

Learning to Detect Bipolar Disorder and Borderline Personality Disorder with Language and Speech in Non-Clinical Interviews

Bo Wang^{1,3}, *Yue Wu*^{2,3}, *Niall Taylor*¹, *Terry Lyons*^{2,3}
Maria Liakata^{3,4}, *Alejo J Nevado-Holgado*¹, *Kate E A Saunders*¹

¹Department of Psychiatry, University of Oxford, UK

²Mathematical Institute, University of Oxford, UK

³The Alan Turing Institute, London, UK

⁴Department of Computer Science, University of Warwick, UK

bo.wang@psych.ox.ac.uk

Abstract

Bipolar disorder (BD) and borderline personality disorder (BPD) are both chronic psychiatric disorders. However, their overlapping symptoms and common comorbidity make it challenging for the clinicians to distinguish the two conditions on the basis of a clinical interview. In this work, we first present a new multi-modal dataset containing interviews involving individuals with BD or BPD being interviewed about a non-clinical topic. We investigate the automatic detection of the two conditions, and demonstrate a good linear classifier that can be learnt using a down-selected set of features from the different aspects of the interviews and a novel approach of summarising these features. Finally, we find that different sets of features characterise BD and BPD, thus providing insights into the difference between the automatic screening of the two conditions.

Index Terms: bipolar disorder, borderline personality disorder, spoken dialogue, computational paralinguistics, path signature

1. Introduction

Bipolar disorder (BD) is a recurrent chronic mental health condition which occurs in approximately 1% of the global population [1]. It is characterised by episodes of low and high mood which cause significant interference with everyday life. Borderline personality disorder (BPD) is characterised by a long-term pattern of constantly variable mood, self-image and behaviour. Although BD and BPD are two very different conditions they share some similar symptoms such as mood instability and impulsive behaviour [2]. A recent study [3] reported the high prevalence of comorbidity between the two conditions, with up to 21.6% of individuals with BD found to have comorbid BPD. As a result they can be difficult to distinguish, but accurate diagnosis is crucial as they require different treatment [4, 5]. Standard diagnostic assessment involves a psychiatrist asking a series of questions about symptoms and the person has to retrospectively describe their account of these symptoms. The success of the assessment also relies on how the psychiatrist interprets both the verbal and non-verbal cues drawn from the person’s responses. In this work, we aim to develop a method that extracts cues automatically from interviews conducted in a non-clinical setting, to assist the existing assessment framework, which is expensive and subjective.

Recent studies have explored data driven approaches to automatically screen patients, incorporating features extracted from multiple modalities in clinical interviews, showing diagnostic value for mental health conditions such as depression and bipolar disorder [6, 7, 8, 9]. [8] finds the performance of

automatic mood detection to be much better in clinical interactions than in personal conversations, and there are significant differences in the features important to each type of interaction. While existing studies of BD have focused on recognising mood episodes, the distinction between BD and BPD remains understudied. In this paper, we aim to bridge this gap by presenting a multi-modal (i.e. speech and text) dataset containing interviews in a non-clinical setting involving individuals with a diagnosis of BD or BPD, and study the automatic assessment of the two mental health conditions.

Motivated to study the interaction between the interviewer and participant during the course of an interview from different aspects (including linguistic complexity, semantic content and dialogue structure), we investigate features extracted from different modalities. Path signatures, initially introduced in rough path theory as a branch of stochastic analysis, has been shown to be successful in a range of machine learning tasks involving modelling temporal dynamics [10, 11, 12]. We propose to apply path signatures for summarising features extracted from each utterance, sentence and speaker-turn into interview-level feature representations, given its ability to naturally capture the order of events. By doing so, we automatically include more non-linear prior knowledge in our final feature set, which leads to effective classification, even with a simple linear classifier.

The contributions of this work are as follows: (1) We present a new non-clinical interview dataset involving BD and BPD patients; (2), We investigate different feature types and propose using path signatures as a novel approach of summarising turn-level features; (3) We demonstrate a good linear model can be learnt for three classification tasks, and provide insights into the distinction between BD and BPD by analysing the importance of the selected features.

