**HACKATHON: AI FOR SAFER ONLINE SPACES FOR WOMEN**

**TASK 3: DETECTING TOXIC OR HARMFUL COMMENTS**

**Model Architecture and Methodology**

**1**. Data Collection and Preprocessing

* Dataset: 1000 comments from /content/final\_labels.csv.
* Detoxify model assigned toxicity scores.
* Toxicity levels:
  + Toxic: Score > 0.5
  + Neutral: Score 0.2 - 0.5
  + Non-toxic: Score ≤ 0.2

2. Model Development

* Pre-trained RoBERTa was fine-tuned for classification.
* Tokenization was performed using RobertaTokenizerFast.
* Data was split into 80% training and 20% testing.

3. Training Details

* Batch size: 8
* Epochs: 3
* Optimizer: AdamW with weight decay

**Evaluation Metrics**

* Accuracy: 85% classification accuracy.
* Toxicity Classification: LABEL\_MAP used for easy interpretation.
* Rephrasing Effectiveness: Toxic comments reworded with suggest\_rephrasing():
  + *"stupid" → "misguided"*
  + *"idiot" → "uninformed person****"***

**Challenges and Solutions.**

* Limited Rephrasing Suggestions: Early versions lacked variety.

**Deployed model:**

<https://huggingface.co/spaces/Charankarnati18/TASK3>

**TASK 2:**

**SUBREDDIT-BASED TOPIC CLASSIFICATION**

**Model Architecture and Methodology**

1. Preprocessing and Data Handling

* Stopwords and special characters were removed.
* Text was normalized and tokenized using DistilBERT tokenizer.
* Subreddit names were mapped to numerical labels.

2. Training Approach

* A DistilBERT model was fine-tuned for multi-class classification.
* The Hugging Face Trainer API was used for training.
* Training parameters:
  + Batch size: 8
  + Epochs: 4
  + Optimizer: AdamW with learning rate decay

**Evaluation Metrics**

* Precision:0.67
* Recall:0.67
* F1-score:0.67
* Confusion Matrix: Assessed misclassification trends.
* Topic Evolution Analysis: Extracted year-month timestamps to track subreddit activity.

**Challenges and Solutions**

**Challenge**: Tracking Trends Over Time: Needed a way to visualize topic distribution.

**Solution**: Used Plotly Express for interactive time-series charts.

**TASK 4:**

**CONTEXT-AWARE MISOGYNY DETECTION**

**Model Architecture and Methodology**

* **Transformer-based Classifier**: The model is built on **DistilBERT**, a lighter version of BERT, fine-tuned on a labeled dataset to classify misogynistic content.
* **Multi-Stage Classification**:
  1. **Step 1**: Binary classification of text into misogyny vs. non-misogyny.
  2. **Step 2**: Contextual analysis to identify sarcasm, jokes, or harmful intent.
  3. **Step 3**: Explainability module highlights misogynistic phrases for interpretability.
* **Training Strategy**:
  1. Learning rate: **2e-5**
  2. Batch size: **16**
  3. Epochs: **5**
  4. Evaluation: **Epoch-wise validation**
* **Hardware Utilization**: Model training and inference are optimized using **GPU acceleration** for efficient processing.

**Evaluation Metrics**

* **Accuracy**: Measures overall correctness in detecting misogynistic content.
* **Precision-Recall**: Evaluates the model's reliability in identifying true misogynistic cases while minimizing false positives.
* **Explainability**: Uses attention mechanisms to highlight problematic words and provide reasons for flagging content.
* **User Engagement**: Monitors the effectiveness of the moderation system, including how users respond to flagged content and warnings.

**Challenges Faced and Solutions Implemented**

1. **Ambiguity in Sarcasm and Jokes**
   * **Challenge**: Distinguishing between misogynistic intent and sarcastic/joking statements.
   * **Solution**: Integrated **RoBERTa for sarcasm detection** and **sentiment analysis** to assess intent.
2. **Data Imbalance**
   * **Challenge**: More non-misogynistic than misogynistic examples in the dataset.
   * **Solution**: Applied **oversampling techniques** and **data augmentation** to balance training samples.
3. **Explainability of Predictions**
   * **Challenge**: Users needed clear justifications for why their content was flagged.
   * **Solution**: Implemented **attention-based word highlighting** to provide transparency.
4. **Real-Time Processing**
   * **Challenge**: High computational demand for real-time moderation.
   * **Solution**: Used **quantized models** and **GPU optimization** to accelerate inference

**TASK 1:**

**PARENT-CHILD CONVERSATION RECONSTRUCTION REPORT**

**Model Architecture and Methodology**

1. Model Selection

* The model used is facebook/bart-base, a transformer-based model optimized for sequence-to-sequence tasks like summarization.
* AutoTokenizer and AutoModelForSeq2SeqLM from Hugging Face's Transformers library are utilized for tokenization and model loading.

2. Data Processing

* The dataset is loaded from final\_labels.csv, which contains conversation text (body) and corresponding summaries (highlight).
* Missing data is removed using dropna() to ensure quality input.
* Text data is processed and tokenized with max\_length=512 for input and 128 for output summaries.

3. Training Methodology

* The Seq2SeqTrainer from Hugging Face is used to fine-tune the model.
* Training parameters:
  + Learning rate: 3e-5
  + Batch size: 4
  + Epochs: 3
  + Weight decay: 0.01
  + Beam search (num\_beams=4) for generating coherent summaries.
* The fine-tuned model is saved for future usage.

**Evaluation Metrics**

* BLEU Score: Measures n-gram overlap to evaluate summary accuracy.
* ROUGE Score: Assesses recall-based similarity with reference summaries.
* Perplexity: Evaluates how well the model predicts the next token in a sequence.
* Semantic Similarity: Uses BERT embeddings with cosine similarity to compare generated and reference summaries.

**Challenges and Solutions**

* Maintaining Context in Summaries: The model sometimes omitted key points. Solution: Beam search tuning improved coherence.
* Handling Long Conversations: Truncation led to loss of information. Solution: Used hierarchical summarization to generate intermediate summaries.
* Overfitting on Small Data: Solution: Implemented dropout regularization and fine-tuned learning rate.

**Deployed Model:**

https://huggingface.co/spaces/Charankarnati18/summarize\_

**CONTRIBUTORS:**

Karnati Charan

Chappidi Pavitra SUbhasri