Quantifying the Language of Schizophrenia in Social Media

Margaret Mitchell
Microsoft
memitc@microsoft.com

Kristy Hollingshead
IHMC
kseitz@ihmc.us

Glen Coppersmith
Qntfy
glen@qntfy.io

Abstract

Analyzing symptoms of schizophrenia has traditionally been challenging given the low prevalence of the condition, affecting around 1% of the U.S. population. We explore potential linguistic markers of schizophrenia using the tweets[1] of self-identified schizophrenia sufferers, and describe several natural language processing (NLP) methods to analyze the language of schizophrenia. We examine how these signals compare with the widely-used LIWC categories for understanding mental health (Pennebaker et al., 2007), and provide preliminary evidence of additional linguistic signals that may aid in identifying and getting help to people suffering from schizophrenia.

1 Introduction

Schizophrenia is a group of mental disorders that affect thinking and emotional responsiveness, documented throughout history (e.g., The Book of Hearts, 1550 BCE). Today it is diagnosed and monitored leveraging self-reported experiences.[2] This may be challenging to elicit from schizophrenia sufferers, as a hallmark of the disease is the sufferer's belief that he or she does not have it (Rickelman, 2004; National Alliance on Mental Illness, 2015). Schizophrenia sufferers are therefore particularly at-risk for not leveraging help (Pacific Institute of Medical Research, 2015). This suggests that techniques that leverage social language shared by schizophrenia sufferers could be greatly beneficial in treatment of the disease. Early identification and monitoring of schizophrenia can increase the chances of successful management of the condition, reducing the chance of psychotic episodes (Häfner and Maurer, 2006) and helping a schizophrenia sufferer lead a more comfortable life.

We focus on unsupervised groupings of the words used by people on the social media platform Twitter, and see how well they discriminate between matched schizophrenia sufferers and controls. We find several potential linguistic indicators of schizophrenia, including words that mark an irrealis mood ("think", "believe"), and a lack of emoticons (a potential signature of flat affect). We also demonstrate that a support vector machine (SVM) learning approach to distinguish schizophrenia sufferers from matched controls works reasonably well, reaching 82.3% classification accuracy.

To our knowledge, no previous work has sought out linguistic markers of schizophrenia that can be automatically identified. Schizophrenia is a relatively rare mental health condition, estimated to affect around 1% of the population in the U.S. (The National Institute of Mental Health, 2015; Perälä et al., 2007; Saha et al., 2005), or some 3.2 million people. Other mental health conditions with a high prevalence rate such as depression[3] have recently received increased attention (Schwartz et al., 2014; De Choudhury et al., 2013b; Resnik et al., 2013; Coppersmith et al., 2014a). However, similar studies for schizophrenia have been hard to pursue, given

[1] A posting made on the social media website Twitter, https://twitter.com/

[2] With the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013).

[3] 16.9% lifetime prevalence rate (Kessler et al., 2005)

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 11–20,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

the rarity of the condition and thus the inherent difficulty in collecting data.

We follow the method from Coppersmith et al. (2014a) to create a relatively large corpus of users diagnosed with schizophrenia from publicly available Twitter data, and match them to Twitter controls. This provides a view of the social language that a schizophrenia sufferer may choose to share with a clinician or counselor, and may be used to shed light on the illness and the effect of treatments.

2 Background and Motivation

There has been a recent growth in work using language to automatically identify people who may have mental illness and quantifying its progression, including work to help people suffering from depression (Howes et al., 2014; Hohman et al., 2014; Park et al., In press; Schwartz et al., 2014; Schwartz et al., 2013; De Choudhury et al., 2013a; De Choudhury et al., 2013b; De Choudhury et al., 2011; Nguyen et al., 2014) and post-traumatic stress disorder (Coppersmith et al., 2014b). Related work has also shown it is possible to aid clinicians in identifying patients who suffer from Alzheimer's (Roark et al., 2011; Orimaye et al., 2014) and autism (Rouhizadeh et al., 2014). The time is ripe to begin exploring an illness that deeply affects an estimated 51 million people.

