WEB SCRAPPING & SENTIMENT ANALYSIS FOR EMIRATES

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

By

1. Sachin Bade (07 / 211101003)
2. Aarohi Pisolkar (46 / 211101009)
3. Arya Sonawane (53 / 221102004)
4. Yash Shirsath (65 / 201101006)

Project Guide

Prof. Juhi Yadav

Department of Artificial Intelligence & Data Science

A. C. Patil College of Engineering
Kharghar, Navi Mumbai

University of Mumbai
(AY 2024 - 25)

CERTIFICATE

This is to certify that the Mini Project entitled

WEB SCRAPPING & SENTIMENT ANALYSIS FOR EMIRATES

is a Bonafide Work of

Sachin Bade - 02
Aarohi Pisolkar - 46
Arya Sonawane - 53
Yash Shirsath - 65

Submitted to the **University of Mumbai** in partial fulfillment of the requirement for the award of the Degree of **"Bachelor of Engineering"** in **"Artificial Engineering and Data Science".**

Prof. Juhi Yadav
(Project Guide)

Prof. Shilpali P. Bansu
(**Head of Department)**

Dr. V. N. Pawar
(**Principal)**

Mini Project Approval

This Mini Project Entitled **“Web Scrapping & Sentiment Analysis for Emirates” by Sachin Bade** (2), **Aarohi Pisolkar** (46), **Arya Sonawane** (53), **Yash Shirsath** (65) is Approved for the Degree of Bachelor of Engineering in Artificial Intelligence and Data Science.

Examiners

1………………………………………
(Internal Examiner Name & Sign)

2………………………………………
(External Examiner name & Sign)

Date: 17/10/2024

Place: A. C. Patil College of Engineering, Kharghar

Contents

Abstract

Acknowledgements

YList of Figures

List of Tables

1] Introduction 9

1.1 Introduction

1.2 Motivation

1.3 Problem Statement & Objectives

1.4 Organization of the Report

2] Literature Survey 11

1.5 Survey of Existing System

1.6 Limitation Existing system or research gap

1.7 Mini Project Contribution

3] Proposed System (E.g. New Approach of Data Summarization) 12

3.1. Introduction
3.2. Architecture/ Framework
3.3. Algorithm and Process Design
3.4. Details of Hardware & Software
3.5. Experiment and Results
3.6. Conclusion and Future work.

4] References 19

Abstract

In the era of digital transformation, customer feedback plays a pivotal role in shaping the strategies of service-oriented businesses, particularly in the airline industry. This project aims to leverage web scraping techniques and natural language processing (NLP) to analyse customer reviews of Emirates airline, providing insights into overall customer sentiment and satisfaction. By developing a comprehensive web scraper, we collected thousands of reviews from various online platforms, extracting valuable data points to assess customer experiences.

The sentiment analysis process involved using established NLP libraries, including VADER and TextBlob, to classify the extracted reviews into positive, negative, and neutral sentiments. Our findings revealed significant patterns in customer feedback, highlighting both the strengths and weaknesses of Emirates airline's services. Visualization tools were employed to present the sentiment distribution effectively, allowing stakeholders to grasp the customer sentiment landscape at a glance. the results of this project not only contribute to the existing body of knowledge in sentiment analysis but also provide actionable insights for Emirates airline to enhance its customer service and operational strategies. This report concludes by discussing the implications of our findings and outlining potential future work, including the integration of real-time sentiment analysis and expansion to other airlines.

Acknowledgment

We extend our heartfelt gratitude to our esteemed college Principal, Dr. V. N. Pawar, for his unwavering support in providing the essential resources for the development of this project. Special thanks go to our Head of the Department, Shilpali Bansu, for her invaluable suggestion of this impactful project topic for departmental purposes. We owe a great debt of thanks to our dedicated Project Guide, Prof. Juhi Yadav, for his mentorship, expert guidance, and constant encouragement throughout this project's journey. His insights and innovative ideas have been instrumental in the project's success. We would also like to express our appreciation to all the faculty members who have been a source of support, knowledge, and motivation during this endeavor. Additionally, our friends and families have played an immeasurable role in our lives, providing unwavering love, support, and understanding that have been our pillars of strength throughout this project. We are profoundly grateful for their presence in our lives. This acknowledgment reflects our sincere appreciation for the contributions and support we've received from everyone who has been part of our project's success.

