Detecting toxic contents in online conversations

Amal Sony	Mohd Sharique Khan	Natansh Negi	Siddu Madhure Jayanna
200261394	200261202	200262834	200263301
asony@ncsu.edu	mkhan8@ncsu.edu	nnegi2@ncsu.edu	smadhur@ncsu.edu

1 Project Proposal

1.1 Dataset

We have picked the Kaggle dataset for the following challenges:

* Toxic Comment Classification Challenge
* Quora Insincere Questions Classification

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
https://www.kaggle.com/c/quora-insincere-questions-classification/data

1.2 Project idea

An existential problem for any major website today is how to handle toxic and divisive content. Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion) and develop a model that's capable of detecting toxicity. We also aim to provide a score of toxicity for these comments if time permits.

1.3 Software

This project will be implemented in Anaconda(Python) / RStudio(R) using respective libraries.

1.4 Papers to read

[1] John Pavlopoulos & Prodromos Malakasiotis & Ion Androutsopoulos (2017) *Deeper attention to abusive user content moderation.* In EMNLP.

[2] Betty van Aken & Julian Risch & Ralf Krestel & Alexander Loser *Challenges for Toxic Comment Classification: An In-Depth Error Analysis.*

[3] Ellery Wulczyn & , Nithum Thain & , Lucas Dixon & *Ex Machina: Personal Attacks Seen at Scale*

1.5 Teammates and Work Division

We will individually do exploratory data analysis to understand our dataset and gain insight on the topic of NLP(Natural Language Processing). Simultaneously we will split the different aspects of data prepossessing (extraction,cleaning,imputation,transformation,etc.) among ourselves.

Amal Sony and Mohd Sharique Khan: Detect toxic comments.
Natansh Negi and Siddu MJ: Classify toxic comments.

1.6 Midterm Milestone

Data Cleaning, Data Preprocessing and exploring Prediction, Optimization techniques.