Sentiment Analysis Using Web Scrapped Data

Wajiha Fatima[*1], Swapnil Singh[*2], Dr. Vibha Srivastava[*3], Er. Aleesha Khan[*4] Students, Department of Information Technology, Shri Ram Swaroop Memorial College of Engineering and Management, Lucknow, U.P., India.

[*3]Professor, Department of Information Technology, Shri Ram Swaroop Memorial College of Engineering and Management, Lucknow, U.P., India.

[*4]Professor, Department of Information Technology, Shri Ram Swaroop Memorial College of Engineering and Management, Lucknow, U.P., India.

Abstract - This study presents a web-scraped data-driven Sentiment Analysis System designed to extract and evaluate public opinions from diverse online platforms. Built with robust data collection techniques, the system gathers real-time textual data, which is then processed using Natural Language Processing (NLP) models to determine sentiment polarity—positive, negative, or neutral. The project emphasizes the structuring of large datasets, efficient data cleaning methods, and the application of machine learning algorithms to ensure accurate sentiment classification. Furthermore, it explores challenges in handling noisy data, managing multilingual content, and optimizing model accuracy. This analysis provides valuable insights for businesses, social researchers, and policymakers to understand public perception and trends.

Keywords: Sentiment Analysis, Web Scraping, Natural Language Processing, Real-Time Data, Machine Learning, Opinion Mining.

1. INTRODUCTION

In the digital age, understanding public opinion has become a vital aspect for businesses, policymakers, and researchers to make informed decisions. Traditional methods of sentiment analysis, such as surveys and interviews, are time-consuming, costly, and limited in scope. With the explosion of social media platforms, news websites, and online forums, vast amounts of public opinions are now available in real-time. However, extracting meaningful insights from this unstructured data remains a challenge for many organizations[1].

To address these challenges, we propose a Sentiment Analysis System powered by web-scraped data and advanced Natural Language Processing (NLP) techniques. The system is designed to:

- Collect real-time textual data from diverse online sources
- Analyze public sentiment through machine learning algorithms
- Categorize opinions into positive, negative, or neutral sentiments
- Generate insightful visualizations for better decision-making

By leveraging modern web scraping technologies and robust NLP models, the proposed system enables efficient sentiment extraction at scale, providing real-time insights into consumer opinions, market trends, and public perception[3].

This study explores the technological stack, system architecture, data preprocessing techniques, sentiment classification algorithms, and the challenges encountered during implementation.

2. METHODOLOGY

1. Technology Stack Overview: The proposed **Sentiment Analysis System** is built using a robust technology stack that enables seamless data collection, processing, and analysis from online sources. This stack ensures real-time sentiment detection and scalable data handling, making it efficient for large-scale applications.

- **Web Scraping Technologies:BeautifulSoup**: Python libraries for extracting structured data from static web pages. Ideal for gathering reviews, comments, and articles.

Selenium: A web automation tool that allows scraping of dynamic websites requiring JavaScript

rendering, enabling real-time data collection from interactive pages.

Processing and Storage:

Pandas & NumPy: Essential for data manipulation, cleaning, and transformation, allowing quick preparation of datasets for analysis.

Natural Language Processing (NLP):

NLTK : Libraries used for text processing, tokenization, lemmatization, and part-of-speech tagging. These tools prepare raw text for sentiment analysis.

VADER: A simple yet powerful sentiment analysis tool for quick polarity detection**.**

- **Backend Framework:**

Flask : Lightweight web frameworks responsible for handling API requests, managing user interactions, and communicating with the database seamlessly.

- **Frontend Technologies:**

React.js: Provides an interactive and dynamic user interface for real-time sentiment visualization and analysis reports.

HTML & CSS: Used for structuring and styling the frontend to ensure a user-friendly experience.

JavaScript: Enhances the interactivity of the web application, allowing smooth user navigation and data visualization.

By integrating these technologies, the system offers a complete solution for web scraping, data processing, sentiment analysis, and visualization. The architecture is designed for scalability, real-time processing, and an interactive user experience, making it a powerful tool for understanding public opinion from online platforms.

System Workflow

The Sentiment Analysis System follows a structured and well-organized workflow to ensure efficient data collection, processing, and sentiment evaluation. By automating key steps, the system minimizes manual intervention, accelerates data handling, and enhances the accuracy of sentiment analysis. Below are the primary steps involved*:*

1. **Data Collection and Web Scraping**: The system begins by scraping real-time data from various online platforms such as news websites, social media, and forums. Web scraping technologies **like** BeautifulSoup**,** Scrapy**,** and **Selenium** are employed to extract user comments, reviews, and posts efficiently.

