Smart Text Insight Tool

Problem Statement

In today's digital world, massive volumes of textual data—such as reviews, feedback, and reports—are generated daily.
Manually reading and interpreting this data is both time-consuming and subjective.

The objective of this project is to **build an intelligent Natural Language Processing (NLP) tool** that can automatically:

- Clean and process raw text,
- Analyze sentiment (Positive, Negative, or Neutral),
- Extract key insights such as word frequency and overall tone,
- Visualize important keywords using graphs and word clouds, and
- Generate concise summaries from lengthy text documents.

This helps users (such as businesses, researchers, or educators) quickly **understand the underlying meaning and sentiment** of textual data without manual effort.

Methodology

The methodology of this project follows the classical **NLP pipeline**, integrated into an interactive **Streamlit web application**.
The process flow is described below:

Input Collection

The system accepts three types of input:

- Manual text entered by the user,
- Uploaded Excel file containing multiple reviews,
- Uploaded PDF document containing textual content.

Once uploaded, the text from all sources is extracted into a single string for processing.

Text Preprocessing

Before performing analysis, the input text undergoes cleaning and normalization using **NLTK** tools.

Steps include:

- **Lowercasing:** Converting all text to lowercase for uniformity.
- **Noise Removal:** Removing numbers, punctuation, and special symbols.
- **Tokenization:** Splitting text into individual words.
- **Stopword Removal:** Removing common non-informative words (like “the”, “is”, “and”).
- **Lemmatization:** Converting words to their root form (e.g., “running” → “run”).

Basic Text Statistics

Once the text is cleaned, the system calculates:

- **Total Words**
- **Unique Words**
- **Average Word Length**

Sentiment Analysis

The **TextBlob** library is used to compute the emotional polarity of the text.

For each text input (or each review):

- **Polarity** → ranges from −1.0 (Negative) to +1.0 (Positive)
- **Subjectivity** → ranges from 0.0 (Objective) to 1.0 (Subjective)

The app then classifies sentiment as:

- Positive
- Negative
- Neutral

For Excel files with multiple reviews, the tool calculates **overall sentiment distribution** and displays a bar chart comparing the number of Positive, Negative, and Neutral reviews.

Visualization

Two main visualizations are used to represent insights visually:

a) Word Cloud

A **Word Cloud** (using the wordcloud library) highlights the most frequently occurring words in the text.
Larger words indicate higher frequency or relevance.

b) Word Frequency Graph

A **Bar Chart** (using matplotlib) displays the top 10–15 most frequent words and their counts.
This gives users a quantitative understanding of word occurrences.

Text Summarization

To avoid manually reading lengthy content, the tool uses **Sumy (LSA Summarizer)** for extractive summarization.

The summarizer selects the most informative sentences from the text based on Latent Semantic Analysis (LSA).
If the text is short, a preview snippet is displayed instead.

Visualization Interface

All the above processes are wrapped into an interactive **Streamlit web interface**, which displays:

- Basic Statistics
- Sentiment Analysis
- Word Cloud Visualization
- Word Frequency Bar Graph
- Text Summary

This interface is hosted on **Google Colab** using pyngrok, which creates a temporary public URL to access the Streamlit app.

Results and Discussion

The **Smart Text Insight Tool** successfully performs end-to-end analysis of textual data.

Key Results:

- **Accurate sentiment classification** for both short and long reviews.
- **Dynamic visualizations** that make it easy to interpret review trends.
- **Automatic summarization** helps in quickly understanding large volumes of text.
- **User-friendly interface** that requires no programming knowledge.

Feature	Result
Text Cleaning	Removes noise and irrelevant words
Sentiment Analysis	Classifies input into Positive, Negative, Neutral
Visualization	Displays insights using Word Cloud and Bar Chart
Summarization	Produces concise 3–4 sentence summaries
Input Flexibility	Accepts manual text, Excel, and PDF files

Example Output Summary

For the sample text:

“The movie was absolutely wonderful and the acting was top-notch!”

Result:

- **Sentiment:** Positive
- **Polarity:** +0.85
- **Word Cloud:** Highlights “wonderful”, “acting”, “top-notch”
- **Summary:** “The overall tone of the review is highly positive, emphasizing the movie’s excellence and performances.”

Visualization Examples (as seen in Streamlit)

- **Bar Graph:** Displays most common words such as “good”, “amazing”, “bad”, etc.
- **Word Cloud:** Graphically shows dominant keywords.
- **Summary Section:** Condensed interpretation of reviews.

Technologies & Libraries Used

Library	Purpose
Streamlit	Building interactive web interface
Pandas	Reading Excel files and managing data
NLTK	Tokenization, stopword removal, lemmatization
TextBlob	Sentiment analysis
Sumy	Automatic text summarization
Matplotlib	Word frequency bar chart visualization
WordCloud	Generating word cloud images
PyMuPDF (fitz)	Extracting text from PDF files
Pyngrok	Hosting Streamlit app in Google Colab

Conclusion

The **Smart Text Insight Tool** demonstrates how Natural Language Processing can be effectively used to analyze textual data without requiring any dataset or machine learning training.
It integrates various NLP techniques — text preprocessing, sentiment analysis, visualization, and summarization — into one cohesive, interactive system.

This project can be extended further to:

- Support multi-language sentiment analysis,
- Provide downloadable summary reports,
- Integrate advanced deep-learning summarization models (like BERT or T5).