UTRECHT UNIVERSITY

METHODOLOGY AND STATISTICS
FOR THE BEHAVIOURAL, BIOMEDICAL AND SOCIAL SCIENCES

MASTER'S THESIS PROPOSAL

Unblackboxing BERT for Text-Mining: Case of Mental Disorder Prediction

---

Student:
Daniel ANADRIA,
2654814

Supervisor:
Dr. Ayoub BAGHERI

November 23, 2022

Word Count: 735

Candidate Journal: Journal of Artificial Intelligence Research

Introduction

Recent studies have used supervised learning to classify individuals into various mental disorder conditions using social media text [7], [29], [30], [15], [12]. However, few have tried to identify which features are driving the prediction of specific disorders [12], [30]. While highly complex algorithms, such as *BERT* (*Bidirectional Encoder Representations from Transformers* [9]), tend to have state-of-the-art performance, this comes at the expense of model interpretability [26], [28]. Such high-performing uninterpretable models are called black-boxes. Unblackboxing BERT in terms of *feature attributions*, measures of individual predictor importance, remains a challenge [1].

Recent developments in unblackboxing BERT include *local* and *global explanations*. Local explanations examine the features important for the outcome of a specific observation in the dataset [26]. An example of a local attribution for BERT is the *TransSHAP* method [18]. Global explanations rank features according to their importance for the performance of the entire method [26]. An example of a global attribution for BERT is the *LRP (AH+LN)* method [1]. Both methods provide explanations that are well-aligned with human expert opinion on some tasks [36], [6], [25], but this has not been investigated on metal disorder prediction. The difference between the two proposed methods is that LRP (AH+LN) looks directly into the black-box relying on gradient information, and TrasSHAP builds a surrogate white-box model that mimics BERT. The similarity is that both LRP (AH+LN) and TransSHAP are based on the methods (Layer-wise Relevance Propagation [4] and Shapley Additive Explanations [20], respectively) which can be adjusted to provide both local and global level explanations.

The present work identifies and addresses several gaps in the literature:

1. To our knowledge, no one has unblackboxed BERT on a mental disorder prediction task. Dinu and Moldovan [10] used BERT for mental disorder prediction and computed feature attributions, however these attributions were not extracted from BERT but from a Naive Bayes Classifier. Therefore, explaining BERT on this specific task remains relevant.
2. Being new techniques, TransSHAP and LRP (AH+LN) have never been adjusted to both levels of explanation, nor directly compared.
3. No one investigated the benefits of combining TransSHAP and LRP (AH+LN) into a single attribution measure suitable for BERT. The present work aims to develop a new attribution method which does this.

The research questions are:

1. Can a combined method, *ExBERT* (*Explainable BERT*), extract feature attributions for both local and global explanations? If so, what are the most important local and global attributions for mental disorder prediction?
2. Can ExBERT local and global explanations outperform LRP (AH+LN) and TransSHAP local and global explanations in the case of mental disorder prediction?

Analytic Strategy

The *Self-reported Mental Health Diagnoses Dataset* [7] is a large collection of Reddit posts from users with one or more of nine different diagnoses and healthy controls. The posts were collected between 2006 and 2017. A subset of users with depression (14,139) and their matched healthy controls will be used for classification of users into depressed and healthy groups. A limitation is that the diagnoses are based on self-report. The project has received an ethical approval (*FETC Registration 22-1869*).

In order to create ExBERT, we will identify several candidate models based on different combinations of TransSHAP and LRP (AH+LN) formulas and compare model performances on local and global levels (Table 2.1).

TABLE 2.1: The Planned Feature Attribution Comparisons

Explanation	Comparison
Local	TransSHAP vs. LRP (AH+LN) vs. ExBERT
Global	TransSHAP vs. LRP (AH+LN) vs. ExBERT

In the local explanation, we focus on the four observations for which BERT had highest classification probabilities (Table 2.2), based on Saarela and Jauhiainen [26]. A selection of sentences which showcase words with high attributions will be shown to demonstrate the similarities and differences between TransSHAP, LRP (AH+LN), and ExBERT explanations.

