Team A: Project Phase Two

October18, 2020

Team Members:

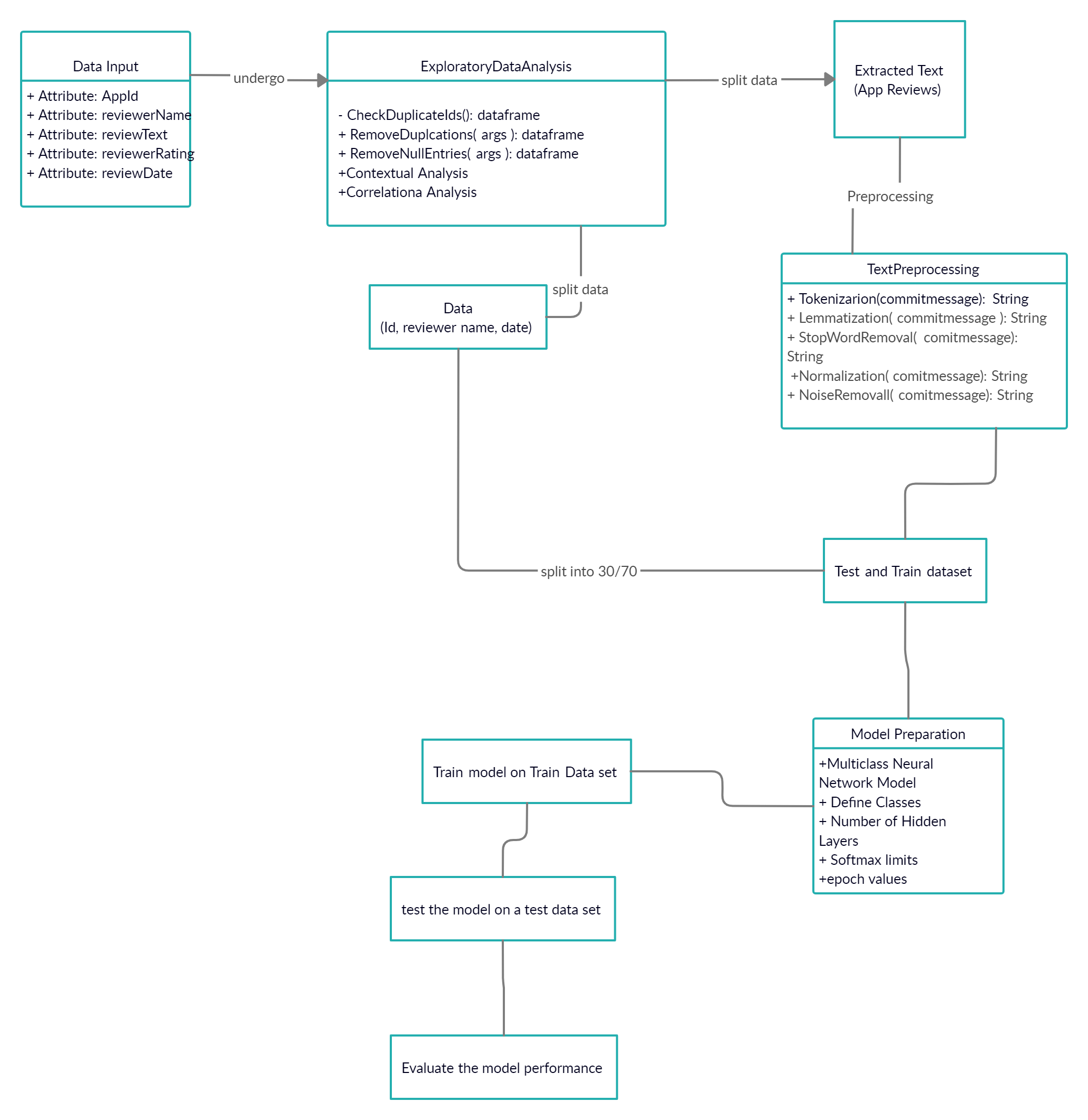
* Goldy Malhotra
* Yuvraj N Singh
* Scotty Williams
* Priyadarshni Suresh Sagar
* Saumya Nagia

Github Repo: <https://github.com/saumya161992/DSCI-644-term-project>

Project Manager: https://trello.com/b/lCrcYPQb/dsci-644-project-team-a

Project Web Page: people.rit.edu/gxm6116/

Architecture:



Milestone Summary:

As the second major phase of the project, this phase primarily involved the development and planning of the next phase with general project management structuring and documentation. This phase consisted of two primary goals, to achieve a well structured project management system with excellent documentation and to plan the foundations of the finished product, a new multi-class model. The first goal was met by compiling together resources to allow for good project structure and organization. A Github repository was created to organize all the files used in the project, including documentation. A Trello project was created to assist in managing and assigning tasks as necessary to the relevant team members. A project web page was created to showcase the project itself in a more user friendly way than a mere proposal paper. The web page contains all the necessary background information on the project and also contains the future goals and objectives of the project. The project web page also includes the planned design and architecture for the final, working product. The architecture itself is made up of several components and phases. The first portion of the architecture focuses on cleaning the testing/training data, including the removal of duplicates and null entries and then partitioning the data into 2 sets, one containing the text reviews for further preprocessing and the other for storage until the training phase. After the text reviews have been processed, by tokenizing, lemmatizing, and normalizing among others, the training/testing dataset is ready. The next phase is model preparation where the number of classes is determined based on the requirements, and the particular parameters of the model itself are set, such as number of hidden layers and softmax limits. The model is then trained using 70% of the dataset and then tested using the remaining 30%. The model’s performance is then evaluated and tweaked as necessary to meet project requirements.