STAT-101: An introductory exploration of detecting Schizophrenia Through Analysis of Tweets

I. INTRODUCTION

We formed a team of four people. Following are the members of our team:

1) Arthita Ghosh
2) Khanh Nguyen
3) Sayantan Sarkar
4) Varun Kumar

We decide to explore different language usages features such as n-grams features, syntatic features, sentiment features etc. The motivation to explore these features comes from the previous studies which indicate that language usage and social expression are telling indicators of mental health. [1][11][12][15]

Broadly, we explore the text and non-text features of the tweets. To determine whether a feature is discriminative or not, we calculate p-values of the features and select only those that are discriminative. We train multiple classifiers on the selected features and also explore the tweets through supervised topic modeling.

II. DATA AND METHODS

A. Data and Resources

The primary dataset for this project is the *Qntfy dataset* which contains approximately *1* million tweets from 274 users. Among them, 137 users self-report being diagnosed with schizophrenia on Twitter. We refer to this group as the *schizophrenia group*. The other 137 users, whose available Twitter history does not self report a schizophrenia diagnosis, are called the *control group*. Each user in the control group is matched with a corresponding user in the schizophrenia group that has the same age and gender.

We also utilize the following additional data resources for feature analysis.

- Linguistic Inquiry Word Count (LIWC): LIWC is a word counting software program that references a dictionary of grammatical, psychological, and content word categories. [3] LIWC has been used to efficiently classify texts along psychological dimensions and to predict behavioral outcomes, making it a text analysis tool widely used in the social sciences.
- AFINN: AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Arup Nielsen. [14]
- Connotation Lexicon: Like AFINN, the Connotation Lexicon assigns scores of $-1, 1$ and 0 to a list of words. [13]

B. Preprocessing

For each user, we combine all the tweets into a single document. For the rest of the document, when we mention feature computation, we mean the above processing method, unless otherwise stated. The tweets contain different kinds of noise like tweets in languages other than English, incorrect spellings. Therefore, we implement a set of common preprocessing tools which are described in the next subsections.

1) Tokenizer: We use the CMU Twitter twokenizer [10] for word tokenization.

2) Spelling Corrections: We build two types of spelling checkers. Both spell-checkers can correct individual words or run in a batch mode.

- *Simple Word Replacement*: Using two lexicon normalization dictionaries [4], [5] which contain commonly misspelt words and their corrections. It is a simple but fast spelling correction system.
- *Probabilistic Single Letter Modification Spell Check*: In this slightly complex, but better spell-checker, simple 1-letter insertions, deletions, replacements of the word are considered and the candidate word with highest probability is returned [6].

3) Splitting Hashtags: Hashtags are important, succinct sources of information, sometimes expressing strong emotions or providing hints about the topics being discussed. But common NLP tools cannot be applied to them without splitting them first. First, we remove the '#' character if present, then we insert dashes before capital letters if they are not surrounded by lower case letters. Finally, the hashtag is split on the dashes and then the smaller chunks are greedily searched for smaller words. This lets the splitter take hints of splitting from capital letters but still be able to split words when capital letters do not separate words. For example, the hashtag #helloworldIBMWow is split into 'hello world IBM wow'

4) Miscellaneous tools: Some utilities to filter users based on folds, gender, age, condition, etc., perform set operations on such groups, and return desired fields.

C. Topic modeling

Topic models are powerful unsupervised tools to discover hidden structures of texts. Based on co-occurrence statistics, a topic model groups words that exhibit certain common themes into clusters, called *topics*. Topic modeling can also be applied in an supervised manner, using bipolar continuous labels to orient topics to be more relevant to a specific response variable of interest. Supervised latent dirichlet allocation (SLDA) [2] associates with each topic a regression variable that measures the correlation of the topic with the response variable. We employ SLDA to discover idosyncratic topics of each group. First of all, we remove authentication information in the tweets such as mentions or URLs to avoid capturing unwanted correlations. Moreover, since topic models are known to perform poorly on short texts such as tweets, we aggregate each person's tweets over one-month periods into longer documents. We then label all documents from the schizophrenic group as positive examples, and documents from the control group as negative examples.

After that, we employ the topic modeling implementations of the Segan Java library [1]. We set the number of topics to be 100 and run the algorithms for 500 iterations. We use the preprocessing and hyperparameter settings as recommended by the library's webpage.

