Text Data Analysis (NLP Basics)

Name : Mohd Huzaifa Ammar
Role : Business Analyst Intern
Date : 02/07/2026

1. Introduction to Text Data Analysis

Text Data Analysis is a core component of **Natural Language Processing (NLP)**, a field of Artificial Intelligence that focuses on enabling computers to understand, interpret, and extract meaningful insights from human language. Unlike structured numerical data, textual data is **unstructured**, making it more complex to analyze using traditional statistical techniques.

Organizations generate massive volumes of text data through:

- Customer reviews
- Feedback forms
- Social media comments
- Support tickets
- Survey responses

Analyzing this data helps businesses understand customer perception, improve products and services, and make data-driven strategic decisions.

---

2. Problem Statement

The objective of this task is to analyze customer review text data in order to:

- Identify overall customer sentiment
- Extract frequently occurring keywords
- Understand key drivers of customer satisfaction and dissatisfaction
- Generate actionable business insights

---

3. Dataset Description

The dataset used in this analysis consists of **customer review text**.
Each record represents a single customer's feedback about a product or service.

Dataset Characteristics:

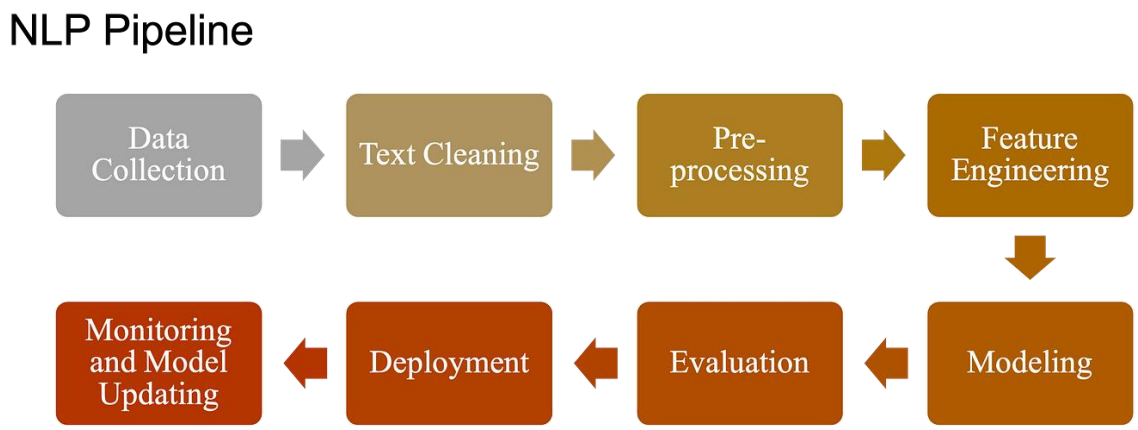
- **Data Type:** Unstructured text
- **Primary Column:** review
- **Source:** Sample dataset (synthetic but realistic)

This type of dataset closely resembles real-world feedback collected by e-commerce platforms, service providers, and SaaS companies.

---

4. Text Preprocessing

Text preprocessing is a **critical step** in any NLP workflow. Raw text contains noise such as punctuation, stopwords, and inconsistent capitalization that can negatively affect analysis results.



4.1 Lowercasing

All text is converted to lowercase to ensure uniformity.
Example:

- "Good" and "good" are treated as the same word

4.2 Tokenization

Tokenization splits sentences into individual words (tokens).
This allows the text to be processed word-by-word instead of as a continuous string.

4.3 Stopword Removal

Stopwords are commonly used words that add little semantic value, such as:

- "is", "the", "and", "was"

Removing stopwords helps focus the analysis on meaningful words.

4.4 Alphabet Filtering

Non-alphabetic characters such as punctuation and numbers are removed to reduce noise.