2. AMoSS Interview Dataset

The original Automated Monitoring of Symptoms Severity (AMoSS) study [13, 10] was a longitudinal study during which a range of wearables in combination with a bespoke smartphone app were used for the daily self-monitoring of mood instability. Among the 139 participants enrolled in the study, 53 had a BD diagnosis, 33 had been diagnosed with BPD and 53 were healthy volunteers. All the diagnoses had been confirmed prior to the study using the structured clinical interview for DSM-IV (the 4th edition of Diagnostic and Statistical Manual of Mental Disorders) and the International Personality Disorder Examination (IPDE) [14]. The majority of the BD participants were euthymic while BPD participants were not in crisis but were still

symptomatic as is the case with chronic experience of the condition. Exclusion criteria for BD and BPD were comorbidity of each diagnosis [15].

62 participants were interviewed halfway through the study to gather qualitative feedback and discuss potential improvement¹. Each participant was interviewed only once. These semi-structured, one-on-one qualitative interviews took place either in person or by telephone, conducted by 2 clinicians and 2 psychology graduates who were involved in the roll out of the AMoSS study. The natural conversations recorded between the interviewer and participant are usually within the scope of: 1) experience using the mood reporting app and the questionnaires in the app; 2) experience using different wearable devices; 3) benefits of taking part in the study and discussion of potential improvement, making the interviews semi-structured.

The AMoSS interview (AMoSS-I) dataset we study here consists of 50 randomly sampled interviews that were initially transcribed by the same interviewers. The audio recordings and manually transcribed text were then aligned by a Sakoe-Chiba Band Dynamic Time Warping based forced speech alignment model in *aeneas*². This was followed by a convolutional neural network (CNN) based noise-robust speech segmentation [16], generating speaker-turn-level alignments in the time domain. The manual transcripts and automatically generated time alignments were then reviewed and improved by three research assistants, where each interview was reviewed twice to ensure quality and consistency. We used *finetuneas*³ as the annotation interface for reviewing. The demographic details of the participants are summarised in Table 1. We can see both BIS-11⁴ and IPDE scores are higher among BD and BPD patients compared to controls. We also observe from the density plots in Figure 1 that distributions of the three user groups are very similar in interview length and number of participant responses.

	BD	BPD	HC
#Interviews	21	17	12
#Room Interviews	14	9	9
#Phone Interviews	7	8	3
Gender (m:f)	7:14	1:16	3:9
Age (years)	44 ± 17	34 ± 21	34 ± 18.5
BIS-11 score	67 ± 16	76 ± 22	48.5 ± 10.25
IPDE score	2 ± 5	16 ± 3	0 ± 0

Table 1: *Demographic characteristics of the three groups: Bipolar disorder (BD), Borderline personality disorder (BPD) and Healthy controls (HC). Appropriate distributions are summarised in the form of the median +/- the interquartile range.*

2.1. Data Preprocessing

Some of the interview recordings show noticeable difference of loudness between two speakers, mainly due to the mode of recording. In order to alleviate the effect of loudness difference, we scale the audio signal for each speaker turn separately,

¹The study protocol was approved by the NRES Committee East of England—Norfolk (13/EE/0288), and all 62 participants had given consent for further analysis on the qualitative interviews.

²<https://github.com/readbeyond/aeneas>

³<https://github.com/ozdefir/finetuneas>

⁴The Barratt Impulsiveness Scale (BIS-11) [17] is a self-report questionnaire designed to assess the personality/behavioral construct of impulsiveness. Higher BIS-11 scores are indicative of higher impulsivity.

