Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Fighting Online Abuse and Bullying

CSCI 5622: Machine Learning

Team :

❑ Hasil Sharma
❑ Ganesh Chandra Satish
❑ Akshit Arora
❑ Paramjot Singh
❑ Ajay Kedia

Project Description & Motivation

Why did we choose this project?

In today's well connected world with platforms such as Twitter, Facebook, Medium etc. available at our disposal it is now easier than ever to voice your opinions and share your thoughts. With these platforms cases of online harassment, cyber bullying etc. are also increasing. All these are a nuisance and precludes people from having a civil discourse. Even among us, there are many people who have bullied by these morons through internet, SMS or any other sort of socializing media. It creates fear in the heart of those who are bullied and keeps them from being able to enjoy life. Project aims to build a machine learning system to understand the nuances and context of abusive language, scoring the level of "incivility" to help content moderators keep things such as abusive language, threats and harassment in check, and enhance the exchange of ideas on the internet.

What is our motivation?

❑ Cyber bullying is at an all time high and it doesn’t seem to be slowing down.

❑ A 2016 report from the **Cyberbullying Research Center** indicates that 33.8% of students between 12 and 17 were victims of cyberbullying in their lifetime. Conversely, 11.5% of students between 12 and 17 indicated that they had engaged in cyberbullying in their lifetime.

❑ A large number of people suicide every year due to cyberbullying. Click Here to checkout.

❑ The motivation of this procedure is to study the accuracy of predicting the level of cyber bullying attack using classification methods and also to examine potential patterns between the linguistic style of each predator by using the comments and their respective labels.

Dataset and Initial Approaches

What is our dataset?

- Conversation AI has made their dataset public and we are planning on using same in our project.
- The data was originally collected from Wikipedia labeled talk where each comment is labeled on gender, ages, education, first-language, year, type of article (namespace), type of attack (quoting-attack, recipient-attack, third party-attack and other attack) and every comment is rated by at most 10 workers.
- Others, https://spring.me/ and https://ask.fm/ , question-and-answer formatted websites that contain a high percentage of bullying content.
- Checkout the samples of our training data: **Click Here**

What are our initial steps?

- Build a system that estimates civility of an online comment.
- Use RNN as baseline
- Since RNN fails for longer sequences, explore more advanced options:
  - Long Short Term Memory
  - Attention Mechanism
  - Gated Recurrent Units
- Explore other approaches like Dynamic Time Warping algorithm, SVD described by Potha, Nektaria and Maragoudakis in their paper ‘Cyberbullying Detection using Time Series Modeling’