can assist in diagnostic procedures and be employed as an early detection mechanism.

The rest of this paper is organized as follows: Section 2 reviews relevant previous research, and 3 describes the dataset that we collected and the features that we use for analysis. In Section 4, we present the analysis of linguistic markers of SZ, and provide a detailed comparison of our findings with prior work. Section 5 presents the results of our machine learning classification of users with SZ. We discuss ethical considerations in Section 6 and conclude in Section 7.

2 Related Work

Some recent research has analyzed Twitter data of self-identified individuals with SZ with promising results. Mitchell et al. (2015) analyzed a variety of linguistic markers of SZ using tweets of self-identified individuals with SZ. Their features included lexicon-based and open-vocabulary approaches, and they discovered several significant signals for SZ. Further, they trained classifiers using these features and obtained an accuracy of 82%.

Coppersmith et al. (2015) used a similar approach to study 10 mental disorders, including SZ, and identified linguistic markers of each. They also leveraged the collected data to explore relationships between linguistic markers of multiple conditions, which is very difficult to analyze without a large-scale corpus. Birnbaum et al. (2017) also analyzed linguistic markers of SZ in Twitter data, and built a classifier to distinguish users with SZ from healthy controls. Importantly, they obtained clinician annotations of the data to validate the approach of annotating social media data based on self-disclosure of mental health conditions.

A limitation of analyzing Twitter data is that posts are constrained in character length so only very short samples of text are available for analysis. Furthermore, the character restrictions imposed by Twitter may affect users' linguistic expression and force users to communicate in ways that differ from their natural way of communicating. An alternative source of social media data are discussion board forums. Discussion board forums are not character-limited, and allow for focused conversations on topics within sub-forums. Lyons et al. (2018) analyzed several discussion board forums dedicated to mental disorders, including Reddit, and used posts from a financial discussion forum as a control. They studied linguistic features related to affective processes and personal pronoun usages, and found that these were effective at distinguishing between individuals with SZ and the control. In our work, we expand on this study by analyzing a larger set of linguistic features. We also collected a control group within the same platform to eliminate confounding factors such as stylistic and topical differences between discussion board forums.

Because all of these studies used overlapping feature sets, and in particular Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2015b) (described in section 3), we had the opportunity to analyze markers of SZ across domains. We compare the results from our study of Reddit data with previously identified markers of SZ in the four studies described in this section.

This analysis allows us to identify some linguistic characteristics of SZ that are domain-independent, and identify differences in markers of SZ across domains.

This work aims to build on the previous studies that have looked at SZ langauge on social media platforms. Specifically, to our knowledge we present the first complete analysis of LIWC features using Reddit data and compare these results with the previous findings of LIWC features of SZ on social media. Additionally, we analyze all Reddit posts of Reddit users claiming to have received a SZ diagnosis, not just those in forums devoted to discussions of SZ, and compare them to a control group of other Reddit users. We also train a machine learning classifier to automatically identify individuals with SZ, which has not been previously explored using Reddit data. This research will add to the current body of knowledge of linguistic characteristics of individuals with SZ and will hopefully help improve diagnoses and bolster early detection of the disorder.

3 Data

3.1 Reddit Corpus

We used the Python Reddit API Wrapper (PRAW) (Boe, 2016) to collect a corpus of Reddit posts from users who stated that they were diagnosed with SZ and a control group of users. We first compiled a list of users with self-disclosures of SZ by visiting subreddits devoted to discussions about SZ. These included: r/schizophrenia, r/schizophrenic, and r/AskReddit under the topic "Any Redditors With Schizophrenia?". We manually inspected the posts to only include contributors with a clear statement of receiving a formal diagnosis of SZ. For example, a user who referred to "my diagnosis of schizophrenia" would be included in the SZ group.

We also collected a random control group of Reddit users, using the r/random subreddit, which takes you to a random subreddit. To ensure a control sample that is more representative of the overall population, every five usernames that were selected came from a different random subreddit. We collected all public Reddit posts from the SZ and CTL users across all subreddits, and removed any users from the CTL group who mentioned suffering from SZ in any of their posts. We collected data from a total of 159 users for each group (318 total) who had posted at least 10 times on Reddit. Users in the SZ group made a total of 66,454 comments, and there were 113,570 comments from the CTL users.

We note that this data is not representative of the general population. For example, Reddit users have been found to be predominantly male and young (under 30) (Finlay, 2014). Our findings are limited to this population, and further research is needed to study the effects of gender and age on linguistic markers of SZ. Another limitation of using anonymous social media data for this work is that it is not externally validated; although the users in the SZ group stated that they were diagnosed with SZ, and the CTL users did not, we do not have clinical information to verify this.