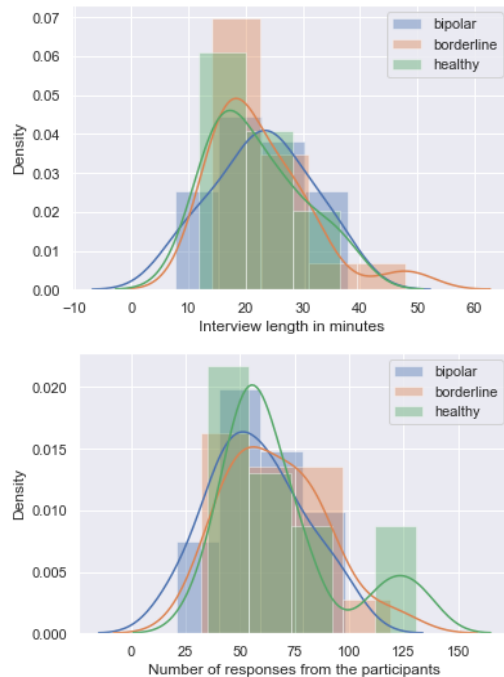


Figure 1: *Density plots of: length of the interview (top), and number of responses from the participants (bottom)*

and make sure each turn is in the range of -1 and 1. We also perform intra-speaker-turn segmentation using an end-to-end voice activity detection (VAD) model [18] trained on the DIHARD Speech Diarisation data [19]. The model extracts domain-independent features during its domain-adversarial multi-task training on DIHARD, showing better performance over the standard VAD models that do not use such domain information. This in turn allows us to extract dialogue features related to pauses in speech appearing within each speaker turn.

3. Feature Extraction

We identify a set of features that are motivated by existing work in automatic mental health assessment and clinical studies of bipolar disorder symptoms. These features are selected to allow studying the interviews from different aspects: lexical diversity and density, syntax, semantic content and dialogue structure.

Linguistic complexity features (LING): Previous studies have shown language disturbances such as pressure of speech in mania and poverty of speech in depression are among the main symptoms of acute episodes in BD [20, 21]. We adopt a set of linguistic complexity measures used in [9], including measurements for lexical diversity such as *moving average type-to-token ratio* (MATTR), *brunet’s index* (BI) and *honore’s statistic* (HS); lexical density such as *number of function words per word* (FUNC/W) and *number of interjections per word* (UH/W); and *mean length sentence* (MLS). These features are shown to be effective in distinguishing patients with schizophrenia and bipolar I disorder. We also use the dependency-based propositional idea density (DEPID), originally proposed in [22] for measuring the rate of propositions or ideas expressed per word in spontaneous speech transcripts.

In addition to the aforementioned measures we extract part-of-speech related variables including: *first person pronouns*

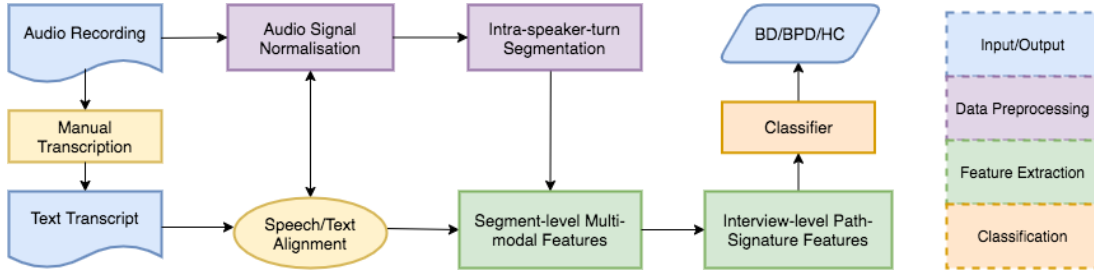


Figure 2: Data pipeline including three stages of data preprocessing (in purple), feature extraction (in green) and classification (in orange).

(it has been shown that people suffering from depression use more first person pronouns in written [23] and spoken language [24]); *swear words*; *speech disfluencies*; *filler words/phrases* such as “okay” and “you know”, using the LIWC-2015 dictionary [25]. Finally, we add the *number of absolutist words per word* (ABS/W) as a feature using a 19-word absolutist dictionary curated by clinical psychologists [26], who found that the online anxiety, depression, and suicidal ideation forums contained more absolutist words than the control ones. Overall we extract 28 linguistic features, denoted as **LING**.

