FINAL PRESENTATION

Cyberbullying Classification

Rachel Chesebro'
Satyen Sabnis
Mamadou Ly II
Jesus Cruz

TABLE OF CONTENTS

01. Introduction , Problem, & Dataset Description

02. GRU Implementation & Evaluation

03. BERT Implementation & Evaluation

TABLE OF CONTENTS

04. LSTM Implementation & Evaluation

05. SHAP Implementation & Evaluation

06. Conclusion & Future Models

INTRODUCTION

01.

Problem & Dataset

THE CHALLENGE OF CYBERBULLYING

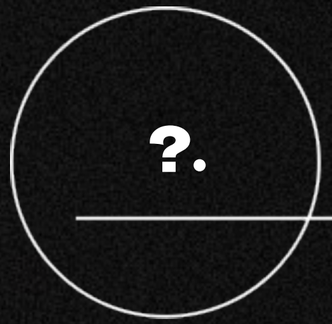
What is Cyberbullying?

- **Cyberbullying** involves repeated behavior aimed at harming someone else, carried out through **digital** platforms. It can range from spreading rumors and posting hurtful comments to more aggressive threats.

How Does It Impact Us?

- The psychological impact on victims can be severe, making early detection crucial.

- Ex: Higher rates of **Depression** and **Anxiety**.
  Increased **Suicide** Attempts (Yale Study)



WHAT IS OUR GOAL?

Make the Internet a **SAFER** place.

OVERVIEW OF OUR DATASET

- Sample of various cyberbullying tweets
- *'cyberbullying_tweets.csv'* has 47,656 example vectors and 2 feature vectors.
- Feature Vectors:

tweet_text — Description of short message

cyberbullying_type — Type of Cyberbullying

- religion 7997
- age 7992
- ethnicity 7959
- gender 7948
- not_cyberbullying 7937
- other_cyberbullying 7823

```
tweet_text          Shutup you idiot. Victim blaming. Acting as if...
cyberbullying_type                                           religion
Name: 20033, dtype: object
```

DATA PREPARATION

- **Data Cleaning**: Removing noise and non-informative text.
- **Text Normalization**: Standardizing text for analysis.
- **Tokenization**: Breaking text into manageable pieces.
- **Handling Missing Data**: Strategies to address data gaps.
- **Refining Data Quality**: Removing outliers in text length.
- **Utilizing Pre-labeled Data**: Handling multi-category labels.

GRU MODEL

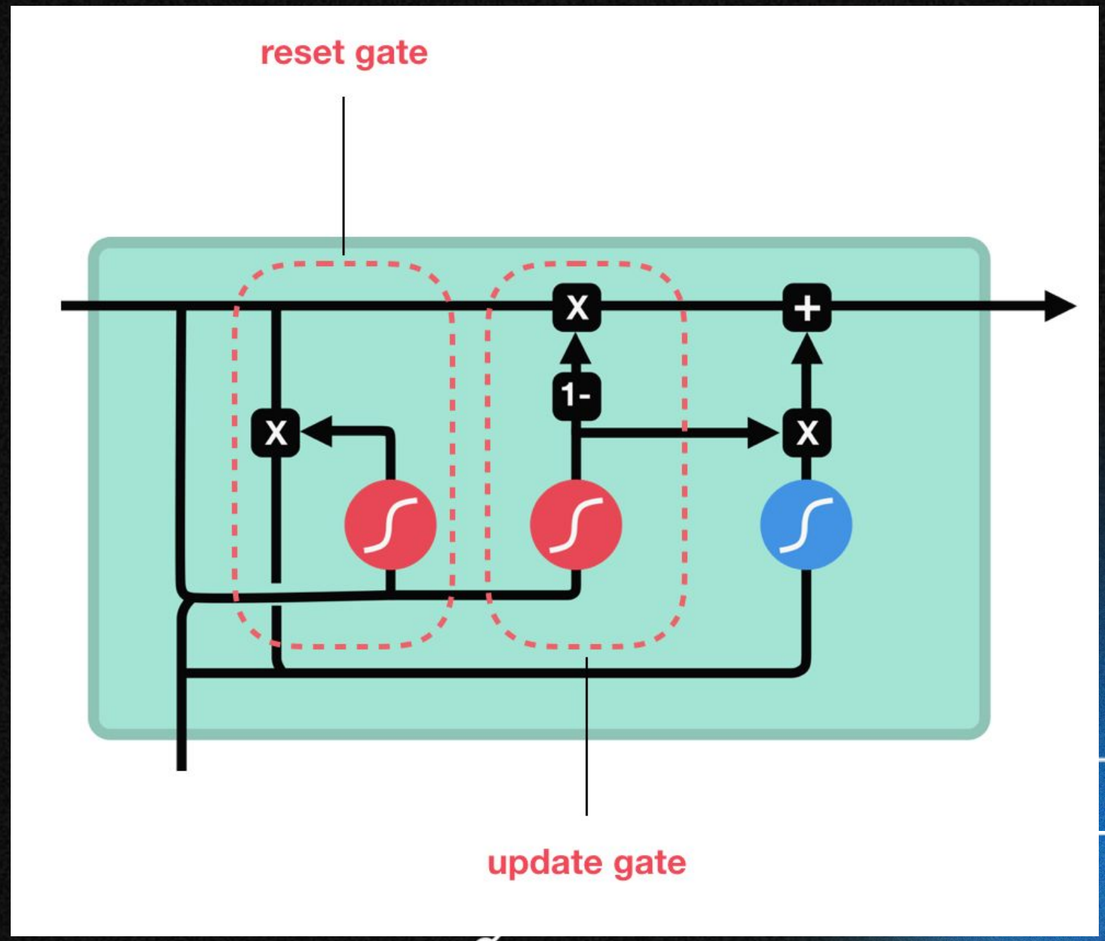
02.

