Evaluation of Interpretable Machine Learning Methods in Depression Diagnosis

Ching Hong So, Guilherme Gryschek, João Rafael Calixto Carvalho, Katja Mariko Wilde, Shweta Prasad Ghaisas

Conclusions

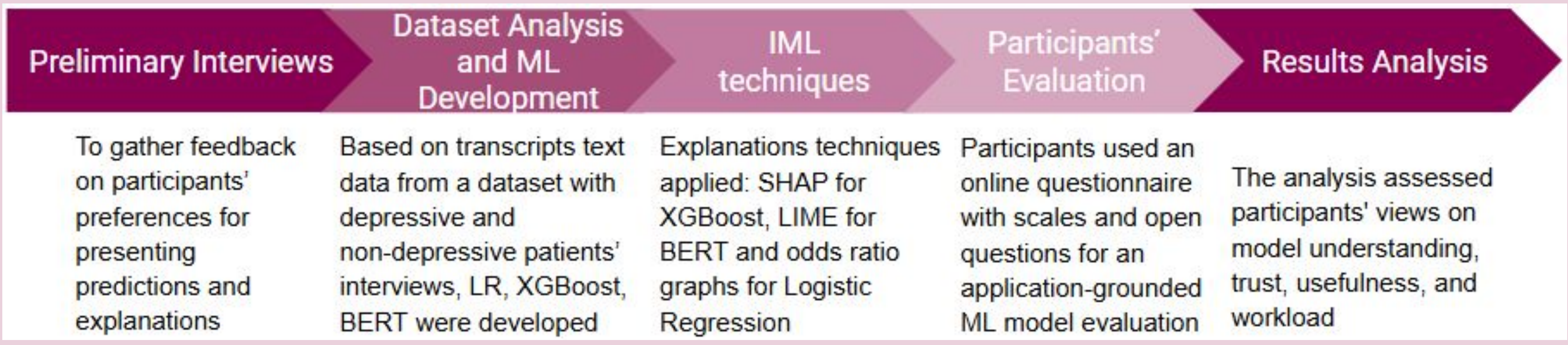
Clinicians favor interpretable machine learning models like Logistic Regression (LR) due to their transparency and alignment with clinical reasoning, making LR the most trusted tool for aiding depression diagnosis. In contrast, complex explanations from SHAP (XGBoost) and incoherent outputs from LIME (BERT) reduced trust and utility. Effective explainability techniques must prioritize clarity, contextual relevance, and ease of use to integrate seamlessly into clinical workflows.

Introduction

Depression is a highly prevalent clinical condition, yet its diagnosis often presents significant challenges for clinicians. Machine learning (ML) has the potential to provide valuable support in identifying these patients, but its effectiveness depends on delivering explanations that are clear, trustful, and useful for clinical use. To address this, interpretation techniques for ML models are designed to ensure their feasibility in the clinical context.

Goals and Methods

The study aimed to evaluate how clinicians understand, trust, and use explanations generated by interpretable machine learning (IML) tools. By assessing various IML methods, we run an application grounded evaluation of the ML models that could enhance clinical decision-making and facilitate integration into routine practice, according to health professionals (*Figure 1*).



Preliminary Interviews: To gather feedback on participants' preferences for presenting predictions and explanations

Dataset Analysis and ML Development: Based on transcripts text data from a dataset with depressive and non-depressive patients' interviews, LR, XGBoost, BERT were developed

IML techniques: Explanations techniques applied: SHAP for XGBoost, LIME for BERT and odds ratio graphs for Logistic Regression

Participants' Evaluation: Participants used an online questionnaire with scales and open questions for an application-grounded ML model evaluation

Results Analysis: The analysis assessed participants' views on model understanding, trust, usefulness, and workload