Semantic content features (CNT): We extract content related to psychological states from the transcripts by applying relevant categories of LIWC [27] such as emotions (e.g. anxiety), social processes (e.g. family) and drives (e.g. reward and risk). We also apply the empathetic concern and personal distress lexica [28], and an optimism lexicon (i.e. positive future-oriented thinking), by taking the sum over all weighted words multiplying their relative frequencies in the speaker turn. Overall we use 19 content features.

Dialogue features (DIAL): We use the set of high-level turn-taking behaviour related features proposed in [7], namely *relative floor control*, *turn hold offset* (i.e. short pauses that are less than half a second), *number of consecutive turns* (separated by longer than half a second pauses), *turn switch offset* and *turn length*, per speaker turn. *Turn switch offset* measures the latency between speaker turns, and it is shown that depressed people take more time to respond to questions by clinicians [29]. While *Relative floor control* measures the percentage of time an individual controlled the conversation floor up to the time of speaking, we also add *relative turn length* that measures the percentage of the length of the current turn relative to the average turn length up to the time of speaking. We compute both features in seconds as well as in number of words, following [30].

Talking over or interrupting others, is one of the characteristics for pressure speech, which has shown to be a significant feature in bipolar mania [31, 32]. We use a LSTM-based overlapping speech detection model proposed in [33] to extract the number and duration of *speech overlaps* in each speaker turn as features. Additionally, we add the *number of words per second* per turn, as a feature representing speaking speed.

We average *turn hold offset* and *speech overlaps* per turn, in addition to *number of consecutive turns*, *relative floor control (time)*, *relative floor control (words)*, *relative turn length (time)*, *relative turn length (words)*, *turn switch offset*, *turn length* and *number of words per second*, which results in 11 features representing each speaker turn.

3.1. Interview-level Feature Representation

The theory of rough paths, developed by Lyons [34], can be thought of as a non-linear extension of the classical theory of controlled differential equations. The signature of a path⁵ (i.e. an ordered data stream) is a collection of n -fold iterated integrals such that every continuous function of the path may be approximated arbitrarily well by a linear function of its signature. Motivated by its ability to naturally capture the order of events and model temporal dynamics, we apply signature transform (SIG), which is the map from a path to its signature, to each type of the turn-level features and generate interview-level fixed-length feature representation.

4. Experiments and Analysis

Following previous work we chose leave-one-participant-out (LOOCV) as the evaluation scheme, and logistic regression as the classification model given our preference over interpretability and the size of our data. For each fold, we first apply signature transform to each type of turn-level features, and keep only the first three levels of the path signature⁶. To avoid overfitting, we conduct feature selection on signature-transformed interview-level features through computing Pearson Correlation Coefficients (PCC) with the IPDE scores on the training data and retain the features with p -values less than 0.001. This results in a small number of features. The selected features are then fed to the classifier for 3 separate binary tasks: (1), BD vs. healthy controls, (2), BPD vs. healthy controls, and (3), BD vs. BPD patients. We conduct three separate experiments, extracting features from the speech of each participant and interviewer respectively, as well as the whole interview (as a sequence of turns) without speaker identification (denoted as ‘Both’)⁷.

4.1. Analysis of the selected features

Five highly ranked and most commonly selected features from each task are briefly summarised in Table 2 as examples. Each interview-level feature is represented as a linear combination of the original turn-level features. We see over half of the selected features are volume integrals, i.e. they are triple integrals of three turn-level features, while the rest are double integrals

⁵We refer the reader to [35] for a rigorous introduction of path signatures, and [36] for a primer on its use in machine learning.

⁶We use `iisignature` Python library, <https://pypi.org/project/iisignature/>, and set the maximal order to which iterated integration is performed in signature to be 3.

