Understanding Cohesion in Writings and Speech of Schizophrenia Patients

Amal Abdullah AlQahtani*§, Efsun Sarioglu Kayi†, and Mona Diab‡*
*Department of Computer Science, The George Washington University, DC, USA
†Department of Computer Science, Columbia University, New York City, USA
‡AWS, Amazon AI, USA
§King Saud University, Riyadh, Saudi Arabia
*amalqahtani@gwu.edu, †ek3050@columbia.edu, ‡diabmona@amazon.com

***Abstract*—Schizophrenia is one of the mental disorders that impacts a person's thinking, speech, and actions. It can reduce a persons ability to process auditory information and make decisions. Analyzing this disorder correctly is important because it might help with different ways of reducing its negative effects on its patients. Linguists and psychiatrists have been investigating language impairments and speech disorder in people with schizophrenia disorder which can be challenging. In this study, we attempt to address this issue by analyzing linguistic features i.e. cohesion in the writings and speech scripts of schizophrenia patients. Our results show that using referential cohesion with text easability or situation model features provides the best performance for speech whereas for writing dataset, readability or a combination of situation model and readability yield the best performance.**

***Index Terms*—Schizophrenia, Machine Learning Algorithms, Binary Classification, Coherence, Cohesion, Coh-Metrix**

I. Introduction

Schizophrenia is a psychotic disorder where the main symptom is that one has an impaired perception of reality [1]. It impairs the normal functioning of the brain in such a way that the manner in which an individual thinks, expresses himself or herself or relates with others becomes distorted [2]. Furthermore, it can significantly impair the functional abilities such as the learning ability and social interactions with others [3].

Currently, more than 21 million people globally, suffer from Schizophrenia [4] and there is a need for a deeper understanding of its conditions. This could be critical in not only assessing the patients, but also in identifying them so that they can receive the appropriate medical care in a timely manner.

Language can play a crucial role in identifying someone's mental illness [5]. Previous studies have shown how language can help in diagnosing and predicting mental illness e.g. identify people who suffer from: depression and anxiety [6]–[10], Alzheimers [11]–[13], post traumatic stress disorder (PTSD) [14], or schizophrenia [15]–[17]. Specifically for schizophrenia, there can be impaired coherence and overall lack of contextual structure [18]. Hence, in this work, we investigate linguistic features related to cohesion for two data sets (1) recorded and transcribed speech; and (2) written essays, with the end goal of identifying and classifying patients with schizophrenia. For this purpose, we trained two machine learning models, namely Support Vector Machine (SVM) and Random Forests (RF) to classify patients and controls. Our results show that among all cohesion features, situation model and readability performed the best for writing dataset and combination of referential cohesion, text easability, and situation model for speech.

II. Related Work

Few studies analyze the linguistic signs of schizophrenia by using natural language processing (NLP) methods. The work of Mitchell et al. [17] and Kayi et al. [16] analyzed several linguistic features of schizophrenia patients' tweets e.g. Latent Dirichlet Allocation (LDA), clustering, and sentiment analysis. The study of [16] additonally examined writings of schizophrenia patients with syntactic and pragmatics features. Another study, conducted by Minor et al. [19], presented lexical analysis to predict whether emotion (*positive* and *negative*) and social word use are associated with metacognition, schizophrenia symptoms, and general functioning. They used Linguistic Inquiry and Word Count (LIWC), a text analysis tool that calculates the percentage of words belonging to different psychological categories, such as emotions, thinking styles, and social relationships [20], to assess the speech content generated by patients and focused on some of the primary psychological processes (e.g., emotions, social) and personal concern categories, i.e. *work* and *achievements*. The results showed that anger and social words have a significant effect on schizophrenia symptoms and metacognition, respectively. Finally, the work of Gupta et al. [21] examined the cohesion in written narratives produced by youth at ultra high risk (UHR). They assessed the cohesion of text in terms of referential cohesion only obtained by Coh-Metrix [22], an automated text analysis tool. The study showed that UHR youth may encounter challenges compared to controls in the use of referential cohesion, and these language disorders are related to symptoms and cognitive function of the patients.

