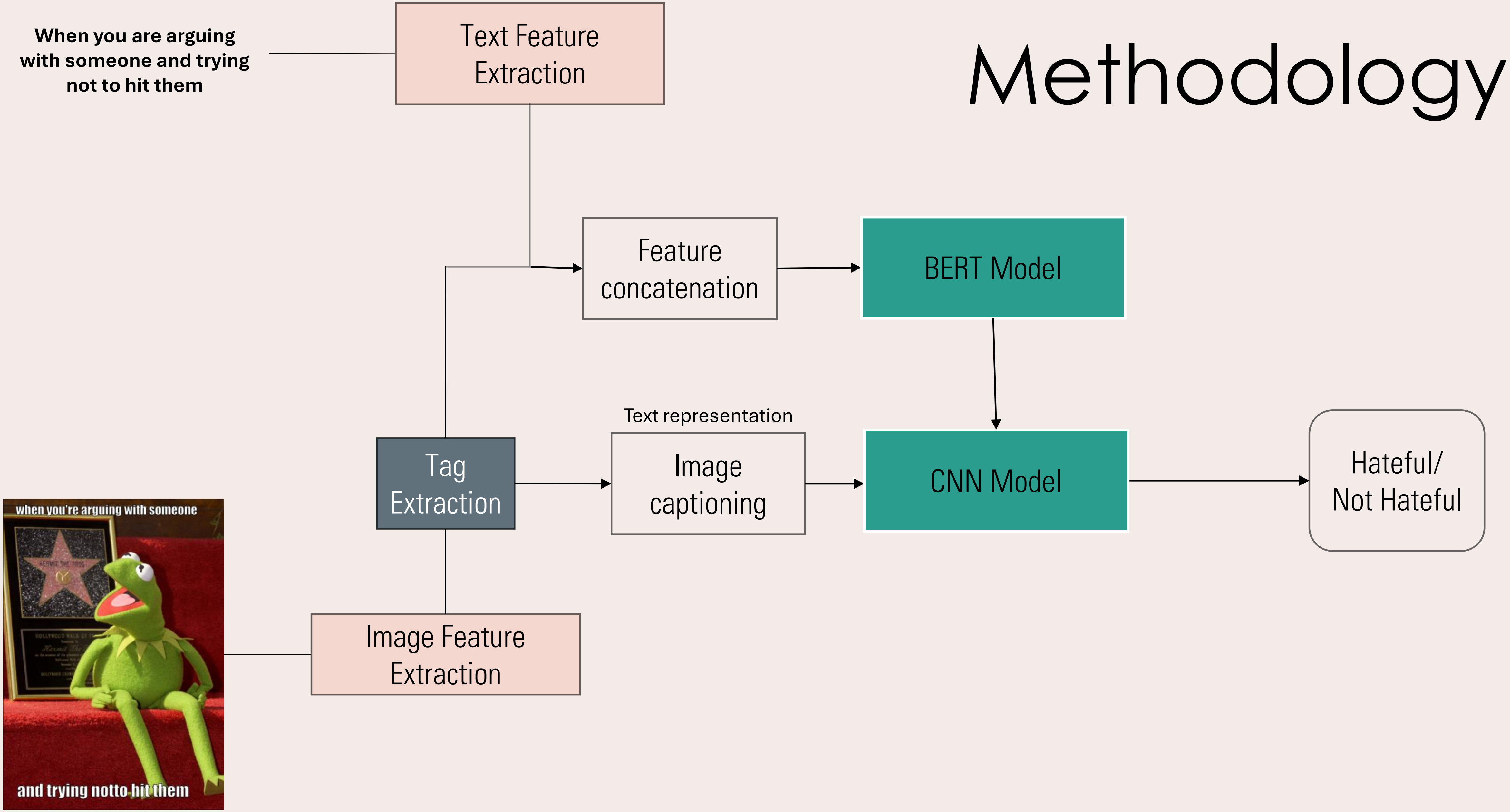
Hateful Meme Detection

Pinal Gajjar
Shilpa Kuppili
Deepshikha Mahato

Methodology



When you are arguing with someone and trying not to hit them

Text Feature Extraction

Feature concatenation

BERT Model

Text representation

Tag Extraction

Image captioning

CNN Model

Hateful/ Not Hateful

Image Feature Extraction

Improvements

Model	Test Accuracy	Precision	F1
VisualBert	0.78	0.64	0.69
Prev VisualBert	0.47	0.35	0.56
RoBERTa	0.66	0.51	0.54
Prev RoBERTa	0.43	0.29	0.60

Summary

- VisualBERT shows the highest accuracy, precision, and F1 score among all models, indicating its effectiveness in detecting hateful memes accurately.

- The previous versions of both VisualBERT and RoBERTa exhibit lower performance across all metrics, highlighting the importance of advancements and updates in model architectures.

- VisualBERT is explicitly designed to handle multimodal inputs by jointly processing text and image data. As the dataset contains complex memes that require understanding both textual content and visual context to detect hateful content accurately, VisualBERT outperformed RoBERTa. Its ability to fuse information from different modalities can lead to a more nuanced understanding of meme content.

Q/A