⁷For the ‘Interviewer’ and ‘Both’ experiments, we increase the p value threshold to 0.002.

which give the *Levy area*. We notice most all of the representative features are from the linguistic category (many are part-of-speech tags), especially for *H vs. BD* and *H vs. BPD*, showing the significance of the structure in the interviews. This is inline with the finding in [8], and extends to nonclinical interviews conducted partially by students.

We also notice the appearances of nonfluencies (Nonflu.) especially in combination with conjunctions (CONJ) in the detection of *H vs. BPD*. The use of the absolute words (ABS) in combination with common adverbs (ADV) and negations (NEG) or article words (e.g. a, an, the) are selected for the two classification tasks involving BPD. Two interview-level features (*DEPID*, *MATTR*, *BI*) and (*BI*, *MATTR*, *MLS*) are shown to be the most commonly selected for the two tasks involving BD.

H vs BD	H vs BPD	BD vs BPD
(<i>DEPID</i> , <i>MATTR</i> , <i>BI</i>)	(<i>Nonflu.</i> , <i>CONJ</i>)	(<i>BI</i> , <i>MATTR</i> , <i>MLS</i>)
(<i>Nonflu.</i> , <i>Verbs</i>)	(<i>ABS</i> , <i>ADV</i> , <i>Articles</i>)	(<i>We</i> , <i>PREP</i>)
(<i>PPRO</i> , <i>CONJ</i> , <i>CONJ</i>)	(<i>WPS</i> , <i>SP_avg</i> , <i>RFC.t</i>)	(<i>PREP</i> , <i>We</i>)
(<i>NEG</i> , <i>AUXV</i> , <i>NEG</i>)	(<i>CONJ</i> , <i>Nonflu.</i>)	(<i>ABS</i> , <i>ADV</i> , <i>NEG</i>)
(<i>PPRO</i> , <i>Swear</i> , <i>Verbs</i>)	(<i>You</i> , <i>Verbs</i> , <i>Nonflu.</i>)	(<i>SOC</i> , <i>DRI</i> , <i>DRI</i>)

Table 2: Top-5 highly ranked and most commonly selected features during LOOCV. Feature selection is based on correlation between each feature and the IPDE scores of the training samples. Features belong to the linguistic category are colored in blue; dialogue features are in green and content features are in red.⁸

4.2. Results and Discussion

The results for the classification tasks are summarised in Table 3. Using the (late) fusion of linguistic, dialogue and content features with signature transform, we obtain a AUROC of 0.810 in H/BD, 0.733 in H/BPD and 0.817 in BD/BPD. We notice the result in H/BPD is significantly worse than the other two tasks. We leave its investigation to future work, we think the fewer data samples and varying recording quality may have contributed to the worse performance. When we model from the speaking segments of the interviewers, we obtain very poor performance. As the purpose of the interviews were merely to understand the individual’s experience of taking part in the AMoSS study rather than establishing their mental state at the time of interview, it is no surprise that features extracted from the interviewers have very weak discriminative power. We also believe having different interviewers impacted negatively on the classifications. Modelling the interviews as a sequence of utterances also resulted in much worse performance than learning from the participants alone.

We also conduct ablation experiments to examine how performance changes after removing each feature type. As seen in Table 4, the linguistic features (LING) are the biggest contribution in all three tasks. As a consequence, we have to increase the p-value feature selection threshold from 0.001 to 0.005 to have any feature for classification, if we remove LING. A sharp performance drop is then observed removing LING in in all three tasks. If we remove both LING and dialogue features (DIAL),

⁸PPRO: personal pronouns; PREP: prepositions; ADV: adverbs; AUXV: auxiliary verbs; CONJ: conjunctions; NEG: negations; Nonflu.: nonfluencies; We: first-person plural; Swear: swear words; MLS: mean length sentence; WPS: number of words per second; SP_avg: average length of short pauses; RFC.t: relative floor control (time); SOC: social processes; DRI: drives.

