

## LAB MANUAL PART A (PART A: TO BE REFERRED BY STUDENTS)

### Experiment No-05

#### A.1 Aim:

Develop Content (text, emoticons, image, audio, video) based social media analytics model for business. (e.g. Content Based Analysis :Topic , Issue ,Trend, sentiment/opinion analysis, audio, video, image analytics)

<b>Lab Objective</b>	To understand the fundamental concepts of social media networks
<b>Lab Outcome</b>	Collect, monitor, store and track social media data

#### A-2 Prerequisite

Data Mining, Data Analytics

#### A.3 OutCome

Students will able to perform exploratory data analysis and visualization on the chosen social media data.

#### A.4 Theory:

**Content-based social media analytics** is a way of analyzing social media data. to gain insight into customer behavior and sentiment. This type of analysis uses machine learning algorithms to extract insights from the text of posts and comments on social media platforms. The model can be used to identify topics, sentiments, and trends in customer conversations, as well as to understand how customers are interacting with brands and products. For example, a business may use a content-based social media analytics model to analyze customer reviews on a product. The model can be used to identify topics, sentiments, and trends in customer reviews. The model can be used to identify positive and negative sentiments in customer reviews, as well as to identify key topics and themes in customer conversations.

#### **Content-Based Analysis:**

**Topic Analysis:** Analyzing the topics discussed in the text, images, and videos based on their relevance to the business. This could include identifying keywords, hashtags, and other related topics to the business.

**Issue Analysis:** Analyzing the issues mentioned in the text, images, and videos that are related to the business. This could include analyzing customer feedback related to products, services, or other aspects of the business.

**Trend Analysis:** Identifying trends associated with the topics and issues discussed in the text, images and videos. This could include analyzing the frequency of certain topics or issues within a given time period.

**Sentiment/Opinion Analysis:** Analyzing the sentiment and opinions expressed in the text, images and videos. This could include analyzing the sentiment of the content and the opinion of the users towards the business.

**Audio Analysis:** Analyzing the audio content associated with the text, images and videos. This could include analyzing the tone, pitch, and volume of the audio content.

**Video Analysis:** Analyzing the video content associated with the text, images and videos. This could include analyzing the visual content, movement, and other elements of the video.

**Image Analysis:** Analyzing the image content associated with the text, images and videos. This could include analyzing the color, size, shape, and other elements of the image.

## PART B (PART B: TO BE COMPLETED BY STUDENTS)

*(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)*

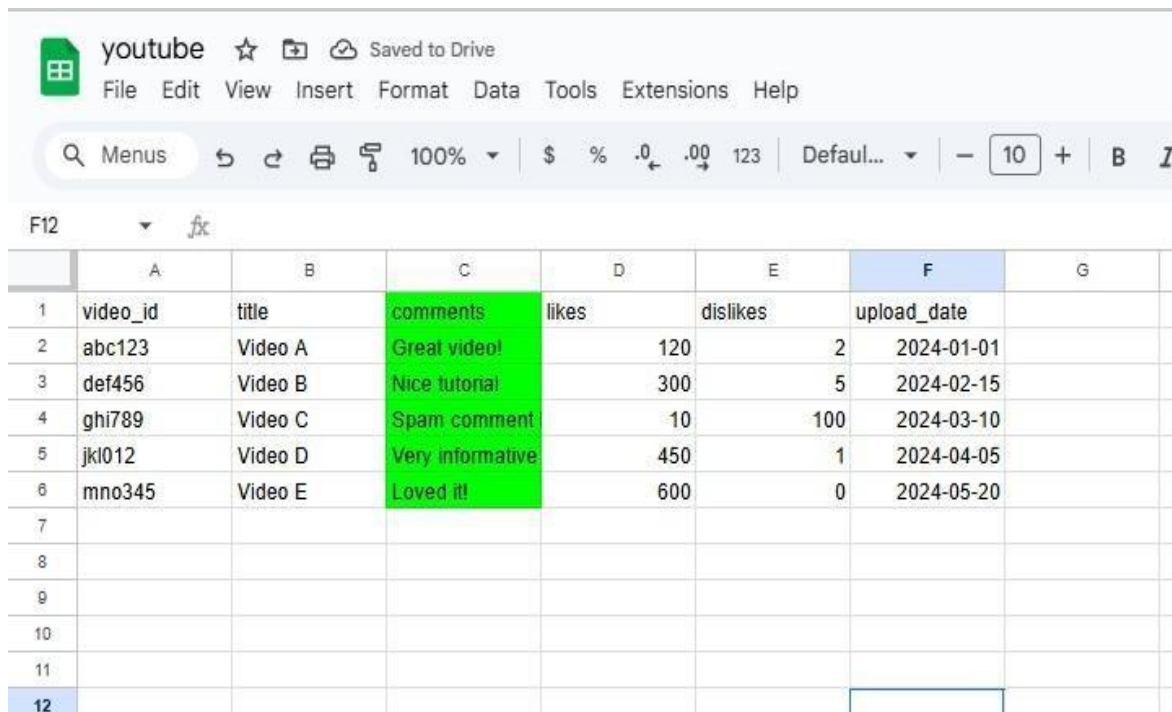
Roll. No: C5	Name: Tavishaa Jaiswal
Class: BE-Comps	Batch: C1
Date of Experiment:	Date of Submission:
Grade:	