D. Classifier pipeline

For classification, we follow the pipeline in Fig. 1. We extract features from non-text meta data of tweets as well as from the texts. From the text of a tweet, we extract general features (such as rhyming and tree features) and also train n-gram models from it. All extracted features are passed through a p-value checking stage, where we discard unimportant features. Finally, we train 3 classifiers using 10-fold cross validation.

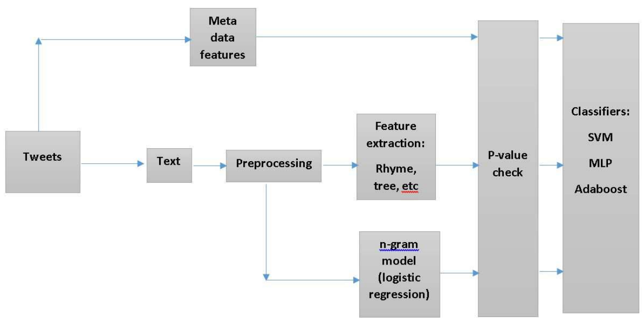

Fig. 1. Block diagram of classifier

III. Feature Analysis

Different groups of features are investigated, some of which are standard features popular in other NLP techniques and some designed specifically for the hopes of detecting schizophrenia. We perform t-test on the features and find a p-value and only include those features in the final classifiers whose p-values are less than 0.05. For each group of features, we report the p-values and plot histograms of each feature for the control and schizophrenic groups in red and blue respectively.

A. Language Models

Language models are mainly used to estimate how likely a given sequence of words is. N-gram model considers previous N words to predict the next word. N-gram models are widely used for language analysis in mental health domain[1]. We choose *2000* most important unigrams based on their TF-IDF values, restricting attention to unigrams appearing in between 2% and 60% of documents. We use scikit-learn machine learning package [19] to implement N-gram features.

[1] https://github.com/vietansegan/segan

B. LIWC Features

The well-known Linguistic Inquiry Word Count (LIWC) is a validated tool for the psychometric analysis of language data [3]. It has been repeatedly used to study language associated with different types of mental disorders like depression. [1][11][12][15] We explore the word usages in all LIWC categories. Each LIWC category provides one feature used by our subsequent machine learning classifiers. From the results, we obeserve that subjects with schizophrenia use fewer adverbs and conjunctions which can be indicative of the 'Poverty of Speech' effect observed in Schizophrenia. Also, the occurence of words related to feeling (perception) is lesser, suggesting that subjects are more detached from the outside world. Control subjects are also seen to be talking more frequently about the future and present than subjects affected by schizophrenia.

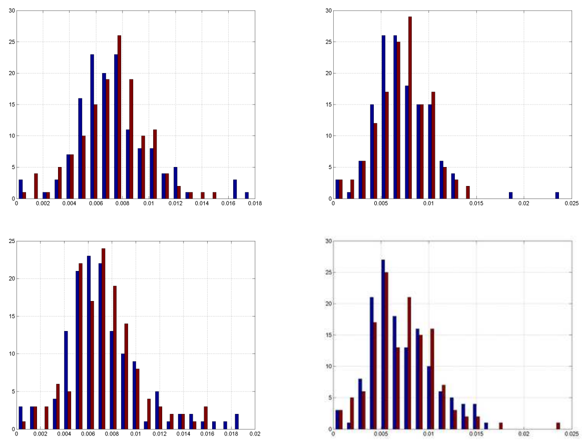

Fig. 2. Histograms of LIWC features. The top row shows occurrence of adverbs and conjunctions and the bottom row shows occurrence of words corresponding to feeling and future tense in the tweets

C. Parse Tree Based Features

Language, the primary mode of communication, draws its contents from thought processes of a person. In schizophrenic patients, abnormality in thoughts (Formal Thought Disorder) and delusions often affect the semantic content of their languages. We analyze these abnormalities with the following features:

1) Frazier Score: Cleaned tweet data for each user is obtained by doing some basic spelling corrections and removing hashtags and special characters. For each user in the schizophrenic group, we take the complete set of tweet data and randomly drop out some of the tweets, to speed up the feature extraction process. The rest are parsed using the pyStatParser, which builds a parse tree assuming a CFG and using the CKY chart parsing algorithm and returns a nltk Tree object [20] with the words in the tweet as nodes. Hence, for each tweet by a particular user, we obtain a parse tree and calculate the Frazier score [17] for each word in the tweet. Frazier Score of a word node is calculated by tracing a path up the tree from the word node (leaf) up to either the root or a node that's not the leftmost child of its parent. We then gather the four following features of the tweet : *sum of word scores in the tweet, average word score per word in the tweet, maximum and minimum word score values for the tweet*. We create a list of these values over all the tweets for the user and then calculate the mean and variance for each of these values for the user. We compute mean and variance of the four above mentioned features for each user.

