**Slide 1: Title Slide**

* **"Hello everyone. I'm Shawn, and together with my teammates, rae, Gaurav, Jheel, we are excited to present our final project 'Identifying and Classifying Toxic Online Behavior'**
* **Here is the agenda**
* **Warning, this presentation includes explicit content, which is no intention to offense and just for study purpose, we advise everybody to watch with caution and feel free to leave if you feel uncomfortable**

**Slide 2: Introduction**

* **nowadays, there are Increasing negative online behavior, including personal insults and hate speech in The internet.**
* **"Our goal is to develop a model that can predict and classify toxic texts to enhance user engagement and mental wellbeing."**

**Slide 3: Dataset Overview**

* **"We used a comprehensive dataset from Wikipedia talk pages, containing 160,000 comments labeled for toxicity types by human raters, including toxic, obscene, threat, insult, and identity hate"**
* **"Here is a summary of our dataset characteristics**
* **For length distribution, 67 in average and 36 for median**
* **For label count, toxic is the most frequent label and threat is least frequent**
* **Majority of comments is non-toxic, which align with the reality**

**Slide 4: Data Preprocessing**

* **"In preprocessing, we performed text normalization and then used NLTK for tokenization."**
* **"For embeddings, we utilized FastText and Word2Vec. Fasttext techniques help in capturing subword information, we are using** one averaged vector for one sentence**, essential for our models to understand and process the text effectively."**
* **warning for explicit content**
* **used Visualization by wordcloud for understanding the toxic pattern**

**Slide 5: Models and Findings**

* **"We implemented several models, Each model served a different purpose and provided varying levels of accuracy into the data."**
* **Logistic Regression: Effective in binary tasks.**
* **Random Forest with OneVSRestClassifier: Beneficial for tackling multi-class classification challenges.**
* **CNN: Comprises two convolutional layers and a classifier, excelling in pattern recognition.**
* **LSTM: Includes one LSTM layer plus a classifier, adept at recognizing long-term textual dependencies.**
* **"Here are some of the results showcasing the effectiveness of these models in classifying different types of toxic behavior.**
* **this is the training and validation loss for models**
* **Notably, our models were able to highlight the very high accuracy in the prediction, it is to be noted LSTM is not performing as expected, due to we did Not use sequential input**

**Slide 6: ZEROSHOT and Findings**

* **We utilized Zero-Shot Classification for its ability to predict new types of toxic behavior without specific training. models include:**
* **General-Purpose Models: Based on the Multi-Genre Natural Language Inference Dataset.**
* **Binary Toxicity Specialized Models: Tailored from the Jigsaw Toxic Comment Classification Dataset.**
* **Data Utilized: Analyzed 200 randomly sampled comments labeled with various forms of toxicity.**
* **Findings indicate that fine-tuned models excel in detecting toxic behavior, though achieving consistent performance across all categories remains challenging, particularly in nuanced areas like threats."**

**Slide 7: Limitations and Lessons Learned**

* **"Our analysis faced several limitations, including the size of our dataset, computational power limitations, and inherent biases in dataset labels."**
* **"From this project, we've learned valuable lessons about the challenges of working with natural language data, especially in handling imbalanced datasets and mitigating model biases."**

**Slide 8: Conclusion and Q&A**

* **"To conclude, our project aimed to address the pressing issue of toxicity in online spaces by developing a predictive model that not only identifies but also classifies different types of toxic behavior."**
* **"We now welcome any questions you might have. Thank you for your attention."**