The term *schizophrenia*, derived from the Greek words for "split mind", was introduced in the early 1900s to categorize patients whose thoughts and emotional responses seemed disconnected. Schizophrenia is often described in terms of symptoms from three broad categories: positive, negative, and cognitive. Positive symptoms include disordered thinking, disordered moving, delusions, and hallucinations. Negative symptoms include a flat affect and lack of ability to begin and sustain planned activities. Cognitive symptoms include poor ability to understand information and make decisions, as well as trouble focusing.

Some symptoms of schizophrenia may be straightforward to detect in social media. For example, the positive symptoms of *neologisms*, or creating new words, and *word salad*, where words and sentences are strung together without a clear syntactic or semantic structure, may be expressed in the text written by some schizophrenia sufferers. Negative symptoms may also be possible to find, for example, a lack of emoticons can reflect a flat affect, or a lower proportion of commonly used terms may reflect cognitive difficulties.

As we discuss below, natural language processing (NLP) techniques can be used to produce features similar to these markers of schizophrenia. For example, *perplexity* may be useful in measuring how unexpected a user's language is, while *latent Dirichlet allocation* (Blei et al., 2003) may be useful in characterizing the difference in general themes that schizophrenia sufferers discuss vs. control users. All NLP features we describe are either automatically constructed or *unsupervised*, meaning that no manual annotation is required to create them. It is important to note that although these features are inspired by the literature on schizophrenia, they are not direct correlates of standard schizophrenia markers.

3 Data

We follow the data acquisition and curation process of Coppersmith et al. (2014a), summarizing the major points here: Social media, such as Twitter, contains frequent public statements by users reporting diagnoses for various medical conditions. Many talk about physical health conditions (e.g., cancer, flu) but some also discuss mental illness, including schizophrenia. There are a variety of motivations for users to share this information on social media: to offer or seek support, to fight the stigma of mental illness, or perhaps to offer an explanation for certain behaviors.[4]

We obtain messages with these self-reported diagnoses using the Twitter API, and filtered via (case-insensitive) regular expression to require "schizo" or a close phonetic approximation to be present; our expression matched "schizophrenia", its subtypes, and various approximations: "schizo", "skitzo", "skitso", "schizotypal", "schizoid", etc. All data we collect are public posts made between 2008 and 2015, and exclude any message marked as 'private' by the author. All use of the data reported in this

[4] Anecdotally, many of the users in this study tend to be talking about a recent diagnosis (looking for information or support) or fighting the stigma of mental illness (by sharing their struggles).

paper has been approved by the appropriate Institutional Review Board (IRB).

Each self-stated diagnosis included in this study was examined by a human annotator (one of the authors) to verify that it appeared to be a genuine statement of a schizophrenia diagnosis, excluding jokes, quotes, or disingenuous statements. We obtained 174 users with an apparently genuine self-stated diagnosis of a schizophrenia-related condition. Note that we cannot be certain that the Twitter user was actually diagnosed with schizophrenia, only that their statement of being diagnosed appears to be genuine. Previous work indicates that inter-annotator agreement for this task is good: $\kappa = 0.77$ (Coppersmith et al., 2014a).

For each user, we obtained a set of their public Twitter posts via the Twitter API, collecting up to 3200 tweets.[5] As we wish to focus on user-authored content, we exclude from analysis all retweets and any tweets that contain a URL (which often contain text that the user did not author). We lowercase all words and convert any non-standard characters (including emoji) to a systematic ASCII representation via Unidecode.[6]

For our community controls, we used randomly-selected Twitter users who primarily tweet in English. Specifically, during a two week period in early 2014, each Twitter user who was included in Twitter's 1% "spritzer" sample had an equal chance for inclusion in our pool of community controls. We then collected some of their historic tweets and assessed the language(s) they tweeted in according to the Chromium Compact Language Detector.[7] Users were excluded from our community controls if their tweets were less than 75% English.[8]

3.1 Age- and Gender-Matched Controls

Since mental health conditions, including schizophrenia, have different prevalence rates depending on age and gender (among other demographic variables), controlling for these will be important when examining systematic differences between schizophrenic users and community controls. In particular, we would like to be able to attribute any quantifiable signals we observe to the presence or absence of schizophrenia, rather than to a confounding age or gender divergence between the populations (Dos Reis and Culotta, 2015). To that end, we estimated the age and gender of all our users (from their language usage) via the tools graciously made available by the World Well-Being Project (Sap et al., 2014). For each user, we applied a hard threshold to the gender prediction to obtain a binary 'Female' or 'Male' label. Then, in order to select the best match for each schizophrenia user, we selected the community control that had the same gender label and was closest in age (without replacement).