Although these previous works have sought to identify schizophrenia patients via linguistics features, none of them analyzed coherence in both written and spoken language. The focus of this paper is to investigate linguistic features of cohesion, i.e. *referential cohesion*, *text easability*, *situation model*,

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00068

and *readability*, to determine whether individuals diagnosed with schizophrenia can be differentiated from controls based on their written text as well as their transcribed speech.

III. Data

We used two datasets for this study. Speech and writing samples were obtained from healthy controls and patients with a diagnosis of schizophrenia. The first dataset called *LabWriting* consists of 188 participants who were native English-speaking and aged $18-50$ years. $N = 93$ were patients who have a diagnosis of schizophrenia form disorder and $N = 95$ were healthy controls. Patients were cognitively aware enough to participate in this study. All participants were asked to write two paragraph-length essays: one about their average Sunday and the second about what makes them the angriest. The total number of writing samples collected from both patients and controls is 373. Detailed information on this dataset can be found in [16].

The second dataset called *LabSpeech* consists of 93 patients and 95 eligible controls. It includes three questions which prompt participants to describe some emotional and social events. Patients and controls were asked to describe (1) a picture, (2) their ideal day, and (3) their scariest experience. The total number of speech scripts samples collected from both patients and controls is 431. Speech data was transcribed to text and a punctuation tool [23] was used to add missing punctuation.

TABLE I
Descriptive analysis of LabWriting and LabSpeech Datasets

Dataset	Descriptive Language Variables	Patient		Control	
		Avg.	Std	Avg.	Std
LabWriting	Num. of paragraphs	1.2	0.8	1.1	0.5
	Num. of sentences	6.1	3.9	7.1	3.8
	Num. of words	109.9	49.8	141.3	37.7
LabSpeech	Num. of paragraphs	1	0	1.1	0
	Num. of sentences	10.5	5,7	13.9	5.9
	Num. of words	219.6	100	276.7	90.7

Table I shows a descriptive analysis of both datasets. The average number of paragraphs in both datasets generated by both classes is close to each other; however, the controls produced more words per text for the same questions in each dataset. This also coincides with previous studies showing that patients might have a hard time when expressing their thoughts and feelings [24] and that they may not be able to express much or follow along due to impoverishment of speech and language [25] [26].

IV. Linguistic Features of Cohesion

The coherence of a text refers to the characteristics of the text that allows a reader to mentally understand it, based on general knowledge and linguistic connections [22]. To analyze the coherence in our datasets, we relied on Coh-Metrix [22] which is a computational tool that measures the cohesion in a text sample and outputs the results in the form of linguistic indices. A subset of those indices that will be analyzed in this paper are focused on Referential Cohesion, Text Easability, Situation Model, and Readability.

A. *Referential Cohesion (RC):*

RC refers to the relationship between words in sentences that are adjacent to each other (local) or within a paragraph (global) in a text sample [22]. In Coh-Metrix, the indices that are output for RC are referred to as co-reference measures. The word relationships (overlaps) in the co-reference measures are determined based on *noun overlap, argument overlap, stem overlap*, and *anaphor overlap*. In general, these co-reference measures analyze the use of nouns, pronouns, verbs, adjectives and adverbs, as well as how they relate to adjacent sentences (local) and the sentences within a paragraph (global) [22]. Examples of words that represent RC are: pronouns (*he, theirs*), demonstratives (*these, it*), comparatives (*more than, fewer, identical*) and so on.

TABLE II
Referential Cohesion (RC) of LabWriting dataset

RC Type	RC Indices	Patient		Control	
		Avg.	Std	Avg.	Std
Local RC	Noun overlap	0.23	0.28	**0.28**	0.27
	Argument overlap	0.62	0.37	**0.70**	0.29
	Stem overlap	0.30	0.30	**0.35**	0.30
	Anaphor overlap	0.60	0.38	**0.69**	0.32
Global RC	Noun overlap	0.20	0.25	**0.25**	0.25
	Argument overlap	0.59	0.35	**0.66**	0.28
	Stem overlap	0.26	0.28	**0.30**	0.27
	Anaphor overlap	0.50	0.37	**0.56**	0.33