Evaluation

OVERVIEW OF GRU

- A type of recurrent neural network architecture that is designed to capture long-range dependencies on sequential data

- GRU architecture consists of **update gate** and **reset gate** which help solve the vanishing gradient problem of a standard RNN
  - Update Gate: Helps model to determine how much of the past information needs to passed along the future
  - Reset Gate: Decides how much of the past information to forget



ABLATION STUDIES

LEARNING RATE	ACCURACY
0.01	0.840
0.001	0.846
0.0001	0.854
0.00005	0.850

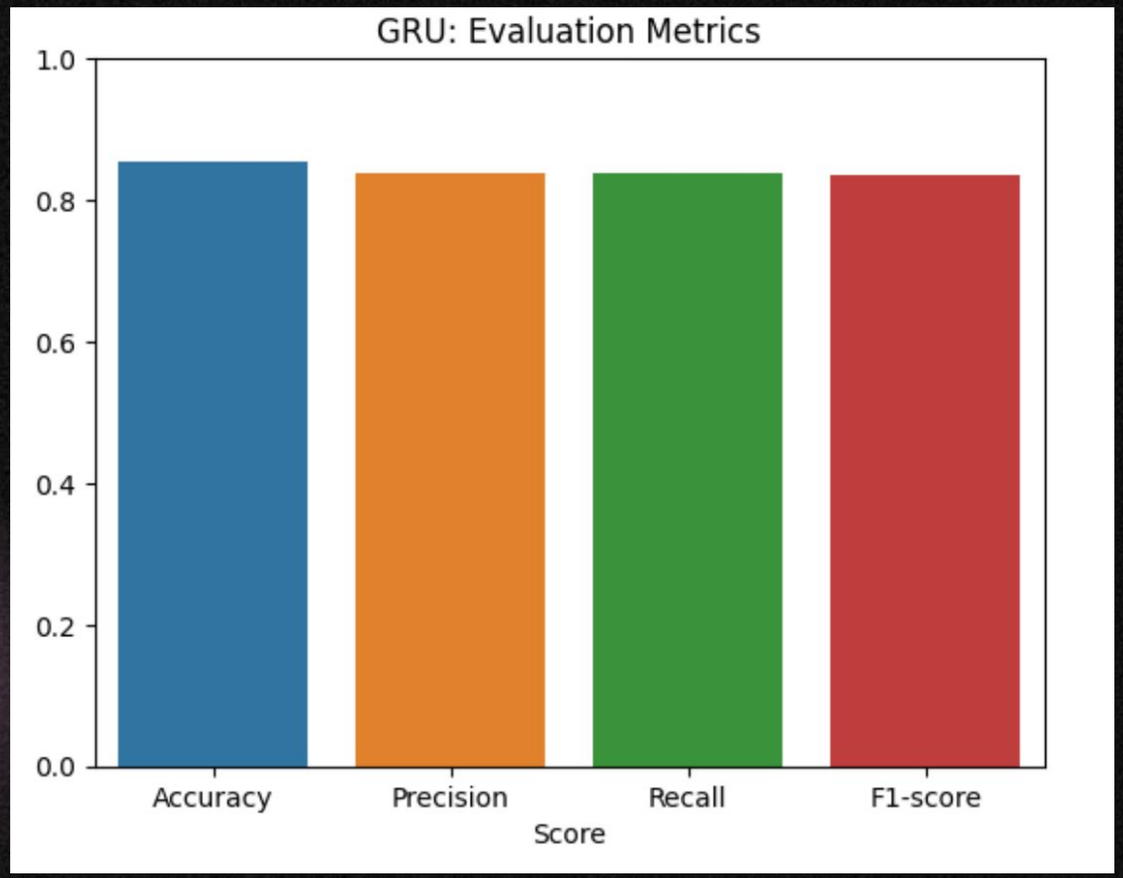
BATCH SIZE	ACCURACY
16	0.856
32	0.854
64	0.851
128	0.846