3.2 LIWC Features

Having collected this dataset, we analyzed linguistic markers of SZ using Linguistic Inquiry and Word Count (Pennebaker et al., 2015b). LIWC is a text analysis program that computes word counts for semantic classes as well as structural features. LIWC relies on an internal dictionary that maps words to psychologically motivated categories. When analyzing a target text, the program looks up the target words in the dictionary and computes frequencies for each of the dimensions. The categories include standard linguistic dimensions (e.g., percentage of words that are pronouns, articles), markers of psychological processes (e.g., affect, social, cognitive words), punctuation categories (e.g., periods, commas), and formality measures (e.g., fillers, swear words). LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker et al., 1997). We extracted a total of 93 features using LIWC 2015. A full description of these features is found in (Pennebaker et al., 2015a).

We selected LIWC to analyze linguistic markers of SZ because these features have been widely studied for this purpose in other domains (such as Twitter), which enables a direct comparison of results across domains.

Category	**Reddit**	**Discussion Forums**	**Twitter**		
Paper	Current	Lyons et. al	(A)	(B)	(C)
Linguistic Processes					
Word count	SZ				
Dictionary words	SZ				
Total function words	SZ		SZ	SZ	
Total pronouns	SZ		SZ		
Personal pronouns	SZ	SZ	SZ		
1st person singular	SZ	SZ	SZ		SZ
1st person plural	CTL	CTL			SZ
2nd person	SZ	SZ			
3rd person singular	CTL	SZ			SZ
3rd person plural		SZ	SZ	SZ	SZ
Impersonal pronouns			SZ		SZ
Articles	CTL		SZ	SZ	SZ
Auxiliary verbs	SZ		SZ	SZ	SZ
Common adverbs	SZ				
Conjunctions	SZ		SZ	SZ	
Negations	CTL				SZ
Other Grammar					
Common verbs	SZ				
Numbers	CTL				
Quantifiers				SZ	SZ
Psychological processes					
Affective processes	SZ	SZ			
Positive emotion	SZ	CTL	CTL		SZ
Negative emotion		SZ	SZ		SZ
Anxiety	SZ	SZ		SZ	
Anger	CTL	SZ			
Sadness		SZ			SZ
Social processes	SZ				
Friends					CTL
Male references	CTL				
Cognitive processes	SZ		SZ	SZ	SZ
Insight	SZ		SZ	SZ	SZ
Causation				SZ	SZ
Discrepancy			SZ		SZ
Tentatitve	SZ		SZ	SZ	SZ
Certainty					SZ
Perceptual processes	SZ				SZ
See	CTL		CTL		
Hear	SZ				SZ
Feel	SZ				SZ
Biological Processes	SZ				SZ
Body					SZ
Health	SZ		SZ	SZ	SZ
Sexual					SZ
Drives	SZ				
Achievement					SZ
Power	CTL				

Reward	SZ				
Time orientations	SZ				SZ
Past focus	SZ				SZ
Present focus	SZ				SZ
Future focus			CTL		
Relativity	CTL		CTL	CTL	
Motion	CTL			CTL	
Space			CTL	SZ	
Personal concerns					
Work					SZ
Leisure	CTL		CTL	CTL	
Home			CTL	CTL	SZ
Money	CTL				
Death			SZ	SZ	
Informal language					
Swear words	CTL				SZ
Assent			CTL	CTL	
Punctuation					
Question marks	CTL				
Exclamation marks	SZ		SZ		
Dashes	CTL				
Other punctuation	CTL				

Table 1: LIWC features that were significantly different between SZ and CTL groups, compared across five studies. "Current" indicates the analysis of Reddit posts conducted in this paper, Lyons et al. (2018) studied some LIWC variables in discussion board posts (including Reddit). The three studies that examined Twitter data are: (A) Mitchell et al. (2015); (B): Coppersmith et al. (2015) ; and (C): Birnbaum et al. (2017). Gray cells indicate categories that were not examined in a study (some are due to differences between LIWC 2015 and 2007 versions).

4 Linguistic Characteristics of SZ and CTL Reddit Comments

To identify linguistic markers of SZ, we compared the frequencies of each LIWC dimension in SZ and CTL users. We averaged the frequencies of the LIWC dimensions across all posts per user so that each user was represented once in the dataset. This was done to avoid skewing the data based on a few users who posted a large number of comments. We used an independent samples t-test to determine whether the difference in mean frequency for each LIWC feature between the SZ and CTL groups was statistically significant. All tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$ (Benjamini and Hochberg, 1995). The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$. Table 1 shows the results of this analysis in the "Reddit" column. "SZ" indicates that the feature was significantly more frequent in posts from users with SZ, and "CTL" indicates that the feature was significantly more frequent in posts from the control group of users.