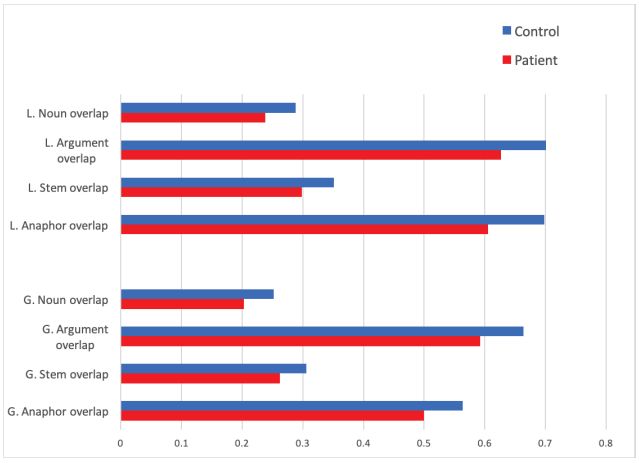

Fig. 1. Mean of the Referential Cohesion Indices for LabWriting Dataset, where L=Local and G=Global

Table II and Table III show the results of RC variables which are represented using means and standard deviations. Means range from 0 (no cohesion) to 1 (highest levels of cohesion). It can be clearly seen that the average values of RC indices in the LabWriting dataset generated by patients are *lower* than controls. However, the average values of RC indices in LabSpeech dataset generated by patients are *higher*

TABLE III
REFERENTIAL COHESION (RC) OF LABSPEECH DATASET

RC Type	RC Indices	Patient		Control	
		Avg.	Std	Avg.	Std
Local RC	Noun overlap	**0.31**	0.26	0.27	0.19
	Argument overlap	**0.63**	0.25	0.60	0.22
	Stem overlap	**0.39**	0.28	0.32	0.20
	Anaphor overlap	**0.60**	0.27	0.59	0.23
Global RC	Noun overlap	**0.27**	0.23	0.21	0.16
	Argument overlap	**0.59**	0.24	0.53	0.21
	Stem overlap	**0.33**	0.25	0.27	0.17
	Anaphor overlap	**0.43**	0.27	0.37	0.24

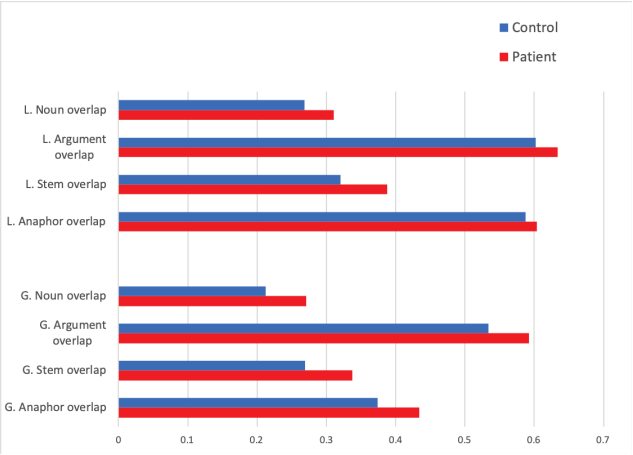

Fig. 2. Mean of the Referential Cohesion Indices for LabSpeech Dataset, where L=Local and G=Global

than controls. This is in line with previous studies [27] [18] that show patients use more self reference and repetitions in speech without needed which affects RC and reduces the overall complexity in the spoken language.

B. *Text Easability (TE):*

The easability of text refers to the overall difficulty of a text sample [22]. In Coh-Metrix, the easability profile of a text sample goes beyond general readability measures, and can be grouped into several categories [22]. The first textual category is *narrativity*, which refers to text that describes a story. Next is *syntactic simplicity*, which refers to the number of words used in a sentence and how difficult or easy it is for the reader to mentally process it. *Word concreteness* refers to text that contain words that are either concrete or text that are too abstract, yielding more difficulty in understanding the text. *Deep cohesion* refers to the extent by which a text sample provides the reader explicit causal and logical inferences that aid in understanding the ideas presented. Next is *verb cohesion*, which refers to the degree to which verbs are connected across multiple sentences in a narrative text sample, allowing for situational understanding by the reader. *Connectivity* refers to the number of explicit logical relations in a text sample, which rely on comparative, additive and adversarial connectives (words such as ”alternatively”, ”additionally”, and ”but”, respectively) to aid understanding. The last textual category is *temporality*, which refers to the texts that consistently focus on cues about tense and aspect of verbs. As a result, a reader is able to fully understand the events described in any such texts.

