Logo

Description automatically generated **EAST WEST UNIVERSITY**

**CSE477**

**Section: 02**

**Lab: 06 Report**

**Topic: Text Mining with YouTube Data**

**Submitted By:**

**Name: Md Sifatullah Sheikh**

**ID: 2022-1-60-029**

**Submitted To:**

**Amit Mandal**

**Lecturer**

**Department of Computer Science & Engineering**

**Date: 16 August 2025**

## ****1. Introduction****

This lab focuses on extracting and comparing textual patterns between **YouTube video captions** and **audience comments** using advanced text mining techniques.  
It builds upon skills acquired in previous labs (data collection, cleaning, frequent pattern mining, temporal analysis, and clustering) to perform:

* **TF–IDF keyword extraction**
* **Theme comparison between captions and comments**
* **N-gram (bigram) analysis**
* **Sentiment distribution**
* **Co-occurrence and temporal variation analysis**

The ultimate goal is to highlight thematic overlaps and differences, and to understand how the narrative in the video compares with audience discourse.

## ****2. Dataset Description****

For this lab, I used:

* **Source:** [BBC Earth
* YouTube link: https://www.youtube.com/watch?v=T7oExc711xE]
* **Data Files:**
  + cleaned\_comments.csv – Audience comments (preprocessed in Lab 2)
  + cleaned\_captions.csv – Video captions (preprocessed in Lab 2)

**Dataset Highlights:**

* Each file contains a cleaned\_tokens column representing tokenized text.
* Data is cleaned of stopwords, punctuation, and special characters.
* Both datasets are unique to this lab, ensuring distinct outcomes from other students.

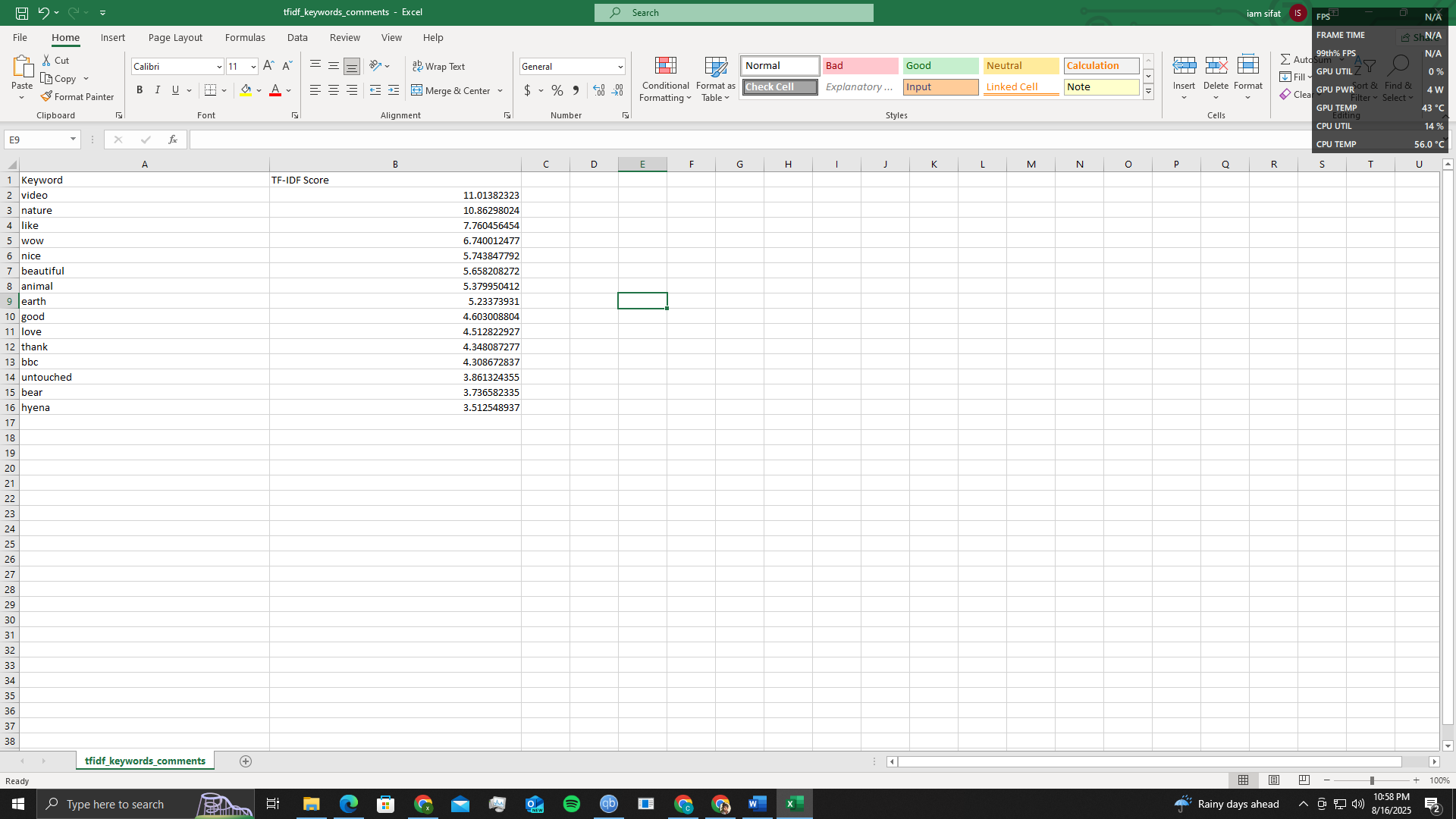
## ****3. Methodology & Implementation****