Date: 17/10/2024

Sign:

List of Figures

Sr. No.	Name of Picture	Page no.
1	Architecture / Framework.	12
2	Review Data Analysis (1) – Pie chart	14
3	Review Data Analysis (2) – Bar chart	14
4	Review Data Analysis (3) – Word Cloud	14

Chapter 1 – Introduction

1.1. Introduction

In today's digital age, the abundance of online reviews significantly influences consumer choices and business strategies. The airline industry, in particular, has witnessed a surge in customer feedback shared through various platforms, including social media, travel websites, and forums. This feedback not only provides insights into customer experiences but also helps airlines gauge their performance and areas needing improvement.

This project focuses on Emirates Airlines, one of the leading airlines in the world, renowned for its exceptional service quality and commitment to passenger satisfaction. By leveraging web scraping techniques, we aim to gather a comprehensive dataset of customer reviews from multiple online sources.

These reviews encompass a wide range of sentiments, reflecting passengers' experiences with Emirates, from their flight experiences to customer service interactions. following the data collection, we will employ sentiment analysis a Natural Language Processing (NLP) technique to classify the sentiments expressed in these reviews. By analysing the sentiment polarity (positive, negative, or neutral), we can extract valuable insights that will contribute to understanding customer perceptions of Emirates Airlines.

The results of this project can provide actionable insights for Emirates Airlines, allowing them to enhance their services and address customer concerns more effectively. Ultimately, this study aims to demonstrate the importance of combining web scraping and sentiment analysis as a powerful tool for businesses to understand customer feedback in real time.

1.2. Motivation

The motivation behind this project arises from the growing importance of online reviews as a critical source of information for consumers in the airline industry. As digital platforms have become the primary medium for sharing travel experiences, understanding the sentiments expressed in these reviews has become essential for airlines aiming to enhance customer satisfaction and loyalty. Emirates Airlines, known for its premium services and expansive global network, faces the challenge of managing vast amounts of customer feedback across various online platforms. Analysing these sentiments can provide valuable insights into customer experiences, enabling Emirates to identify strengths and weaknesses in their service offerings and ultimately improve overall passenger satisfaction.

Moreover, this project aims to showcase the practical applications of Natural Language Processing (NLP) by integrating web scraping and sentiment analysis techniques. By automating the process of extracting and analysing online reviews, the project not only highlights the efficiency of data-driven approaches but also emphasizes how technology can transform unstructured data into actionable insights. This is particularly relevant for Emirates Airlines, as real-time feedback can guide strategic decisions and service improvements. In essence, the project's motivation lies in bridging the gap between customer feedback and business enhancement, illustrating how effective sentiment analysis can lead to better customer experiences and a competitive advantage in the airline industry.

1.3. Problem Statement and Objective

- **Problem Statement**: -

The increasing volume of online reviews poses a significant challenge for airlines, particularly Emirates, in effectively understanding and responding to customer sentiments. The vast array of unstructured data generated across multiple platforms makes it difficult to derive meaningful insights. Traditional methods of analyzing customer feedback are often time-consuming and inefficient, leading to a gap in understanding customer experiences and expectations. Without a systematic approach to scrape and analyze these reviews, Emirates risks missing critical information that could inform service improvements and enhance customer satisfaction.

- **Objective**: -
- **Web Scraping:** To develop a robust web scraping solution that automatically collects customer reviews from various online platforms related to Emirates Airlines, ensuring comprehensive data coverage.
- **Sentiment Analysis:** To implement Natural Language Processing techniques for sentiment analysis of the collected reviews, categorizing sentiments as positive, negative, or neutral, and thereby providing an overview of customer perceptions.
- **Insight Generation:** To analyse the sentiment data to identify trends, common themes, and areas of concern expressed by customers, enabling Emirates to understand customer satisfaction levels and pinpoint areas for improvement.
- **Real-time Feedback Mechanism:** To create a framework that allows Emirates to continuously monitor online sentiments and receive real-time feedback, facilitating timely responses to customer concerns and enhancing overall service quality.
- **Reporting and Visualization:** To present the findings through insightful reports and visualizations, making it easier for stakeholders at Emirates Airlines to interpret sentiment trends and inform strategic decision-making.

Chapter 2 - Literature Survey

2.1. Survey of Existing System

The existing systems for web scraping and sentiment analysis predominantly rely on conventional methodologies and tools that often fall short in handling the complexity and scale of data available from online reviews. Traditional sentiment analysis approaches typically use predefined lexicons or rule-based systems that struggle to capture the nuances of human language, especially in domains as dynamic as airline services.