We found significant differences between the SZ group and the CTL group for many of the LIWC features. These differences spanned various linguistic domains including linguistic processes, grammar, psychological processes, and punctuation. In addition to showing the results of our analysis of Reddit posts, Table 1 shows a comparison of our results with four other studies that examined LIWC features and SZ in social media data: one study (Lyons et al., 2018) used data from Reddit and other online discussion forums (but only examined personal pronouns and affective processes), and 3 studies examined Twitter data: (A) Mitchell et al. (2015), (B) Coppersmith et al. (2015), and (C) Birnbaum et al. (2017).

Many of our findings were in line with previous research on other social media platforms, while some of the markers that we identified differed from previous studies. We identified several markers of SZ in our Reddit corpus that have not been previously noted. These include an increased association between users with SZ and the following features: Word count, Dictionary words, Common adverbs, Verbs, Reward, and Drives. Additionally, unlike previous social media studies, we found diminished expression among the following features: 3rd person singular, Articles, Negations, Anger, Male references, Power, Money, Swear words, Question marks, Dashes, and Other punctuation. It is not surprising that there are discrepancies between this study and others. This type of analysis has not been previously conducted on data taken exclusively from Reddit, and the majority of these features were not analyzed in the discussion forum data by (Lyons et al., 2018). There is a substantial domain mismatch between Reddit and Twitter data, and markers of SZ that have been observed in Twitter data may not generalize to other domains, while other markers that we have observed in the Reddit may not have been observed in previous work with Twitter data due to the character constraints that platform places on users' posts.

On the other hand, some of the findings regarding association between specific LIWC features and SZ are more robust and have been replicated in multiple studies. When comparing results from the five studies that looked at SZ language and social media, at least 3 out of the 5 studies reported increased frequency among users with SZ in the following features: Total function words, Personal Pronouns, 1st person pronouns, 3rd person plural, Articles, Auxiliary verbs, Conjunctions, Negative emotion, Anxiety, Cognitive processes, Insight, Tentative, and Health. Other findings that have been replicated multiple times relate to diminished expression of certain LIWC features among users with SZ in comparison with control users. Three of the five studies found that users with SZ used words associated with the features Relativity and Leisure significantly less than control groups.

4.1 Discussion

The present results are consistent with past studies that have found that users with SZ use words associated with health issues, anxiety, negative emotion and use of 1st person singular pronouns more than control groups. An emphasis on health related matters, expressions of negative emotions, and a focus on one's self are understandable for people suffering from a serious mental illness. It is also somewhat understandable that users with SZ use leisure related words significantly less than controls, since individuals suffering from mental illness appear to be less focused or interested in leisure activities (Thornicroft et al., 2004). However, some of the linguistic features that have been found elevated among users with SZ in multiple studies are not as intuitive, such as usage of 3rd

person plural pronouns, Insight words, and Tentative words.

The robust findings of usage of 3rd person plural pronouns may be related to SZ symptomatology. For example, relative excessive use of pronouns such as "they" and "them" may reflect a disaffiliativeness from others that is reflected in symptoms of social anhedonia. Further support for this line of reasoning comes from our finding and findings by Lyons et al. (2018) that members of the SZ group used 1st person plural pronouns such as "we" and "us" less than the CTL group, which may also be an indication of social disaffiliation and withdrawal.

Additionally, the use of 3rd person plural pronouns may reflect positive symptoms common to the disorder (Bentall et al., 2001; APA, 2013). Previous researchers have posited that the increased usage of 3rd person plural pronouns among SZ patients may be a reflection of an externalizing bias, paranoid thinking, and a focus on outside groups (Fineberg et al., 2015; Lyons et al., 2018). The decreased usage of 1st person plural pronouns may also reflect social withdrawal due to paranoid suspicions that result in social anxiety and subsequent isolation.

All of the studies reported here that looked at tentative language in social media data and SZ found that users with SZ used tentative words like "perhaps" and "maybe" significantly more than CTL users. Tausczik and Pennebaker (2010) suggest that tentative language is suggestive of difficulty processing events and forming events into stories and may indicate uncertainty or insecurity about a topic. Use of tentative language may be a manifestation of an impaired sense of agency and diminished self-presence reportedly associated with people with SZ (Jeannerod, 2009; Sass and Parnas, 2003). The increased usage of 1st person pronouns may also be a marker of a hyperreflexivity (exaggerated self-consciousness) experienced by individuals with SZ, as described by Sass and Parnas (2003).

