**NAME –** Pranav S

**SRN –** PES2UG23CS430

**Section** - G

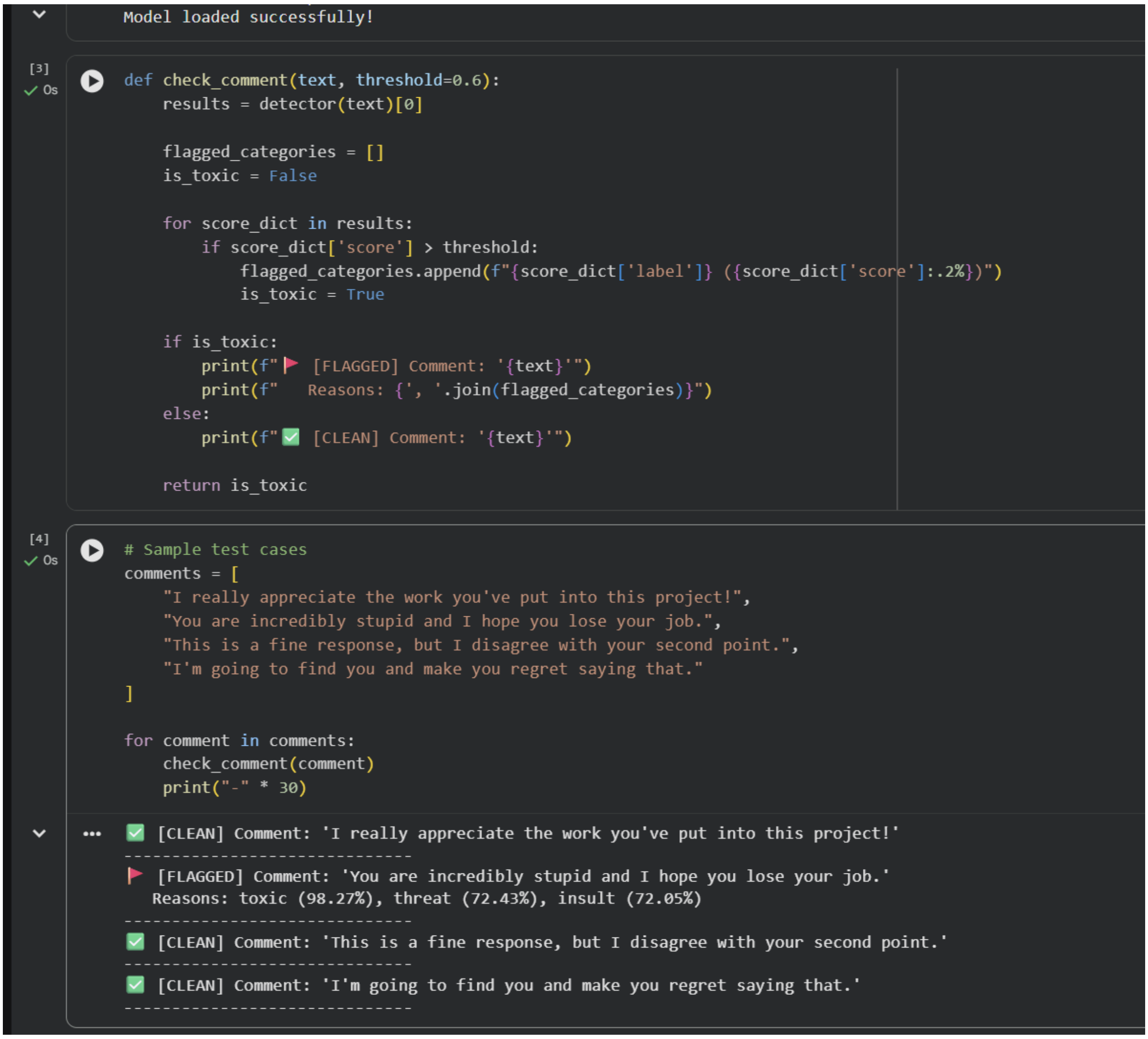
**Problem statement(39) - Cyberbullying Detector**

- **Goal**: Flag toxic comments on a forum.
- **Tech**: text-classification (toxicity model).

**Short documentation** - I've developed a **Multi-Label Toxicity Classifier** using the Hugging Face transformers framework to automate forum moderation. At its core, the system utilizes a pre-trained **BERT-based model** ($unitary/toxic-bert$) that analyzes the semantic context of a comment rather than relying on simple keyword blacklists. It maps input text against six distinct categories—toxic, severe toxic, obscene, threat, insult, and identity hate—providing a granular understanding of how a message violates community standards.

The implementation features a custom **threshold-based logic** ($P > 0.6$) to balance safety with free expression, ensuring only high-confidence violations are flagged. I built a processing function that parses the model's raw probability outputs into a human-readable format, identifying the specific nature of the abuse and its confidence score. This creates a scalable, real-time solution for detecting cyberbullying that can distinguish between general profanity and targeted identity-based harassment.

## Sample output/ screen shots –



```
Model loaded successfully!
```

```python
def check_comment(text, threshold=0.6):
    results = detector(text)[0]

    flagged_categories = []
    is_toxic = False

    for score_dict in results:
        if score_dict['score'] > threshold:
            flagged_categories.append(f"{score_dict['label']} ({score_dict['score']:.2%})")
            is_toxic = True

    if is_toxic:
        print(f"🚩 [FLAGGED] Comment: '{text}'")
        print(f"   Reasons: {', '.join(flagged_categories)}")
    else:
        print(f"✅ [CLEAN] Comment: '{text}'")

    return is_toxic
```

```python
# Sample test cases
comments = [
    "I really appreciate the work you've put into this project!",
    "You are incredibly stupid and I hope you lose your job.",
    "This is a fine response, but I disagree with your second point.",
    "I'm going to find you and make you regret saying that."
]

for comment in comments:
    check_comment(comment)
    print("-" * 30)
```

```
✅ [CLEAN] Comment: 'I really appreciate the work you've put into this project!'
------------------------------
🚩 [FLAGGED] Comment: 'You are incredibly stupid and I hope you lose your job.'
   Reasons: toxic (98.27%), threat (72.43%), insult (72.05%)
------------------------------
✅ [CLEAN] Comment: 'This is a fine response, but I disagree with your second point.'
------------------------------
✅ [CLEAN] Comment: 'I'm going to find you and make you regret saying that.'
------------------------------
```