**About Supernova**: An organization which founding principle is to enhance leadership and organizational skills for North American youth through real life projects. In 2021, we started a AI and machine learning branch of…………..

**Project Guardian Lion Background:** A sad reality is that hate and racism has always been a prevalent part in North American societies, from hate against Black people originating from the era of slavery and Jim Crow to the hate against Asian amplified due to the COVID-19 pandemic. In recent times, the rise of the social media platforms has allowed hate speech to go unchecked with the protection of the anonymity of the internet. This is where Project Guardian Lion comes in.

**Goal:** To create an artificial intelligence that can monitor the feeds of social media platforms and report racially motivated hate speech.

**Means:** The application will use Natural Language Processing in order to achieve this end. Initially, data will be compiled from various social media sites, namely Twitter, Reddit, YouTube, etc. This data will serve to train the model to understand sentiment of the text in general. Then, the data will be refined in order to understand what constitutes as racially motivated hate speech. The model will need to be trained to understand context, a task which would require a substantial amount of training and data. The model will run periodically and report suspicious post. During that process, the model will also collect more data and refine itself.

**Concerns:** This will likely be the most difficult part of the project, as it can be incredibly difficult to distinguish between benign and hate speech online. For example, if the model is trained to recognize the phrase “Asian Hate”, then the phrase “Stop Asian Hate” would also be flagged. This is a big problem for us as most social media posts are short, usually the longer text you have, the more accurate the model can be. This difference would be easy for a human to detect, but hard for a model. We are also aware of the data usage laws and regulations. Model bias, source data bias.

**Why it Matters:** As noted above, most companies already have moderators to regulate and ban hate speech. However, due to the fact that there are a large amount of posts, a lot of hate speech can go unnoticed. Overall, this project can represent an undertaking in natural language processing and help to make the internet a better place.

**Reporting examples: (assume bi-weekly??)**

Project status reporting: XXXXX

Social media platforms scraped:

Twitter: 1 Mil

YouTube: 2K

Reddit: 1 Mil

Hate comment identified:

Twitter: 5

YouTube: 2

Reddit: 1

Total: 8

Data analysis:

Hate comment radiant average: 30 re-posts Neg, positive, neutual, etc.

Compare Asian hate and other racism speech

Moderator bias again Asian or other race hate speech.