Figure 1 - Methods Flowchart

Results and Discussion

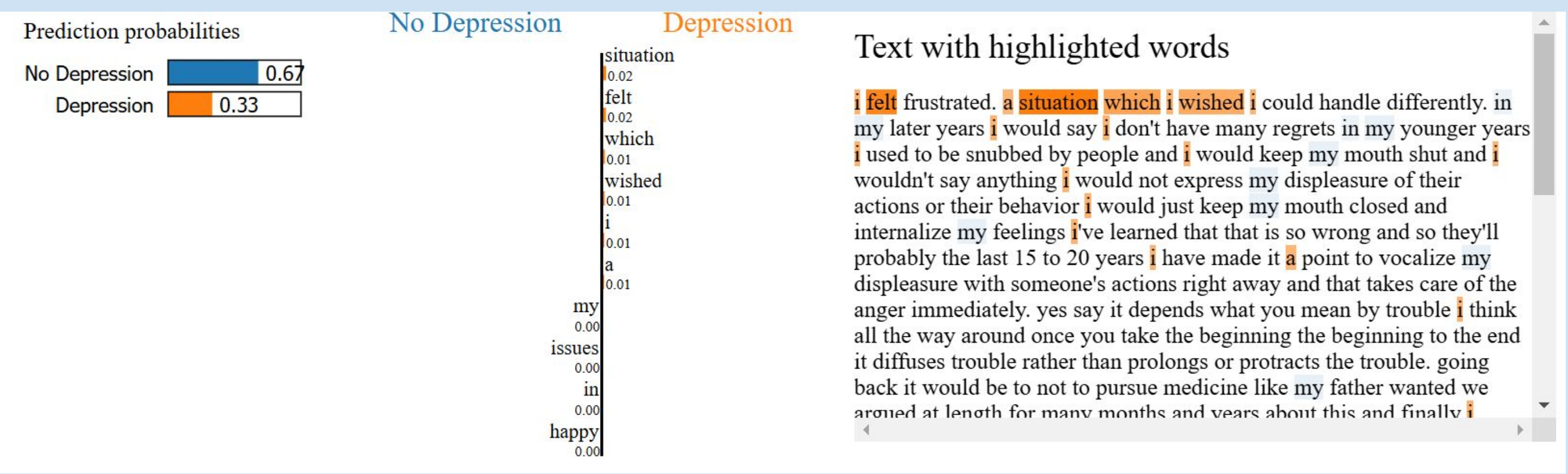
We evaluated perspectives of seven clinicians on various ML models (*Table 1*) and explainability techniques used for diagnosing depression (*Figures 2-4*).

Table 1 - Final Models

Models	Accuracy	F1 Score	Feature Engineering	Explanation Method
BERT¹	0.92	0.91	No	LIME*
LG²	0.59	0.47	Yes	Inherited
XGBoost³	0.67	0.37	Yes	SHAP**

*1.Bidirectional Encoder Representations from Transformers; 2. Logistic Regression; 3. Extreme Gradient Boosting; *Local Interpretable Model-Agnostic Explanations; **Shapley Additive exPlanations*

Our findings revealed that while clinicians valued interpretability, their trust and satisfaction varied depending on the technique and the visualization it offered (*Figure 5*). BERT/LIME, despite its high accuracy, was perceived as unreliable and difficult to trust due to its lack of coherent explanations, particularly when the highlighted words did not align with the context. In contrast, Logistic Regression (LR) was favored for its clear, mathematical interpretation, which clinicians found reliable, accurate, and useful for decision-making. XGBoost/SHAP was helpful for global feature ranking, but its complexity posed challenges for local interpretability, making it less helpful for clinicians.



Text with highlighted words

i felt frustrated. a situation which i wished i could handle differently. in my later years i would say i don't have many regrets in my younger years i used to be snubbed by people and i would keep my mouth shut and i wouldn't say anything i would not express my displeasure of their actions or their behavior i would just keep my mouth closed and internalize my feelings i've learned that that is so wrong and so they'll probably the last 15 to 20 years i have made it a point to vocalize my displeasure with someone's actions right away and that takes care of the anger immediately. yes say it depends what you mean by trouble i think all the way around once you take the beginning the beginning to the end it diffuses trouble rather than prolongs or protracts the trouble. going back it would be to not to pursue medicine like my father wanted we argued at length for many months and years about this and finally i