2) Yngve Score: Yngve Score [16] is calculated in a similar way. We build parse trees on individual tweets with words as leaf nodes. Means and variances are calculated for each user for the sum of word scores, average word Score, maximum and minimum scores. In order to calculate the Yngve score of an individual word node in a parse tree, we sum edge scores from the word up to the root of the parse tree. However, an important thing to note here is that the NLTK tree object numbers the outgoing edges from a tree node starting with 0 for the leftmost node whereas in case of the Yngve score calculation we have to number the edges starting from the rightmost edge being numbered as 0. So we had to modify the parse tree returned by pyStatparser in order to calculate the Yngve Score. It essentially gives us the size of pushdown stack at each word in a top down left to right parser.

Table IV shows the p-values of the features for the two tree based features. Each cell shows the p-value of the mean and variance of the corresponding feature type (sum, average, min and max). Fig. 3 shows the histograms for both Frazier (top row) and Yngve (bottom row) features for the mean and variance of sum of scores.

TABLE I
TREE BASED FEATURES P-VALUES

	Frazier	Yngve
sum of scores	1.84e−5, 1.47e−5	5.53e−4, 0.813
average of scores	0.0051, 0.379	0.025, 0.875
min score	0.836, 0.281	0.0463, 0.613
max score	0.522, 0.665	0.279, 0.389

D. Part Of Speech (POS) Tag Based Features

The propositional idea density (P-density) of a text is approximated by the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words. The Computerized Propositional Idea Density Rater (CPIDR 3.2)[9] provides an automated method for calculating Propositional Idea Density. We have used the CPIDR3 application to estimate the propositional density in the collection of all the tweets for each user. The measure of P-density assigns a score based on the proportion of words within a text that contribute semantically to its overall meaning. Through some studies, it has been observed that in conditions like Aphasia or Dementia, the P-density of written language is significantly lowered[7][8]. We investigate the effect of Schizophrenia on the P-density of tweet data and find that the p-value is 0.0009 for the mean score. Fig. 5 shows that the CPIDR scores of the schizophrenic group is slightly higher, which may be indicative of their jumpy thought process as CPIDR quantifies the density of key ideas in a text.

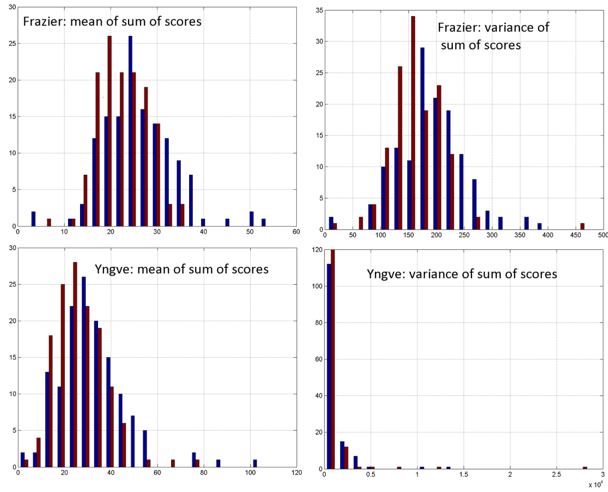

Fig. 3. Histograms of tree features. The top row shows Frazier features and the bottom row shows Yngve features for mean and variance of sum of scores

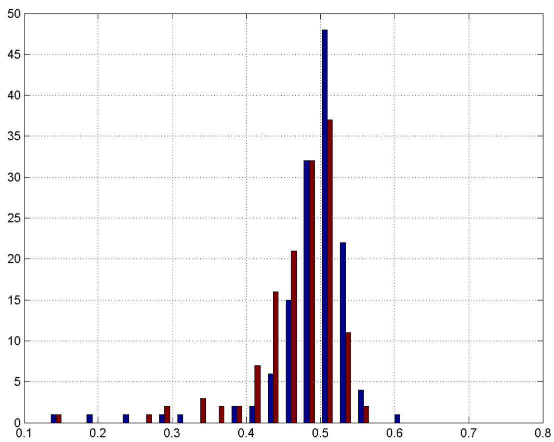

Fig. 4. Histogram showing CPIDR score mean feature for schizophrenic (blue) and control (red) groups. The slightly higher mean for the blue histogram, may be indicative of a jumpy thought process

