Notes: This is just me putting down everything I can think of. Probably not every part of this will be used. Also, I made this under the general banner of stopping hate speech simply because it’s easier to edit, and we can change it to be more specific later. Finally, this is just a rough draft. Feel free to edit any part of it.

**Background:** A sad reality is that hate and racism has always been a prevalent part in American society, from hate against African Americans originating from the era of slavery and Jim Crow to the hate against Asian Americans originating due to the COVID-19 pandemic. In recent times, the rise of the internet has allowed hate speech to go unchecked with the protection of the anonymity of the internet. This is where <Insert Project Name Here> comes in.

**Goal:** To create an artificial intelligence that can monitor the feeds of social media websites and report any hate speech of any kind.

**Means:** The bot will use natural language processing in order to achieve this end. Initially, data will be compiled from various social media sites, namely Facebook, Twitter, Reddit, Quora, TicTok, and Tumblr. This data will serve to train the bot to understand language in general. Then, the data for the bot will be refined in order to let it be able to understand what constitutes as hate speech. This will likely be the most difficult part of the project, as it can be incredibly difficult distinguish between benign and hate speech online. For example, if the bot is trained to recognize the phrase “Asian Hate”, then the phrase “Stop Asian Hate” would also be flagged. This difference would be easy for a human to detect, but hard for a bot. In order to overcome this shortcoming, the bot will need to be trained to understand context, a task which would require a substantial amount of training and data. After this step is completed, the bot can undergo some final debugging, and be ready for field testing.

**Concerns:** One major concern about this project is that it violates privacy. This project may alienate some potential users due to it appearing too much like Big Brother style surveillance. This concern in legitimate, and in the end, there will always be a tradeoff between security and privacy. However large social media companies do not condone hate speech, and the policy of most companies is that any hate speech will get banned by human moderators. This project will simply automate that process, so in the end there will be no change in the status quo. In addition, the bot should try to be as transparent as possible about what it is doing. In doing so, this may get more people on board with this project.

**Why it Matters:** As noted above, most companies already have human moderators to regulate and ban hate speech. However, due to the fact that there are a lot more posts than moderators, a lot of hate speech can go unnoticed. This project can greatly help this cause by automating said process, both stopping hate speech and taking some pressure off the human moderators. Overall, this project can represent an undertaking in natural language processing and help to make the internet a slightly better place.