Figure 2 - LIME for BERT Model local interpretability

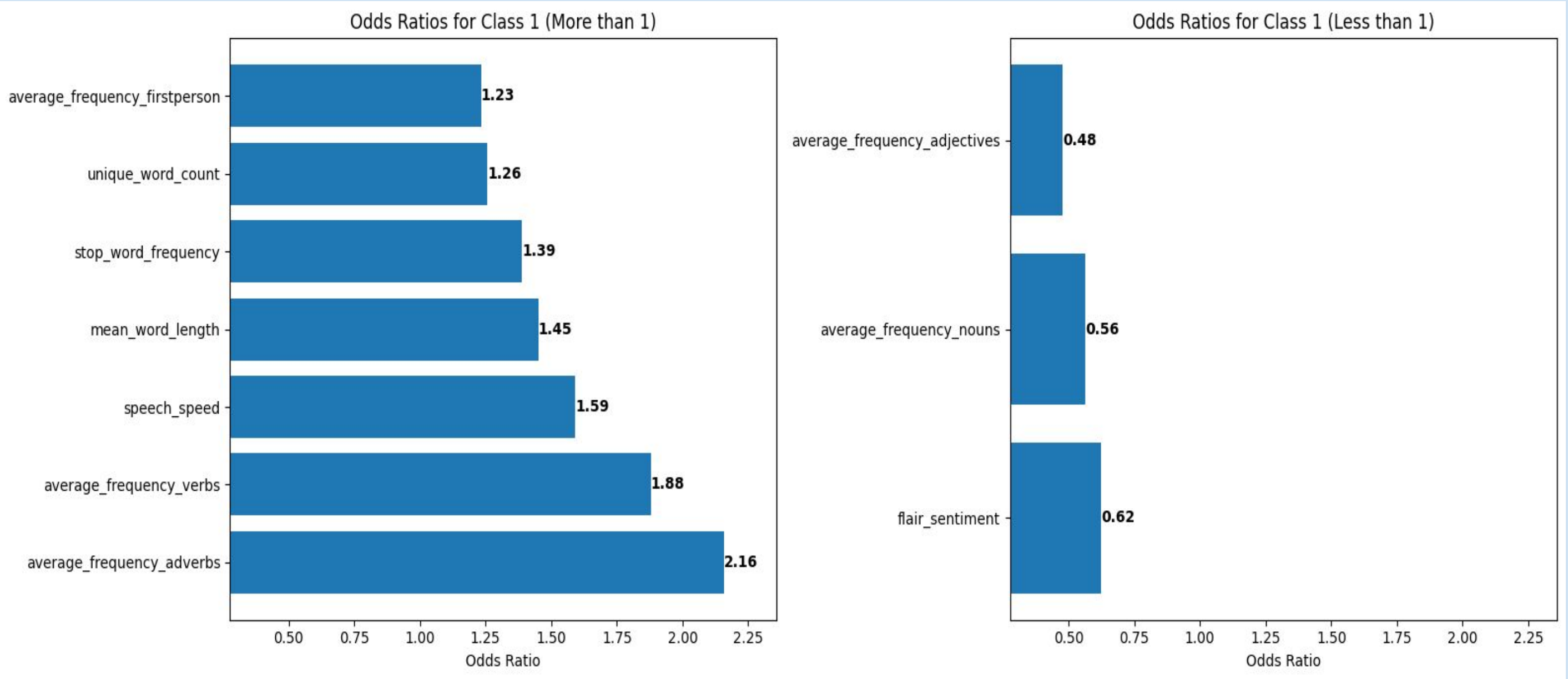


Figure 3 - LR odds ratio graphs for global interpretability

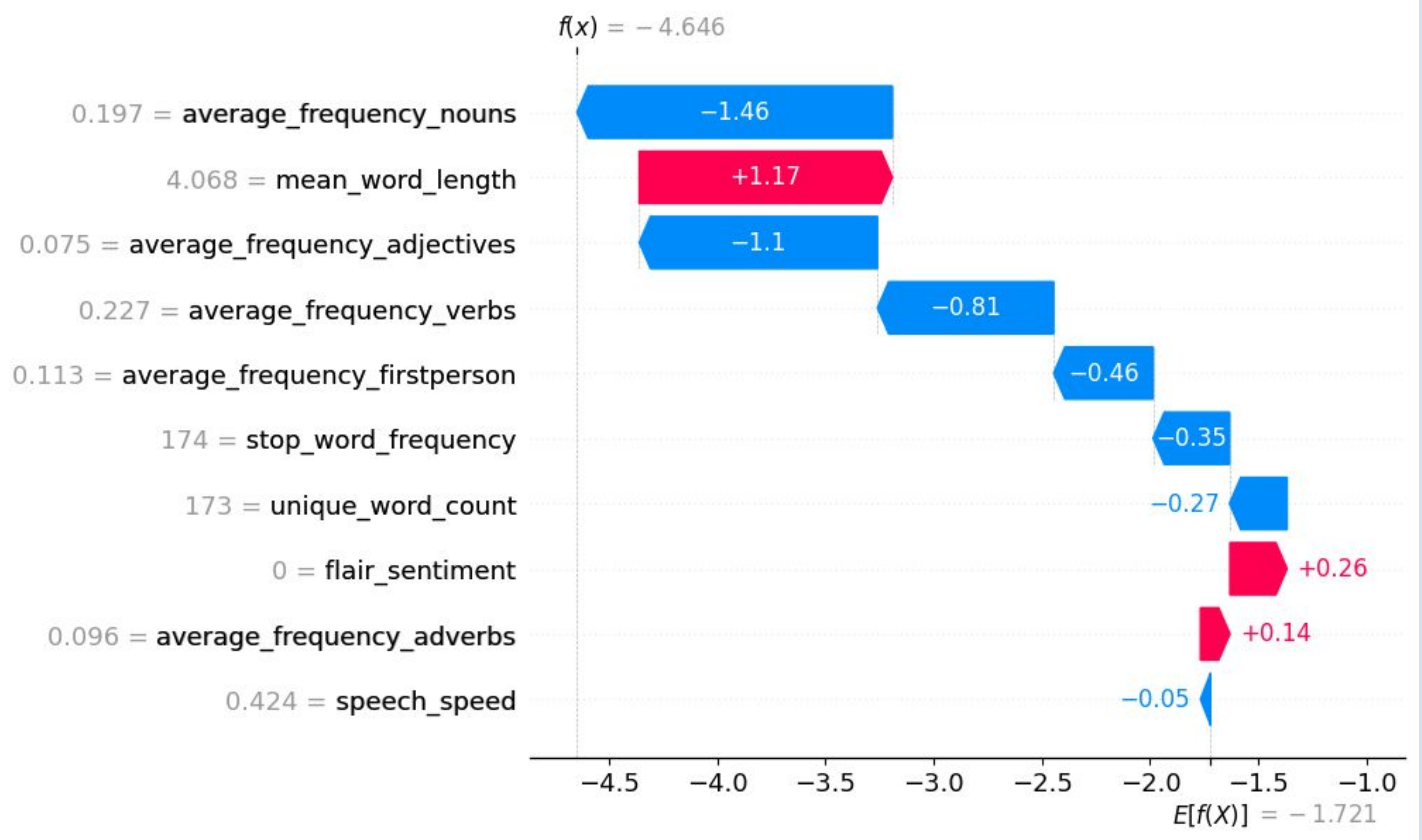


Figure 4 - SHAP for XGBoost local interpretability

