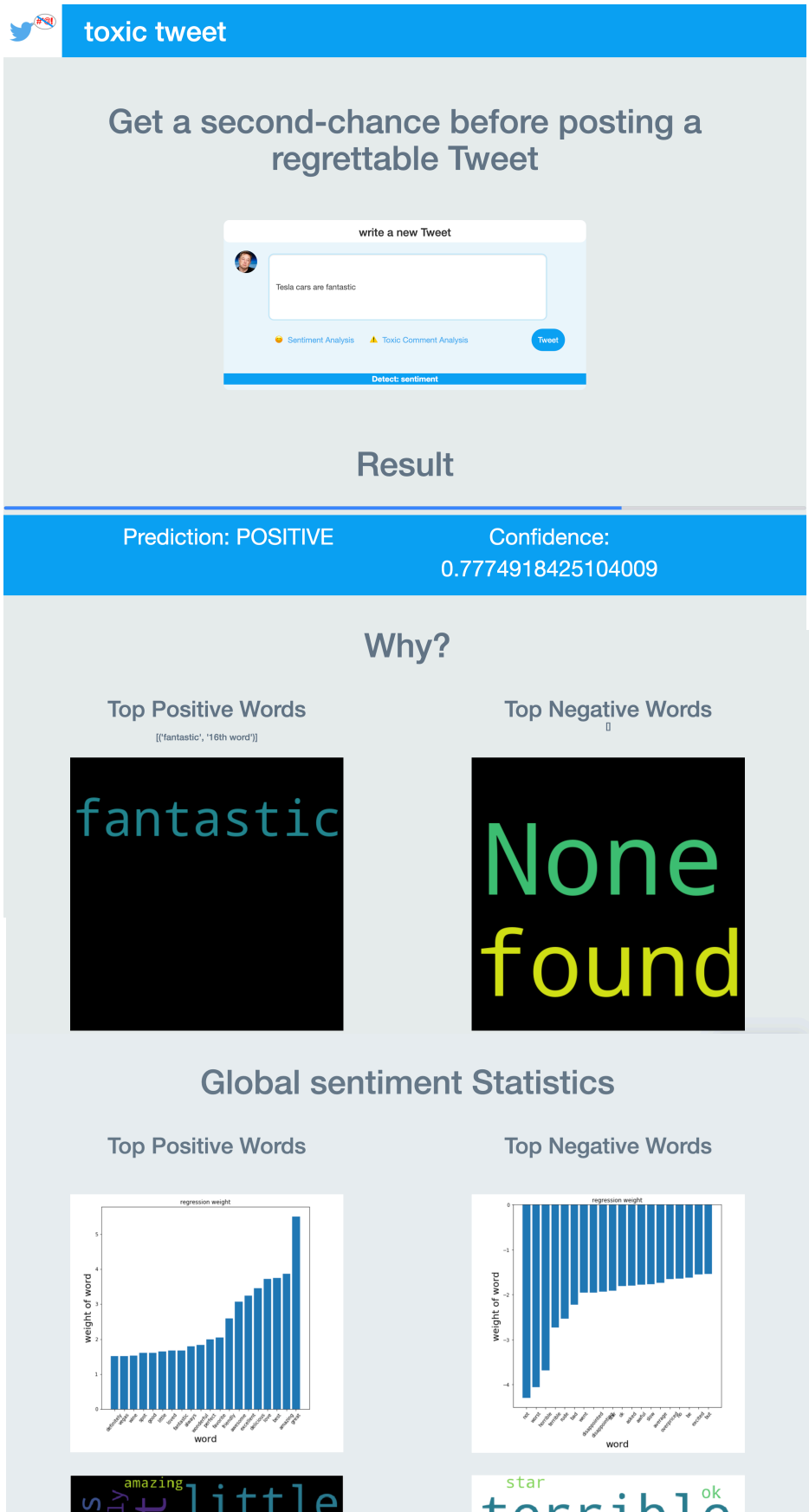
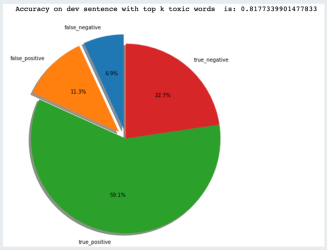
CSE156 Final Project – Toxic Tweet