EPOCHS	ACCURACY
5	0.856
10	0.854
20	0.851

NUMBER OF UNITS IN GRU LAYER	ACCURACY
32	0.829
64	0.854
128	0.845
256	0.856

- Accuracy stayed relatively around the **same** when testing different hyperparameters
- Lower learning rates, lower batch sizes, 5-10 epochs, and 256 units led to better accuracy

FINAL MODEL EVALUATION METRICS



	precision	recall	f1-score	support
not_cyberbullying	0.68	0.58	0.63	1394
gender	0.90	0.86	0.88	1488
religion	0.94	0.96	0.95	1596
age	0.97	0.97	0.97	1536
ethnicity	0.98	0.97	0.98	1610
other_cyberbullying	0.55	0.70	0.61	988
accuracy			0.85	8612
macro avg	0.84	0.84	0.84	8612
weighted avg	0.86	0.85	0.86	8612

- Achieves around 84-85% score for most metrics accuracy, precision, recall, and f1-score. Struggles in predicting **not_cyberbullying** and **other_cyberbullying** targets

BERT MODEL

03.

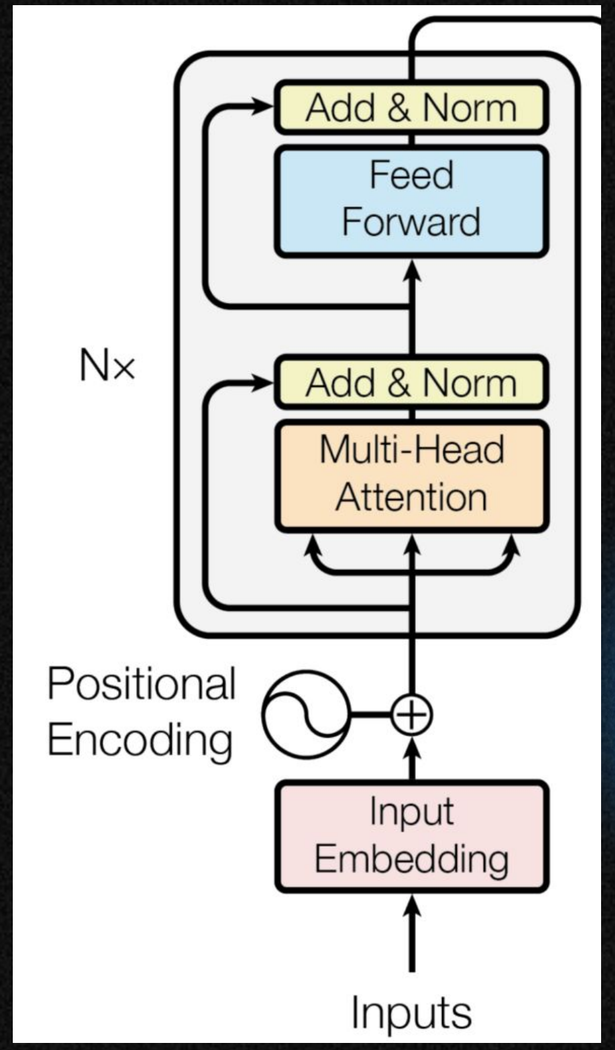
Evaluation

OVERVIEW OF BERT

- **BERT** (Bidirectional Encoder Representation from Transformers) is a transformer based model introduced by Google in 2018