TABLE IV
AVERAGE PERCENTILE OF TEXT EASABILITY (TE) INDICES

TE Indices	LabWriting		LabSpeech	
	Patient	*Control*	*Patient*	*Control*
Narrativity	80%	**87%**	83.9%	**85.5%**
Syntactic Simplicity	**30%**	27%	26.4%	**32.6%**
Word Concreteness	66%	**68%**	**67.1%**	58.8%
Deep Cohesion	71%	**81%**	40.8%	**47.5%**
Verb Cohesion	**65%**	57%	66.7%	**72.7%**
Connectivity	**21%**	12%	**10.4%**	8.6%
Temporality	54%	54%	**58.2%**	56.5%

Table IV show the average percentile of TE indices of both datasets. *Narrativity* generated by the control cohort is *higher* than those found in the patient cohort. Narrativity relates to word familiarity, world knowledge, and oral language. This implies that narrative text by healthy participants was well developed with characters, events, places, and things that are familiar to the reader. In addition, this is in line with a previous study showed that when schizophrenia patients provided narratives, they were unable to describe the proper sequence of events, causing their speech to be incoherent [28]. In terms of *deep cohesion*, the analysis shows that the patients in both datasets, have *lower* scores compared to controls. When the text is low in deep cohesion, this means that the text contains many relationships, but does not contain enough causal and intentional connectives. On the other hand, the average percentiles of *connectivity* for patients is *higher* in both datasets; this means that patients used more explicit adversarial and comparative connectives to express their relations in text [22].

C. *Situation Model (SM):*

SM refers to the degree of causal verbs in a text sample that evokes mental images in the reader [22]. In Coh-Metrix, these causal verbs (e.g. *make, allow, require*) are analyzed by measuring the relationships between causal particles and causal verbs, which represent *causal cohesion*. Other SM analyses measure *intentional cohesion* which is the ratio of intentional particles and actions and *temporal cohesion* which is the average of repetition tense and repetition score [22]. Additional phrases that fall within this category are: *I am feeling sick, I used to like*.

Table V shows the average values and standard deviations of SM indices of the *LabWriting* and *LabSpeech* datasets. *Casual*, *intentional*, and *temporal cohesion* in both datasets and generated by patients are *lower* than control. The text is less coherent when it has many causal verbs but fewer causal particles that help in indicating how the events and actions are connected [22]. Hence, when reading each sentence, the reader needs to determine the relationships between casual events and actions [22] [29].

TABLE V
SITUATION MODEL (SM) RESULTS FOR BOTH DATASETS

SM Indices	LabWriting		LabSpeech	
	Patient	Control	Patient	Control
	Avg. (SD)	Avg. (SD)	Avg. (SD)	Avg. (SD)
Causal verbs	28.9 (21.0)	26.3 (17.8)	24.1 (14.5)	25.5 (13.4)
Causal content	46.7 (26.7)	43.4 (20.0)	34.4 (17.6)	37.1 (26.6)
Intentional content	18.6 (17.0)	**19.9** (16.0)	12.8 (11.5)	**14.2** (9.3)
Causal cohesion	0.65 (0.97)	**0.76** (0.99)	0.44 (0.52)	**0.51** (0.71)
Intentional cohesion	1.45 (1.79)	**1.70** (1.86)	1.02 (1.06)	**1.14** (1.16)
Temporal cohesion	0.49 (0.98)	**0.66** (0.74)	0.80 (0.42)	**0.85** (0.11)

D. *Readability (Read*):*

Readability refers to the level of difficulty in understanding written text [22]. Historically, several readability formulas have been developed to assess the readability of texts. Using Coh-Metrix, we used two of such formulas for this work. The first one is the *Flesch Reading Ease* formula (FRE), whose output is a number between 0 and 100. The higher the score, the greater the ease of reading the input text is. The second readability formula is the *Flesch-Kincaid Grade Level* formula (FKGL), whose output is a score representing a U.S. grade-school level ($0-12$), where the score 12 means the text is harder to read. Both formulas require at least 200 words for proper, meaningful analysis. Detailed information about the formulas can be found in [22]. Table VI shows the average values and standard deviations of Read* indices for both datasets. The average of FRE and FKGL for patients in Labwriting dataset indicates that the text written by patient is easier to read comparing to the ones written by control, and this means that patients write short sentences and and the number of syllables in words is less than 2.