3.2 Drawbacks of a Balanced Dataset

We use a balanced dataset here for our analysis (an equal number of schizophrenia users and community controls). This 50/50 split makes the machine learning and analysis easier, and will allow us to focus more on emergent linguistics that are related to schizophrenia than if we had examined a dataset more representative of the population (more like 1/99). Moreover, we have not factored in the cost of false negatives or false positives (how should the consequences of misclassifying a schizophrenia user as non-schizophrenic be weighed against the consequences of misclassifying a non-schizophrenic user as schizophrenic?). All our classification results should be taken as validation that the differences in language we observe are relevant to schizophrenia, but only one step towards applying something derived from this technology in a real world scenario.

3.3 Concomitance

Often, people suffering from mental illness have a diagnosis for more than one disorder, and schizophrenia is no exception. Of our 174 users with a genuine self-statement of diagnosis of a schizophrenia-related condition, 41 also state a diagnosis of at least one other mental illness (30%), while 15 of those state that they have a diagnosis of more than one other mental illness (11%). The vast majority of these concomitances are with bipolar (25 users), followed by depression (14), post traumatic stress disorder (8) and generalized anxi-

[5] This is the maximum number of historic tweets permitted by the API.

[6] `https://pypi.python.org/pypi/Unidecode`

[7] `https://code.google.com/p/cld2/`

[8] A similar exclusion was applied to the schizophrenia users, but in practice none fell below the 75% threshold.

ety disorder (6). These comorbidity rates are notably lower than the generally accepted prevalence rates, which may be due to one of several factors. First, we rely on stated diagnoses to calculate comorbidity, and the users may not be stating each of their diagnosed conditions, either because they have not been diagnosed as such, or they choose to identify most strongly with the stated diagnosed conditions, or they simply ran out of space (given Twitter's 140-character limit). Second, we are analyzing Twitter users, which consists of only a subset of the population, and the users that choose to state, publicly, on Twitter, their schizophrenia diagnosis, may not be an accurate representation of the population of schizophrenia sufferers. The noted concomitance of schizophrenia and bipolar disorder is frequently labeled as "schizoaffective disorder with a bipolar subtype", with some recent research indicating shared impairments in functional connectivity across patients with schizophrenia and bipolar disorders (Meda et al., 2012). It is worth keeping in mind throughout this paper that we examine all subtypes of schizophrenia together here, and further in-depth analysis between subtypes is warranted.

4 Methods

We first define features relevant to mental health in general and schizophrenia in particular, and explore how well each feature distinguishes between schizophrenia-positive users and community controls. We then design and describe classifiers capable of separating the two groups based on the values for these features in their tweets. We reflect on and analyze the signals extracted by these automatic NLP methods and find some interesting patterns relevant to schizophrenia.

4.1 Lexicon-based Approaches

We used the Linguistic Inquiry Word Count (LIWC, Pennebaker et al. (2007)) to analyze the systematic language differences between our schizophrenia-positive users and their matched community controls. LIWC is a psychometrically validated lexicon mapping words to psychological concepts, and has been used extensively to examine language (and even social media language) to understand mental health. LIWC provides lists of words for categories such as FUTURE, ANGER, ARTICLES, etc. We treat each category as a feature; the feature values for a user are then the proportion of words in each category (e.g., the number of times a user writes "I" or "me", divided by the total number of words they have written is encoded as the LIWC "first person pronoun" category).

4.2 Open-vocabulary Approaches

In addition to the manually defined lexicon-based features described above, we also investigate some open-vocabulary approaches. This includes latent Dirichlet allocation (LDA) (Blei et al., 2003), Brown clustering (Brown et al., 1992), character n-gram language modeling (McNamee and Mayfield, 2004), and perplexity.[9] We now turn to a brief discussion of each approach.

