SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Identification of Toxic Comments in Online Platforms

Mehvish Saleem
Ramanpreet Singh
Ehsan Montazeri

Introduction

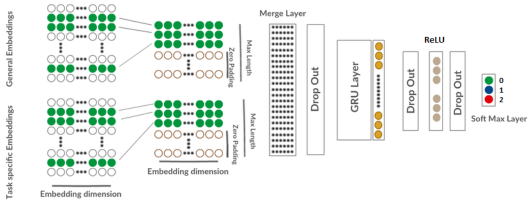
- Identifying toxicity in multiple online communities
- Categorizing different types of toxicities
- Comparing different communities

Motivation

- Toxicity in social interactions is very common
- Can have multiple repercussions such as low self esteem, health problems, depression and isolation
- Automatic toxicity detection can help platform moderators to remove toxic comments and block users

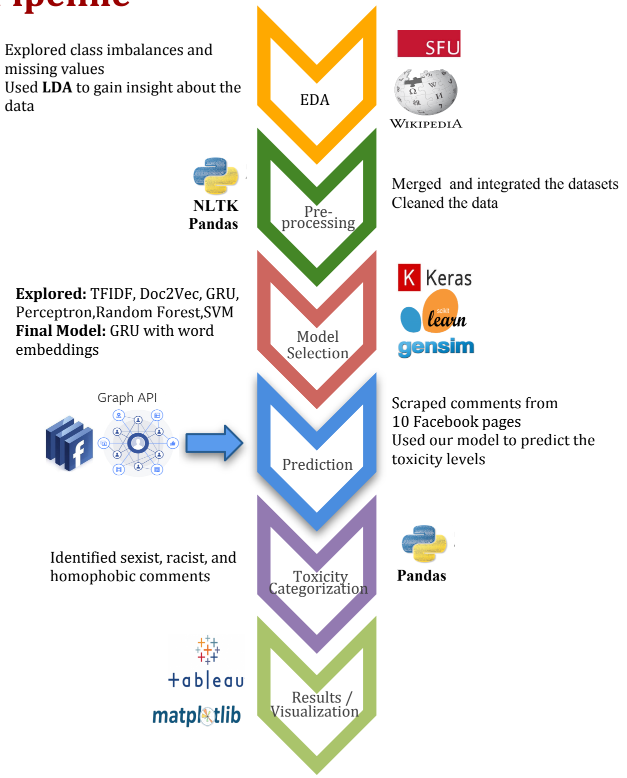
Model

- Experimented with various NLP techniques (Doc2Vec, bag of words) and multiple Machine Learning models such as Naive Bayes, Random Forest, SVM and Perceptron
- Selected Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU) for toxicity classification. The model takes word embeddings as input



Architecture

Pipeline

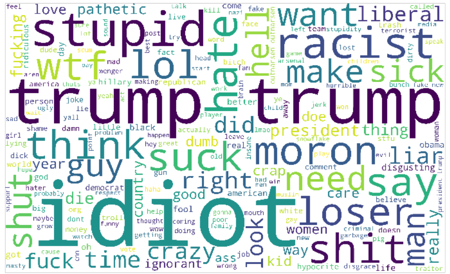


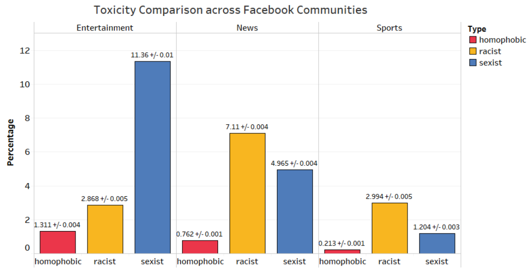
Results

Precision	Recall	F1 Score
0.89	0.94	0.91

Analysis

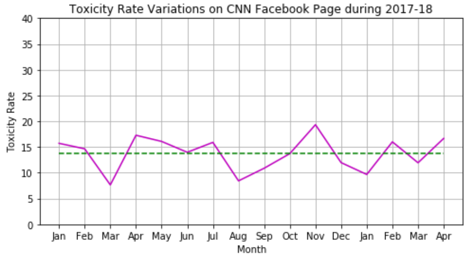
Category	Total	Toxicity	Rate
Entertainment	189,452	8,962	4.91
News	193,769	31,886	16.46
Sports	190,986	9,385	4.71





- Identified different types of toxicities in multiple Facebook communities
- Entertainment was found to be more sexist than racist or homophobic
- News had a higher percentage of racist comments as compared to the other two types
- Sports had a relatively higher percentage of racism than sexism
- 95% confidence interval were computed as shown on the plot

- Toxicity rate on CNN Facebook page was analyzed for each month of 2017 and 2018
- 14% of comments were found to be toxic on average
- Slight fluctuations were observed in each month, the highest being in November, 2017



Future Work

- Conducting supervised learning for toxicity categorization
- Comparing amount of toxicity across multiple platforms (e.g. Twitter Vs. Facebook)
- Identifying bots, trolls, and spammers