2. **Data Preprocessing:** Collected data is often unstructured and noisy. The system cleans and preprocesses this data using **Pandas** and **NumPy**, performing tasks like text normalization, removal of stop words, and handling of missing values. This ensures that the data is well-structured for sentiment analysis.

3. **Sentiment Analysis:** After preprocessing, the cleaned text is fed into sentiment analysis models. Initially, tools like **VADER** perform quick sentiment scoring. It is used to classify sentiments as *Positive*, *Negative*, or *Neutral*. This multi-layered analysis ensures higher accuracy and contextual understanding.

4. **Admin Dashboard:** An interactive admin panel, built with **React.js**, allows administrators to manage data collection sources, view sentiment trends, and export analysis reports. Real-time analytics are displayed for quick decision-making and performance monitoring.

3.MODELING AND ANALYSIS

1. Features and Functionalities:

1.1 User-Centric Features: The **Sentiment Analysis System** includes features that make it easy for users to understand public opinions from online platforms:

- **Real-Time Sentiment Tracking:** Instantly shows how people feel about topics by collecting data from social media, news, and forums.
- **Interactive Dashboard:** Displays
- sentiment trends and keyword analysis in a simple, visual format.

1.2 Customizable Data Sources: Allows users to choose specific websites or keywords for focused analysis.

- **Admin Features:** For administrators, the system provides simple tools to manage data collection and keep the platform running smoothly:
- **Manage Data Sources:** Admins can add or remove data sources as needed.
- **User Management:** Control user access and monitor activity for better security.
- **Analytics Dashboard:** Provides insights into data collection and system performance

2. Level-0 Data Flow Diagram (DFD): The Data Flow Diagram (DFD) provides us with a visual guide of how data moves around in our motion-controlled media actuator framework.

1. **Entities:**

- **User:** The main actor that interacts with the Sentiment Analysis System. It also sends requests to analyze sentiment from web data and receives the sentiment analysis results.
- **Web Server:** It is the source of data for analysis.It stores web content that is scraped by the system for further processing.

2.Process: It collects raw text data from the specified web pages on the Web Server.It cleans and structures the scraped data (e.g., removing HTML tags, handling missing values, tokenizing text). It applies algorithms to extract sentiment information (e.g., Positive, Negative, Neutral) from the processed data.

3. Data Flows: The user initiates a request for sentiment analysis based on specific keywords or data. Then the system scrapes data from the Web Server as part of the analysis process. The server responds with the raw data from the web pages.After processing and analysis, the sentiment results are returned to the user.

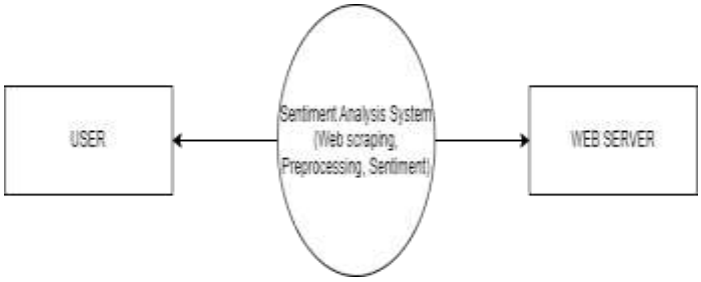

Figure 1 – DATA Flow Diagram – 0LEVEL

3.Use Case Diagram:

A **Use Case Diagram** represents the interactions between different users (actors) and the system. It provides an overview of the system's functionality from the user's perspective and helps in defining system requirements.

The **Use Case Diagram for the Sentiment Analysis using Web Scrapped data** includes:

1.**Scrape Web Data**: This process represents the collection of real-time data from various online platforms like news websites, social media, and forums.Tools like **BeautifulSoup**, **Selenium** are used to extract user reviews, comments, and posts for analysis.

2. Preprocess Data: The raw data collected from the web is often unstructured and noisy.This step involves cleaning the data—removing special characters, stop words, and performing tokenization to make it suitable for sentiment analysis.

3. **Analyse Sentiment:** In this phase, the processed data is analyzed to understand the emotional tone behind user comments and reviews. Libraries like is used to determine whether the sentiment is *Positive, Negative, or Neutral.*

4. **Classify Sentiment:** This step involves categorizing the analyzed sentiments into specific labels (e.g., *Positive, Negative, Neutral*).It helps in organizing data and allows for deeper insights into public opinion.

5. **Visualise Result:** Finally, the classified sentiments are presented visually on a dashboard.Visualization tools display real-time trends, keyword analysis, and overall sentiment distribution for easy interpretation.