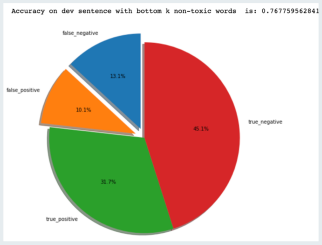
By Nicolas Carmont Zaragoza (A15677088) and Shiang Hu (A53267858)
TEAM ID 46

Part 1.) Sentiment Analysis



Classification Performance





Dev Examples

Correct

Overconfident

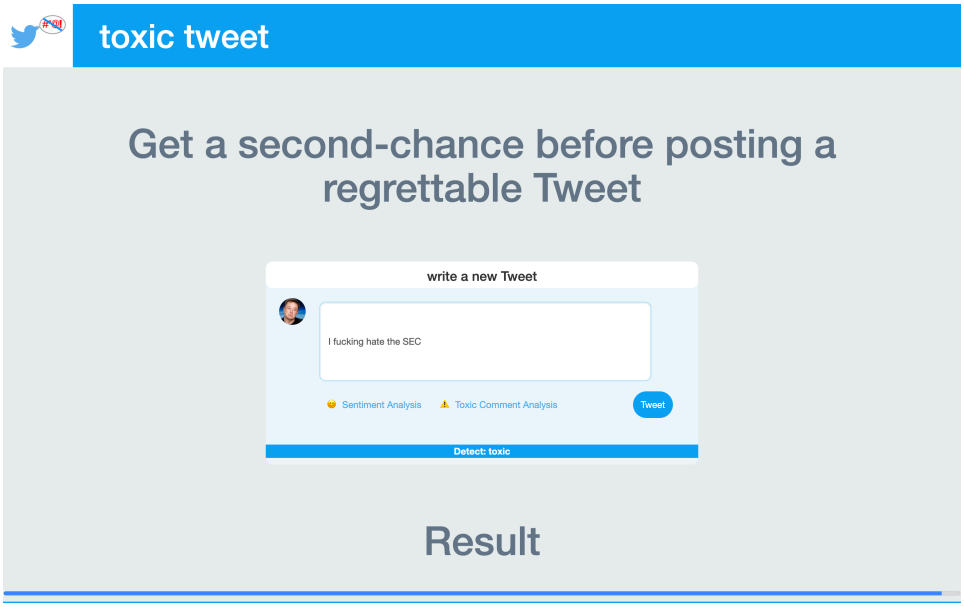
Datasets Used

For Sentiment Analysis we used a dataset containing review data with a vocabulary of 9880 words with sentiment labeled as POSITIVE or NEGATIVE

For Toxic Comment Analysis we used a dataset of 160k Wikipedia comments marked as either TOXIC or NON-TOXIC. Where Toxic is defined as containing hate-speech, profanity, insult, adultery or discrimination

Part 2.) Toxic Comment Analysis

Using 160k Wikimedia TOXIC/NON-TOXIC labeled comments dataset



Result

Prediction: TOXIC

Confidence: 0.9805655164762537

Why?

Top Toxic Words

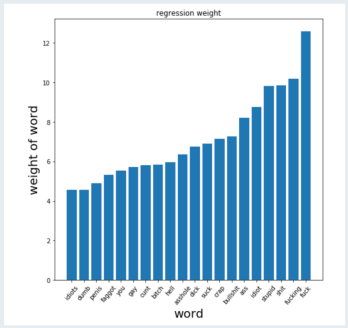
[('fucking', '127th word'), ('hate', '685th word'), ('sec', '373th word')]

Top Non-Toxic Words

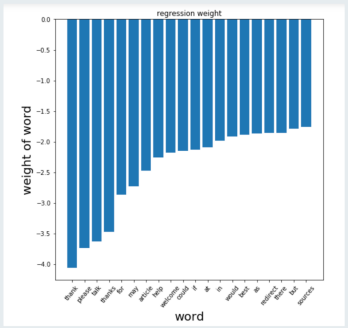
[]

