**ACHARYA INSTITUTE OFTECHNOLOGY**

Affiliated to Visvesvaraya Technological University, Belagavi, Govt. of Karnataka.

Approved by AICTE, NewDelhi

**Department of**

 **Computer Science & Engineering (Data Science)**

REPORT ON

**TEXT ANALYTICS, WEB MINING AND**

**SOCIAL MEDIA ANALYTICS**

Subject Name: Business Analytics Subject Code: BAD714B

Submitted By:

Pranav BJ  
(1AY22CD038)

Submitted To:

Ms. Surbhi,

Assistant Professor,

Department of Computer Science & Engineering

(Data Science)

**ACHARYA INSTITUTE OFTECHNOLOGY**

Affiliated to Visvesvaraya Technological University, Belagavi, Govt. of Karnataka.

Approved by AICTE, NewDelhi

**Department of**

**Computer Science & Engineering (Data Science)**

Table of Contents

|  |  |  |
| --- | --- | --- |
| SL No | Content | Page No |
| 1 | Text Analytics | 1 |
| 2 | Web Mining Analytics | 4 |
| 3 | Social Media Analytics | 6 |

**PART A – TEXT ANALYTICS REPORT**

**1. Corpus Creation**

* Ten smartphone product reviews were collected from online shopping platforms (Amazon/Flipkart style reviews). The reviews covered various aspects of smartphone performance including camera quality, battery life, processor speed, and overall user satisfaction.
* A screenshot of a computer

  AI-generated content may be incorrect.The data was saved into a CSV file named reviews.csv, with each row representing one unique customer review with columns for Review\_ID and Review\_Text.

**2. Text Preprocessing**

The collected corpus underwent the following Natural Language Processing (NLP) preprocessing steps using Python and NLTK:

**2.1 Tokenization**

Each review was converted to lowercase and split into individual tokens (words) using NLTK's word\_tokenize() function.A screenshot of a computer

AI-generated content may be incorrect.

**2.2 Stop-word Removal**

Common English stop-words such as "is", "the", "and", "this", "for" etc., were removed using NLTK's stopwords corpus to retain only meaningful content-bearing terms.

**Results:** After stop-word removal, **661 tokens** remained, representing a **30% reduction** in noise while preserving semantic content.

**2.3 Lemmatization**

Words were transformed into their base or dictionary forms using WordNet Lemmatizer to reduce vocabulary redundancy:"camera -> Camera

A black screen with white lines

AI-generated content may be incorrect.

**3. Word Frequency Analysis**

After preprocessing, all tokens were combined and analyzed using Python's Counter() function to identify the most frequently occurring terms across the entire corpus.

A black screen with white lines

AI-generated content may be incorrect.

**4. Term–Document Matrix (TDM)**

* A TF-IDF Vectorizer was used to represent the text data numerically and build a Term–Document Matrix (TDM).

A black screen with white text

AI-generated content may be incorrect.

* This matrix shows how important each word is to each review by assigning numerical weights based on word frequency and uniqueness.

**5. Knowledge Extraction – Sentiment Analysis**

Sentiment classification was performed using NLTK’s VADER Sentiment Analyzer.  
Each review was assigned a compound sentiment score and labeled as:

* Positive: Score > 0
* Neutral: Score = 0
* Negative: Score < 0

**Sentiment Distribution**

From the 20 collected comments, the following sentiment distribution was observed:

A screenshot of a computer

AI-generated content may be incorrect.