In contrast to earlier social media data we found that the SZ group used punctuation significantly less frequently than the CTL group. The discrepancy between this work and previous work using Twitter data may be due to differences between these two platforms. The character restrictions Twitter places on posts may discourage usage of proper punctuation to preserve space for content words. However, Reddit posts that do not have these restrictions may reflect more natural language of users and allow for additional observations such as differences in punctuation usage. In line with the hypothesis put forth by Fineberg et al. (2015) our finding that users with SZ use punctuation significantly less than CTL users may reflect more disorganized use of language, a prominent symptom of schizophrenia (Covington et al., 2005; APA, 2013).

5 Automatic Identification of Users with Schizophrenia

Having identified many differences in language usage between Reddit users with SZ and the control group, we trained a machine learning classifier to automatically distinguish between the groups, using the LIWC features. We used the scikit-learn (Pedregosa et al., 2011) implementation of a Logistic Regression model using the default parameters. The model was trained and evaluated using stratified 5-fold cross-validation. We averaged the LIWC features across all comments per user and trained the model to determine whether the aggregated LIWC features were from the posts of a user from the SZ group or the CTL group. The random baseline is 50%, since the data is balanced across groups.

The average performance of the classifier across 5 folds was 81.56% accuracy, and the standard deviation was 2.29. The top 10 LIWC dimensions for the SZ and CTL classes, obtained from the logistic regression coefficients, are shown in Table 2. Some of these weighted features were consistent with our statistical analysis of LIWC features. For example, the Health category was highly predictive of SZ, as was the Tentative dimension. Intuitively, Sadness was the strongest (negative) predictor of the control group, and 3rd person singular was also a useful (negative) predictor of the control group.

These findings suggest that linguistic features are useful for automatically identifying social media users with self-described SZ on a large, public, anonymous social media site. The classifier achieved strong performance, 31.56% better than a random baseline. However, although a balanced data set is useful for analyzing linguistic indicators of SZ and for evaluating the machine learning classification results, we note (as do Mitchell et al. (2015)) that this setup is not representative of

Control (CTL)		Schizophrenia (SZ)	
Weight	Feature	Weight	Feature
-1.2748	Sadness	1.6105	Health
-1.1109	Quotation mark	1.0717	Interrogatives
-0.8715	3rd person singular	1.0614	Tentative
-0.7956	Feel	0.9825	Hear
-0.7949	Articles	0.9426	Colon
-0.7302	Nonfluencies	0.9304	Death
-0.6705	Adjectives	0.8021	Biological processes
-0.6329	See	0.7642	1st person singular
-0.6214	Motion	0.6975	Parentheses
-0.6182	Present focus	0.6478	Verbs

Table 2: Top weighted features from the logistic regression classifier for the SZ and CTL groups.

the true distribution of SZ and healthy individuals (only 1% have SZ).

6 Ethical Considerations

Detecting mental health conditions using linguistic features extracted from social media has the potential to enhance detection of disorders for early intervention and improve outcomes for individuals suffering from mental illness. However, there are several important ethical concerns with this line of research, and necessary precautions must be taken. First, is the issue of informed consent. Although social media posts are publicly available, users are typically unaware of the research being conducted and do not explicitly provide consent for their data to be mined for sensitive information. Additionally, individuals with mental illness, and especially young individuals, are a sensitive, at risk population and extra caution must be taken when collecting and analyzing their data to ensure they remain anonymous and unidentifiable.

Submitting to IRB review and obtaining IRB approval or exemption for any study with this population is critical. Extreme caution must be taken to protect this sensitive data, and collected corpora should not be shared without IRB approval. Further, if data is shared with specific parties, the data should be anonymized so that identifying information is not disclosed. As data mining for mental health research becomes more popular and prevalent, it is important to be aware of these ethical considerations and to take the necessary precautions to protect the studied population. For further guidance in this area, Benton et al. (2017) have compiled an excellent review of ethical considerations for social media health research.

7 Conclusion

We collected a corpus of Reddit users claiming to have received a diagnosis of SZ and used natural language processing and statistical techniques to analyze and compare language from their posts and those of a control group comprised of random Reddit users. We identified several linguistic markers of SZ, and compared these findings with previous research on linguistic markers of SZ using data from other social media platforms. This work is useful for identifying markers of SZ that are robust across domains. Finally, we trained a machine learning classifier that identified self-described SZ sufferers on Reddit with over 80% accuracy, using linguistic features. These findings contribute toward the ultimate goal of identifying high risk individuals and providing early intervention to improve overall treatment outcomes.