- Designed to understand context of words by considering the **surrounding words** (i.e reads the entire sequence of words at once)

- In our project we use the pretrained ‘bert-base-uncased’ model
  - We utilize the **Trainer** class that the transformer package provides which we use to optimize for training
  - Makes it easier to start training without manually writing your own training loop



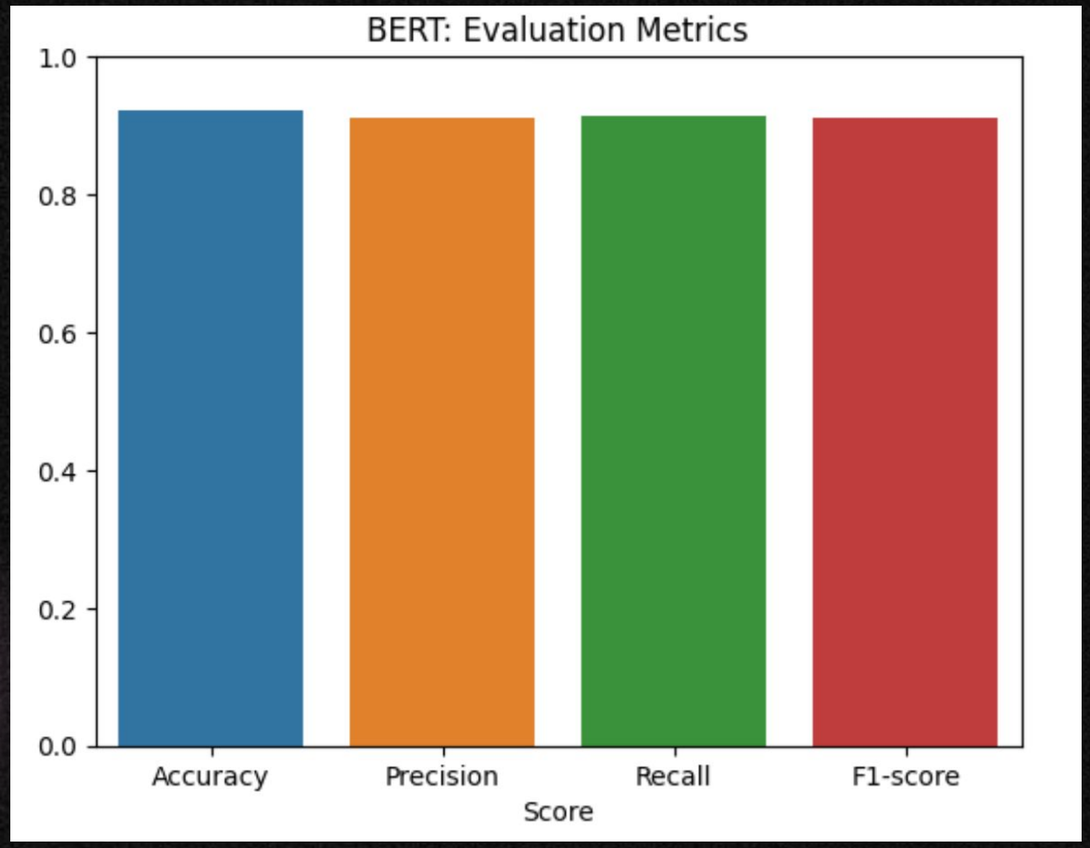
ABLATION STUDIES

Epochs = 4

LEARNING RATE	ACCURACY
6e-6	0.832
2e-5	0.923
5e-5	0.931
1e-4	0.879

- Changing learning rates drastically changed the accuracy of the model
- Low Learning rate (6e-5) possibly led to overfitting data leading to lower accuracy
- Optimal learning rate was 5e-5

FINAL MODEL EVALUATION METRICS



	precision	recall	f1-score	support
not_cyberbullying	0.84	0.74	0.79	1395
gender	0.95	0.93	0.94	1449
religion	0.97	0.98	0.97	1621
age	0.99	0.99	0.99	1600
ethnicity	1.00	0.99	0.99	1599
other_cyberbullying	0.70	0.87	0.78	948
accuracy			0.92	8612
macro avg	0.91	0.91	0.91	8612
weighted avg	0.93	0.92	0.92	8612

- Improves on GRU model, achieves around a 92% score for most metrics. Handles classifying **not_cyberbullying** and **other_cyberullying** targets better