**B.1. Students are required to scrape google reviews of any local business and perform**

- 1. topic modeling**
- 2. sentiment analysis**
- 3. Issue analytics by analyzing negative reviews**

**B.2 Input and Output:**

Input:



The screenshot shows a Google Sheets interface with a spreadsheet titled 'youtube'. The spreadsheet has columns A through G. Column A is 'video\_id', B is 'title', C is 'comments', D is 'likes', E is 'dislikes', F is 'upload\_date', and G is empty. The data is as follows:

	A	B	C	D	E	F	G
1	video_id	title	comments	likes	dislikes	upload_date	
2	abc123	Video A	Great video!	120	2	2024-01-01	
3	def456	Video B	Nice tutorial!	300	5	2024-02-15	
4	ghi789	Video C	Spam comment	10	100	2024-03-10	
5	jkl012	Video D	Very informative	450	1	2024-04-05	
6	mno345	Video E	Loved it!	600	0	2024-05-20	
7							
8							
9							
10							
11							
12							

## Python Code:

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from gensim import corpora
from gensim.models import LdaModel
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Download necessary NLTK resources
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('punkt_tab') # This line was added to download the punkt_tab tokenizer

# Load dataset
data = pd.read_csv('/content/drive/My Drive/SMA/youtube.csv')

# Preprocess data
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

def preprocess_text(text):
    text = re.sub(r'\W', ' ', str(text)) # Remove non-word characters
    text = re.sub(r'\s+[a-zA-Z]\s+', ' ', text) # Remove single characters
    text = re.sub(r'^[a-zA-Z]\s+', ' ', text) # Remove single characters at start
    text = re.sub(r'\s+', ' ', text, flags=re.I) # Remove multiple spaces
    text = re.sub(r'^b\s+', '', text) # Remove prefixed 'b'
    text = text.lower() # Convert to lowercase
    tokens = word_tokenize(text) # Tokenize text
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop_words] # Lemmatization and remove stopwords
    return tokens

# Apply preprocessing
data['comments'] = data['comments'].apply(preprocess_text)

# Create dictionary and corpus
dictionary = corpora.Dictionary(data['comments'])
corpus = [dictionary.doc2bow(review) for review in data['comments']]

# Create dictionary and corpus
dictionary = corpora.Dictionary(data['comments'])
corpus = [dictionary.doc2bow(review) for review in data['comments']]

# Train LDA model
num_topics = 3 # Adjust as needed
lda_model = LdaModel(corpus, num_topics=num_topics, id2word=dictionary, passes=15)

# Print topics
for idx, topic in lda_model.print_topics(num_topics):
    print(f"Topic {idx}: {topic}")

# Generate Word Cloud for each topic
for i in range(num_topics):
    words = dict(lda_model.show_topic(i, 20))
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate_from_frequencies(words)

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.title(f"Topic {i+1}")
    plt.show()
```