1. **Web Scraping Techniques:-** Existing web scraping solutions utilize libraries such as BeautifulSoup, Scrapy, and Selenium to extract data from web pages. However, many of these systems require manual intervention and are limited to static content extraction. Dynamic websites, which frequently change their structure or use JavaScript for rendering content, present a challenge for these traditional scraping methods. As a result, they may miss valuable customer feedback due to incomplete or outdated data collection strategies.
2. **Sentiment Analysis Models:-** Sentiment analysis in existing systems often employs machine learning techniques such as logistic regression or support vector machines (SVM) with limited success in accurately classifying sentiments. More advanced models, like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, offer improved accuracy but require substantial computational resources and extensive labelled datasets for training. Furthermore, many systems fail to incorporate domain-specific lexicons, which can significantly enhance sentiment classification accuracy in specific sectors like aviation.
3. **Integration and Real-time Processing:-** Many existing solutions lack a seamless integration of web scraping and sentiment analysis. This disconnection can lead to delays in processing customer feedback and hinder the ability to derive actionable insights in real time. Moreover, traditional systems often do not provide mechanisms for continuous monitoring, making it difficult for airlines to adapt promptly to changing customer sentiments and preferences.
4. **User Interface and Reporting:-** Existing systems often fall short in providing user-friendly interfaces for visualizing sentiment analysis results. Many reports are generated in text-heavy formats that are difficult for stakeholders to interpret quickly. Effective data visualization techniques, such as dashboards and interactive graphs, are rarely employed, limiting the utility of the sentiment analysis outcomes for strategic decision-making.

2.2. Limitation Existing System or Research Gap

Despite advancements in web scraping and sentiment analysis, several limitations exist in current systems:

1. **Static Data Collection**: Many systems struggle to capture dynamic content that changes frequently, leading to gaps in real-time insights from customer feedback.
2. **Lack of Domain-Specific Analysis**: Existing sentiment analysis models often fail to account for industry-specific terminology, resulting in inaccuracies in sentiment classification within the airline context.
3. **Limited Multilingual Support**: Current systems have difficulty handling reviews in multiple languages, which is essential for a global airline like Emirates.

4. **Underutilization of Advanced Techniques**: While deep learning methods can enhance sentiment analysis, many existing solutions do not leverage these due to resource limitations.
5. **Absence of Real-Time Analysis**: Most systems provide post-hoc analysis, which delays responses to customer feedback, reducing their effectiveness.
6. **Ineffective Data Visualization**: Current tools often lack user-friendly visual representations of sentiment data, making it challenging for stakeholders to derive actionable insights.
7. **Ethical and Privacy Concerns**: The ethical implications of web scraping and data use are frequently overlooked, raising concerns about user privacy and data ownership.

Addressing these gaps is crucial for developing a more effective web scraping and sentiment analysis system tailored to Emirates Airlines.

2.3. Mini Project Contribution

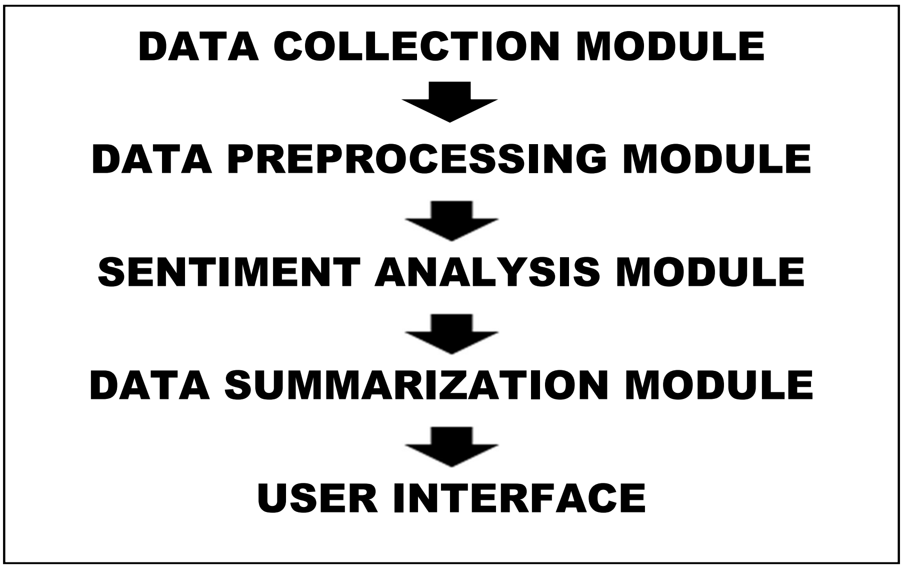
This mini project on "Web Scraping and Sentiment Analysis for Emirates" makes several key contributions to the field of natural language processing and customer feedback analysis. Firstly, it implements a system that enables real-time sentiment analysis by scraping reviews from various online platforms, allowing Emirates to promptly gauge customer sentiments and respond to feedback, thereby enhancing service quality and customer satisfaction. Secondly, the project utilizes customized natural language processing techniques tailored to the unique language and terminology of the airline industry, improving sentiment accuracy and providing deeper insights into customer perceptions specific to Emirates Airlines.