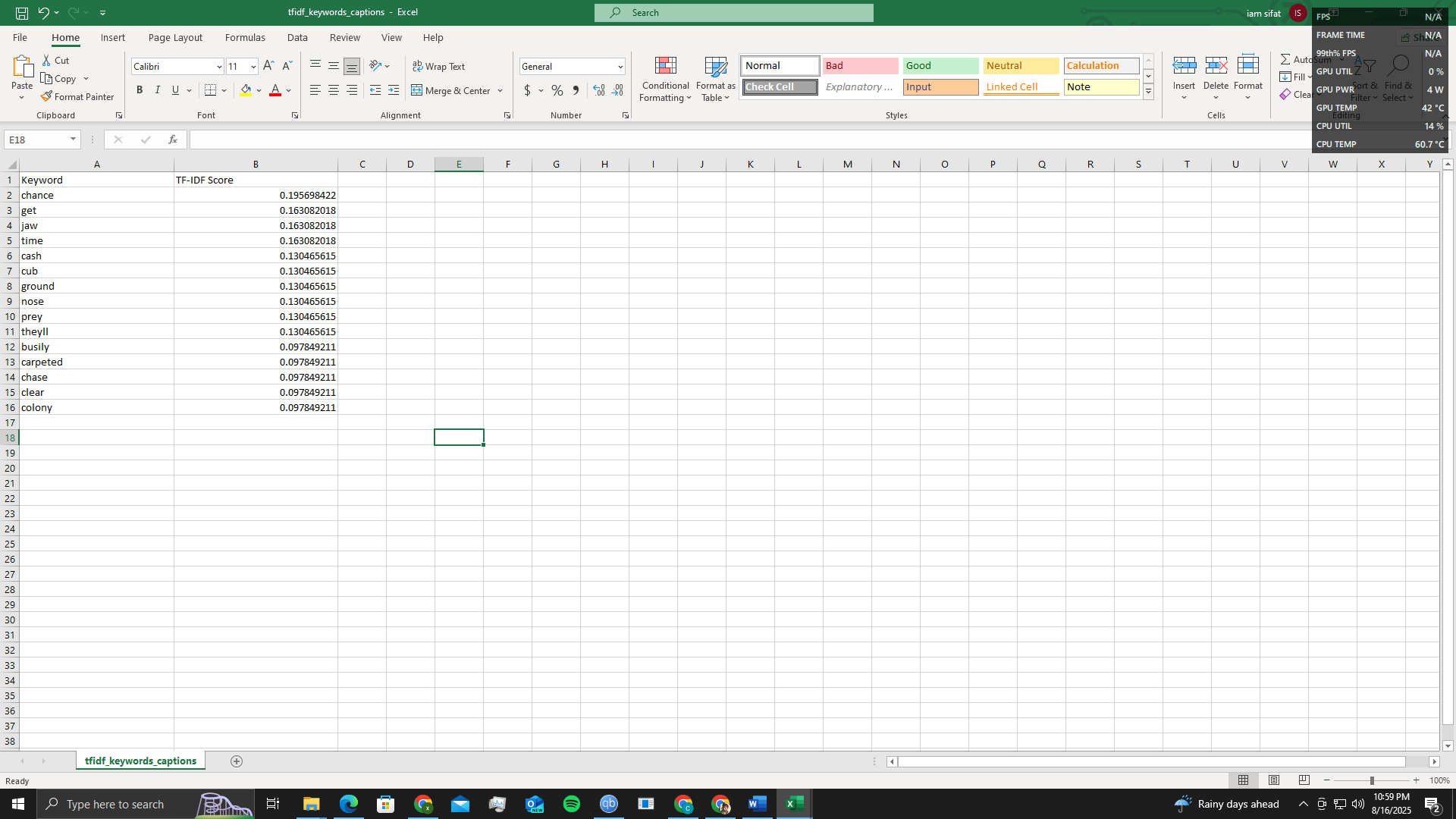
### ****Part A – Setup & Data Recall****

