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 analyzed results of choosing variable eigenvalues for more than one class and fine tune it for better performance
- We used XGBoost library for the same
- We had a meeting with TA to discuss about our conclusions

Outcomes of the tasks performed:

- On comparing the accuracy of Random Forest classifier for variable eigenvalues, we came to the conclusion that 13 eigen values gives highest accuracy
- For fine-tuning, we first trained the XGboost by taking 3 and 5 eigenvalues, it took around 5 - 6 hours for training, therefore, we'll be training it over 13 eigenvalues this week.
- We also studied that since our dataset has variable number of features generated for each comment and since PCA required fixed no. of features, we would not be using it for dimensionality reduction

Tasks to be performed in the upcoming week:

- Prepare for End-semester expo

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