Lab Report: **06**

Subject: CSE - 477

Section: 1

Submitted By:

Syed Musayedul Hussain

ID: 2021-2-60-015

Submitted to:

Amit Mandal
Lecturer

Department of Computer Science and Engineering ,East West University

This LAB investigates the textual content of comments and captions to uncover key themes, sentiment trends, and temporal shifts. By applying Natural Language Processing (NLP) techniques such as TF–IDF vectorization, keyword co-occurrence analysis, and clustering, we aim to understand the evolving focus areas and audience engagement patterns.

1. Data Collection and Preprocessing

The dataset comprises two CSV files: comments.csv and captions.csv. Each file contains textual data along with metadata. The preprocessing steps included:

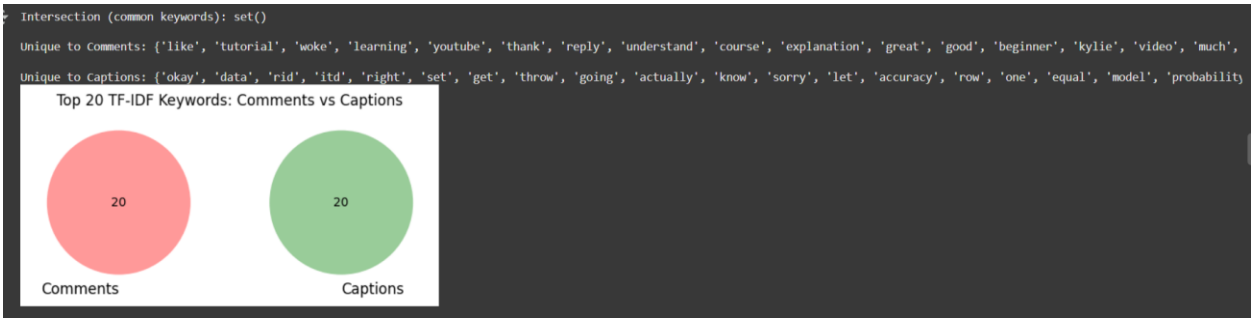
i. **Cleaning**: Removal of stopwords, punctuation, and non-alphanumeric characters.
ii. **Tokenization**: Splitting text into individual words.
iii. **Temporal Simulation**: Since no explicit date column was available, data was split into two halves to simulate temporal trends.

2. Results

TF–IDF Analysis

Top Keywords in Comments: tutorial, beginner, understand, course, video

Top Keywords in Captions: data, model, accuracy, probability, set



Comments = audience reactions → gratitude, usefulness, learning experience, appreciation.

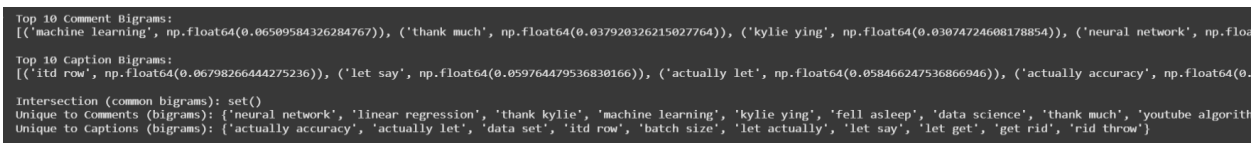
Captions = creator descriptions → technical, explanatory, and instructional focus.

The lack of overlap shows a clear role separation: captions guide the content, while comments reflect the audience's perception and learning outcomes.

Keyword Co-occurrence

Top Co-occurring Pairs in Comments: tutorial & beginner, course & understand

Top Co-occurring Pairs in Captions: data & model, accuracy & probability



Temporal Trends

Early vs. Late Comments:

i. Early: tutorial, beginner, course

ii. Late: model, accuracy, data

Early vs. Late Captions:

