Jonathan Wiseman

April 5, 2020

CS 495

Milestone 3 Report

**Goals**

1. Finalize model
2. Consolidate single-word comments

**Progress:** <https://github.com/jonwiseman/draftsense>

The project’s overarching goals have been accomplished. As of this milestone, there is a functional Support Vector Machine model that can predict with ~85% accuracy if a comment is positive or negative. Of course, there are a number of caveats to this statement. Comments that were deemed jokes or irrelevant to the thread were not included in the model’s training; that is, only comments that were clearly positive or negative were included.

Information about which player’s were the most positively, negatively, and jokingly received is available in the Exploratory Data Analysis notebook. The GenSim implementation for sentence embedding was chosen, mostly because the other two (Google’s BERT and another sent2vec implementation) are difficult to use or require downloading massive pre-trained models and datasets. With GenSim, the ideal vector size is difficult to determine; however, the highest accuracy was achieved with a vector size of 63. Logistic Regression, Naïve Bayes, and Random Forest models were also trained; however, a Support Vector Machine performed the best. Figure 1 shows how each model’s accuracy score changed with vector size:

A close up of a logo

Description automatically generated

**Figure 1: Vector Size vs. Accuracy for Models**

Finally, text processing in the form of consolidating single-word comments was performed. Prior to this step, there were ~300 different one-word comments; after applying this cleaning, the number moved down to ~140.

**Deliverables**

1. “Exploratory Data Analysis”.ipynb: EDA notebook with added text cleaning
2. “Gensim Sentence Embedding”.ipynb: notebook with model creations and vector size experimentation