**6. Interpretation of Results  
A screenshot of a computer

AI-generated content may be incorrect.**

* **Preprocessing Effectiveness:** Text preprocessing successfully transformed unstructured customer reviews into clean, analyzable tokens. Tokenization split sentences into 941 words, stop-word removal reduced noise by 30%, and lemmatization standardized word variations, resulting in 334 unique meaningful terms ready for analysis.
* **Feature Identification:** The word-frequency analysis revealed that "phone" (62 occurrences), "battery" (11), and "camera" (11) are the most discussed product attributes, indicating these features heavily influence customer purchasing decisions and satisfaction levels.
* **Numerical Representation:** The TermDocument Matrix successfully converted textual reviews into numerical vectors using both CountVectorizer and TF-IDF approaches. TF-IDF proved particularly effective in highlighting discriminative terms like "battery", "camera", and "quality" while downweighting common words.
* **Sentiment Patterns:** Sentiment analysis using TextBlob achieved a balanced distribution with 50% positive, 30% negative, and 20% neutral reviews. Positive reviews consistently used words like "excellent", "amazing", "great", while negative reviews emphasized "worst", "poor", "disappointed", demonstrating clear lexical polarization.
* **Business Insights:** The text mining pipeline revealed that while half the customers are satisfied (positive sentiment), 30% express strong dissatisfaction primarily related to quality control and performance issues. Product features requiring improvement include battery life consistency and build quality.
* **Overall Value:** This comprehensive text analytics workflow successfully transformed 10 unstructured smartphone reviews into actionable insights. The combination of preprocessing, TDM construction, and sentiment classification enabled identification of key product strengths (camera, value) and weaknesses (quality control, performance), providing manufacturers with data-driven feedback for product improvement strategies.

**PART B – WEB MINING REPORT**

**Web Content Mining**

**1. Website Selection**

For this task, the official Python.org website homepage was chosen for comprehensive web content analysis. Python.org serves as the official portal for the Python programming language, containing documentation, downloads, community resources, and news updates.

The objective was to extract web content structure, analyze HTML elements, identify hyperlink patterns, and examine textual content through automated web scraping and natural language processing techniques.

URL Analyzed: [https://www.python.org](https://www.python.org/)

Website Type: Educational/Technical Documentation Portal

Content Focus: Programming language documentation, community resources, software downloads  
A screenshot of a computer

AI-generated content may be incorrect.

**2. Content Extraction**

The webpage content was extracted using Python with the following libraries:

* Requests – HTTP library for fetching webpage HTML content with custom headers
* BeautifulSoup (bs4) – HTML/XML parser for DOM traversal and element extraction
* NLTK – Natural Language Toolkit for text tokenization and stop-word filtering
* Pandas – Data manipulation library for tabular result storage and analysis
* Matplotlib – Visualization library for generating charts and graphs
* Regular Expressions (re) – Pattern matching for text cleaning and preprocessing

**3. Extraction Process:**

* HTTP Request: Sent GET request to [https://www.python.org](https://www.python.org/) with custom User-Agent headers
* Response Status: Received HTTP 200 (Success) with 51,159 characters of HTML content
* HTML Parsing: BeautifulSoup parsed the raw HTML into navigable DOM structure
* Content Storage: Raw HTML saved to webpage\_source.html for documentation

Total Content Retrieved: 51,159 characters of HTML markup

A screen shot of a computer

AI-generated content may be incorrect.  
A screenshot of a computer screen

AI-generated content may be incorrect.

**PART C – SOCIAL MEDIA ANALYTICS REPORT**

### ****1. Data Collection****

A total of **20 user-generated comments** related to iPhone 16 Launch were collected from public social media discussions. These comments reflect diverse opinions about pricing, design, performance, and new features.

### ****2. Preprocessing & Sentiment Extraction****

Each comment was cleaned and then processed using **VADER SentimentIntensityAnalyzer**, which assigns a compound polarity score. Based on this score, every comment was categorized into **Positive, Negative, or Neutral** sentiment.

### ****3. Sentiment Distribution****

The sentiment results showed a mix of reactions:

* Positive comments appreciated the **camera upgrades**, **new colors**, and **design refinements**.
* Negative feedback mainly focused on the **high price**, **lack of major innovation**, and **heating issues**.
* Neutral comments expressed balanced opinions without clear sentiment leaning.

### ****4. Visualization & Interpretation****

Pie charts and bar graphs were used to present the sentiment distribution visually. Positive comments formed the largest portion, followed by negative comments, indicating that although the overall reaction was favorable, **pricing remains a major concern among users**.

### ****5. Engagement Trend Observation****

Comments praising visual improvements (camera/design) generated the highest engagement in terms of likes and replies. Negative comments related to high pricing also gained traction, showing **strong price sensitivity and user frustration**.

### ****6. Overall Insight****

The overall sentiment trend suggests that the **iPhone 16 is well-received for its performance and design**, but the **premium pricing continues to divide user opinions**. This mixed sentiment highlights both strong brand loyalty and growing consumer expectations.









