Deployment Architecture

Christian Lavelle

The trained model will be deployed as an engine for category prediction during the software ticket submission process. In a research demo, users will interface with the model through a Space hosted on huggingface.co. This simple form will include a title and body, which will serve as model input. The model's category predictions will determine whether to present the user with additional, contextually relevant form fields.

Quality

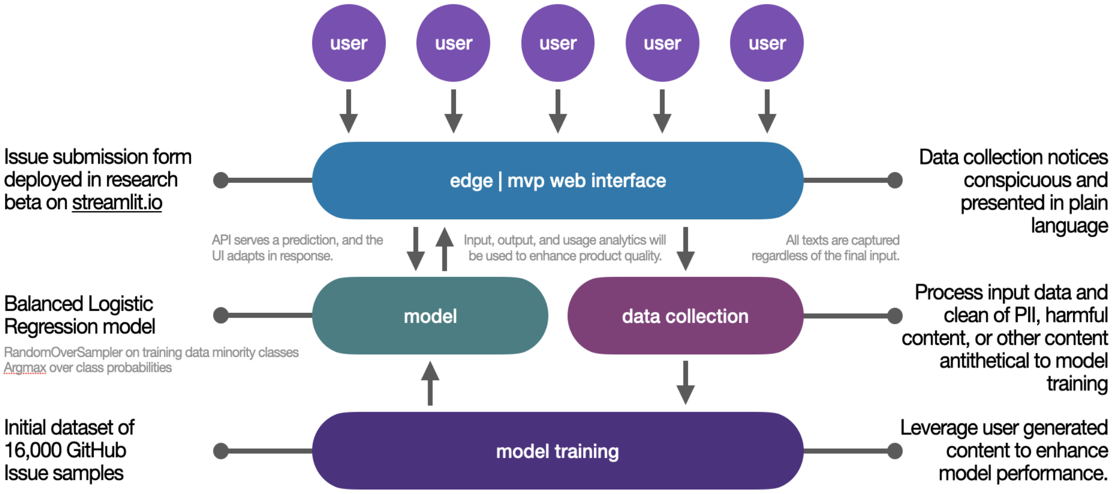
Users will be presented with the ability to select a correct category if the model output is not desirable. This feedback will be used to enhance model performance over time. Current performance is as follows:

Accuracy: 0.885
Macro-F1: ~0.608

Per-class metrics:

- bug: P=0.924, R=0.968, F1=0.945, support=2657
- documentation: P=0.629, R=0.598, F1=0.613, support=102
- enhancement: P=0.545, R=0.497, F1=0.520, support=157
- help wanted: P=0.571, R=0.312, F1=0.403, support=77
- other: P=0.704, R=0.447, F1=0.547, support=197

Production Lifecycle



Issue submission form deployed in research beta on streamlit.io

Data collection notices conspicuous and presented in plain language

API serves a prediction, and the UI adapts in response.

Input, output, and usage analytics will be used to enhance product quality.

All texts are captured regardless of the final input.

Balanced Logistic Regression model

RandomOverSampler on training data minority classes
Argmax over class probabilities

Process input data and clean of PII, harmful content, or other content antithetical to model training

Initial dataset of 16,000 GitHub Issue samples

Leverage user generated content to enhance model performance.

Analytics and Continuous Improvement

All user generated text will be captured, regardless of final input, to help develop a deeper understanding of related strings, and semantic context for software issues. This data, along with model performance will be used to enhance the model throughout the beta lifecycle. Separately, research will continue to see if leveraging LLMs to generate synthetic data for the underrepresented classes will benefit the training process.