Latent Dirichlet Allocation LDA operates on data represented as "documents" to infer "topics". The idea behind LDA is that each document can be viewed as a mixture of topics, where each topic uses words with different probabilities (e.g., "health" would be likely to come from a *psychology* topic or an *oncology* topic, but "schizophrenia" is more common from the former). LDA infers these topics automatically from the text – they do not have labels to start with, but often a human reading the most frequent words in the topic can see the semantic relationship and assign one.

In our case, all tweets from a user make up a "document", and we use collapsed Gibbs sampling to learn the distribution over topics for each document. In other words, given a specific number of topics k (in our work, k=20), LDA estimates the probability of each word given a topic and the probability of each topic given a document. Tweets from a user can then be featurized as a distribution over the topics: Each topic is a feature, whose feature value is the probability of that topic in the user's tweets.

The LDA implementation we use is available in the MALLET package (McCallum, 2002).

Brown Clustering Words in context often provide more meaning than the words in isolation, so we use methods for grouping together words that occur in similar linguistic constructions. Brown clustering is

[9] http://en.wikipedia.org/wiki/Perplexity

a greedy hierarchical algorithm that finds a clustering of words that maximizes the mutual information between adjacent clusters; in other words, words that are preceded by similar words are grouped together to form clusters, and then these clusters are merged based on having similar preceding words, and then these clusters are further merged, etc. Each word is therefore associated to clusters of increasing granularity. We define all leaf clusters[10] as features, and the feature value of each for a user is the proportion of words from the user in that cluster. The Brown clustering implementation we use is currently available on github,[11] and is used with default parameter settings, including a limit of 100 clusters.

Character n-grams Character n-gram language models are models built on sequences (n-grams) of characters. Here, we use 5-grams: for all the tweets a user authored, we count the number of times each sequence of 5 characters is observed. For example, for this sentence we would observe the sequences: "for e", "or ex", "r exa", " exam", and so on. The general approach is to examine how likely a sequence of characters is to be generated by a given type of user (schizophrenic or non-schizophrenic).

To featurize character n-grams, for each character 5-gram in the training data, we calculate its probability in schizophrenic users and its probability in control users. At test time, we search for sets of 50 sequential tweets that look "most schizophrenic" by comparing the schizophrenic and control probabilities estimated from the training data for all the 5-grams in those tweets. We experimented with different window sizes for the number of tweets and different n for n-grams; for brevity, we report only the highest performing parameter settings at low false alarm rates: 5-grams and a window size of 50 tweets. An example of this can be found in Figure 1, where one schizophrenic and one control user's score over time is plotted (top). To show the overall trend, we plot the same for all users in this study (bottom), where separation between the schizophrenics (in red) and control users (in blue) is apparent. The highest score from this windowed analysis becomes the feature value.

Note that this feature corresponds to only a subset of a user's timeline. For schizophrenia sufferers, this is perhaps when their symptoms were most severe, a subtle but critical distinction when one considers that many of these people are receiving treatment of some sort, and thus may have their symptoms change or subside over the course of our data.

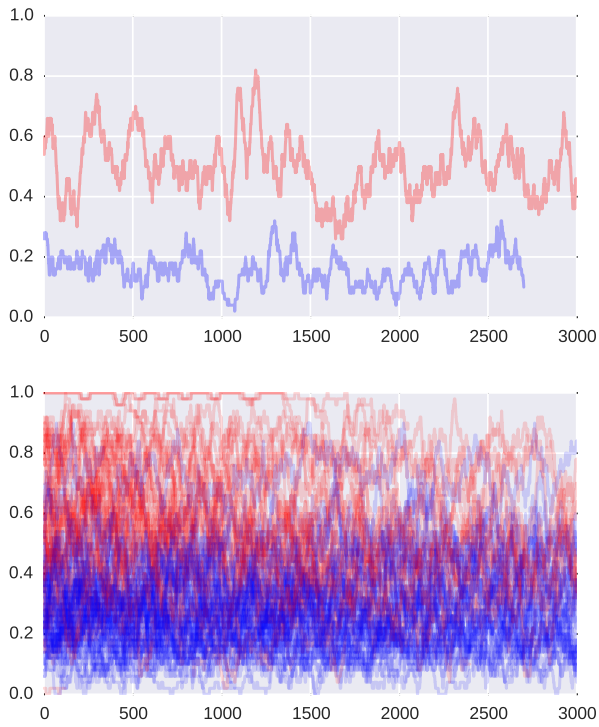

Figure 1: Timeseries of schizophrenia-like tweets for each user, x-axis is the tweets in order, y-axis denotes the proportion of tweets in a window of 50 tweets that are classified as *schizophrenia-like* by the CLMs. Top: Example plots of one schizophrenia (red) and one control user (blue). Bottom: All users.