TABLE 2.2: The Four Observations for Local Feature Attribution Comparisons

	True Depressed	True Healthy
Predicted Depressed	True Positive	False Positive
Predicted Healthy	False Negative	True Negative

In the global condition, correlation between the three attribution methods will be computed, similar to Wu and Ong [35].

The validity of local and global explanations of the three methods will be evaluated following the methodology of Ullah et al. [34]. For each method and level of explanation, a subset of most influential features will be selected and used as input for training simple classifiers. The performance of these classifiers will then be examined in terms of accuracy, precision, and F1-Score. If the classifiers with only the subset of most important features perform equally well or better than BERT with all the features, the noise has been successfully removed from the data. Then the feature subset truly drives the prediction of depression and generalizes well across algorithms.

The analysis relies on: `Python 3`, `R 4.2` (`ggplot2`), `Git`, and `SURFdrive`.

References and Additional Readings (*)

[1] Ameen Ali et al. *XAI for Transformers: Better Explanations through Conservative Propagation*. arXiv:2202.07304 [cs]. June 2022. DOI: 10.48550/arXiv.2202.07304. URL: http://arxiv.org/abs/2202.07304 (visited on 10/01/2022).

[2] *Marco Ancona et al. *Towards better understanding of gradient-based attribution methods for Deep Neural Networks*. arXiv:1711.06104 [cs, stat]. Mar. 2018. DOI: 10.48550/arXiv.1711.06104. URL: http://arxiv.org/abs/1711.06104 (visited on 10/05/2022).

[3] *Plamen P. Angelov et al. "Explainable artificial intelligence: an analytical review". en. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.5 (2021), p. 13. DOI: 10.1002/widm.1424. URL: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1424 (visited on 09/06/2022).

[4] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". en. In: *PLOS ONE* 10.7 (July 2015). Publisher: Public Library of Science, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140 (visited on 09/09/2022).

[5] *Anindya Bhaduri et al. "On the usefulness of gradient information in surrogate modeling: Application to uncertainty propagation in composite material models". en. In: *Probabilistic Engineering Mechanics* 60 (Apr. 2020), p. 103024. ISSN: 0266-8920. DOI: 10.1016/j.probengmech.2020.103024. URL: https://www.sciencedirect.com/science/article/pii/S0266892020300096 (visited on 10/02/2022).

[6] Moritz Böhle et al. "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification". In: *Frontiers in Aging Neuroscience* 11 (2019). ISSN: 1663-4365. URL: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00194 (visited on 09/29/2022).

[7] Arman Cohan et al. *SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions*. arXiv:1806.05258 [cs]. July 2018. DOI: 10.48550/arXiv.1806.05258. URL: http://arxiv.org/abs/1806.05258 (visited on 09/08/2022).

[8] *John Danaher. "The Threat of Algocracy: Reality, Resistance and Accommodation". en. In: *Philosophy & Technology* 29.3 (Sept. 2016), pp. 245–268. ISSN: 2210-5441. DOI: 10.1007/s13347-015-0211-1. URL: https://doi.org/10.1007/s13347-015-0211-1 (visited on 09/23/2022).

[9] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs] version: 2. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 09/14/2022).

[10] Anca Dinu and Andreea-Codrina Moldovan. "Automatic Detection and Classification of Mental Illnesses from General Social Media Texts". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sept. 2021, pp. 358–366. URL: https://aclanthology.org/2021.ranlp-1.41 (visited on 09/25/2022).

[11] *Santiago González-Carvajal and Eduardo C. Garrido-Merchán. *Comparing BERT against traditional machine learning text classification*. arXiv:2005.13012 [cs, stat]. Jan. 2021. DOI: 10.48550/arXiv.2005.13012. URL: http://arxiv.org/abs/2005.13012 (visited on 09/07/2022).