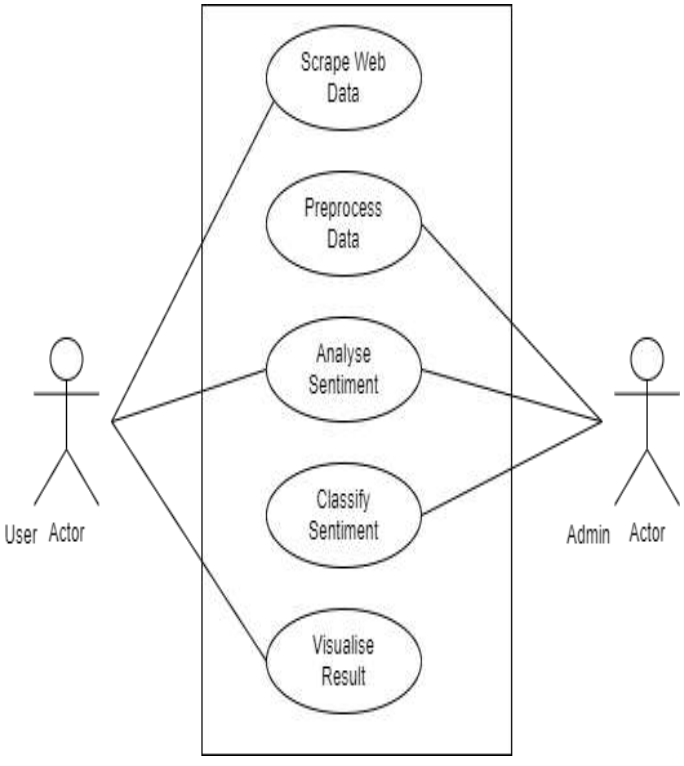

Figure 2 – Use Case Diagram

4.RESULT AND DISCUSSION

The Sentiment Analysis using Web Scrapped data system was designed to make understanding public opinion easier and faster. After testing, the system showed that it could quickly collect and analyze large amounts of data from social media, news sites, and forums[5]. Thanks to powerful web scraping tools and advanced language processing, it accurately identified whether sentiments were positive, negative, or neutral[2].

However, there is still room for improvement. Adding more data sources and optimizing the scraping process would make it even better. Future plans include introducing AI-driven predictions, expanding to multilingual analysis, and making the platform mobile-friendly so users can access insights on the go[6].

5. CONCLUSIONS

The **Sentiment Analysis using Web-Scraped Data** project showcases how we can harness modern web technologies to understand what people are feeling and saying online[7]. By using web scraping tools like **BeautifulSoup** and **Selenium**, the system collects real-time information from social media, news websites, and blogs[4]. This means we can keep track of public opinion as it changes, making it easier for businesses, researchers, and policymakers to understand what people care about and how they feel about certain topics.

The project also uses **VADER** for sentiment analysis, which helps to quickly identify whether the mood of the collected text is *positive*, *negative*, or *neutral*. This makes it simple for organizations to react to public feedback, improve their marketing strategies, and enhance customer satisfaction.

In short, this system provides an effective and real-time way to understand people's opinions online, turning raw data into meaningful insights that can drive better decisions.

6.REFERENCES

[1] Vijayaragavan Pichiyana , S Muthulingamb , Sathar , Sunanda Nalajala , Akhil , Manmath Nath Das "Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis", 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023).

[2] M. Devika, "Sentiment analysis: a comparative study on different approaches",Procedia Comput. Sci.(2016).

[3] Saurabh Sahu , Km Divya , Dr. Neeta Rastogi , Puneet Kumar Yadav , Dr. Yusuf Perwej, "Sentimental Analysis on Web Scraping Using Machine Learning Method", Journal of Information and Computational Science1, Volume 12 , Issue 8 – 20221, ISSN: 1548-7741.

[4] R. T. Rajan and S. K. Paul,(2021) "Web Scraping: A Comprehensive Review," Journal of Web Engineering, vol. 20, no. 3-4, pp. 185-204.

[5] A. Singh and B. R. Gupta,(2018) "Web Scraping for Business Intelligence: A Survey," International Journal of Business Intelligence Research, vol. 15, no. 2, pp. 65-82.

[6] L. M. Fernandez and C. S. Will liams,(2020) "Natural Language Processing Techniques for Text Extraction," Journal of Computational Linguistics, vol. 25, no. 1, pp. 39-54.

[7] R. G. Thomas and S. J. Anderson,(2022) "Future Directions in Web Scraping and NLP for Data Extraction and Analysis," Journal of Future Technology, vol. 30, no. 4, pp. 78-92