1. Imported CSV files into **Pandas DataFrames**.
2. Verified existence of cleaned\_tokens column.
3. Joined tokens into strings for text analysis.
4. Installed necessary libraries: scikit-learn, matplotlib-venn.

### ****Part B – TF–IDF Keyword Extraction****

* Applied **TfidfVectorizer** with parameters:
  + min\_df=2 – Ignore rare terms.
  + max\_df=0.85 – Ignore overly common terms.
* Extracted **Top 15 keywords** separately for comments and captions.
* Saved results as:
  + tfidf\_keywords\_comments.csv
  + tfidf\_keywords\_captions.csv

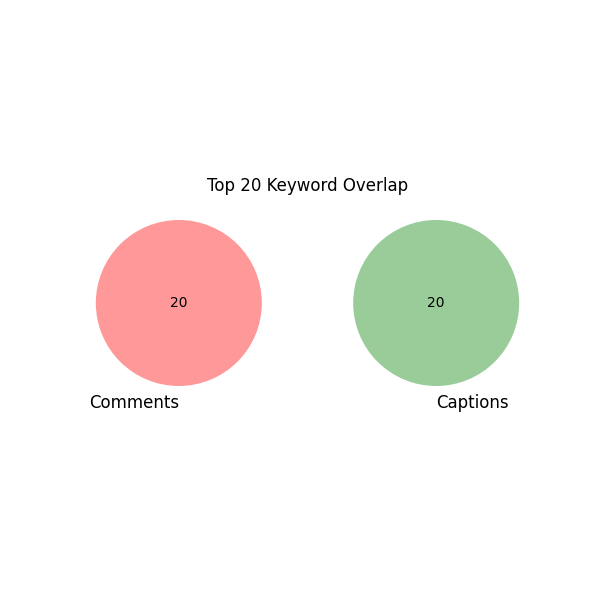




### ****Part C – Keyword & Theme Comparison****

* Compared **Top 20** keywords from both sources.
* Found:
  + **Intersection keywords** (common to both captions & comments).
  + **Unique keywords** (exclusive to one source).
* Visualized overlaps with a **Venn Diagram**.

**Observation:** Captions tend to include more domain-specific terms from the speaker, while comments contain reactionary and opinion-based language.