[12] Xiaobo Guo, Yaojia Sun, and Soroush Vosoughi. "Emotion-based Modeling of Mental Disorders on Social Media". In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. WI-IAT '21. New York, NY, USA: Association for Computing Machinery, Dec. 2021, pp. 8–16. ISBN: 978-1-4503-9115-3. DOI: 10.1145/3486622.3493916. URL: https://doi.org/10.1145/3486622.3493916 (visited on 09/24/2022).

[13] *Keith Harrigian, Carlos Aguirre, and Mark Dredze. "Do Models of Mental Health Based on Social Media Data Generalize?" en. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 3774–3788. DOI: 10.18653/v1/2020.findings-emnlp.337. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.337 (visited on 10/07/2022).

[14] *Andreas Holzinger et al. "Explainable AI Methods - A Brief Overview". en. In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18,*

2020, Vienna, Austria, Revised and Extended Papers. Ed. by Andreas Holzinger et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 13–38. ISBN: 978-3-031-04083-2. DOI: 10.1007/978-3-031-04083-2_2. URL: https://doi.org/10.1007/978-3-031-04083-2_2 (visited on 10/02/2022).

[15] Yen–Hao Huang et al. "LiBRA: A Linguistic Bipolar Disorder Recognition Approach". In: *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. Aug. 2021, pp. 228–235. DOI: 10.1109/IRI51335.2021.00037.

[16] *Uday Kamath and John Liu. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. en. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-83355-8 978-3-030-83356-5. DOI: 10.1007/978-3-030-83356-5. URL: https://link.springer.com/10.1007/978-3-030-83356-5 (visited on 10/07/2022).

[17] *Maximilian Kohlbrenner et al. *Towards Best Practice in Explaining Neural Network Decisions with LRP*. arXiv:1910.09840 [cs, stat]. July 2020. URL: http://arxiv.org/abs/1910.09840 (visited on 09/29/2022).

[18] Enja Kokalj et al. "BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers". In: *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Online: Association for Computational Linguistics, Apr. 2021, pp. 16–21. URL: https://aclanthology.org/2021.hackashop-1.3 (visited on 09/26/2022).

[19] *I. Elizabeth Kumar et al. "Problems with Shapley-value-based explanations as feature importance measures". en. In: *Proceedings of the 37th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, Nov. 2020, pp. 5491–5500. URL: https://proceedings.mlr.press/v119/kumar20e.html (visited on 10/01/2022).

[20] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. arXiv: 1705.07874 [cs, stat]. Nov. 2017. DOI: 10.48550/arXiv.1705.07874. URL: http://arxiv.org/abs/1705.07874 (visited on 09/09/2022).

[21] *Christoph Molnar. *Interpretable Machine Learning*. en. Google-Books-ID: jBm3DwAAQBAJ. Lulu.com, 2020. ISBN: 978-0-244-76852-2.

[22] *Grégoire Montavon et al. "Layer-Wise Relevance Propagation: An Overview". en. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019,

pp. 193–209. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_10. URL: https://doi.org/10.1007/978-3-030-28954-6_10 (visited on 09/28/2022).

[23] *P. Jonathon Phillips et al. *Four Principles of Explainable Artificial Intelligence*. en. Tech. rep. 8312. National Institute of Standards and Technology, 2020. URL: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf (visited on 09/06/2022).

[24] *Nicholas Proferes et al. "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics". en. In: *Social Media + Society* 7.2 (Apr. 2021). Publisher: SAGE Publications Ltd, p. 20563051211019004. ISSN: 2056-3051. DOI: https://doi.org/10.1177/20563051211019004. URL: https://doi.org/10.1177/20563051211019004 (visited on 10/10/2022).

[25] Gabrielle Ras et al. "Explainable Deep Learning: A Field Guide for the Uninitiated". en. In: *Journal of Artificial Intelligence Research* 73 (Jan. 2022), pp. 329–396. ISSN: 1076-9757. DOI: 10.1613/jair.1.13200. URL: https://www.jair.org/index.php/jair/article/view/13200 (visited on 09/08/2022).