Perplexity The breadth of language used (to include vocabulary, topic areas, and syntactic construction) can be measured via perplexity – a measurement based on entropy, and roughly interpreted as a measurement of how predictable the language is. We train a trigram language model on one million randomly selected tweets from the 2014 1% feed, and then use this model to score the perplexity on all the tweets for each user. If a user's language wanders broadly (and potentially has the *word salad* effect sometimes a symptom of schizophrenia), we would expect a high perplexity score for the user. This gives us a single feature value for the perplexity feature for each user.

[10] I.e., the most granular clusters for each word.

[11] https://github.com/percyliang/brown-cluster

Cond.	Topic	Top Words
Sch	2	don('t) (I')ve (I')ll feel people doesn('t) thing didn('t) time twitter won('t) make kind woman things isn('t) bad cat makes
Sch	9	don('t) love fuck fucking shit people life hell hate stop gonna god wanna die feel make kill time anymore
Sch	12	people don('t) le world mental schizophrenia (I')ve god jesus schizophrenic illness health care paranoid medical truth time life read
Sch	18	people work today good years time make call long find made point thought free twitter back thing days job
Con	6	lol shit nigga im tho fuck ass ain('t) lmao don('t) good niggas gotta bitch smh damn ya man back
Con	7	game rochester football girls basketball final boys billsmafia win rt valley team season sectional north play miami st soccer
Con	11	great love time hope today day rt support custserv big happy awesome amazing easy trip toronto forward orleans hear
Con	19	lol dd love don('t) today day good happy time ddd miss hate work night back (I')ll birthday tomorrow tonight

Table 1: LDA topics with statistically significant differences between groups. The condition with the highest mean proportion is given in column 1, where Sch=schizophrenia and Con=control.

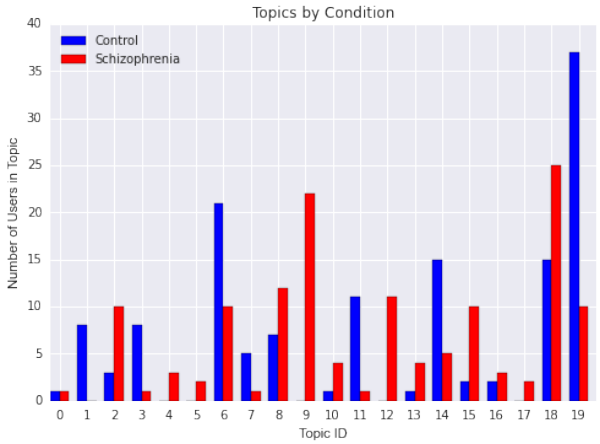

Figure 2: LDA topic prevalence by condition, shown by the number of users with each identified topic as their maximum estimated probability topic (t).

5 Results

5.1 Isolated Features

We examine differences in the language between schizophrenia sufferers and matched controls by mapping the words they use to broader categories, as discussed above, and measuring the relative frequencies of these categories in their tweets. Different approaches produce different word categories: We focus on LIWC vectors, topics from latent Dirichlet allocation (LDA), and clusters from Brown clustering. We compare whether the difference in the relative frequencies of each category is significant using an independent sample t-test,[12] Bonferroni-corrected.

[12]We assume a normal distribution; future work may examine how well this assumption holds.