```

[23] #Visualize topics with word clouds
def visualize_topics(lda_model):
    for i in range(num_topics):
        plt.subplot(1, num_topics, i+1) #This line and the next 3 lines were improperly indented
        topic_words = dict(lda_model.show_topic(i, topn=20))
        wordcloud = WordCloud(width=400, height=300, background_color='white').generate_from_frequencies(topic_words)
        plt.imshow(wordcloud)
        plt.axis('off')
        plt.title('Topic {}'.format(i+1))
    plt.show() #This line should be outside the loop

visualize_topics(lda_model)

# Print topic-word distribution print("Topic-Word Distribution:")
for idx, topic in lda_model.print_topics(-1): print('Topic: {} \nWords: {}'.format(idx, topic))

```

## Output:

```

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data] Package punkt_tab is already up-to-date!
Topic 0: 0.200*"video" + 0.200*"spam" + 0.200*"great" + 0.200*"comment" + 0.050*"informative" + 0.050*"loved" + 0.050*"nice" + 0.050*"tutorial"
Topic 1: 0.235*"tutorial" + 0.235*"nice" + 0.234*"informative" + 0.059*"loved" + 0.059*"comment" + 0.059*"spam" + 0.059*"great" + 0.059*"video"
Topic 2: 0.361*"loved" + 0.092*"informative" + 0.091*"comment" + 0.091*"great" + 0.091*"video" + 0.091*"spam" + 0.091*"tutorial" + 0.091*"nice"

```

Topic 1

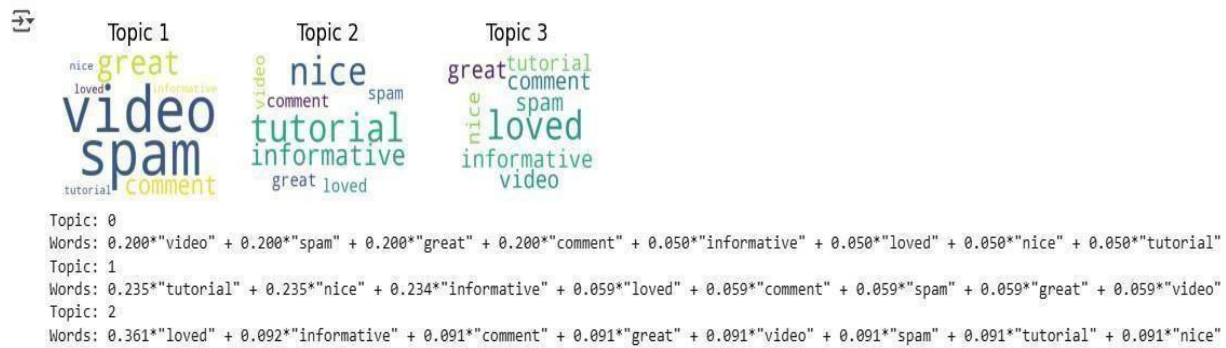


Topic 3

nice **loved** tutorial  
great comment  
informative  
video spam



# Word Cloud



## B.3 Observations and learning:

In our experiment aimed at developing a Content-Based Social Media Analytics Model for business, we focused on analyzing various types of content including text, emoticons, images, audio, and video to extract valuable insights.

Through topic analysis, we identified recurring themes and subjects prevalent in user-generated content, allowing businesses to tailor their strategies accordingly. Additionally, sentiment and opinion analysis provided a deeper understanding of customer attitudes and preferences towards products or services, enabling companies to adapt their marketing approaches effectively.

Our model also incorporated advanced techniques such as audio, video, and image analytics, allowing businesses to glean insights from multimedia content shared on social media platforms.

By leveraging these analytical tools, businesses can gain a comprehensive understanding of their target audience, track emerging trends, and make data-driven decisions to enhance their online presence and drive business growth.

## B.4 Conclusion:

In conclusion, the experiment aimed to develop a comprehensive Content Based Analysis model for social media analytics tailored for business needs. Throughout the experiment, we delved into various facets of content analysis including text, emoticons, images, audio, and video, aiming to extract valuable insights for businesses.

By implementing techniques such as sentiment analysis, topic modeling, and trend detection, we strived to uncover significant patterns, sentiments, and emerging topics within the vast landscape of social media content.

Furthermore, integrating advanced analytics methods for audio, video, and image data allowed for a more holistic understanding of user-generated content.

The culmination of this experiment underscores the potential for businesses to leverage

content-based analytics to gain actionable intelligence, inform decision-making processes, and enhance their overall social media strategies in an increasingly digital and competitive market environment.



## **B.5 Question of Curiosity**

**(To be answered by student based on the practical performed and learning/observations)**

Q1. What is EDA? Explain its importance?

EDA stands for Exploratory Data Analysis. It is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA allows analysts and data scientists to understand the structure of the data, uncover patterns, detect anomalies, and formulate hypotheses for further investigation.

Here are some key aspects of EDA:

1. **Data Summary:** EDA provides summary statistics and visualizations that offer an overview of the dataset, including measures of central tendency, dispersion, and distribution of the variables.
2. **Data Visualization:** Visualizations such as histograms, scatter plots, box plots, and heatmaps are used to explore relationships between variables, identify trends, and detect outliers or anomalies.