TABLE VI
READABILITY (READ*) RESULTS FOR BOTH DATASETS

Read* Indices	LabWriting		LabSpeech	
	Patient	Control	Patient	Control
	Avg. (SD)	Avg. (SD)	Avg. (SD)	Avg. (SD)
FRE (0-100)	69.9 (22.9)	67.4 (20.9)	78.3 (16.6)	80.2 (11.7)
FKGL (0-12)	10.5 (10.7)	11.2 (9.9)	8.5 (6.1)	7.6 (3.7)

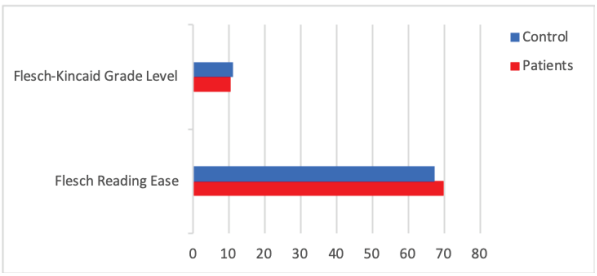

Fig. 3. Mean of the Read* Indices for LabWriting Dataset

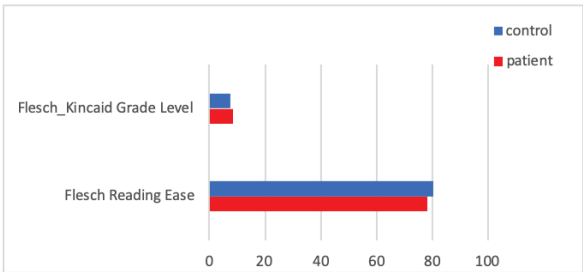

Fig. 4. Mean of the Read* Indices for LabSpeech Dataset

V. EXPERIMENT DESIGN

We frame the problem as a supervised binary classification task that requires a model to differentiate between patients and healthy controls. In addition to bag of words representation of the text where entries are weighted by term frequency and inverse document frequency, we added cohesion scores either as numerical or categorical features. For read*, a threshold value was used to convert it to binary categories whereas all others were used as their original values. For classifiers, we relied on Scikit-learn package [30] using Python. Two popular machine learning algorithms, i.e., Support Vector Machines (SVM) with linear kernel [31], and Random Forest (RF) are used for training the classifiers. We conducted the experiments with 70% for training and 30% for testing. We report F-score and Area Under Curve (AUC) value which is calculated as the area under the receiver operating characteristics curve (ROC).

A. Analysis of Classification Results

Table VII and Table VIII illustrate results for the LabWriting and LabSpeech datasets, respectively. We report several combinations of feature sets and their effect on the performance where the top one is shown in bold. Overall, SVM performs better than RF and using cohesion features in addition to text and speech improves classification performance. For LabWriting dataset, the best performing features according to F-Score are *Text+SM*, *Text+SM+Read**, and *Text+RC+SM* for SVM and *Text+Read** for RF. According to AUC, the best performing feature is *Text+RC* for RF, and a combination of different features *Text+SM+Read** performs the best for SVM. For LabSpeech dataset, the best performing features according to F-Score are the combination of features: *Text+RC+TE* for SVM, and *Text+RC+SM* for RF classifier. According to

both F-Score and AUC, the best performing features are the combination of different features: *Text+RC+SM*.

