**Tweet Tamers: AI-Powered Social Media Moderation for Safer Online Interactions**  
**Team Members:** Raviteja Pola, Gilbert Nat-Narh, Adil Ahmed, Jm Quilario  
**Presentation Date:** July 21, 2025

**Introduction**  
With the growing influence of social media in daily life, the proliferation of harmful and toxic content poses a significant threat to users' mental well-being. Manual moderation at scale is neither efficient nor consistent. To address this, *Tweet Tamers* proposes an automated approach to classify user-generated content by emotional tone and toxicity level, enabling platforms to proactively shield vulnerable audiences from potentially harmful posts.

**Dataset and Preprocessing**  
We utilized the *uglyWords1.csv* dataset, which contains a collection of tweets with two target labels:

* **Emotion:** Categorized from -2 to 2, representing a range from *Very Sad* to *Very Happy*
* **Caution (Toxicity):** Ranging from -1 (toxic) to 1 (non-toxic)

For modeling, we remapped the Emotion labels from {-2, -1, 0, 1, 2} to a normalized scale of {0, 1, 2, 3, 4}, and the Caution labels from {-1, 0, 1} to {0, 1, 2}, respectively. The dataset was imbalanced, so we applied upsampling technique to ensure uniform representation across all classes. The dataset was then split into 70% training, 15% validation, and 15% testing.

**Model Architecture**  
Our model leverages the **Google Universal Sentence Encoder (USE)** to capture semantic representations of input text. The classification architecture consists of the following layers:

* A **KerasHub Layer** for USE embeddings
* A **Dense** layer with 255 units and ReLU activation
* A **Dropout** layer (rate = 0.2) to prevent overfitting
* A **Dense** layer with 64 units and ReLU activation
* A final **Dense** output layer with 5 units for multi-class emotion classification

The model was compiled using the **Nadam** optimizer, **Categorical Focal Crossentropy** loss (with logits), and **categorical accuracy** as the evaluation metric.

**Performance Evaluation**  
After training the emotion detection model for 10 epochs, it achieved the following metrics:

* **Emotion Detection Accuracy:** 90.44%
* **Loss:** 0.0435

For the Caution (toxicity) classification task, the model achieved:

* **Caution Detection Accuracy:** 79%
* **Loss:** 0.1253

**Proposed Application**  
Our envisioned system integrates this model into a real-time content filtering pipeline for social media. When a user submits a post, the AI model analyzes its cautious tone. If the detected emotion falls under the “Toxic” categorie and the user is identified as a minor, the system will automatically hide the content. Additional features include:

* A warning label: *“Content hidden for your safety”*
* Optional guardian notifications
* An override mechanism for verified adults or moderators

**Conclusion**  
*Tweet Tamers* demonstrates the potential of AI in making online spaces safer, particularly for young users. By detecting emotional distress and toxicity in user content, our system promotes mental well-being and offers a scalable, proactive approach to content moderation.