References

APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378. ACM.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Richard P Bentall, Rhiannon Corcoran, Robert Howard, Nigel Blackwood, and Peter Kinderman.

2001. Persecutory delusions: a review and theoretical integration. *Clinical psychology review*, 21(8):1143–1192.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. # whywetweetmh: understanding why people use twitter to discuss mental health problems. *Journal of medical Internet research*, 19(4).

Max Birchwood, Patrick McGorry, and Henry Jackson. 1997. Early intervention in schizophrenia. *The British Journal of Psychiatry*, 170(1):2–5.

Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research*, 19(8).

Bryce Boe. 2016. Python Reddit API Wrapper (PRAW). https://praw.readthedocs.io/en/v6.1.1/. Accessed: 2019-03-10.

Andreas Carlborg, Kajsa Winnerbäck, Erik G Jönsson, Jussi Jokinen, and Peter Nordström. 2010. Suicide in schizophrenia. *Expert review of neurotherapeutics*, 10(7):1153–1164.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Michael A Covington, Congzhou He, Cati Brown, Lorina Naçi, Jonathan T McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. 2005. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research*, 77(1):85–98.

SK Fineberg, S Deutsch-Link, M Ichinose, T McGuinness, AJ Bessette, CK Chung, and PR Corlett. 2015. Word use in first-person accounts of schizophrenia. *The British Journal of Psychiatry*, 206(1):32–38.

S Craig Finlay. 2014. Age and gender in Reddit commenting and success. *Journal of Information Science Theory and Practice*, pages 18–28.

Philip D Harvey and Mark R Serper. 1990. Linguistic and cognitive failures in schizophrenia: A multivariate analysis. *Journal of Nervous and Mental Disease*.

Elizabeth Highton-Williamson, Stefan Priebe, and Domenico Giacco. 2015. Online social networking in people with psychosis: a systematic review. *International Journal of Social Psychiatry*, 61(1):92–101.

Marjolijn Hoekert, René S Kahn, Marieke Pijnenborg, and André Aleman. 2007. Impaired recognition and expression of emotional prosody in schizophrenia: review and meta-analysis. *Schizophrenia research*, 96(1-3):135–145.

Kahyee Hor and Mark Taylor. 2010. Suicide and schizophrenia: a systematic review of rates and risk factors. *Journal of psychopharmacology*, 24(4_suppl):81–90.

Andrew Hutchinson. 2018. Reddit now has as many users as twitter, and far higher engagement rates. https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/. Accessed: 2019-03-10.

Marc Jeannerod. 2009. The sense of agency and its disturbances in schizophrenia: a reappraisal. *Experimental Brain Research*, 192(3):527.

Simon Kemp. 2014. Social, digital & mobile in 2014. *We Are Social Singapore*, 28.

Gina R Kuperberg. 2010. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Amanda Lenhart, Maeve Duggan, Andrew Perrin, Renee Stepler, Harrison Rainie, Kim Parker, et al. 2015. *Teens, social media & technology overview 2015*. Pew Research Center [Internet & American Life Project].

Minna Lyons, Nazli Deniz Aksayli, and Gayle Brewer. 2018. Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87:207–211.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

NIMH. 2019. Schizophrenia. https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml. Accessed: 2019-03-10.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015a. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. 1997. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, 72(4):863.

JW Pennebaker, CK Chung, M Ireland, A Gonzales, and RJ Booth. 2015b. Liwc. *Austin, Texas; 2007. LIWC2007: Linguistic inquiry and word count [software program for text analysis] URL: http://liwc. wpengine. com/[accessed 2017-02-27].*

Bonnie L Rickelman. 2004. Anosognosia in individuals with schizophrenia: toward recovery of insight. *Issues in Mental Health Nursing*, 25(3):227–242.

S Rochester and JR Martin. 1979. Jr, 1979 crazy talk: A study of the discourse of schizophrenic speakers.

Louis A Sass and Josef Parnas. 2003. Schizophrenia, consciousness, and the self. *Schizophrenia bulletin*, 29(3):427–444.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Graham Thornicroft, Michele Tansella, Thomas Becker, Martin Knapp, Morven Leese, Aart Schene, José Luis Vazquez-Barquero, EPSILON Study Group, et al. 2004. The personal impact of schizophrenia in europe. *Schizophrenia research*, 69(2-3):125–132.

Theo Vos, Ryan M Barber, Brad Bell, Amelia Bertozzi-Villa, Stan Biryukov, Ian Bolliger, Fiona Charlson, Adrian Davis, Louisa Degenhardt, Daniel Dicker, et al. 2015. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(9995):743–800.