E. Rhyme Based Features

It has been observed that Schizophrenic subjects sometimes say words together because they rhyme (clang association), even if they do not make much sense. An example from the dataset is *'Lol I tweet lots and lots of loony verses, I think I've created Universes. :-)'*. We generate a rhyming score for each user. Our hypothesis is that in tweets too schizophrenic subjects would manifest this tendency to rhyme. To generate a rhyming score vector for a tweet of n words, we generate rhyming scores for all $n(n-1)/2$ pairs of words. The rhyming score of two words are calculated by converting the words to a list of syllables using NLTK and then finding longest runs of matching syllables, with extra weight being given if the match occurs at the end of the word (so that the words 'Wanda' and 'sundry' have a slightly less score than 'alright' and 'tonight', as the later pair is considered a better rhyme due to the matching at the end of the words than the former, even though both pairs match only 1 syllable). Then the mean, variance and percentage of non zero scores are computed for each user and used as features. The process is conducted on texts with hashtags and text with hashtags separated out.

TABLE II
RHYMING FEATURES P-VALUES

	Mean	Variance	% of non-zero scores
Hashtags present	3.2e−6	2.02e−7	5.37e−6
Hashtags separated	1.21e−4	9.33e−6	4.79e−5

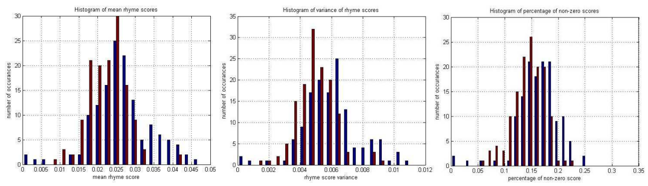

Fig. 5. Histograms of rhyme features (mean, variance and percentage of non-zero scores) for control (red) and schizophrenic (blue) groups. The higher mean for the schizophrenic group shows their tendency to rhyme.

As seen from Table II, splitting out hashtags increases the p-values, indicating that the rhymes are mostly contained in the tweet text rather than in the hashtags. The small p-values are also supported by the clearly shifted distributions that can be seen in Fig. 5. The 1st histogram which depicts the mean rhyme score, clearly shows that the schizophrenic population (blue) has a slightly higher mean, while the 3rd histogram which depicts the number of non zero rhyming scores also shows that the schizophrenic subjects (blue) score higher. These results confirm our aforementioned hypothesis.

F. Simple Sentiment Analysis

A simple sentiment/connotation detection strategy is to get pre-defined scores for words and then calculate frequency, mean and variance of scores for the whole corpus. Here negative words have smaller scores and positive numbers have larger scores. We use two lexicons, namely the 'Connotation Lexicon' [13] that assigns values of -1, 0 or 1 and 'AFINN' [14], which assigns integer values from -5 to 5. The mean and variance of the number of time each word score occurs is used as features.

The AFINN based sentiment analysis does not produce good p-values or histograms, therefore we focus on the 'Connotation lexicon' based analysis here. Table III shows the p-values from which we see a somewhat small p-value for the first 6 features. Interestingly, the p-value for the 'frequency of neutral words' is smallest, indicating that the two groups use neutral words at different rates. Fig. 6 shows that the schizophrenic group (blue) uses neutral words with a slightly higher rate than the control group. Perhaps this is indicative of the flat affect with which schizophrenic subjects are known to speak with.