TABLE VII
CLASSIFICATION PERFORMANCE OF LABWRITING DATASET

	SVM		RF	
Features	**F-Score**	**AUC**	**F-Score**	**AUC**
Text only	0.66	0.78	0.64	0.71
Text+RC	0.69	0.74	0.66	**0.72**
Text+TE	0.65	0.71	0.65	0.69
Text+SM	0.72	0.77	0.64	0.68
Text+Read*	0.67	0.75	**0.67**	0.68
Text+RC+TE	0.66	0.69	0.64	0.66
Text+RC+SM	0.72	0.72	0.64	0.68
Text+RC+Read*	0.70	0.75	0.59	0.63
Text+TE+SM	0.68	0.73	0.59	0.62
Text+TE+Read*	0.65	0.70	0.65	0.68
Text+SM+Read*	**0.72**	**0.79**	0.60	0.63
ALL	0.65	0.69	0.60	0.65

TABLE VIII
CLASSIFICATION PERFORMANCE OF LABSPEECH DATASET

	SVM		RF	
Features	**F-Score**	**AUC**	**F-Score**	**AUC**
Text only	0.70	0.79	0.64	0.63
Text+RC	0.73	0.78	0.69	0.70
Text+TE	0.71	0.79	0.69	0.65
Text+SM	0.71	0.78	0.57	0.63
Text+Read*	0.73	0.77	0.69	0.67
Text+RC+TE	**0.74**	0.78	0.63	0.65
Text+RC+SM	0.73	0.78	**0.71**	**0.72**
Text+RC+Read*	0.72	0.78	0.66	0.66
Text+TE+SM	0.71	0.77	0.66	0.68
Text+TE+Read*	0.73	**0.79**	0.66	0.65
Text+SM+Read*	0.70	0.76	0.66	0.70
ALL	0.73	0.78	0.65	0.65

B. Discussion the results:

Coh-Metrix is a tool that analyze text in terms of cohesion and readability. The results of the cohesion features examined in this study, (RC, SM, TE, and Read*), provide signals to understand the linguistic features of cohesion for schizophrenia patients. Specifically, the mean values of SM indices shows that writing and speech produced by control group are more coherent than the patients. This feature also performs the best in predicting the class of the groups in LabWriting dataset where the F-score was improved by 6% comparing to baseline model (using text only). The following text from *LabWriting* dataset was an answer written by a patient to describe his/her Sunday (it has been paraphrased to preserve anonymity).

> *My Sunday is chilly My sunday is harsh My sunday is dead My sunday is soulless*

The overall cohesion of this text is low which is confirmed by the low scores of text easability indices. For instance, the text does not contain enough verbs, that can convey actions and thoughts, or pronouns which effects the narrativity. In addition, the word concreteness score is low because the numbers of concrete words that refer to things a one can see, hear, or feel e.g. (*mask, forest*) is very low or not even used at all. Finally, the text is poor in terms of deep cohesion since it does not have enough connectives that that can help to tie the events, ideas and information in the text together and make it easier for the reader. To sum up, low cohesion scores might make it more difficult to fully understand the text.

Based on the findings of this study, this tool can provide an accurate assessment on over 200 measures of cohesion which can be used to analyze and better understand the text written by patients with schizophrenia.

VI. CONCLUSION

Patients with schizophrenia have different cognitive symptoms, some of which involve problems with concentration and memory, which in return may lead to disorganization in speech or behavior. Diagnosing this disorder early and correctly is extremely important as it may help alleviate the negative effects on its patients. Even though, previous works have investigated language impairments and speech disorder in people with schizophrenia disorder, availability of recordings of spoken language as well as writings provides an opportunity to systematically analyze the language use by patients.

Among the linguistic features of cohesion which were investigated for this study, we found that a combination of features such as referential cohesion, text easability, and situational model features provide the biggest boost in classification performance for LabSpeech dataset. For LabWriting dataset, readability and situation model for SVM performs the best performance, and a combination of features such as RC and Read* for RF have the best performance.

In the future, we will explore other features of cohesion such as connectives, which create cohesive connections between ideas and clauses and show how the text is organized. We also plan to collect more data from social media such as Reddit for a similar analysis in this study. Finally, we plan to expand our analysis to other related mental health disorders.

ACKNOWLEDGMENT

We would like to thank the Coh-Metrix team for granting access to the tool and providing valuable support to us on using it for the analyses performed in this study. We also would like to thank Michael Compton for granting access to Writing and Speech datasets.

