Human-centered Machine Learning Project Proposal 2025 @ UU

Julia Baas
j.n.baas@students.uu.nl
Utrecht University
Utrecht, the Netherlands

Bar Melinarskiy
b.melinarskiy@students.uu.nl
Utrecht University
Utrecht, the Netherlands

Ali Nesaei
a.nesaei@students.uu.nl
Utrecht University
Utrecht, the Netherlands

Youssef Ben Mansour
y.benmansour@students.uu.nl
Utrecht University
Utrecht, the Netherlands

ACM Reference Format:
Julia Baas, Bar Melinarskiy, Ali Nesaei, and Youssef Ben Mansour. 2025. Human-centered Machine Learning Project Proposal 2025 @ UU. In *Proceedings of Utrecht University (INFOMHCML'2025)*. Utrecht University, 1 page.

1 INTRODUCTION

In this project, we aim to predict user ratings of medications based on review texts, with a focus on how model choice and feature representation affect both performance and interpretability. In doing this, we continue the work of Cynthia Rudin [1], who advocates the use of inherently interpretable models over black-box models. We will compare a range of models across the interpretability spectrum: from transparent models like Linear Regression, to moderately interpretable ones like GBMs, to black-box models such as fine-tuned BERT, as well as intrinsically interpretable deep learning models like SelfExplain [2] or Proto-LM [3].

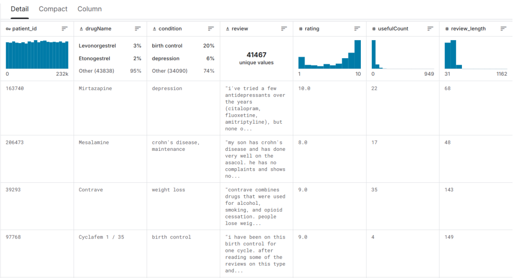
We will employ both interpretable text features (TF-IDF and Word2Vec) and opaque representations (BERT embeddings), utilizing explainability methods including SHAP and LIME to elucidate the decision-making processes of complex models.

2 DATASET

Our project will utilize a Kaggle drug review dataset [4] containing user reviews of drug treatments. It comprises six key features: patient identification, drug name, associated medical condition (e.g., birth control, anxiety), free text review, numerical rating (1-10 scale where 10 is most helpful), review date, and usefulness votes. A glimpse of it can be seen in Fig. 1. The dataset is already split into train (196,397 samples) and test (65,784 samples).

3 EXPERIMENTAL SETUP

To explore how different models and features affect both prediction and interpretability, we train several approaches on the drug review dataset. Our methods include Linear Regression, Gradient Boosting Machines, fine-tuned BERT, and newer interpretable models like SelfExplain and Proto-LM. For each, we experiment with text features such as TF-IDF, Word2Vec, and BERT embeddings. We measure performance using RMSE and MAE, while interpretability is addressed through inherently transparent models and post-hoc explanation techniques such as SHAP and LIME. All models are trained on the training set and evaluated on the test set, with key parameters tuned using grid search. Additionally, we will explore whether the explanation patterns uncovered through these interpretability methods can be leveraged to perform sentiment analysis, offering further insight into the relationship between textual features and user ratings.



patient_id	drugName	condition	review	rating	usefulCount	review_length
0 232k	Levonorgestrel 3% Etonogestrel 2% Other (43838) 95%	birth control 20% depression 6% Other (34090) 74%	41467 unique values	1 10	0 949	31 1162
163740	Mirtazapine	depression	"i've tried a few antidepressants over the years (citalopram, fluoxetine, amitriptyline), but none o...	10.0	22	68
206473	Mesalamine	crohn's disease, maintenance	"my son has crohn's disease and has done very well on the asacol. he has no complaints and shows no...	8.0	17	48
39293	Contrave	weight loss	"contrave combines drugs that were used for alcohol, smoking, and opioid cessation. people lose weig...	9.0	35	143
97768	Cyclafem 1 / 35	birth control	"i have been on this birth control for one cycle. after reading some of the reviews on this type and...	9.0	4	149

Figure 1: Drug review Dataset

4 POTENTIAL PROBLEMS

The biggest risk of our plan is the workload involved in both extracting and fine-tuning the features, as well as implementing the deep learning model Self-Explaining or Proto-LM. The complexity of implementing this model is difficult to assess based solely on the available code and documentation, so this is something that we will discover along the way. If we are unable to get the model working as intended, we will shift to a more practical, interpretable alternative that remains suitable for our analysis goals.

REFERENCES

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[2] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. Selfexplain: A self-explaining architecture for neural text classifiers. *arXiv preprint arXiv:2103.12279*, 2021.

[3] Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models. *arXiv preprint arXiv:2311.01732*, 2023.

[4] Mohamed Abdelwahab Ali. drug-review, 2025.

INFOMHCML'2025, 2025, Utrecht, the Netherlands
© 2025 Utrecht University
ACM ISBN xxxxxxx.
https://doi.org/xxxxxxx