Cond.	Topic	Top Words
Sch	01011111111	but because cause since maybe bc until cuz hopefully plus especially except
Sch	01011111110	if when sometimes unless whenever everytime someday
Sch	010000	i
Sch	010100111	know think thought care believe guess remember understand forget swear knew matter wonder forgot realize worry imagine exist doubt kno realized decide complain
Sch	010111010	of
Con	0001001	lol haha omg lmao idk hahaha wtf smh ugh o bruh lmfao ha #askemma tbh exactly k bye omfg hahahaha fr hahah btw jk
Con	01011011010	today
Con	0010111	! << >>
Con	01011010100	back home away checked forward asleep stuck button stream rewards closer messages anywhere apart swimming inspired dong tricks spree cv delivered tuned increased
Con	00001	" rt #nowplaying

Table 2: Example Brown clusters with statistically significant differences between groups. The condition with the highest mean proportion is given in column 1, where Sch=schizophrenia and Con=control.

LIWC vectors We did not make predictions about which LIWC categories might show deviations between our schizophrenia and control users, but instead examine all the LIWC categories (72 categories, corrected $\alpha = 0.0007$). We find that the language of schizophrenia users had significantly more words from the following major categories: COGNITIVE MECHANISMS, DEATH, FUNCTION WORDS, NEGATIVE EMOTION, and in the following subcategories: ARTICLE, AUXILIARY VERBS, CONJUGATIONS, DISCREPANCIES, EXCL, HEALTH, I, INCL, INSIGHT, IPRON, PPRON, PRO1, PRONOUN, TENTATIVE, and THEY. Schizophrenia users had significantly fewer words in the major categories of HOME, LEISURE, and POSITIVE EMOTION, and in the subcategories of ASSENT, MOTION, RELATIVE, SEE, and TIME.

Latent Dirichlet Allocation We find that the difference between the two groups is statistically significant for 8 of the 20 topics, i.e., the relative frequency of the topic per user is significantly different between groups (corrected $\alpha = 0.0025$). Significant topics and top words are shown in Table 1, with the condition with the highest mean proportion shown in the leftmost column and indicated by color: red for schizophrenia (Sch) and blue for control (Con) topics. We then find the topic t with the maximum estimated probability for each user. To see the prevalence of each topic for each condition, see Figure 2, where each user is represented only by their LDA topic t.

Brown Clustering To narrow in on a set of Brown clusters that may distinguish between schizophrenia sufferers and controls, we sum the relative frequency of each cluster per user, and extract those clusters with at least a 20% difference between groups. This yields 29 clusters. From these, we find that the difference between most of the clusters is statistically significant (corrected $\alpha = 0.0017$). Example significant clusters and top words are shown in Table 2.

Perplexity We find this to be only marginally different between groups (p-value = 0.07872), suggesting that a more in-depth and rigorous analysis of this measure and its relationship to the *word salad* effect is warranted.

5.2 Machine Learning

In Section 4, we discussed how we featurized LIWC categories, LDA topics, Brown clusters, Character Language Models, and perplexity. We now report machine learning experiments using these features. We compare two machine learning methods: Support Vector Machines (SVM) and Maximum Entropy (MaxEnt). All methods are imported with default parameter settings from python's scikit-learn (Pedregosa et al., 2011).

As shown in Table 3, the character language model ('CLM') method performs reasonably well at classifying users in isolation, and the features based on the distribution over Brown clusters ('BDist') performs well in a maximum entropy model. An SVM model with features created from LIWC categories and a distribution over LDA topics ('LIWC+TDist') works best at discovering schizophrenia sufferers in our experiments, reaching 82.3% classification accuracy on our balanced test set. Featurizing the distribution over topics provided by LDA increases classification accuracy over using linguistically-informed LIWC categories alone by 13.5 percentage points.

The CLM method performed surprisingly well, given its relative simplicity, and outperformed the LIWC features by nearly ten percentage points when used in isolation, perhaps indicating that the open-vocabulary approach made possible by the CLM is more robust to the type of data we see in Twitter. Combining the LIWC and CLM features, though, only gives a small bump in performance over CLMs alone. Given the fairly distinct distribution of LDA topics by condition as shown in Figure 2, we expected that the ID of the LDA topic t would serve well as a feature, but found that we needed to use the distribution over topics (TDist) in order to perform above chance. This topic distribution feature was the best-performing individual feature, and also performed well in combination with other features, thus seeming to provide a complementary signal. Interestingly, while the CLM model out-performed the LIWC model, the combination of LIWC and TDist features outperformed the combination of CLM and TDist features, yielding our best-performing model.

