

# CSE156 Final Project – Toxic Tweet

By Nicolas Carmont Zaragoza (A15677088) and Shiang Hu (A53267858)

TEAM ID 46

## Part 1.) Sentiment Analysis

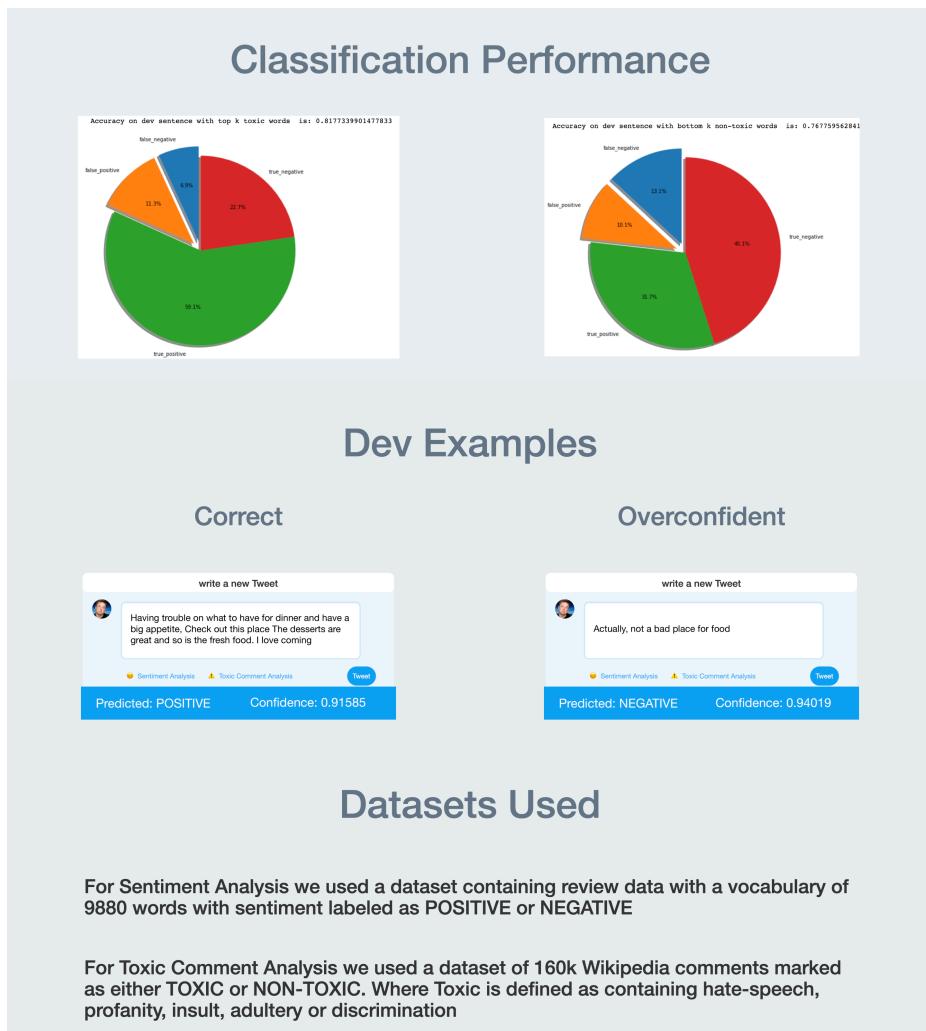
The screenshot shows a web-based application for sentiment analysis. At the top, there's a blue header bar with a Twitter logo and the text "toxic tweet". Below this is a main content area with a light gray background.

In the center of the main area, there's a box labeled "write a new Tweet" containing a placeholder image of a user profile and the text "Tesla cars are fantastic". Below this box are three small icons: a sun-like icon for "Sentiment Analysis", a triangle-like icon for "Toxic Comment Analysis", and a blue "Tweet" button. A blue progress bar at the bottom of this box is labeled "Detect: sentiment".

Below the main input area, the word "Result" is centered. Underneath it, there are two sections: "Prediction: POSITIVE" and "Confidence: 0.7774918425104009".

Further down, the word "Why?" is displayed above two large, black rectangular boxes. The left box is titled "Top Positive Words" and contains the word "fantastic" in white. The right box is titled "Top Negative Words" and contains the words "None found" in white.

At the bottom, the heading "Global sentiment Statistics" is shown above two smaller sections. The left section is titled "Top Positive Words" and features a histogram of regression weights for positive words. The right section is titled "Top Negative Words" and also features a histogram of regression weights for negative words.