Chapter 3 - Proposed System

3.1. Introduction

The proposed system focuses on a novel approach to web scraping and sentiment analysis for Emirates Airlines. As customer feedback increasingly drives improvements in service quality, the need for efficient sentiment analysis of online reviews has become paramount. This system automates the scraping of reviews from various platforms, analyses the sentiments expressed, and summarizes the findings to provide actionable insights for the airline. this approach is motivated by the growing volume of customer feedback, which can be challenging to analyse manually. By utilizing advanced natural language processing techniques, the system enhances sentiment detection accuracy and enables quicker decision-making. Ultimately, this innovation supports Emirates Airlines in improving customer satisfaction and developing targeted strategies based on real-time insights.

3.2. Architecture/ Framework


DATA COLLECTION MODULE
DATA PREPROCESSING MODULE
SENTIMENT ANALYSIS MODULE
DATA SUMMARIZATION MODULE
USER INTERFACE

3.3. Algorithm and Process Design

The algorithm and process design for the web scraping and sentiment analysis project consists of the following key components:

1. **Web Scraping Algorithm**:
   - **Input**: Target website URLs.
   - **Process**: Utilize the **requests** library to send HTTP requests, and **Beautiful Soup** to parse HTML content and extract reviews and metadata.
   - **Output**: Structured dataset (CSV or JSON) containing scraped reviews.
2. **Data Preprocessing Steps**:
   - **Input**: Raw review data.

    - **Process**: Clean and normalize text (remove HTML tags, convert to lowercase), tokenize the text, remove stop words, and apply stemming or lemmatization.
    - **Output**: Pre-processed text ready for sentiment analysis.
3. **Sentiment Analysis Algorithm**:
    - **Input**: Pre-processed text data.
    - **Process**: Use a sentiment analysis model (lexicon-based or machine learning) to classify reviews as positive, negative, or neutral.
    - **Output**: Sentiment labels and scores for each review.
4. **Data Summarization Process**:
    - **Input**: Sentiment analysis results.
    - **Process**: Aggregate sentiment scores, identify key themes, and create visualizations to present findings.
    - **Output**: Summary report highlighting customer sentiment trends.

This structured approach ensures effective analysis of customer sentiments toward Emirates Airlines, providing valuable insights for stakeholders.