5.3 Analysis of Language-Based Signals: LDA and Brown Clustering

In the previous section, we examined how well the signals we define discriminate between schizophrenia sufferers and controls in a balanced dataset. We now turn to an exploratory discussion of the language markers discovered with the unsupervised NLP techniques of LDA and Brown clustering, in the hopes of shedding some light on language-based differences between the two groups.

Features	SVM	MAXENT
Perplexity (ppl)	52.0	51.4
Brown-Cluster Dist (BDist)	53.3	72.3
LIWC	68.8	70.8
CLM	77.1	77.2
LIWC+CLM	78.2	77.2
LDA Topic Dist (TDist)	80.4	80.4
CLM+TDist+BDist+ppl	81.2	79.7
CLM+TDist	81.5	81.8
LIWC+TDist	**82.3**	**81.9**

Table 3: Feature ablation results on 10-fold cross-validation. We find that LIWC categories combined with the distribution over automatically inferred LDA topics (TDist) works well for this classification task.

Refer to Tables 1 and 2. Both LDA and Brown clustering produce groups of related words, with different views of the data. We find that both methods group together words for laughing – "haha", "lol", etc. – and these discriminate between schizophrenia sufferers and controls. In LDA, this is Topic 6; in Brown clustering, this is Cluster 0001001.[13] Controls are much more likely to ask someone to retweet ("rt"), pulled out in both methods as well (Topics 7 and 11; Cluster 00001). The two approaches produce word groups with time words like "today" and "tonight" that discriminate between schizophrenia sufferers and controls differently; the word "today" in particular is found in a topic and in a cluster that is more common for controls (Topic 19 and Cluster 01011011010).

LDA pulls out positive sentiment words such as "love", "awesome", "amazing", "happy", "good", etc. (Topics 11 and 19), and topics with these words are significantly more common in controls. It also finds groups for negated words like "don't", "didn't", "won't", etc. (Topic 2), and this is significantly more common in the language of schizophrenia sufferers. Both decreased occurrence of positive sentiment topics and increase of negated word topics is suggestive of the *flat affect* common to schizophrenics. Topic 12 contains a group of words specific to mental health, including the words "mental", "health", and "medical", as well as, interestingly, "schizophrenia" and "schizophrenic" – unsurprisingly occurring significantly more under the schizophre-

[13] Brown clustering is an unsupervised learning process, so the labels just indicate the hierarchical structure of the clusters; for example, Cluster 01 is the parent of Clusters 010 and 011.

nia condition. Recall that we remove the original diagnosis tweet from our analysis, but this topic indicates much more talk about the condition. One wonders whether this might extend to other mental health conditions, and whether the stigma of discussing mental health is reduced within the anonymity provided by the Internet and social media. Figure 2 furthermore indicates that only schizophrenia sufferers have this Topic 12 as their LDA topic t.

Brown clustering pulls out the first person pronoun 'I' as a main cluster, and we find that this is significantly more frequent in schizophrenia sufferers than in controls. This is comparable to the LIWC category 'I', which we also find to be proportionally higher in the language of schizophrenia sufferers. Interestingly, Brown clustering pulls out words that mark *hedging* and *irrealis moods* in English (Cluster 010100111). This is found in phrases such as "I think", "I believe", "I guess", etc. We find that this cluster is significantly more common in the language of schizophrenia sufferers, perhaps related to the dissociation from reality common to the disorder. We also find a Brown cluster for *connectives* (words like "but", "because", "except") in Cluster 01011111111; and this is also significantly more common in schizophrenia sufferers. The use of an exclamation point (Cluster 0010111) also differs between schizophrenia sufferers and controls. Note that markers $<<$ and $>>$ are also common in this cluster. This is an artifact of our text processing of emojis; in other words, both emojis and exclamation points are significantly less likely in the language of schizophrenics. This is potentially another reflection of the *flat affect* negative symptom of schizophrenia.

6 Conclusion

Given its relative rarity compared to other mental health conditions like depression or anxiety disorders, schizophrenia has been harder to obtain enough data to leverage state-of-the-art natural language processing techniques. Many such techniques depend on large amounts of text data for adequate training, and such data has largely been unavailable. However, we can discover a sufficient amount of schizophrenia sufferers via publicly available social media data, and from here we can begin to explore text-based markers of the illness. This comes with a notable caveat: These users battling schizophrenia may be different in some systematic ways from the schizophrenic population as a whole – they are Twitter users, and they are speaking publicly about their condition. This suggests that replication of these findings in more controlled settings is warranted before hard conclusions are drawn.