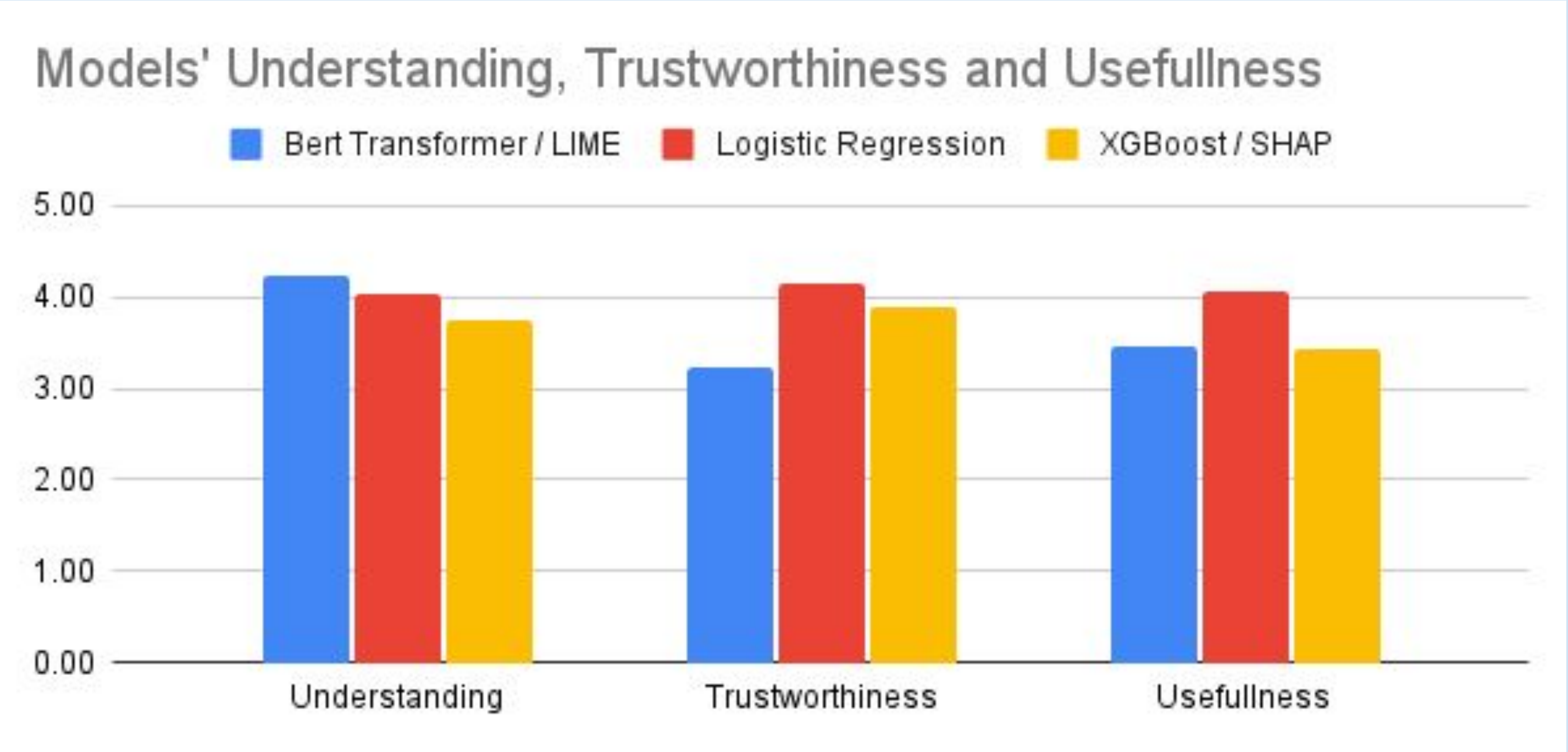


Figure 5 - Average values on each evaluation dimension for each model

Team Contact Details

- Ching Hong So: chinghongso@gmail.com
- Guilherme Gryschek: ggryschek@gmail.com
- João Rafael Calixto Carvalho: joaocalixto17@gmail.com
- Katja Mariko Wilde: katjawi00@gmail.com
- Shweta Prasad Ghaisas: shwetaghaisas@gmail.com

Machine Learning for Health Informatics • Theme 3

SCAN HERE to access the study report for more details

Stockholm University

Karolinska Institutet