REFERENCES

[1] E. Palmer, J. Gilleen, and D. Strelchuk, "Insight into schizophrenia and its relationship with clinical symptoms:: a meta-analysis involving 20 515 patients," in *Schizophrenia Bulletin*, vol. 43, no. 1, 2017, p. S98.

[2] M. F. Green, W. P. Horan, and J. Lee, "Social cognition in schizophrenia," *Nature Reviews Neuroscience*, vol. 16, no. 10, p. 620, 2015.

[3] A. E. Goldman, "A comparative-developmental approach to schizophrenia," in *Schizophrenia*. Routledge, 2017, pp. 106–127.

[4] S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang *et al.*, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, p. 421, 2014.

[5] National Institute of Mental Health, Office of Behavioral and Social Sciences Research. (2016) Computer scientists learn the language patterns of psychiatric disorders. Accessed 18 July 2019. [Online]. Available: https://obssr.od.nih.gov/computer-scientists-learn-the-language-patterns-of-psychiatric-disorders/

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.

[7] T. Nalabandian and M. Ireland, "Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or friends," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 62–73.

[8] C. Howes, M. Purver, and R. McCabe, "Linguistic indicators of severity and progress in online text-based therapy for depression," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, 2014.

[9] E. Hohman, D. Marchette, and G. Coppersmith, "Mental health, economics, and population in social media," in *Proceedings of the Joint Statistical Meetings*, 2014.

[10] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language." *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[11] R.-P. Filiou, N. Bier, A. Slegers, B. Houzé, P. Belchior, and S. M. Brambati, "Connected speech assessment in the early detection of alzheimers disease and mild cognitive impairment: a scoping review," *Aphasiology*, pp. 1–33, 2019.

[12] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimers disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[13] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimers disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[14] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in twitter," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[15] J. Zomick, S. I. Levitan, and M. Serper, "Linguistic analysis of schizophrenia in reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 74–83.

[16] E. S. Kayi, M. Diab, L. Pauselli, M. Compton, and G. Coppersmith, "Predictive linguistic features of schizophrenia," *arXiv preprint arXiv:1810.09377*, 2018.

[17] M. Mitchell, K. Hollingshead, and G. Coppersmith, "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*, 2015, pp. 11–20.

[18] S. Deutsch-Link, "Language in schizophrenia: What we can learn from quantitative text analysis," 2016.

[19] K. S. Minor, K. A. Bonfils, L. Luther, R. L. Firmin, M. Kukla, V. R. MacLain, B. Buck, P. H. Lysaker, and M. P. Salyers, "Lexical analysis in schizophrenia: how emotion and social word use informs our understanding of clinical presentation," *Journal of psychiatric research*, vol. 64, pp. 74–78, 2015.

[20] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[21] T. Gupta, S. J. Hespos, W. S. Horton, and V. A. Mittal, "Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis," *Schizophrenia research*, vol. 192, pp. 82–88, 2018.

[22] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior research methods, instruments, & computers*, vol. 36, no. 2, pp. 193–202, 2004.

[23] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *Interspeech 2016*, 2016.

[24] A. M. Kring and J. M. Caponigro, "Emotion in schizophrenia: where feeling meets thinking," *Current directions in psychological science*, vol. 19, no. 4, pp. 255–259, 2010.

[25] N. C. Andreasen, "Negative symptoms in schizophrenia: definition and reliability," *Archives of general psychiatry*, vol. 39, no. 7, pp. 784–788, 1982.

[26] G. R. Kuperberg, "Language in schizophrenia part 1: an introduction," *Language and linguistics compass*, vol. 4, no. 8, pp. 576–589, 2010.

[27] K. Hong, A. Nenkova, M. E. March, A. P. Parker, R. Verma, and C. G. Kohler, "Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls," *Psychiatry research*, vol. 225, no. 1-2, pp. 40–49, 2015.

[28] J. Phillips, "Schizophrenia and the narrative self," *The self in neuroscience and psychiatry*, pp. 319–335, 2003.

[29] D. S. McNamara, Y. Ozuru, A. C. Graesser, and M. Louwerse, "Validating coh-metrix," in *Proceedings of the 28th annual conference of the cognitive science society*, 2006, pp. 573–578.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.