Subject	AUROC		
	H/BD	H/BPD	BD/BPD
Participant	0.810	0.733	0.817
Interviewer	0.304	0.473	0.231
Both	0.494	0.431	0.657

Table 3: Classification results for three binary tasks: *H vs. BD*, *H vs. BPD* and *BD vs. BPD*, using logistic regression. Results shown are average AUROCs across all interviews.

the results get even worse. If we exclude both DIAL and content features (CNT), we still get reasonably good performance without significant drop in AUROC. It’s also worth noticing the ineffectiveness of the content features (CNT) apart from in BD vs. BPD, possibly due to the semi-structured nature of the interviews and the questions asked fall within the same scope. Given that the majority of BD and BPD participants were clinically stable this may also account for the relatively poor distinction between the groups using CNT features that are related to the psychological states from responses by the participants.

Features	H vs BD	H vs BPD	BD vs BPD
All	0.810**	0.733**	0.817**
All-CNT	0.810**	0.733**	0.787**
All-DIAL	0.768**	0.733**	0.811**
All-LING	0.625*	0.578*	0.669*
All-LING-CNT	0.642*	0.703*	0.604*
All-LING-DIAL	0.442*	0.429*	0.550*
All-CNT-DIAL	0.768**	0.733**	0.763**

Table 4: Feature ablation results (AUROC) for each task, e.g. the final row in the table shows the result for using linguistic features (LING) only. p-value used for feature selection: **<0.001; *<0.005.

5. Conclusions

In this paper, we demonstrate the potential of using features extracted from language and speech in non-clinical interviews to assist the assessment of bipolar disorder BD and borderline personality disorder BPD, which is challenging for clinicians to distinguish. We first presented a non-clinical interview dataset, named AMoSS-I, conducted partially by psychology graduates, for the task of detecting BD and BPD. We demonstrated good performance in three classification tasks using down-selected features and a new way of summarising these features based on path signatures. Lastly, we showed the importance of linguistic features in all three tasks and the benefits of feature fusion from different modalities. For future work, we plan to learn acoustic features, and investigate the effect of acoustic properties of the interviews and the impact of recording environments.

6. Acknowledgements

This work was supported by the MRC Mental Health Data Pathfinder award to the University of Oxford [MC_PC_17215], by the NIHR Oxford Health Biomedical Research Centre and by the The Alan Turing Institute under the EPSRC grant EP/N510129/1. We would also like to thank Zakaria Aldeneh and Kairit Sirts for sharing code and thoughts, and to Priyanka Panchal and Rota Silva for conducting the interviews. The views expressed are those of the authors and not necessarily those of the NHS, NIHR or the Department of Health.