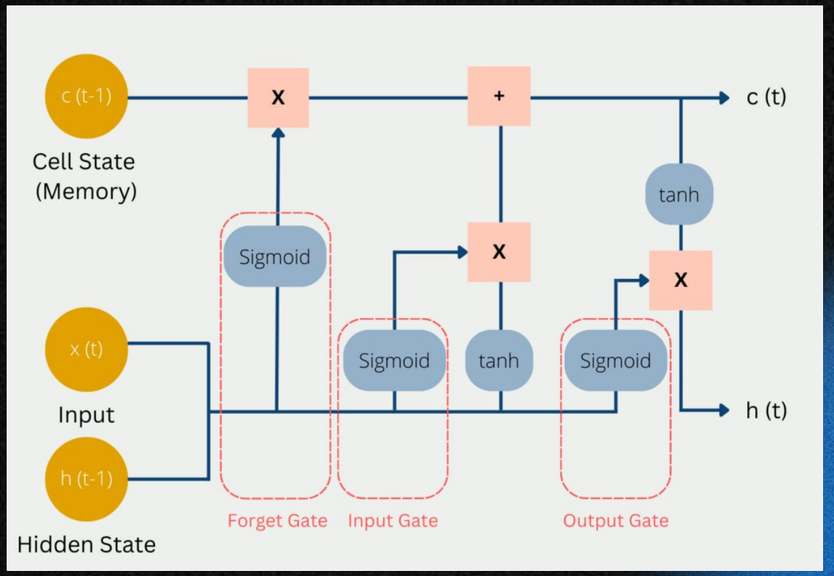
04.

LSTM

Implementation & Evaluation

OVERVIEW OF LSTM

- **LSTM** (Long Short-Term Memory) is another type of RNN for sequential data



- Similar to GRU model but with an additional element:

Forget Gate

Determines <u>which</u> information to forget

vs

Reset Gate

Determines <u>how</u> <u>much</u> information to forget

ABLATION STUDIES

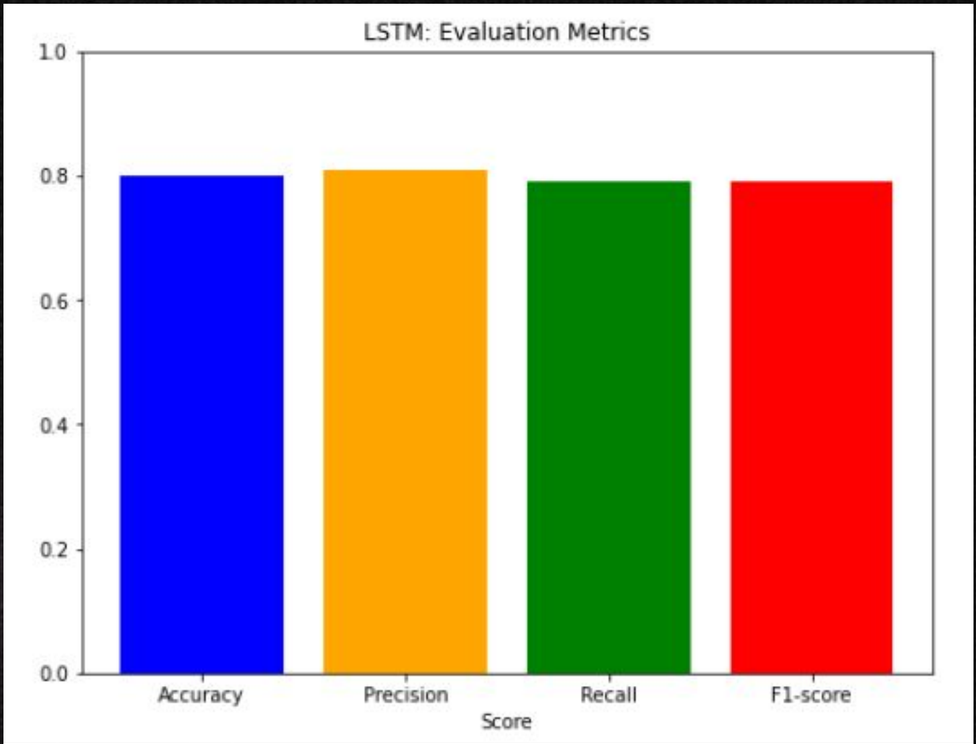
EPOCHS	ACCURACY
5	0.817
10	0.802
20	0.794