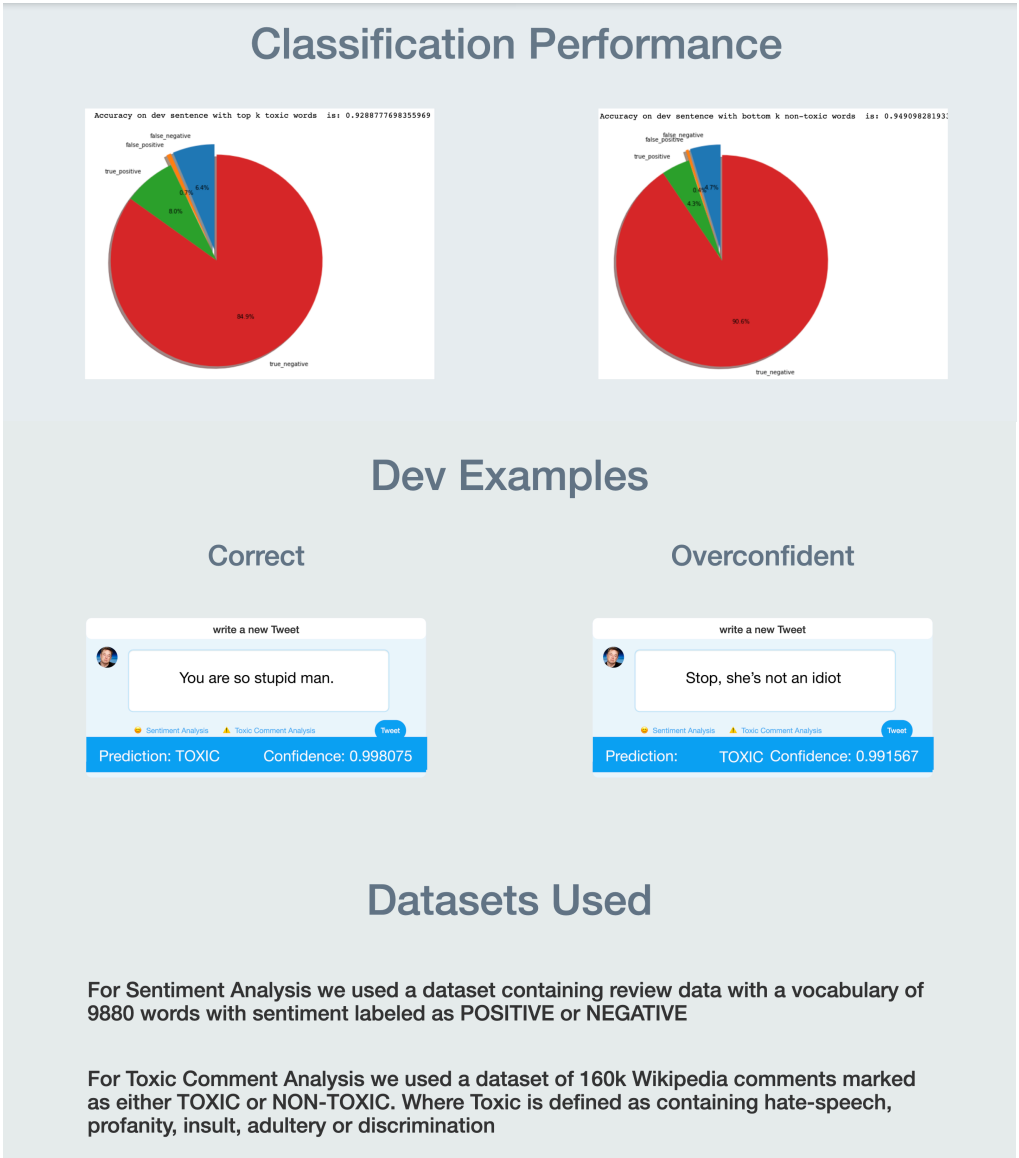
Global toxic Statistics

Top Toxic Words



Top Non-Toxic Words





Part 3.) Creativity

- Interesting and Useful Application of idea to warn people before posting potentially regrettable toxic tweets
- Interactive website UI with twitter theme, user input and dynamic visualization and effects
- Model optimizations: used Bigrams and TF-IDF vectorizer to improve our logistic regressor
- Nice and varied data visualization including word coefficient bar charts, TP + TN pie charts, classified word clouds and textual explanation data.