ASPECT-BASED SENTIMENT ANALYSIS ON INDIAN NEWS ARTICLES (Github)

A) Web Scraping
   1) Source: The Hindu website (past 24 hours) from this link.
   2) Tried Scraping the Times of India from this link, Extracted the articles links. problem faced in extraction of content section.
   3) Scraped URLs of **126 news articles**.
   4) For each article, extracted the title, full content, and publication date.
   5) Stored all data in a structured Excel file.

B) Preprocessing
   1) Combined article title and content for uniform input.
   2) Cleaned text (removal of nulls, extra whitespaces).

C) Article Classification (Zero-Shot Learning)
   1) **Model:** facebook/bart-large-mnli via Hugging Face pipeline.
   2) **Labels Used:** "crime", "fraud", "murder", "scam", "non-crime".
   3) Assigned each article a top label and score.
   4) Marked article as is_crime_related = True if any of the 4 crime-specific **labels had a score ≥ 0.6.**
   5) Filtered down to **45 crime-related articles**.
   6) Stored both
      i) All classified articles → all_classified_articles_zero_shot.xlsx
      ii) filtered crime-related articles → filtered_crime_articles_zero_shot.xlsx

D) NER and Sentiment Analysis
   1) **NER Model:** spaCy en_core_web_sm
   2) **Sentiment Model:** Hugging Face pipeline("sentiment-analysis", model="cardiffnlp/twitter-roberta-base-sentiment"). Classifies into: **positive**, **neutral**, or **negative**
   3) For each crime-related article,identified named PERSON entities.
   4) Extracted all sentences mentioning each person.
   5) **Applied sentiment model on each sentence mentioning the person**
   6) Extracted age if patterns like "22-year-old" were found.
   7) Aggregated sentiment per person per article based on majority vote
   8) Final Data Structure:
      i) name, age (if present)
      ii) matched_sentence, sentiment_label, sentiment_score
      iii) article_url, article_title, article_date
   9) Results:
      i) Total profiles extracted: **301**
      ii) Negative sentiment profiles: **62**
   10) Saved results to:
      i) All profiles → all_profiles_output.xlsx
      ii) Negative-only → negative_profiles_output.xlsx

E) Potential Improvements:
   1) Replace the current zero-shot and sentiment models with ones fine-tuned on legal, crime, or news datasets for better accuracy.
   2) The recognition of person names using spaCy's NER can be improved, possibly by using a more advanced or fine-tuned NER model.