3. **Pattern Recognition:** EDA helps in identifying patterns or trends in the data that can provide valuable insights into underlying relationships or phenomena. This can include trends over time, correlations between variables, or clusters of similar data points.
4. **Identifying Outliers:** Outliers are data points that deviate significantly from the rest of the data. EDA techniques help in identifying outliers, which may represent errors in the data or interesting phenomena worth investigating further.
5. **Data Cleaning and Preprocessing:** EDA often reveals missing values, inconsistencies, or errors in the data that need to be addressed before further analysis. Understanding these issues is crucial for data cleaning and preprocessing.
6. **Hypothesis Generation:** Through EDA, analysts can generate hypotheses or conjectures about relationships within the data. These hypotheses can guide more formal statistical analysis or modeling efforts.
7. **Feature Selection:** EDA aids in understanding the importance and relevance of different features or variables in the dataset. This knowledge can inform feature selection or dimensionality reduction techniques for building predictive models.

### Importance of EDA:

1. **Understanding the Data:** EDA helps in gaining familiarity with the dataset, its structure, and the relationships between variables. This understanding is essential for making informed decisions and drawing reliable conclusions from the data.
2. **Quality Assurance:** EDA helps in identifying data quality issues such as missing values, outliers, or inconsistencies, allowing for data cleaning and ensuring the reliability of subsequent analyses.
3. **Insight Generation:** EDA uncovers patterns, trends, and relationships in the data that may not be immediately apparent. These insights can lead to new discoveries or inform decision-making processes in various domains.
4. **Formulating Hypotheses:** EDA provides a basis for generating hypotheses about relationships or phenomena in the data. These hypotheses can guide further investigation or experimentation, leading to a deeper understanding of the underlying processes.
5. **Communicating Findings:** Visualizations and summaries produced during EDA can effectively communicate findings to stakeholders, facilitating decision-making and driving actionable insights from the data.

Overall, EDA is a crucial first step in the data analysis process, laying the groundwork for subsequent analyses and ensuring that data-driven decisions are based on a solid understanding of the underlying data.

### Q2. What is the importance of visualization?

Visualization plays a crucial role in data analysis and interpretation across various domains. Here are some key reasons highlighting its importance:

1. **Understanding Complex Data:** Visualization provides an intuitive way to understand complex datasets by representing data visually through charts, graphs, maps, and other graphical elements. Visualizations help to simplify complex relationships and patterns in the data, making it easier for analysts and stakeholders to grasp important insights.
2. **Identifying Patterns and Trends:** Visualizations enable analysts to identify patterns, trends, and correlations within the data that may not be apparent from raw numbers alone. By visually representing data, trends over time, spatial relationships, and clusters of data points can be easily identified, leading to deeper insights and informed decision-making.

3. **Communication and Presentation:** Visualizations are powerful tools for communicating findings and insights to stakeholders, clients, and decision-makers. Visual representations of data are often more compelling and easier to understand than tables or textual descriptions, making it easier to convey complex information effectively.
4. **Exploratory Data Analysis (EDA):** Visualizations play a central role in exploratory data analysis (EDA), helping analysts to explore the data, detect outliers, and formulate hypotheses. Interactive visualizations allow users to dynamically explore different aspects of the data, enabling more flexible and insightful analysis.
5. **Decision Support:** Visualizations support decision-making processes by providing stakeholders with actionable insights derived from data analysis. By visualizing key metrics, trends, and relationships, decision-makers can make more informed decisions across various domains such as business, finance, healthcare, and science.