- Less epochs, better accuracy
- Otherwise, accuracy fluctuates at different hyperparameters

BATCH SIZE	ACCURACY
16	0.807
32	0.803
64	0.804
128	0.807

NUMBER OF UNITS IN LSTM LAYER	ACCURACY
32	0.812
64	0.806
128	0.808
256	0.810

FINAL MODEL EVALUATION METRICS



	precision	recall	f1-score	support
not_cyberbullying	0.54	0.50	0.52	1448
gender	0.83	0.84	0.84	1530
religion	0.95	0.92	0.94	1594
age	0.97	0.93	0.95	1544
ethnicity	0.97	0.97	0.97	1615
other_cyberbullying	0.50	0.58	0.54	1319
accuracy			0.80	9050
macro avg	0.80	0.79	0.79	9050
weighted avg	0.81	0.80	0.80	9050

- Worst performing model, achieves around a 80% score for most metrics. Similar issues classifying **not_cyberbullying** and **other_cyberullying** targets

05.

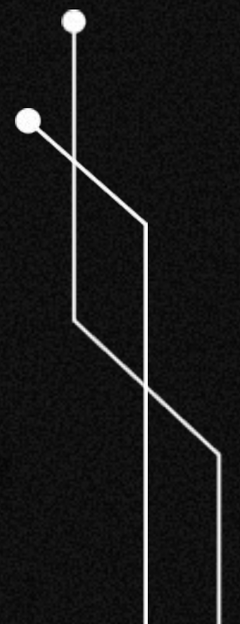
SHAP

Implementation & Evaluation

OVERVIEW OF SHAP

- SHAP (SHapley Additive exPlanations) is a method for explaining the output of machine learning models.

- Model interpretability is crucial for understanding why models make specific predictions, especially for complex models like BERT.



- Key Features:
  - Local & Global Interpretability: Provides explanations for individual predictions & overall model behavior.

  - Model Agnosticism: Applicable to a wide range of machine learning models.

  - Consistency: Meets desirable properties for explanations.
  - Visualizations: Provides various visualization techniques for interpretation.

IMPLEMENTATION OF SHAP

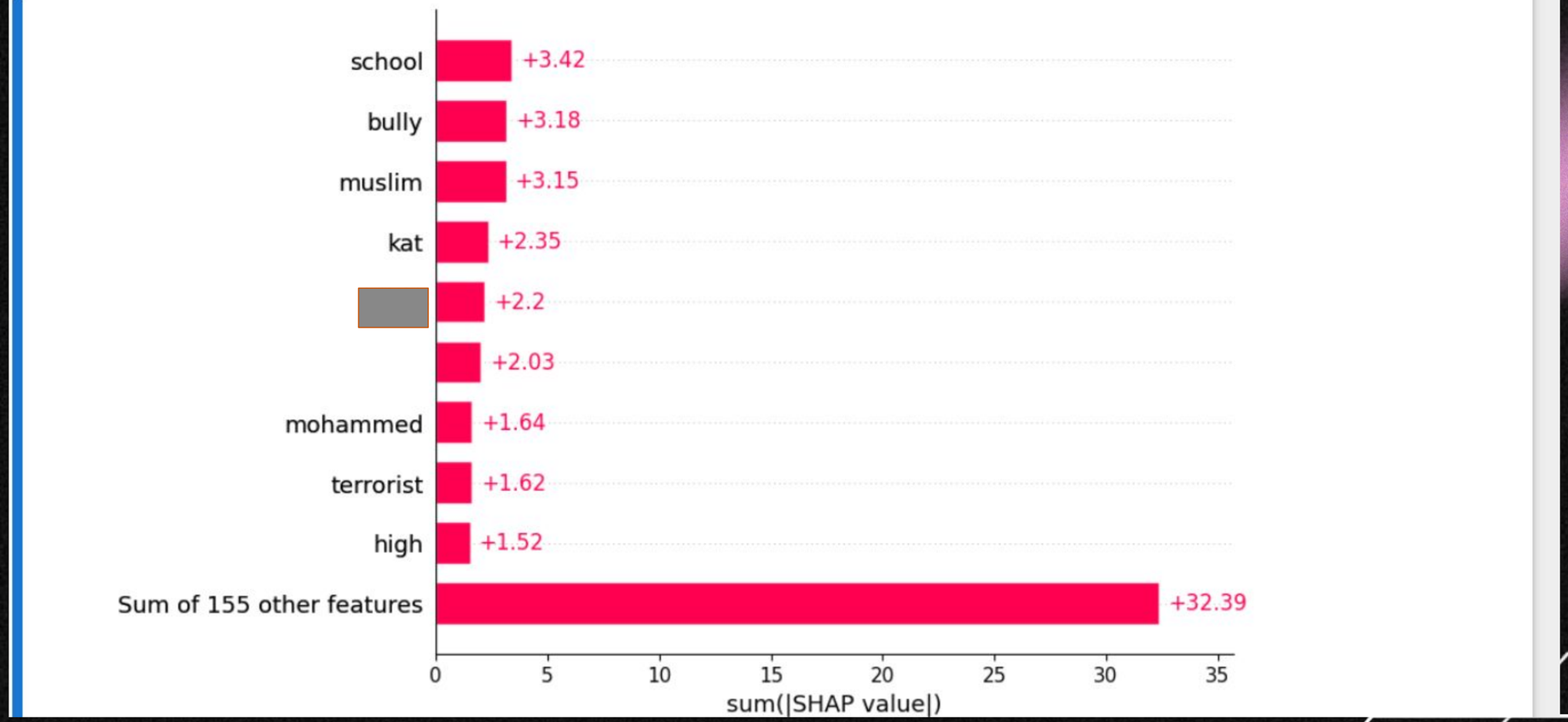
```python
import shap

# Define the prediction function
def f(x):
    inputs = tokenizer(x, padding=True, truncation=True, max_length=100, return_tensors="pt")
    outputs = model(**inputs)[0]
    return outputs.cpu().detach().numpy()

# Create the SHAP explainer
explainer = shap.Explainer(f, tokenizer)

# Explain model predictions on dataset
dataset_texts = list(X_test.values)
shap_values = explainer(dataset_texts)
```