TABLE III
CONNOTATION SCORES P-VALUES

features	p-values
frequency of '1'	0.0018
frequency of '-1'	0.0123
frequency of '0'	0.00031029
frequency of 'unknown'	0.00064892
mean score	0.0018
score variance	0.000023416
max score	0.5719

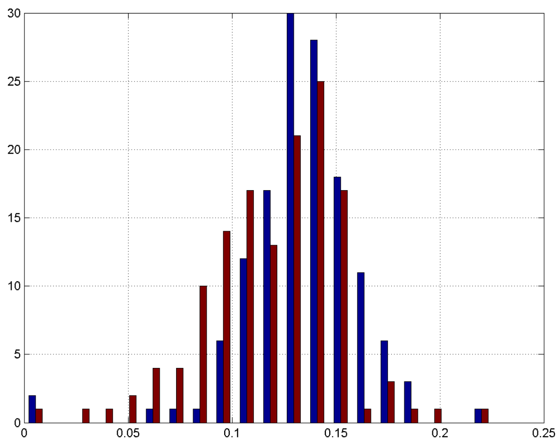

Fig. 6. Histogram of a AFINN sentiment feature (frequency of neutral words) for control (red) and schizophrenic (blue) groups. The slightly more frequency of neutral words in schizophrenic group may indicate flat affect.

G. Emoticon Analysis

The top 400 most frequent emoticons are manually annotated with numbers indicating their sentiment. Then each tweet is assigned a score based on its emoticon content. For features, the mean, variance of scores and the number of emoticons are considered. However, the p-values are large as shown in Table IV and the histograms in Fig. 7 did not show any significant difference for the two groups.

TABLE IV
EMOTICON FEATURES P-VALUES

mean	variance	number
0.8271	0.6045	0.1568

H. Neologisms

Given that schizophrenic subjects may sometimes use new words, a feature that calculated the percentage of new or unknown words used in a tweet is considered. However the p-value or histogram plots did not reveal any interesting pattern.

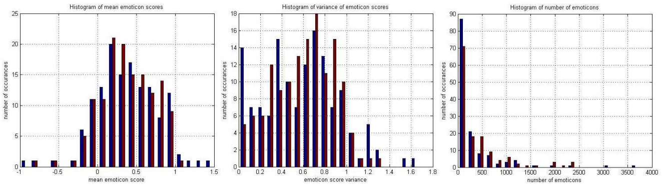

Fig. 7. Histograms of emoticon scores features (mean, variance and count) for control (red) and schizophrenic (blue) groups

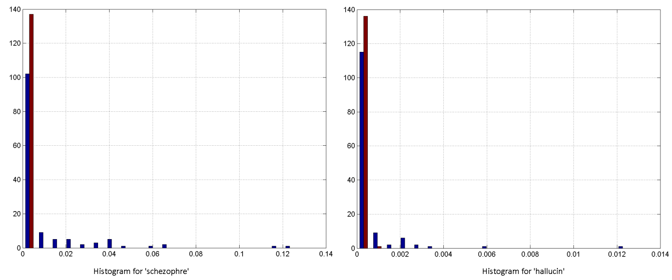

Fig. 8. Histograms of Schizophrenia specific words frequency features (frequency of 'schizophre-' and 'hallucin-') for control (red) and schizophrenic (blue) groups. The blue group shows a significant probability mass in the higher frequency region.

I. Non-text features from tweets

Other than the actual text of the tweet, some additional meta-data is present in the dataset. Information that indicates a person's social life (retweets, number of followers, etc.) could be a potential indicator of someone's mental state. Therefore, the following features are explored: retweet counts, favorite count, follower and friend counts, status counts etc. The mean, variance, minimum, maximum and percentage of zero values are extracted from these counts for each user.

The p-values of most of these features are not promising. It seems that unlike depression or other mental illnesses, where users tend to retract from social interactions (manifesting in the form of less friends, less tweets or followers), there is no pronounced change in the online presence of schizophrenic subjects.

J. Detect Schizophrenia specific words

In this set of features we find the relative frequency of certain schizophrenia related words like those containing the terms 'schizophre-', 'insane', 'paranoia', 'hallucin-', 'confus-', 'medic-', 'symptom' etc. Naturally, these words occur much less in the control group and is very discriminative as can be seen in Table V and Fig. 8. However this feature somewhat defeats the purpose of our study as one might assume that user tweet such words only after they have been diagnosed. The histograms in Fig. 8 show that for non-zero counts, the blue (schizophrenic) group has significantly more probabilistic mass compared to the red plot (control).

