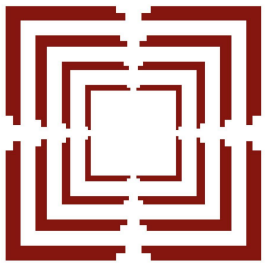

Ahmedabad
University

CSE 523 Machine Learning

Progress Report:

Toxic Comments Classification

Group Details:

NaN-Prediction Pending

Sr. No.	Name	Enrollment Number
1	Aneri Dalwadi	AU1940153
2	Kairavi Shah	AU1940177
3	Nandini Bhatt	AU1940283
4	Mananshi Vyas	AU1940289

Tasks performed in the week:

- We implemented Apache Spark queries for data cleaning and did data preprocessing with R.
- We also plot visualization of frequency of words in the dataset.

Outcomes of the tasks performed:

- We were able to generalize our dataset in tidy text format for converting them into vectors.
- The tokenized words were then classified into various documents for each of the comments.
- We also got an idea of the term frequency for each of the documents. We'll be using this information to apply a tf-idf approach to detect the stop words from the dataset and remove them.

Tasks to be performed in the upcoming week:

- We would be implementing pre-trained models on our preprocessed dataset to get multiclass classification.
- For mid-term presentation, we'll show the result of this classification as well as try to predict the accuracy of each model.

References:

1. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/dat
2. https://smltar.com/
3. https://www.tidytextmining.com/tidytext.html
4. https://arxiv.org/ftp/arxiv/papers/1903/1903.06765.pdf