[26] Mirka Saarela and Susanne Jauhiainen. "Comparison of feature importance measures as explanations for classification models". en. In: *SN Applied Sciences* 3.2 (Feb. 2021), p. 272. ISSN: 2523-3971. DOI: 10.1007/s42452-021-04148-9. URL: https://doi.org/10.1007/s42452-021-04148-9 (visited on 09/09/2022).

[27] *Wojciech Samek et al. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications". In: *Proceedings of the IEEE* 109.3 (Mar. 2021). Conference Name: Proceedings of the IEEE, pp. 247–278. ISSN: 1558-2256. DOI: 10.1109/JPROC.2021.3060483.

[28] Iqbal H. Sarker. "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions". In: *Sn Computer Science* 2.6 (2021), p. 420. ISSN: 2662-995X. DOI: 10.1007/s42979-021-00815-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8372231/ (visited on 09/07/2022).

[29] Ivan Sekulić, Matej Gjurković, and Jan Šnajder. *Not Just Depressed: Bipolar Disorder Prediction on Reddit*. arXiv:1811.04655 [cs]. Mar. 2019. DOI: 10.48550/arXiv.1811.04655. URL: http://arxiv.org/abs/1811.04655 (visited on 09/08/2022).

[30] Ivan Sekulić and Michael Strube. "Adapting Deep Learning Methods for Mental Health Prediction on Social Media". en. In: *arXiv:2003.07634 [cs]* (Mar. 2020). arXiv: 2003.07634.

DOI: 10.18653/v1/D19-5542. URL: http://arxiv.org/abs/2003.07634 (visited on 03/15/2021).

[31] *Yi-han Sheu. "Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research". en. In: *Frontiers in Psychiatry* 11 (2020). ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.551299. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.551299 (visited on 09/07/2022).

[32] *Chi Sun et al. "How to Fine-Tune BERT for Text Classification?" en. In: *Chinese Computational Linguistics*. Ed. by Maosong Sun et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 194–206. ISBN: 978-3-030-32381-3. DOI: 10.1007/978-3-030-32381-3_16.

[33] *Xiaofei Sun et al. *Interpreting Deep Learning Models in Natural Language Processing: A Review*. arXiv:2110.10470 [cs]. Oct. 2021. DOI: 10.48550/arXiv.2110.10470. URL: http://arxiv.org/abs/2110.10470 (visited on 10/04/2022).

[34] Ihsan Ullah et al. *Explaining Deep Learning Models for Structured Data using Layer-Wise Relevance Propagation*. arXiv:2011.13429 [cs]. Nov. 2020. DOI: 10.48550/arXiv.2011.13429. URL: http://arxiv.org/abs/2011.13429 (visited on 09/28/2022).

[35] Zhengxuan Wu and Desmond C. Ong. *On Explaining Your Explanations of BERT: An Empirical Study with Sequence Classification*. arXiv:2101.00196 [cs]. Jan. 2021. DOI: 10.48550/arXiv.2101.00196. URL: http://arxiv.org/abs/2101.00196 (visited on 09/14/2022).

[36] Yinchong Yang et al. "Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks". In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. ISSN: 2575-2634. June 2018, pp. 152–162. DOI: 10.1109/ICHI.2018.00025.

[37] *Shaomin Zheng and Meng Yang. "A New Method of Improving BERT for Text Classification". en. In: *Intelligence Science and Big Data Engineering. Big Data and Machine Learning*. Ed. by Zhen Cui et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 442–452. ISBN: 978-3-030-36204-1. DOI: 10.1007/978-3-030-36204-1_37.

[38] *Yilun Zhou et al. *Do Feature Attribution Methods Correctly Attribute Features?* en. arXiv: 2104.14403 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2104.14403. URL: http://arxiv.org/abs/2104.14403 (visited on 09/09/2022).