TABLE V
SCHIZOPHRENIA RELATED WORDS FREQUENCIES P-VALUES

schizophre	$5.9e{-7}$	insane	$1.6e{-3}$
hallucin	$1.55e{-4}$	confus	$6.02e{-5}$
medic	$1.0e{-4}$	symptom	$3.5e{-3}$

IV. Results

A. Exploratory analyses

1) Supervised topic modeling: Table VI shows the top 5 positive and negative topics scored by SLDA. After examining these topics, we identify several different patterns in the structure and semantics of the languages used by the two groups. The most positive topics contain words related to religion (topic 1), offensive topic (topic 2), and other controversial topics (topic 4 about Armenian and Hitler). If a topic does not fall into those categories (e.g. topic 3 and 5), its top words appear to be fairly random to the extend that we hardly recognize a common theme between them. On the other hand, top words of the most negative topics exhibit much stronger coherence. They are mostly about sports, arts and politics, which are arguably creative and socially engaging topics. Moreover, the top words are rarely offensive or taboo but either neutral or positive. For instance, topic 9 carries a jovial attitude, with a lot of "laugh" words (e.g. "lmfao", "lolol") and words about friendship (e.g. bestfriend, ommf). We find it interesting that the most negative topic is about politics and journalism. Since those tend to be favorite topics of well-spoken, extrovert people, this finding provides evidence to support that people possessing eloquent linguistic skills are unlikely to be diagnosed with schizophrenia.

2) Correlation among features: It is expected that some features might be correlated with each other. To check that, heat maps of correlations between the different features are plotted. A sample heat map is shown in Fig. 9.

The strongly correlated group of features in the centre are for the rhyming features. The mean rhyme score is expected to increase with the number of non-zero scores. Also the rhyming features are present for both hashtags retained and for hashtags split and expanded. These features are expected to correlate strongly.

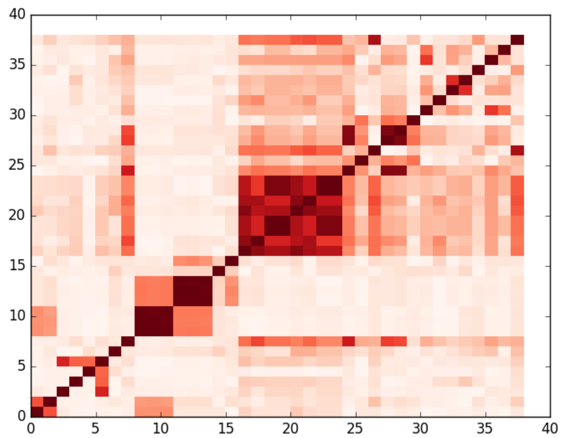

Fig. 9. Heat map of feature correlations

Topic	Score	Topic words
1	1.980	christ, please_retweet, the_bible, satan, evolution, jesus_christ, christians, atheist, praise_jesus, please_pray
2	1.943	your_tweet, favs, favorited, penis, rat, fucken, schizophrenic, sext, sant, knife
3	1.615	daisy, wiz_khalifa, nebraska, kratom, cigarette, www, druitt, soccer, owh, pug
4	1.593	dan, now_following, for_following, armenian, israel, lmfao, pal, par, tbh, hitler
5	1.556	yeh, gerard, frank, mcr, emo, ray, anime, dont_have, youre, peach
6	-0.938	toronto, finale, rob_ford, fave, the_blog, yoga, designer, fashion_week, runway, presentation
7	-1.057	sound_art, exhibition, devotion, barriers, gig, sound_through, museum, gallery, jeph_jerman
8	-1.093	lebron, bills, miami, baseball, hip_hop, bulls, kobe, nba, spurs, nfl
9	-1.107	lmfao, lls, bestfriend, oomf, homework, lolol, lololol, rft, hampton, uvalde
10	-1.139	romney, bush, gop, fox_news, novel, congress, arizona, brian_williams, president_obama, flickr

TABLE VI
TOPIC 5 MOST POSITIVE/NEGATIVE TOPICS DISCOVERED BY SLDA.

B. Classification

1) Protocol and Metrics: The given dataset is split into 10 folds, each with approximately 14 schizophrenic subjects and 14 control subjects. We test our classifiers by cross-validating on these folds. Specifically, models are trained and evaluated 10 times, where each time a different fold is used for validation, while the rest are used for training. Before training the classifiers during each of the 10 iterations, we also randomly shuffle the training data and normalize the features by z-scoring them.

To evaluate the models, we report the accuracy and F-1 score of each fold and finally report the mean of these metrics across the 10 folds.

