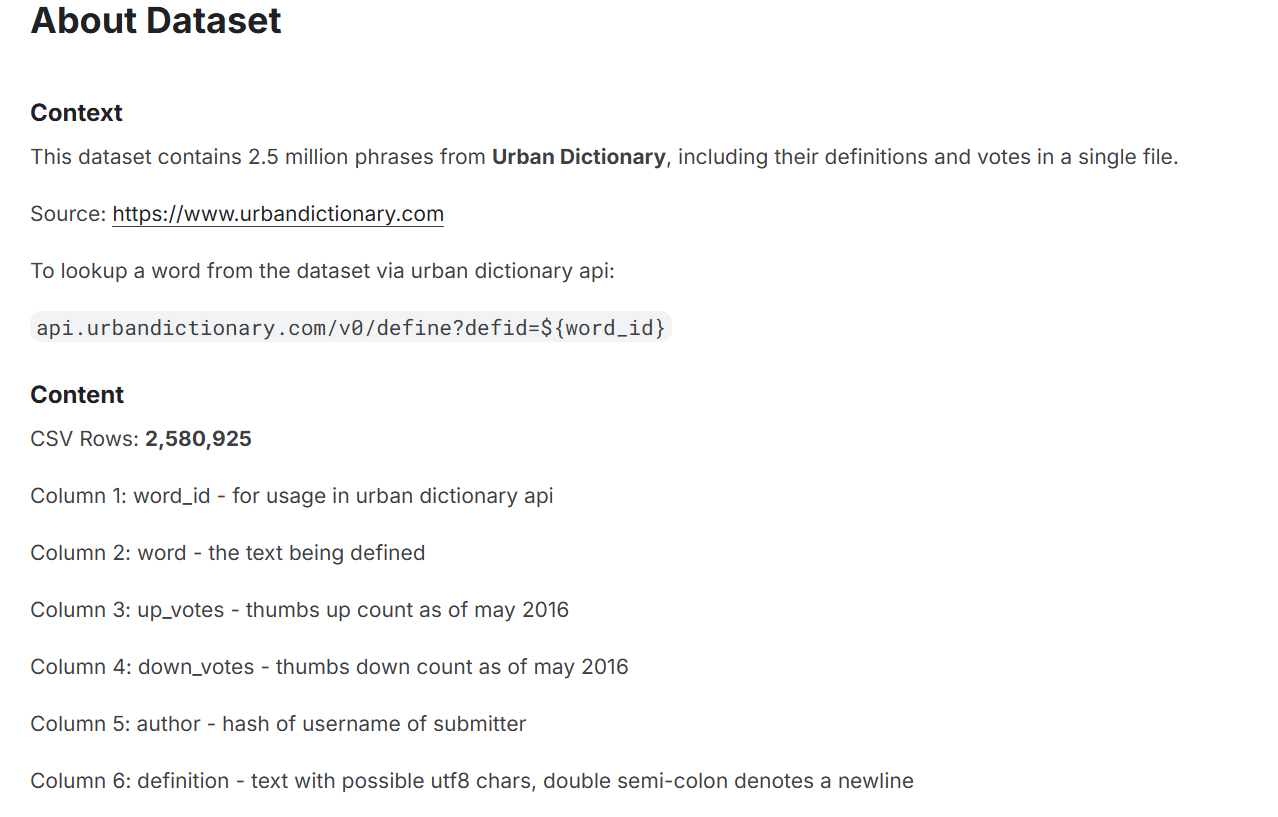
1. **What is the problem you are investigating? Why is it interesting?**

**Problem Being Investigated:**  
This project investigates how multimodal deep learning techniques can be effectively utilized to automatically detect and interpret slang, phrasal verbs, and cultural references within video clips, audio clips, and textual inputs. The objective is to specifically explore the combined effects of video and audio modalities, with a strong focus on modality fusion methods. Traditional language detection approaches typically rely solely on textual data, often missing the nuanced meanings conveyed through vocal intonation, facial expressions, or contextual visual cues. By integrating multiple modalities—textual, audio, and eventually advanced video fusion techniques—this research aims to significantly improve the accuracy and contextual understanding of these informal language elements. However, due to practical timing constraints, the initial phase of the project will primarily focus on fusing audio and textual inputs, establishing a robust multimodal foundation. Incorporation of advanced video modality fusion techniques will be pursued only if time permits, setting the stage for future expansion.

**Why Is It Interesting?**  
This research addresses an important, yet underserved, learning need for non-native English speakers who rely heavily on multimedia content such as TV shows, movies, and online videos to learn conversational English. Slang, phrasal verbs, and cultural references are frequently used in everyday communication but remain challenging for learners due to their informal, contextual, and culturally bound nature. Developing an automated multimodal solution to accurately detect, highlight, and explain these language features would greatly enhance learners' comprehension and retention, transforming passive entertainment into an active language acquisition process. Ultimately, this approach could evolve into a practical educational tool—a Chrome extension—that enables seamless, real-time learning experiences, allowing users to enjoy authentic English content without pausing frequently to take notes or look up unfamiliar phrases.

**2. What dataset will you be using?**

**Urban Dictionary** Dataset (Download Website: https://www.kaggle.com/datasets/therohk/urban-dictionary-words-dataset)



Sample data： A screenshot of a computer

AI-generated content may be incorrect.

**MELD Dataset (Download website: https://github.com/declare-lab/MELD/)**

MELD (Multimodal EmotionLines Dataset) is a publicly available multimodal dataset containing approximately 13,000 utterances extracted from the "Friends" TV series. The dataset provides audio clips, textual transcripts, speaker identification, and annotated emotion labels (such as happiness, sadness, anger, surprise, etc.). It captures rich emotional dialogues in a natural conversational setting, making it highly suitable for multimodal emotion and sentiment analysis tasks, especially when combining audio and textual modalities.

A screenshot of a computer

AI-generated content may be incorrect.

**3. Deep Learning Approach:**

The project uses a staged multimodal deep learning approach:

* **Stage 1 (Initial MVP): Pure text slang classification.**
  + Pre-trained transformer-based language models (**RoBERTa/DistilBERT**).
  + Fine-tune on labeled slang and non-slang text data for binary slang classification.
* **Stage 2 (Extended phase): Audio emotion detection.**
  + Pre-trained audio feature extractor (**Wav2Vec2.0/WavLM**).
  + Train or fine-tune audio classifiers to detect emotional intent in speech segments corresponding to slang phrases.
* **Stage 3 (Final phase, if time allows): Multimodal fusion (Transformer-based Cross-attention).**
  + Develop a multimodal transformer model integrating textual and audio embeddings.
  + Experiment with different multimodal fusion methods (early fusion, late fusion, attention-based fusion) to identify optimal integration.

**4. Anticipated Challenges:**

* **Dataset Quality and Size:**
  + Public datasets may not fully align with slang-focused multimodal analysis.
  + Manual labeling of slang phrases and emotional intent is time-consuming.
* **Multimodal Fusion Complexity:**
  + Effectively integrating audio and text modalities is nontrivial.
  + Optimal fusion architecture (attention-based fusion, concatenation, early/late fusion) needs exploration.
* **Real-time inference constraints (future extension):**
  + Ensuring the model is lightweight and fast enough for real-time Chrome extension deployment.