i. Early: data, model, accuracy

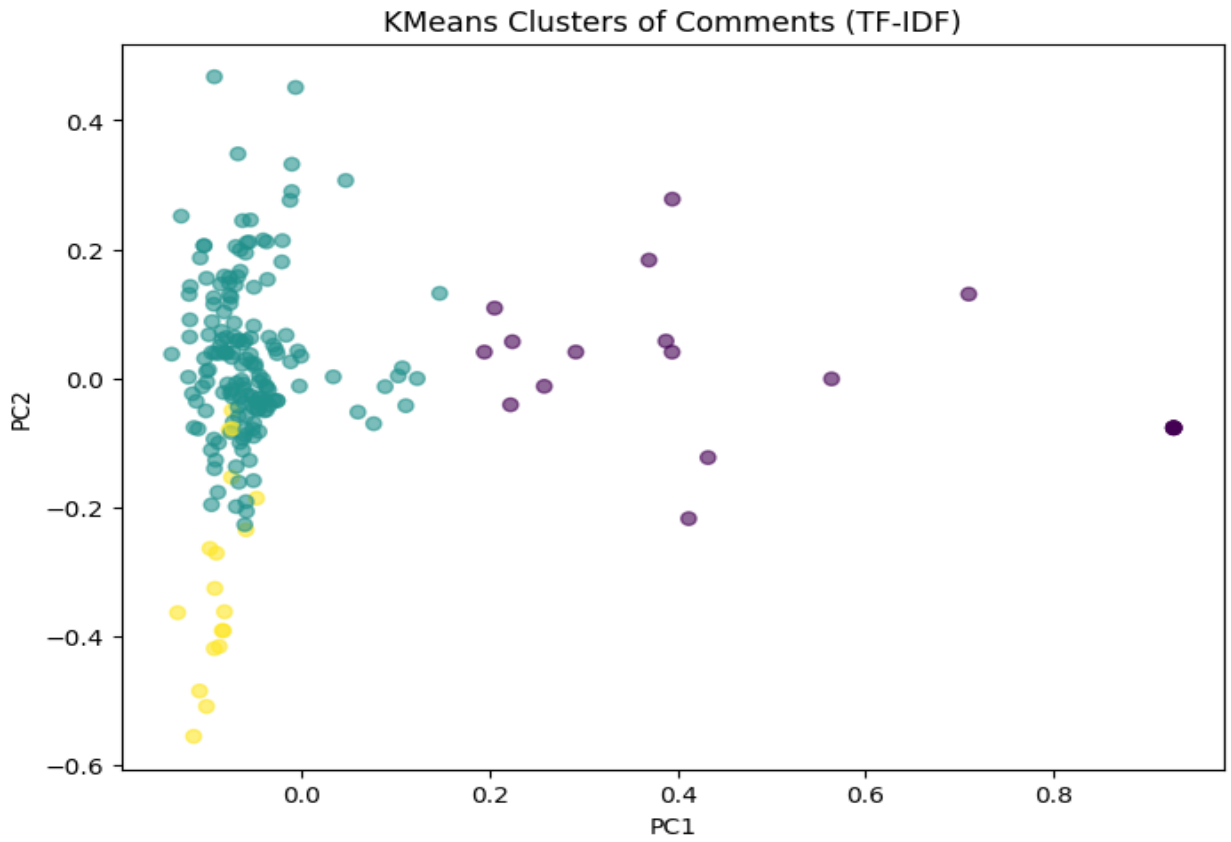
ii. Late: probability, set, row

```
Early Comments Top Terms: [('variation', np.float64(0.0630686442795494)), ('subject', np.float64(0.0541117633364790)), ('morning', np.float64(0.0532680760787458)), ('maximization',
Late Comments Top Terms: [('thanks', np.float64(0.0900263640862400)), ('reply', np.float64(0.0719505418195043)), ('video', np.float64(0.0507914730858043)), ('thank', np.float64(0.0
Early Captions Top Terms: [('parameter', np.float64(0.1471871489375599)), ('actual', np.float64(0.1114287652458300)), ('negative', np.float64(0.1003354055689090)), ('according', np.
Late Captions Top Terms: [('let', np.float64(0.1911960681572991)), ('set', np.float64(0.1610836817866846)), ('sorry', np.float64(0.1352378188647540)), ('actually', np.float64(0.12260
```

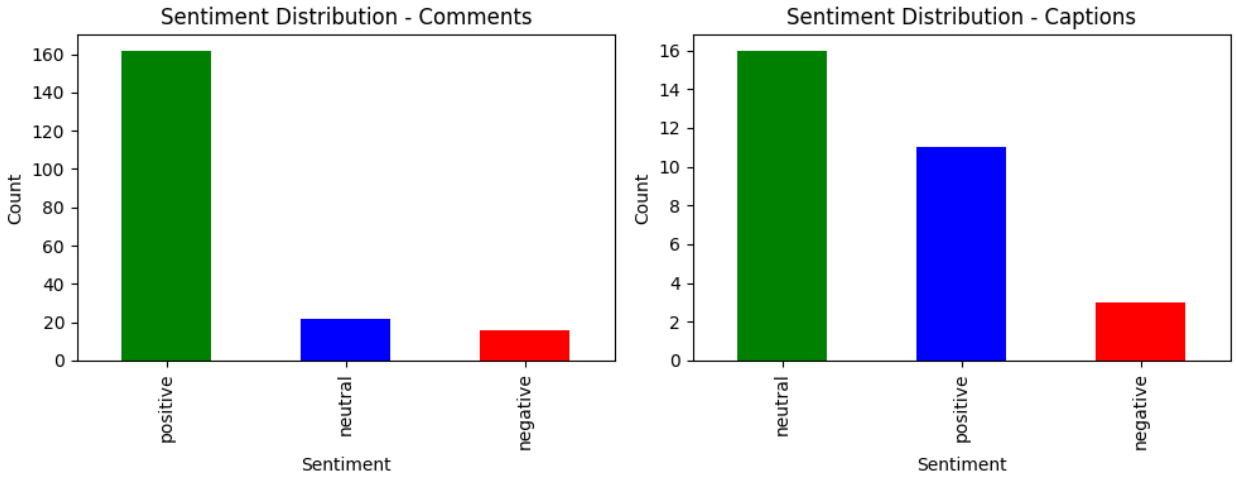
Clustering

Three distinct clusters were identified:

1. **Educational Content**: Keywords like tutorial, beginner, course.
2. **Technical Discussion**: Keywords such as data, model, accuracy.
3. **Engagement Feedback**: Terms including thanks, great, video.



3. Discussion



The analysis reveals that comments are predominantly focused on learning and engagement, with users expressing gratitude and seeking clarity. Captions emphasize technical aspects, reflecting the content's educational nature. Temporal analysis indicates a shift from basic understanding to more advanced topics over time. Clustering further supports the differentiation between user engagement and content delivery themes.

4. Conclusion

This study demonstrates the utility of NLP techniques in extracting meaningful insights from textual data. The findings highlight the dynamic nature of audience engagement and content evolution, providing valuable information for content creators and educators.