INTERPRETATION OF SHAP

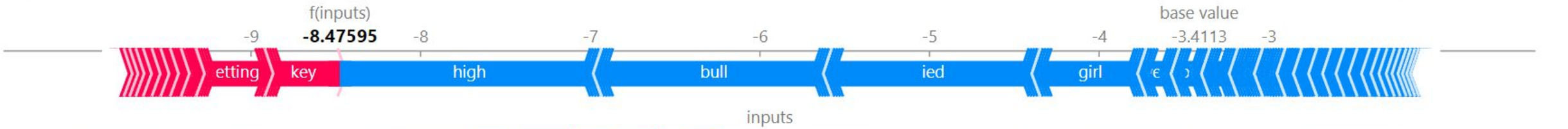
- Words or tokens with **high positive SHAP values** indicate a strong connection to the prediction of cyberbullying behavior. Examples: *“bullying,” “harassment.” and “hate.”*

- Words or tokens with **negative SHAP values** indicate a negative connection, suggesting a lower likelihood of cyberbullying. Examples: *“positive,” “support.” and “community.”*

```
[38]: shap.plots.bar(shap_values.abs.sum(0))
```



```
# plot the first sentence's explanation
shap.plots.text(shap_values[4])

#https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/text.htmlhttps://shap.readthedocs.io/en/latest/example_notebooks/api_examples/
```



happened mcr fandom defended one year old girl getting high key bullied suppoing trump amp boy get shit got kind disgusting thing said u one place fandom

CONCLUSION

06.

Future Steps

FINAL COMPARISON

GRU MODEL

EPOCHS	ACCURACY
5	0.856
10	0.854
20	0.851

BERT MODEL

Epochs = 4

LEARNING RATE	ACCURACY
6e-6	0.832
2e-5	0.923
5e-5	**0.931**
1e-4	0.879

LSTM MODEL

EPOCHS	ACCURACY
5	0.817
10	0.802
20	0.794

PATHWAYS FOR FUTURE

- **Expanding** our dataset.
  - More varied examples of cyberbullying could enhance our model's robustness and accuracy.

- More **Complex** Architectures.
  - Larger Transformer Models could yield better results.

- Integrating our model in **real-world** applications.
  - Using our model in social media platforms where it could work in real time.

- Leveraging LLMs through **prompt engineering**.
  - Open up new opportunities for non-traditional ML approaches.

QUESTIONS?