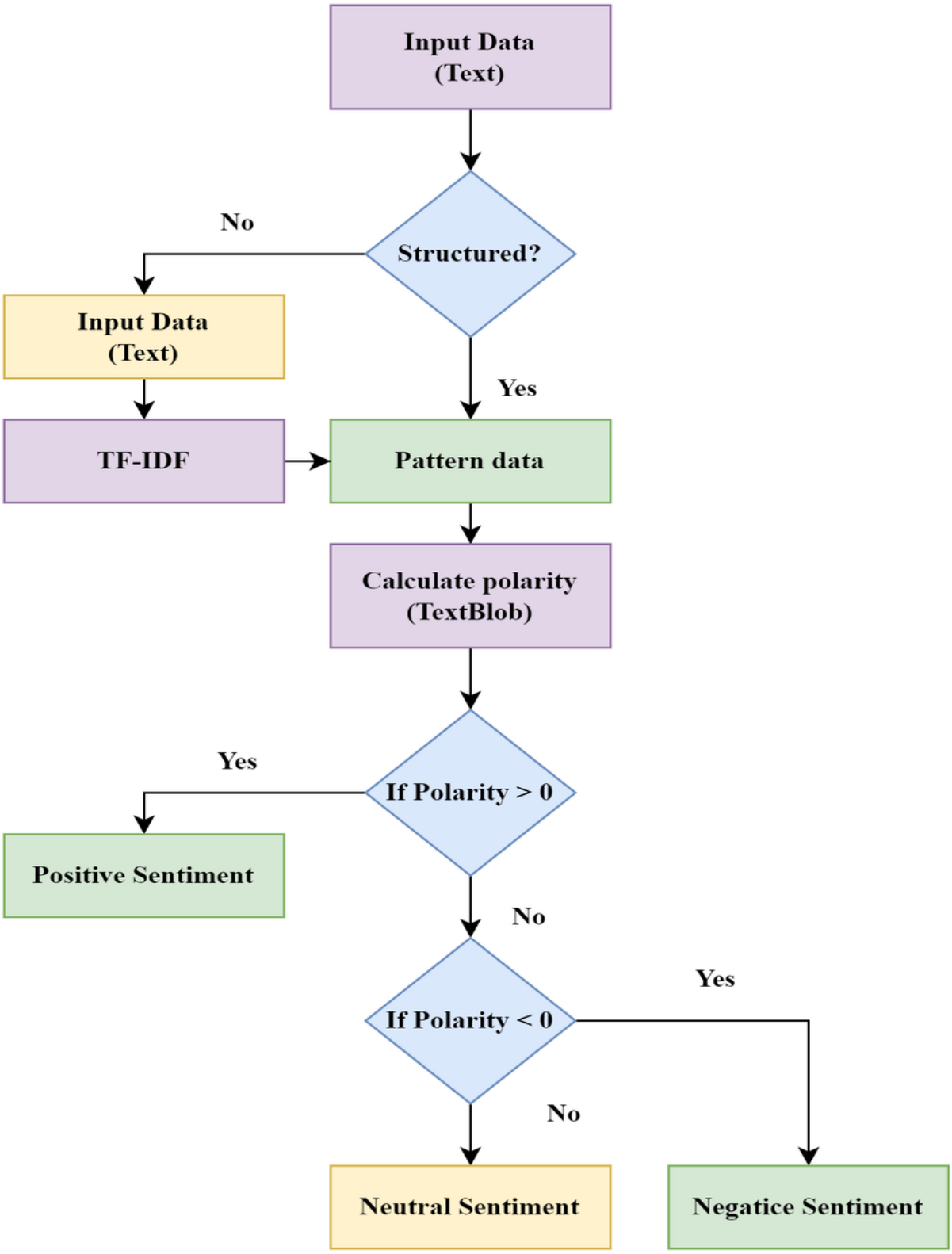
Importance of Text Preprocessing

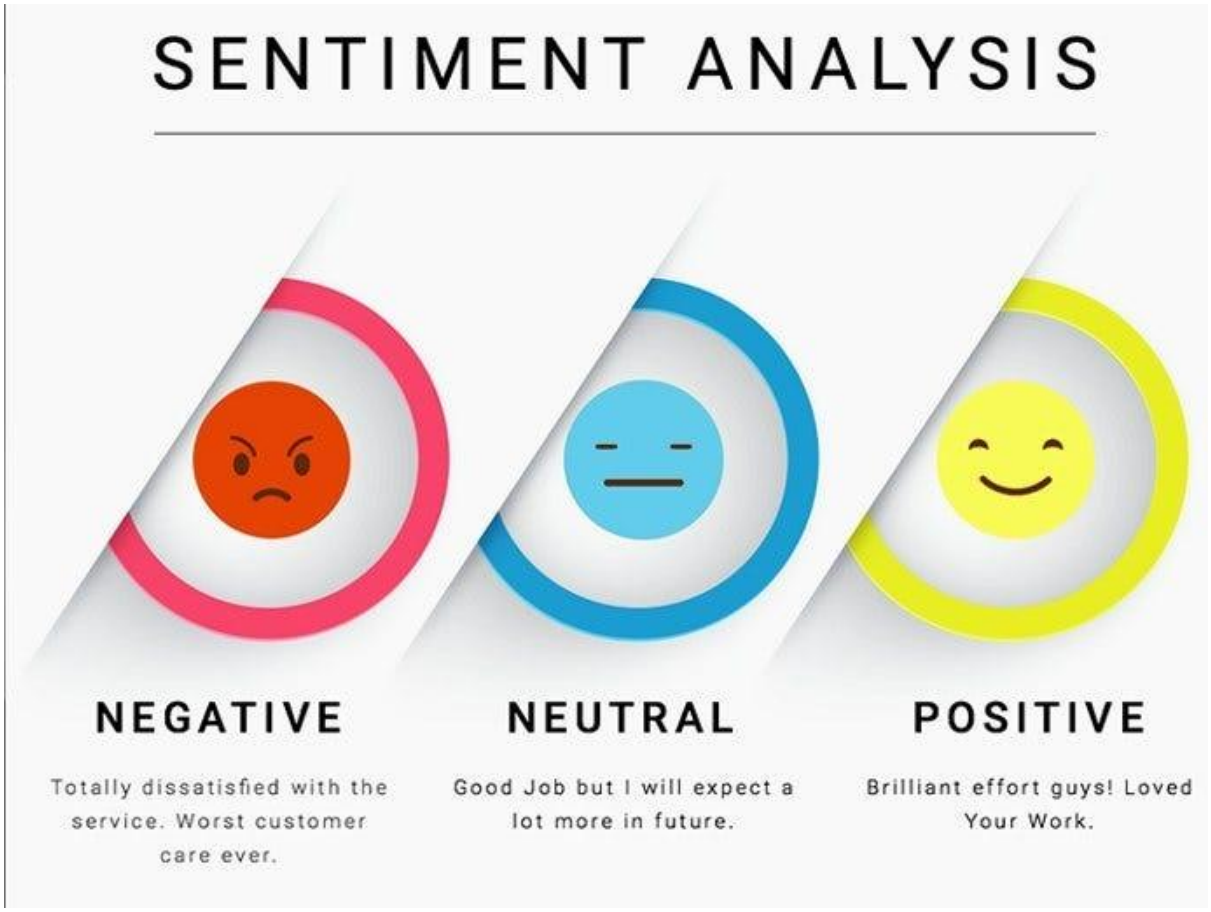
- Improves model accuracy
- Reduces dimensionality
- Enhances keyword extraction
- Ensures reliable sentiment analysis

---

5. Sentiment Analysis

Sentiment analysis is the process of identifying the emotional tone behind a piece of text. It helps determine whether a customer's opinion is **positive, negative, or neutral**.





5.1 Tool Used: VADER Sentiment Analyzer

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based sentiment analysis tool specifically designed for:

- Short texts
- Customer reviews
- Social media content

5.2 Sentiment Scoring

VADER generates a **compound score** ranging from:

- **+1** → Extremely positive sentiment
- **0** → Neutral sentiment
- **-1** → Extremely negative sentiment

5.3 Sentiment Classification Rules

Compound Score Range	Sentiment Label
≥ 0.05	Positive
≤ -0.05	Negative

Compound Score Range	Sentiment Label
Otherwise	Neutral

These thresholds are industry-standard and recommended by VADER documentation.