We use 3 different types of classifiers:

- *Adaboost*: an Adaboost classifier begins by fitting a decision tree classifier of a maximum depth of 5 on the training data and thereafter it fitted multiple copies of the same classifier on the data with additional weight on the wrongly classified samples. We use the Adaboost implementation from the *sklearn* Python package. We set the maximum depth of the trees to be 5. The maximum number of estimators at which boosting is terminated is 500.
- *Multi Layer Perceptron*: Lasagne/Theano [21] MLP with one hidden layer of 75 units.
- *Support Vector Machine*: *sklearn*'s SVM with RBF kernel.

C. Baseline

To construct a baseline, we build a unigram model with 2000 most important unigrams based on their TF-IDF values. Our unigram model yields 75.3% F1-score on cross-validation. A trigram model with 2000 trigrams works a little better and yields 77% F1-score. We also explore higher n-gram models like 4-gram and 5-gram models but do not observe any improvement in the prediction accuracy.

D. Classification Results

The best performance that we observed is 80.64% accuracy and 80.63% F1 score with SVM. Even with schizophrenia related word frequency features removed (section III.J), since they might not be present in other datasets, the performance is similar. Results of furthur experiments are summarized in Fig. 10. There are 6 bars in each bundle, representing the accuracy and F1 scores for SVM, Multilayer perceptron (MLP) and Adaboost (AB). The last group shows the best performing classifier, which uses all the available features. To understand the importance of each feature, we train the models using only one set of features at a time. Based on Fig. 10 we conclude that n-grams model features, LIWC features and schizophrenia specific word features are most important. Each individually gives performance above 70%. Features based on parse trees and rhyming are second best, which offers between 60% to 70% accuracy. The sentiment/connotation and non-text features are the worst performing features individually, providing only around 60% accuracy. Removing correlations may let us decrease model complexity, as we can train with a smaller number of features. Hence PCA is tried, with 99% variance retained, however performance dropped significantly (by srounf 10%). Therefore we skipped PCA in furthur analyses.

Another observation is that the accuracy and F1 metrics correlate with each other very strongly. We also see that all three classifiers perform roughly similarly for the same set of features.

V. DISCUSSION AND FUTURE WORK

We have tried out a variety of features that can be extracted from tweets which can be potential indicators of schizophrenia.

- Simple models based on n-grams and LIWC features perform quite well.
- AFINN based sentiment analysis show some interesting results about the way Schizophrenia patients use language and the how their emotions (or the lack of it) are expressed through language. For example, AFINN sentiment analysis indicates that subjects affected by Schizophrenia use more neutral words which may suggest 'flatness in emotions' through the usage of language.
- Rhyming features in tweets are indicative of the 'Clanging or Clang Association' effect seen in subjects with Schizophrenia. As a classifier feature, rhyming features

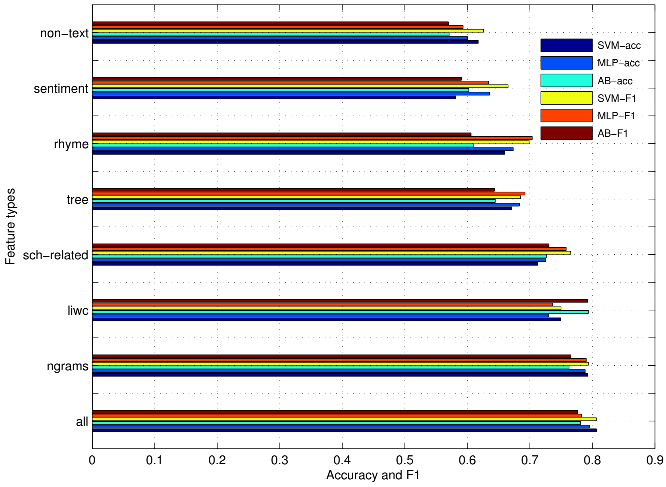

Fig. 10. Performance of 3 classifiers (SVM, MLP and Adaboost) for 2 metrics (accuracy and F1) trained individually on certain feature groups and when trained all together

perform quite well, acheiving around 65% accuracy without the help of any other feature.

- POS tag based Propositional Density studies show how their though process works tangentially and its manifestation in their tweets. Our analysis indicates a slightly higher density of Propositions in the tweets by subjects with Schizophrenia which may be the effect of the certain symptoms of thought disorder in patients with Schizophrenia for example 'Blocking', 'Tangentiality','Derailment' and 'Flight of Ideas'.

