**CS 4800**

**Undergraduate Research in Computer Science**

**Name \_\_David O‘Brien\_\_\_\_\_ Student No. \_\_\_535913\_\_\_\_\_**

**Major \_\_\_\_\_Computer Science\_\_\_\_\_\_ Classification Fr \_\_\_ So \_\_\_\_ Jr \_\_\_\_ Sr \_X\_**

**Credit Hours \_\_1\_\_\_\_ Semester \_\_\_\_Spring 2021\_\_\_\_\_\_**

**Faculty Sponsor \_\_\_Dr. Eugene Wallingford\_\_\_\_**

**Student Responsibilities: Briefly describe the research project you wish to carry out. Be specific in defining the goals of your project. You may attach additional pages to this form if needed.**

**\*See Next Page\***

**Faculty Responsibilities: Briefly describe the faculty sponsor’s role in this project and how they will evaluate the work.**

**\_\_David OBrien\_\_ \_2-5-2021\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_**

**Student Date Faculty Date Head Date**

**Student Responsibilities:**

This Undergrad Research Project will result in two different components: the Classifier and the Software Interface. Both will contribute to their overall system design and play an important role in pulling back the curtain on Data Science Technical Debt. The goal of the Research Project is to be fed a directory containing a software repository, mine out all of the existing comments in the repository, then classify each instance as containing Technical Debt (TD), Data Science Specific Technical Debt (DSTD), or containing no Technical Debt at all.

**Classifier:**

The Classifier’s design will be evaluated throughout the semester. The two leading ideas involve using multiple Classifier to fulfill separate duties of the system.

**Idea 1:** The Classifier would consist of two neural networks, one to classify if a comment indicated either DSTD or normal TD, then a second neural network that would classify a TD comment as DSTD or TD.

**Idea 2:** The Classifier would instead be an ensemble of Classifiers, each one trained on a different domain of AI (example: a Classifier trained specifically on comments from Natural Language Processing, another only trained on Computer Vision comments) and the Classifier each “vote” on whether an unknown comment contains TD, DSTD, or none.

Depending on the idea chosen, the training data’s design could vary slightly. However, the training data will be the result of extracting comments from popular software repositories involving Machine Learning / Artificial Intelligence tools and applications. The following comments will be labelled according to whether they contain traditional or ML-specific Technical Debt by the student and may be asked the training data be looked over by either the faculty supervisor or other students in the Computer Science Department.

The Classifier will be experimented on throughout the semester. Testing different model types (such as K-Nearest-Neighbors, Random Forest Classifier, etc…), parameters, weights, and libraries available to my disposal. The two architectures listed above will be experimented with varieties of training data in hopes of achieving the best performance possible.

**Software Interface:**

The Software Interface’s goal is relatively simple and will not be as time demanding as the above Classifier. However, the interface’s output design will be experimented upon for the best form for user clarity. The Software Interface will simply prompt the user for a directory to an existing software repository on the current machine.

Afterwards, the entire software repository is mined for its existing comments, then will pass it to the decided neural network described above. Then, depending on the output from the Classifier, it will output to the user through an output file where all TD and DSTDs exist within the software system indicated by the user. This will result in an advancement towards maintainable and extendable Machine Learning Software Systems by supplying users a tool to display the technically indebted areas of their data science programs.