

# Capstone Proposal

Christian Lavelle

## Problem Statement

GitHub Issues provides developers with the means of tracking bugs, feature requests, and discussions within a repository. Developers will document issues with a title and description, and are provided the ability to assign tags, which help convey categorical information throughout the issue lifecycle.

This project aims to expedite the issue submission process by leveraging machine learning to predict relevant tags for the issue. This reduces time spent manually querying, browsing, or assigning tags. With well-executing classification algorithms, the software development lifecycle would further benefit from the reduction in custom, error-laden, or otherwise unhelpful tags.

The model will be implemented at the time of issue submission. After a developer has added a title and description, the model will suggest tags and present confidence scores to the user.

## Data

Helpdesk Tickets - High Quality Github Issue

<https://www.kaggle.com/datasets/tobiasbueck/helpdesk-github-tickets/data>

The dataset above includes ~16,000 samples, with a rich feature set. Features such as post-submission replies, issue closure timestamp, and other data points not available at the time of issue submission have been removed.

Features such as word counts, TF-IDF and BERT encodings, have been engineered and included in the model training. These features are later computed at execution time prior to predicting issue tags.

## Approach

The tags for each issue in the dataset act as the targets. There's a significant imbalance to the tags applied, with the 'bug' tag significantly outnumbering all other classes. Balancing the dataset will enhance prediction accuracy for minority classes.

Once a model is finalized, the artifacts will be leveraged to implement a service for tag prediction. This service will receive the user-input title and description, calculate the necessary features, and return up to 3 tags (and confidence scores) for the user to select.