If more data is available in future, interesting subgroup studies based on gender or age group of Schizophrenia patients can be carried out.

REFERENCES

[1] Glen Coppersmith, Mark Dredze and Craig Harman *Quantifying mental health signals in twitter.* ACL 2014 (2014): 51. http://anthology.aclweb.org/W/W14/W14-32.pdf#page=63

[2] Mcauliffe, Jon D., and David M. Blei. *Supervised topic models.* Advances in neural information processing systems. 2008. http://papers.nips.cc/paper/3328-supervised-topic-models.pdf

[3] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn *The Development and Psychometric Properties of LIWC2015* UT Faculty/Researcher Works, 2015 https://utexas-ir.tdl.org/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?sequence=3

[4] Text Normalization Data Fei Liu http://www.hlt.utdallas.edu/~yangl/data/Text_Norm_Data_Release_Fei_Liu/

[5] Han, Bo and Cook, Paul and Baldwin, Timothy *Automatically Constructing a Normalisation Dictionary for Microblogs* http://dl.acm.org/citation.cfm?id=2390948.2391000

[6] P. Norvig *How to Write a Spelling Corrector* http://norvig.com/spell-correct.html

[7] Lucy Bryant , Elizabeth Spencer , Alison Ferguson , Hugh Craig , Kim Colyvas and Linda Worrall *Propositional Idea Density in aphasic discourse* Aphasiology, 27:8,992-1009, DOI: 10.1080/02687038.2013.803514 http://dx.doi.org/10.1080/02687038.2013.803514

[8] Elizabeth Spencera*, Alison Fergusona, Hugh Craiga, Kim Colyvasb, Graeme J. Hankeycd and Leon Flickere *Propositional idea density in older mens written language: Findings from the HIMS study using computerised analysis* Clinical Linguistics and Phonetics Volume 29, Issue 2, 2015 http://www.tandfonline.com/doi/full/10.3109/02699206.2014.956263

[9] Cati Brown, Tony Snodgrass, Susan J. Kemper, Ruth Herman, and Michael A. Covington *Automatic Measurement of Propositional Idea Density from Part-of-Speech Tagging* PMC2423207 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2423207/

[10] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. *Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments* InProceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 2011 Jun 19 (pp. 42-47). Association for Computational Linguistics.

[11] Philip Resnik, Anderson Garron, Rebecca Resnik *Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students* Proceedings of the 2013 Conference on Empirical Methods in Natural. Association for Computational Linguistics, 2013 http://www.aclweb.org/anthology/D13-1133

[12] Jennifer Alvarez-Conrad, Lori A. Zoellner and Edna B. Foa *Linguistic predictors of trauma pathology and physical health* Applied Cognitive Psychology 15.7 (2001): S159-S170 http://onlinelibrary.wiley.com/doi/10.1002/acp.839/abstract

[13] Feng, Song and Kang, Jun Sak and Kuznetsova, Polina and Choi, Yejin *Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning* Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) http://www3.cs.stonybrook.edu/~ychoi/connotation/

[14] F. Å. Nielsen *AFINN* Informatics and Mathematical Modelling, Technical University of Denmark http://www3.cs.stonybrook.edu/~ychoi/connotation/

[15] Yla R. Tausczik and James W. Pennebaker *The psychological meaning of words: LIWC and computerized text analysis methods* Journal of Language and Social Psychology, 29(1):2454, 2010 http://jls.sagepub.com/content/29/1/24

[16] Yngve VH. *A model and an hypothesis for language structure.* Proceedings of the American philosophical society. 1960 Oct 17;104(5):444-66.

[17] L. Frazier. *Syntactic complexity.* In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, Natural Language Parsing. Cambridge University Press, Cambridge, UK, 1985

[18] Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. *Spoken language derived measures for detecting mild cognitive impairment.* Audio, Speech, and Language Processing, IEEE Transactions on. 2011 Sep;19(7):2081-90.

[19] *Scikit-learn: Machine learning in Python* http://scikit-learn.org/stable/about.html

[20] Bird, Steven, Edward Loper and Ewan Klein *Natural Language Processing with Python. OReilly Media Inc.* http://www.nltk.org/

[21] Theano Development Team *Theano: A Python framework for fast computation of mathematical expressions* http://deeplearning.net/software/theano/