7. References

- [1] I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," *The Lancet*, vol. 387, no. 10027, pp. 1561–1572, 2016.
- [2] C. J. Ruggero, M. Zimmerman, I. Chelminski, and D. Young, "Borderline personality disorder and the misdiagnosis of bipolar disorder," *Journal of psychiatric research*, 2010.
- [3] M. Fornaro, L. Orsolini, S. Marini, D. De Berardis, G. Perna, A. Valchera, L. Ganança, M. Solmi, N. Veronese, and B. Stubbs, "The prevalence and predictors of bipolar and borderline personality disorders comorbidity: systematic review and meta-analysis," *Journal of Affective Disorders*.
- [4] N. C. C. for Mental Health (UK *et al.*, "Borderline personality disorder: treatment and management." British Psychological Society, 2009.
- [5] N. C. C. for Mental Health *et al.*, "Bipolar disorder: the assessment and management of bipolar disorder in adults, children and young people in primary and secondary care," *Leicester UK: The British Psychological Society*, 2014.
- [6] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *19th Annual Conference of the International Speech Communication Association, Interspeech 2018*, vol. 2522, 2018, pp. 1716–1720.
- [7] Z. Aldeneh, M. Jaiswal, M. Picheny, M. McInnis, and E. M. Provost, "Identifying mood episodes using dialogue features from clinical interviews," *arXiv preprint arXiv:1910.05115*, 2019.
- [8] K. Matton, M. G. McInnis, and E. M. Provost, "Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder," *Proc. Interspeech 2019*, pp. 1438–1442, 2019.
- [9] R. Voleti, S. Woolridge, J. M. Liss, M. Milanovic, C. R. Bowie, and V. Berisha, "Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder," in *Proc. Interspeech*, 2019.
- [10] I. P. Arribas, G. M. Goodwin, J. R. Geddes, T. Lyons, and K. E. Saunders, "A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder," *Translational psychiatry*, vol. 8, no. 1, pp. 1–7, 2018.
- [11] B. Wang, M. Liakata, H. Ni, T. Lyons, A. J. Nevado-Holgado, and K. Saunders, "A path signature approach for speech emotion recognition," in *20th Annual Conference of the International Speech Communication Association, Interspeech*. ISCA, 2019, pp. 1661–1665.
- [12] P. Kidger, P. Bonnier, I. P. Arribas, C. Salvi, and T. Lyons, "Deep signature transforms," in *Advances in Neural Information Processing Systems*, 2019, pp. 3099–3109.
- [13] A. Tsanas, K. Saunders, A. Bilderbeck, N. Palmius, M. Osipov, G. Clifford, G. Goodwin, and M. De Vos, "Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder," *Journal of affective disorders*.
- [14] A. W. Loranger, N. Sartorius, A. Andreoli, P. Berger, P. Buchheim, S. Channabasavanna, B. Coid, A. Dahl, R. F. Diekstra, B. Ferguson *et al.*, "The international personality disorder examination: The world health organization/alcohol, drug abuse, and mental health administration international pilot study of personality disorders," *Archives of General Psychiatry*, 1994.
- [15] N. M. McGowan, G. M. Goodwin, A. C. Bilderbeck, and K. E. Saunders, "Circadian rest-activity patterns in bipolar disorder and borderline personality disorder," *Translational psychiatry*, 2019.
- [16] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [17] J. H. Patton, M. S. Stanford, and E. S. Barratt, "Factor structure of the barratt impulsiveness scale," *Journal of clinical psychology*, vol. 51, no. 6, pp. 768–774, 1995.
- [18] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-end Domain-Adversarial Voice Activity Detection," 2020. [Online]. Available: <https://arxiv.org/abs/1910.10655>
- [19] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.
- [20] F. K. Goodwin and K. R. Jamison, *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford University Press, 2007, vol. 1.
- [21] L. Weiner, N. Doignon-Camus, G. Bertschy, and A. Giersch, "Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.
- [22] K. Sirts, O. Pigué, and M. Johnson, "Idea density for predicting alzheimer's disease from transcribed speech," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017, pp. 322–332.
- [23] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.
- [24] J. Zimmermann, T. Brockmeyer, M. Hunn, H. Schauenburg, and M. Wolf, "First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients," *Clinical psychology & psychotherapy*, vol. 24, no. 2, pp. 384–391, 2017.
- [25] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," *Tech. Rep.*, 2015.
- [26] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.
- [27] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, 2010.
- [28] J. Sedoc, S. Buechel, Y. Nachmany, A. Buffone, and L. Ungar, "Learning word ratings for empathy and distress from document-level user responses," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 1664–1673.
- [29] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, Semdial*, 2013, pp. 160–169.
- [30] T. Meshorer and P. A. Heeman, "Using past speaker behavior to better predict turn transitions," in *Proc. Interspeech*, 2016.
- [31] N. C. Andreasen, "Thought, language, and communication disorders: I. clinical assessment, definition of terms, and evaluation of their reliability," *Archives of general Psychiatry*, vol. 36, no. 12, pp. 1315–1321, 1979.
- [32] —, "Thought, language, and communication disorders: II. diagnostic significance," *Archives of general Psychiatry*, vol. 36, no. 12, pp. 1325–1330, 1979.
- [33] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- [34] T. J. Lyons, "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.
- [35] T. Lyons, "Rough paths, signatures and the modelling of functions on streams," in *Proceedings of the International Congress of Mathematicians*, 2014, pp. 163–184.
- [36] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," *arXiv preprint arXiv:1603.03788*, 2016.