6. **Detecting Anomalies and Outliers:** Visualizations help in identifying anomalies, outliers, and unexpected patterns in the data that may indicate errors, fraud, or other unusual phenomena. Visual representations of data make it easier to spot outliers and investigate the underlying causes.
7. **Comparative Analysis:** Visualizations facilitate comparative analysis by allowing users to compare different datasets, groups, or categories visually. By visualizing data side by side, users can identify differences, similarities, and trends across various dimensions, leading to deeper insights and better decision-making.
8. **Data Exploration and Discovery:** Visualizations support data exploration and discovery by enabling users to interactively explore different aspects of the data, drill down into details, and uncover hidden patterns or relationships. Interactive visualizations empower users to ask ad-hoc questions and gain new insights from the data in real-time.

Overall, visualization is an essential tool in the data analysis toolkit, enabling analysts, stakeholders, and decision-makers to gain insights, communicate findings, and make informed decisions based on data-driven evidence.

Q3. Explain the steps involved in EDA?

Exploratory Data Analysis (EDA) involves several steps to understand, summarize, and visualize the main characteristics of a dataset. Here are the typical steps involved in EDA:

1. **Data Collection:**

- Obtain the dataset from various sources such as databases, files, APIs, etc.
- Ensure that the dataset is properly documented, including information about the variables, data types, and any known issues or limitations.

## 2. Data Cleaning:

- Handle missing data: Identify and handle missing values in the dataset through imputation, deletion, or other techniques.
- Handle duplicates: Detect and remove duplicate records from the dataset if necessary.
- Handle outliers: Identify and handle outliers that may skew the analysis or model performance.
- Standardize or normalize data: Scale numerical features to a similar range if needed.

## 3. Data Transformation:

- Feature engineering: Create new features from existing ones to capture additional information or improve model performance.
- Data encoding: Convert categorical variables into numerical representations through techniques like one-hot encoding or label encoding.
- Data aggregation: Aggregate data at different levels (e.g., daily, monthly) for better analysis or modeling.

## 4. Data Visualization:

- Univariate analysis: Visualize individual variables using histograms, box plots, bar plots, etc., to understand their distribution and characteristics.
- Bivariate analysis: Explore relationships between pairs of variables using scatter plots, line plots, correlation matrices, etc.
- Multivariate analysis: Visualize relationships between multiple variables simultaneously using techniques like pair plots, heatmaps, and 3D plots.
- Time series analysis: Visualize temporal patterns and trends in time-series data using line plots, seasonal decomposition, autocorrelation plots, etc.

## 5. Statistical Analysis:

- Descriptive statistics: Compute summary statistics such as mean, median, standard deviation, etc., to describe the central tendency and variability of the data.
- Inferential statistics: Conduct hypothesis testing and confidence interval estimation to make inferences about the population based on sample data.
- Correlation analysis: Measure the strength and direction of linear relationships between variables using correlation coefficients.

## 6. Exploratory Modeling:

- Build simple models or prototypes to explore relationships between variables and test hypotheses.
- Use techniques like linear regression, logistic regression, or clustering to gain insights into the data.
- Evaluate model performance using appropriate metrics and validation techniques.

## 7. Interpretation and Reporting:

- Interpret the findings from the analysis in the context of the problem domain.
- Communicate the results effectively through reports, dashboards, or presentations.
- Provide actionable insights and recommendations based on the analysis to stakeholders.

These steps are iterative and may require revisiting earlier steps as new insights are gained or data issues are identified. EDA plays a crucial role in understanding the data, uncovering patterns and relationships, and informing subsequent analysis or modeling tasks.