By applying a wide range of natural language processing techniques to users who state a diagnosis of schizophrenia, age- and gender-matched to community controls, we discovered several significant signals for schizophrenia. We demonstrated that character n-grams featurized over specific tweets in a user's history performs reasonably well at separating schizophrenia sufferers from controls, and further, featurizing the distribution over topics provided by latent Dirichlet allocation increases classification accuracy over using linguistically-informed LIWC categories alone by 13.5 percentage points in an SVM machine learning approach. Moreover, the features produced by these unsupervised NLP methods provided some known, some intuitive, and some novel linguistic differences between schizophrenia and control users.

Our cursory inspection here is only capturing a fraction of the insights into schizophrenia from text-based analysis, and we see great potential from future analyses of this sort. Identifying quantifiable signals and classifying users is a step towards a deeper understanding of language differences associated with schizophrenia, and hopefully, an advancement in available technology to help those battling with the illness.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (5th Edition)*. Arlington, VA: American Psychiatric Publishing.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2011. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th ACM International Conference on Web Science*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

H. Häfner and K. Maurer. 2006. Early detection of schizophrenia: current evidence and future perspectives. *World Psychiatry*, 5(3):130–138.

Elizabeth Hohman, David Marchette, and Glen Coppersmith. 2014. Mental health, economics, and population in social media. In *Proceedings of the Joint Statistical Meetings*.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Ronald C. Kessler, Olga Demler, Richard G. Frank, Mark Olfson, Harold Alan Pincus, Ellen E. Walters, Philip Wang, Kenneth B. Wells, and Alan M. Zaslavsky. 2005. Prevalence and treatment of mental disorders, 1990 to 2003. *New England Journal of Medicine*, 352(24):2515–2523.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu. [Online; accessed 2015-03-02].

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Shashwath A. Meda, Adrienne Gill, Michael C. Stevens, Raymond P. Lorenzoni, David C. Glahn, Vince D. Calhoun, John A. Sweeney, Carol A. Tamminga, Matcheri S. Keshavan, Gunvant Thaker, et al. 2012. Differences in resting-state functional magnetic resonance imaging functional network connectivity between schizophrenia and psychotic bipolar probands and their unaffected first-degree relatives. *Biological Psychiatry*, 71(10):881–889.

National Alliance on Mental Illness. 2015. Schizophrenia. http://www.nami.org/Learn-More/Mental-Health-Conditions/Schizophrenia. [Online; accessed 2015-03-10].

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Pacific Institute of Medical Research. 2015. Common schizophrenia symptoms. http://www.pacificmedresearch.com/common-schizophrenia-symptoms/. [Online; accessed 2015-03-10].

Greg Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and Martin E. P. Seligman. In press. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, and Matthieu Perrot Édouard Duchesnay. 2011. scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX.

Jonna Perälä, Jaana Suvisaari, Samuli I. Saarni, Kimmo Kuoppasalmi, Erkki Isometsä, Sami Pirkola, Timo Partonen, Annamari Tuulio-Henriksson, Jukka Hintikka, Tuula Kieseppä, et al. 2007. Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Archives of General Psychiatry*, 64(1):19–28.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural*, pages 1348–1353.

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

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey A. Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech & Language Processing*, 19(7):2081–2090.

Masoud Rouhizadeh, Emily Prud'hommeaux, Jan van Santen, and Richard Sproat. 2014. Detecting linguistic idiosyncratic interests in autism using distributional semantic models. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Sukanta Saha, David Chant, Joy Welham, and John McGrath. 2005. A systematic review of the prevalence of schizophrenia. *PLoS medicine*, 2(5):e141.

Maarten Sap, Greg Park, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and H. Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Richard E. Lucas, Megha Agrawal, Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

The National Institute of Mental Health. 2015. Schizophrenia. http://www.nimh.nih.gov/health/topics/schizophrenia. [Online; accessed 2015-03-04].