3.4. Details of Hardware & Software

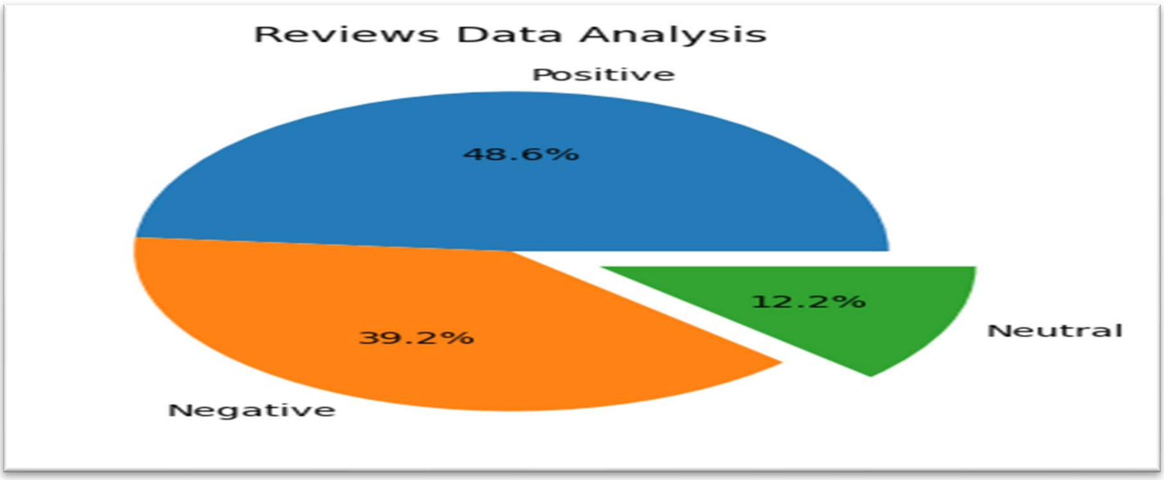
Hardware Requirements: -

1. **Processor**: Minimum Intel Core i5 or equivalent.
2. **RAM**: At least 8 GB for smooth operation, especially when handling larger datasets.
3. **Storage**: A minimum of 100 GB of available disk space to store project files, datasets, and outputs.
4. **Network**: Reliable internet connectivity for web scraping and accessing online resources.

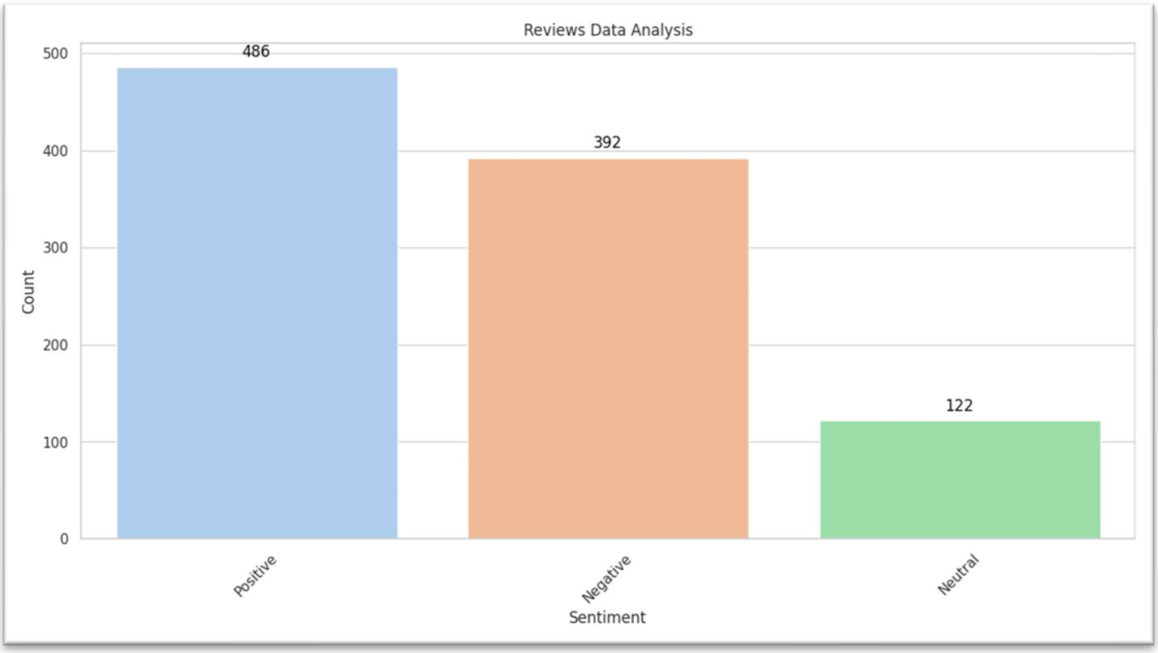
Software Requirements: -

1. **Operating System**: Windows, macOS, or Linux.
2. **Python**: Version 3.6 or higher for compatibility with various libraries.
3. **Libraries**:
    - **Requests**: For sending HTTP requests and handling responses.
    - **Beautiful Soup**: For parsing HTML and XML documents to extract data.
    - **Pandas**: For data manipulation and analysis, particularly in handling structured data.
    - **NLTK or SpaCy**: For natural language processing tasks, including text cleaning, tokenization, and sentiment analysis.
    - **Matplotlib or Seaborn**: For data visualization to present the analysis results effectively.
4. **Integrated Development Environment (IDE)**:
    - **Jupyter Notebook**: For an interactive coding environment and documentation.
    - **VS Code or PyCharm**: For a robust development experience with additional features.