---

6. Sentiment Distribution Analysis

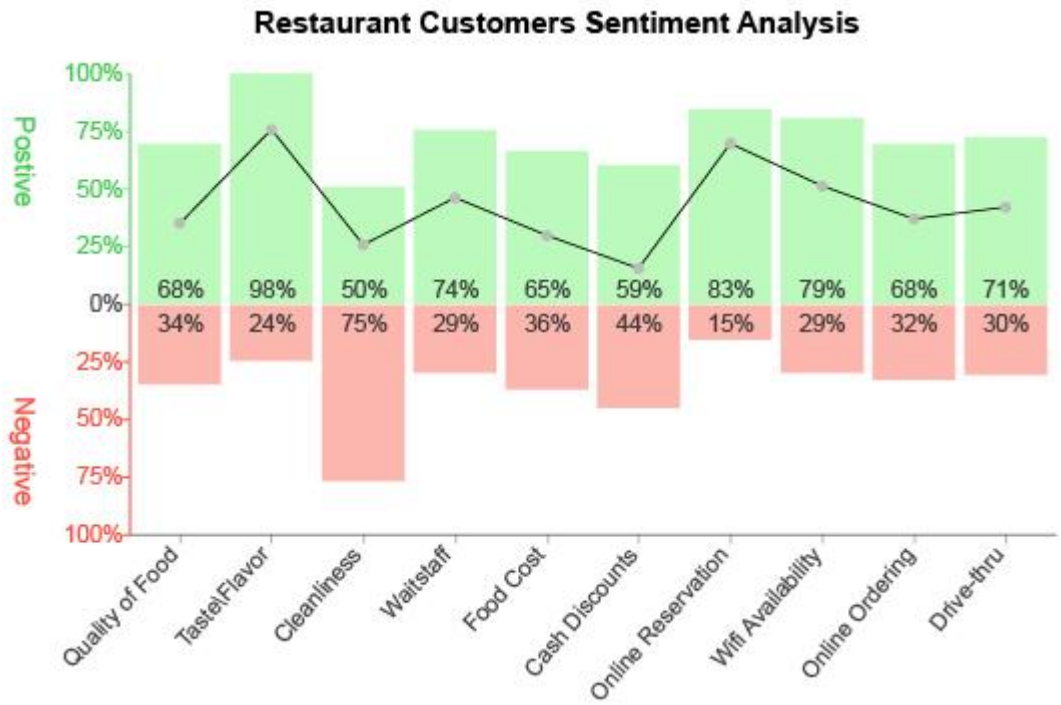
After assigning sentiment labels to each review, a sentiment distribution analysis is performed using a **bar chart**.

Purpose:

- Understand overall customer mood
- Identify the proportion of satisfied vs dissatisfied customers
- Detect potential risk areas

Interpretation:

- A higher number of positive reviews indicates strong customer satisfaction
- Negative reviews highlight areas requiring immediate attention
- Neutral reviews indicate scope for engagement and improvement



---

7. Keyword Extraction

Keyword extraction identifies frequently occurring words in customer reviews.

Technique Used:

- Word frequency analysis after text preprocessing

Why Keyword Extraction is Important:

- Reveals recurring customer concerns
- Highlights product strengths and weaknesses
- Supports product development and marketing strategies

---

8. Word Cloud Visualization

A **word cloud** is a visual representation of text data where:

- Larger words indicate higher frequency
- Smaller words indicate lower frequency

Benefits:

- Quick visual understanding of dominant themes
- Easy to communicate insights to non-technical stakeholders

Observations:

Frequently occurring words such as:

- “quality”
- “delivery”
- “support”
  indicate what customers care about most.

---

9. Keyword Frequency Analysis

In addition to the word cloud, a structured keyword frequency table is generated showing:

- Top 10 most common keywords
- Their respective frequencies

Advantages:

- Quantitative measurement of keyword importance
- Enables prioritization of improvement areas
- Useful for reporting and dashboards

---

10. Actionable Business Insights

Key Insights:

- Overall sentiment is predominantly positive, indicating healthy customer satisfaction
- Negative sentiment is mainly associated with delivery delays and packaging issues
- Product quality and customer support are major contributors to positive feedback

Business Recommendations:

1. Improve logistics and packaging processes to reduce negative experiences
2. Maintain and strengthen customer support operations
3. Emphasize product quality in marketing campaigns
4. Continuously monitor customer sentiment using automated NLP systems

---

11. Conclusion

This analysis demonstrates how NLP techniques can transform unstructured text data into valuable business insights. By applying systematic text preprocessing, sentiment analysis, keyword extraction, and visualization techniques, organizations can gain a deeper understanding of customer opinions and make informed, data-driven decisions.