## Part 2.) Toxic Comment Analysis

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

**toxic tweet**

**Get a second-chance before posting a regrettable Tweet**

**Result**

## Result

## Prediction: TOXIC

Confidence:  
0.9805655164762537

## Why?

## Top Toxic Words

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

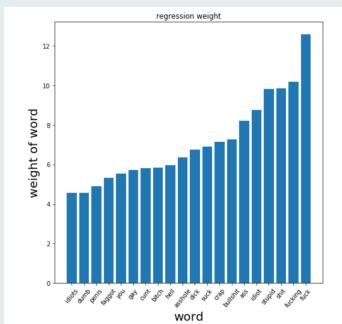
hate  
fucking  
sec

## Top Non-Toxic Words

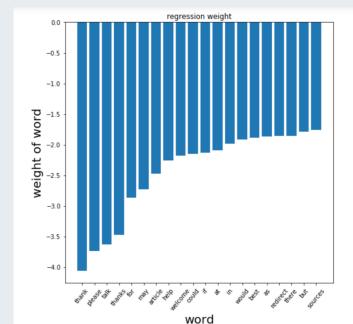
None  
found

# Global toxic Statistics

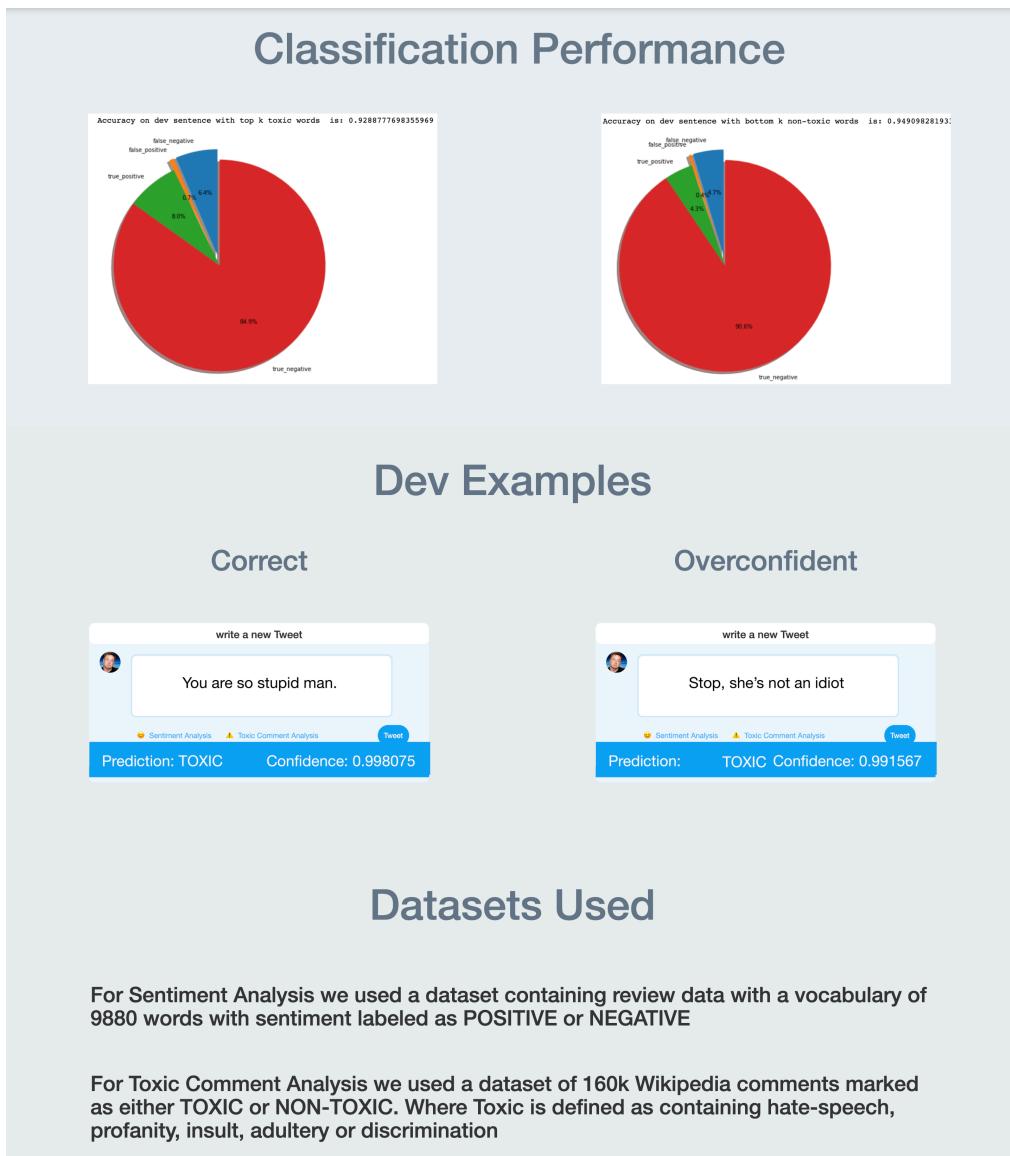
## Top Toxic Words



## Top Non-Toxic Words



dick bitch cunt  
crap penis asshole  
fucking dumb gay  
ass suck stupid  
idiot fuck  
faggot hell  
shit bullshit



## 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.