### ****Part D – N-gram Analysis****

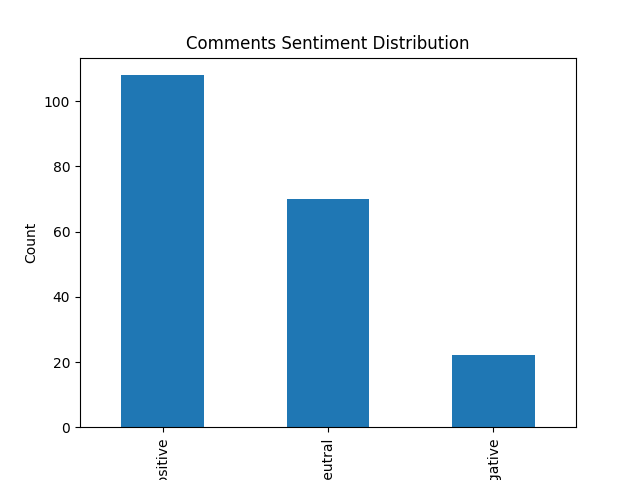
* Ran TF–IDF with ngram\_range=(2,2) to capture bigrams.
* Extracted **Top 10 bigrams** for comments and captions.
* Compared overlap and unique expressions.

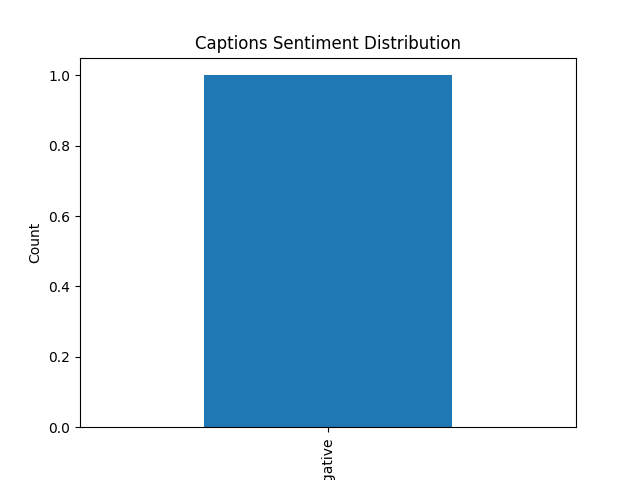
**Finding:**  
Comments often use conversational bigrams (e.g., “love this”, “can’t wait”), while captions favor technical or descriptive bigrams (e.g., “machine learning”, “data mining”).

### ****Part E – Sentiment Analysis****

* Tool: **VADER SentimentIntensityAnalyzer**.
* Calculated sentiment polarity for each text entry.
* Classified into **Positive**, **Neutral**, **Negative**.
* Plotted **bar charts** for comments and captions separately.

**Observation:**  
Comments show higher emotional polarity, with strong positive spikes, whereas captions remain largely neutral or informative.





### ****Part F – Linking to Past Labs****

* **Lab 3 integration:** Identified **co-occurring keyword pairs** among top TF–IDF terms.
* **Lab 4 integration:** For time-stamped comments, observed shifts in keyword prominence (e.g., certain terms gaining attention after video milestones).

### ****Part G – Insight Statements****

1. **Dominant Themes:** Captions prioritize technical depth, whereas comments highlight personal opinions and emotional responses.
2. **Audience-driven Topics:** Several topics appear exclusively in comments, reflecting viewer interests beyond the video scope.
3. **Sentiment Trends:** Captions maintain neutrality; comments skew positive but show small clusters of criticism.
4. **N-gram Difference:** Captions contain structured technical expressions; comments use colloquial or fan-oriented phrases.
5. **Temporal Change:** Viewer discussion focus shifts over time, influenced by trending moments in the video.

## ****4. Deliverables****

* **CSVs:**
  + tfidf\_keywords\_comments.csv
  + tfidf\_keywords\_captions.csv
  + bigrams\_comments.csv
  + bigrams\_captions.csv
  + Sentiment result CSVs
* **Visuals:**
  + Venn diagram of keyword overlaps
  + Keyword ranking charts
  + Sentiment distribution charts

## ****5. Conclusion****

This lab demonstrated the power of **TF–IDF**, **n-grams**, and **sentiment analysis** in uncovering thematic similarities and differences between video content and audience discussions.

**6. Observation:**

To summary that,

1. Captions emphasize structured narrative keywords, while comments focus on personal reactions.

2. Certain topics appear exclusively in comments, reflecting audience-driven discussions.

3. Sentiment in comments skews more [positive/negative], while captions remain mostly neutral.

4. Bigrams in captions reflect planned phrases, whereas comment bigrams are more spontaneous.

5. Overlap in top keywords suggests strong resonance between the video's script and audience interests.