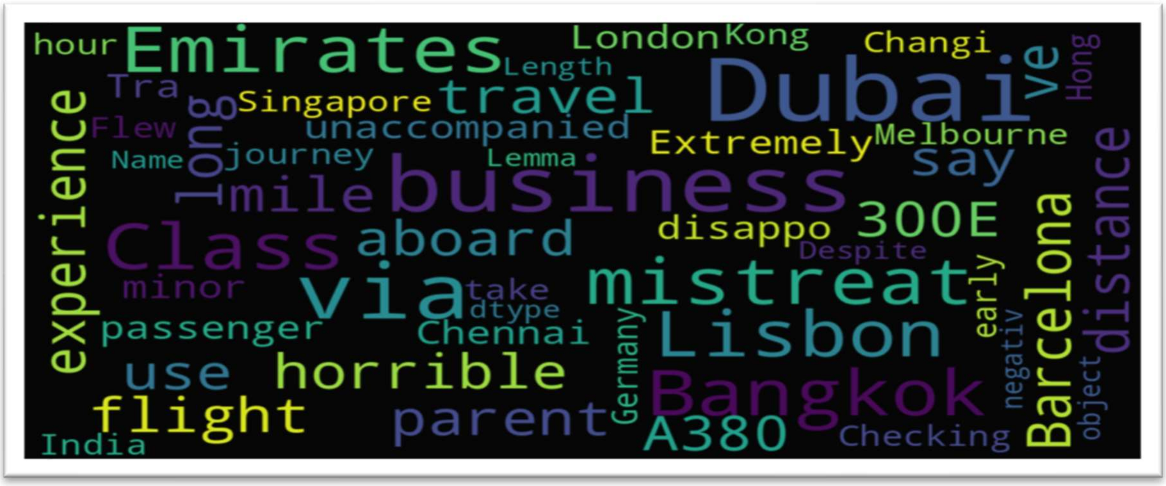
3.5. Results



Review Data Analysis (1) – Pie Chart



Review Data Analysis (2) – Bar Chart



Review Data Analysis (3) – Word Cloud

3.6. Conclusion and Future Work

In this project, we successfully implemented a web scraping and sentiment analysis system for Emirates airline reviews. By leveraging various Python libraries, we efficiently extracted valuable insights from user-generated content on social media and review platforms. The sentiment analysis revealed important trends in customer opinions, allowing us to better understand user experiences and perceptions of Emirates services. Overall, our approach not only demonstrates the potential of natural language processing in the airline industry but also provides a framework that can be adapted for similar projects in different sectors.

6. Future Work:-

1. **Expanded Data Sources**: Incorporating additional platforms, such as travel blogs and forums, could provide a more comprehensive view of customer sentiments.
2. **Real-Time Analysis**: Implementing real-time data scraping and analysis can help monitor sentiments as they evolve, enabling Emirates to respond proactively to customer feedback.
3. **Multilingual Support**: Developing capabilities to analyse reviews in multiple languages would broaden the scope of the project, capturing sentiments from a diverse customer base.
4. **Advanced Sentiment Analysis Techniques**: Exploring deep learning models for more nuanced sentiment classification could improve the accuracy of our analysis.
5. **Dashboard for Visualization**: Creating an interactive dashboard to present insights and trends visually would facilitate easier interpretation and decision-making for stakeholders.

By pursuing these avenues, we can enhance the utility and impact of our sentiment analysis system, ultimately contributing to improved customer satisfaction and operational efficiency in the airline industry.

- **References:-**

1. **Beautiful Soup Documentation**. (n.d.). Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/
2. **Scrapy Documentation**. (n.d.). Retrieved from https://docs.scrapy.org/en/latest/
3. P. (2021). "Sentiment Analysis with Python". Real Python. Retrieved from https://realpython.com/python-sentiment-analysis/
4. A. (2020). "A Comprehensive Guide to Sentiment Analysis in Python". Towards Data Science. Retrieved from https://towardsdatascience.com/a-comprehensive-guide-to-sentiment-analysis-in-python-d3688d39e6af
5. **Natural Language Toolkit (NLTK) Documentation**. (n.d.). Retrieved from https://www.nltk.org/book/
6. **SpaCy Documentation**. (n.d.). Retrieved from https://spacy.io/usage
7. A. (2020). "How Airlines Use Machine Learning and AI to Improve Customer Experience". Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2020/09/how-airlines-use-machine-learning-and-ai-to-improve-customer-experience/