Tweet Sentiment Extraction Project

# Objective

Extract the portion of a tweet (selected\_text) that reflects the sentiment (positive, negative, or neutral) using Natural Language Processing (NLP) techniques.

# Data

* Train samples
* Test samples
* Columns: text, sentiment, selected\_text

# Preprocessing

* Lowercasing
* Removing links, mentions, punctuation, and numbers
* Stopword removal
* Lemmatization

# Baseline Modeling

- If sentiment = neutral → return full text

- If sentiment ≠ neutral:

* If selected\_text in text → return it
* Else → return first 3 words as a simple approximation

# Evaluation (Jaccard Score)

* Average Jaccard Score: 0.989
* Good predictions: e.g. 'I love it' → 'love it'
* Bad predictions: e.g. 'this is annoying' → 'this is'

# Future Improvements

* Use BERT or deep NLP models for contextual span extraction
* Enhance preprocessing (e.g., handle negations, emojis, spelling correction)
* Visualize attention weights for model interpretability