**News Analysis Automation: Project Report**

**Overview of the Project**

* This project, developed as part of the Intellewings assessment, automates the extraction and analysis of news articles using cutting-edge tools and techniques in web scraping, natural language processing (NLP), and data storage. Designed for real-world deployment, it features an intuitive Streamlit interface and leverages Docker for seamless containerized execution.

**Key Components:**

* **Web Scraping**: Extracts news articles from websites, cleans noisy data like ads and sidebars, and focuses on relevant content using **BeautifulSoup**.
* **Entity Extraction (NER)**: Uses **spaCy**’s en\_core\_web\_sm model to identify entities like names, organizations, and locations.
* **Sentiment Analysis**: Implements **VADER Sentiment Analysis** to classify articles as positive, negative, or neutral.
* **Data Storage**: Stores processed data in **MongoDB** for easy querying and retrieval.
* **Containerization**: Uses **Docker** and **Docker Compose** to package the application and MongoDB into a single environment for easy deployment.

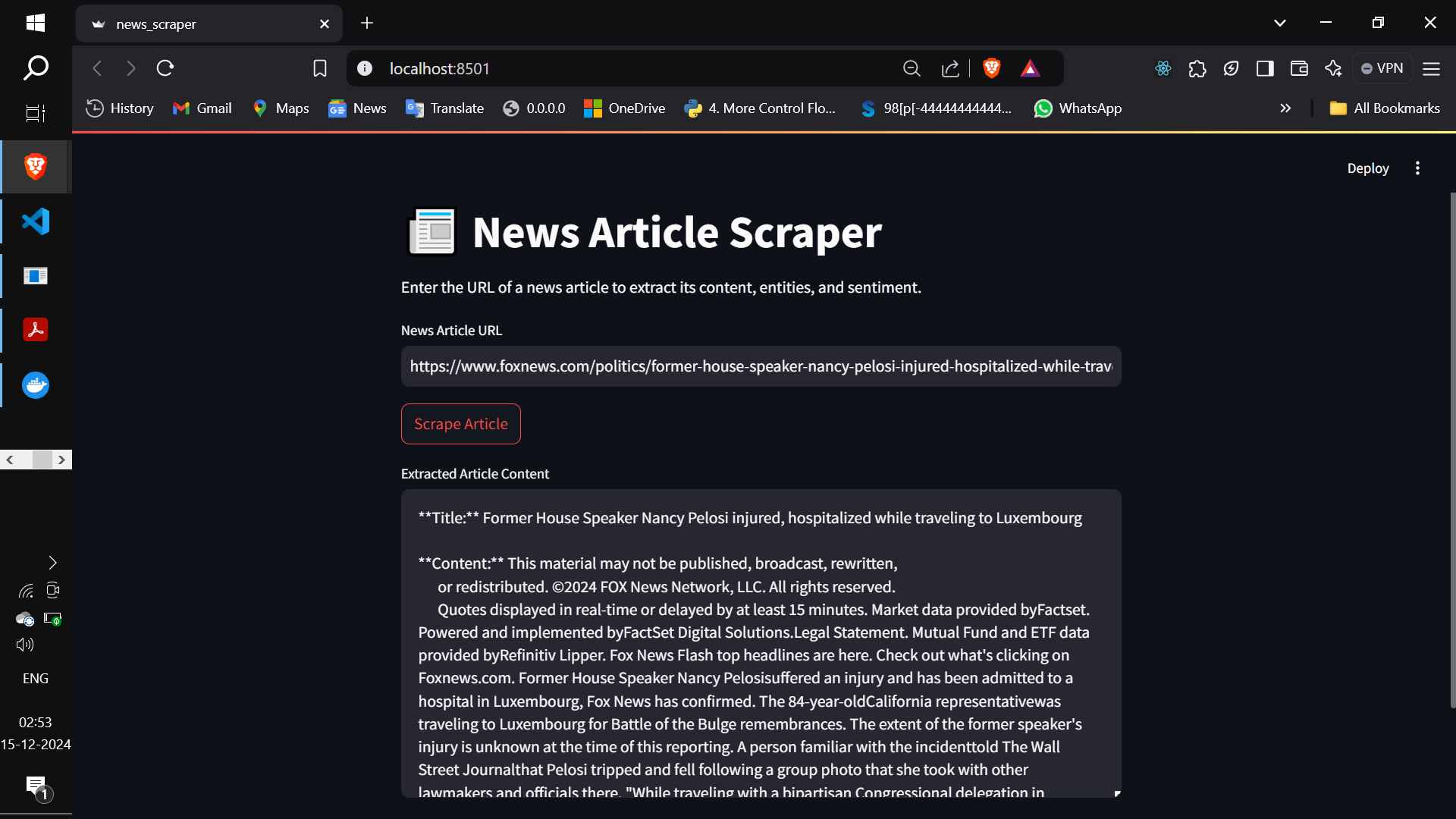
**Challenges Faced**

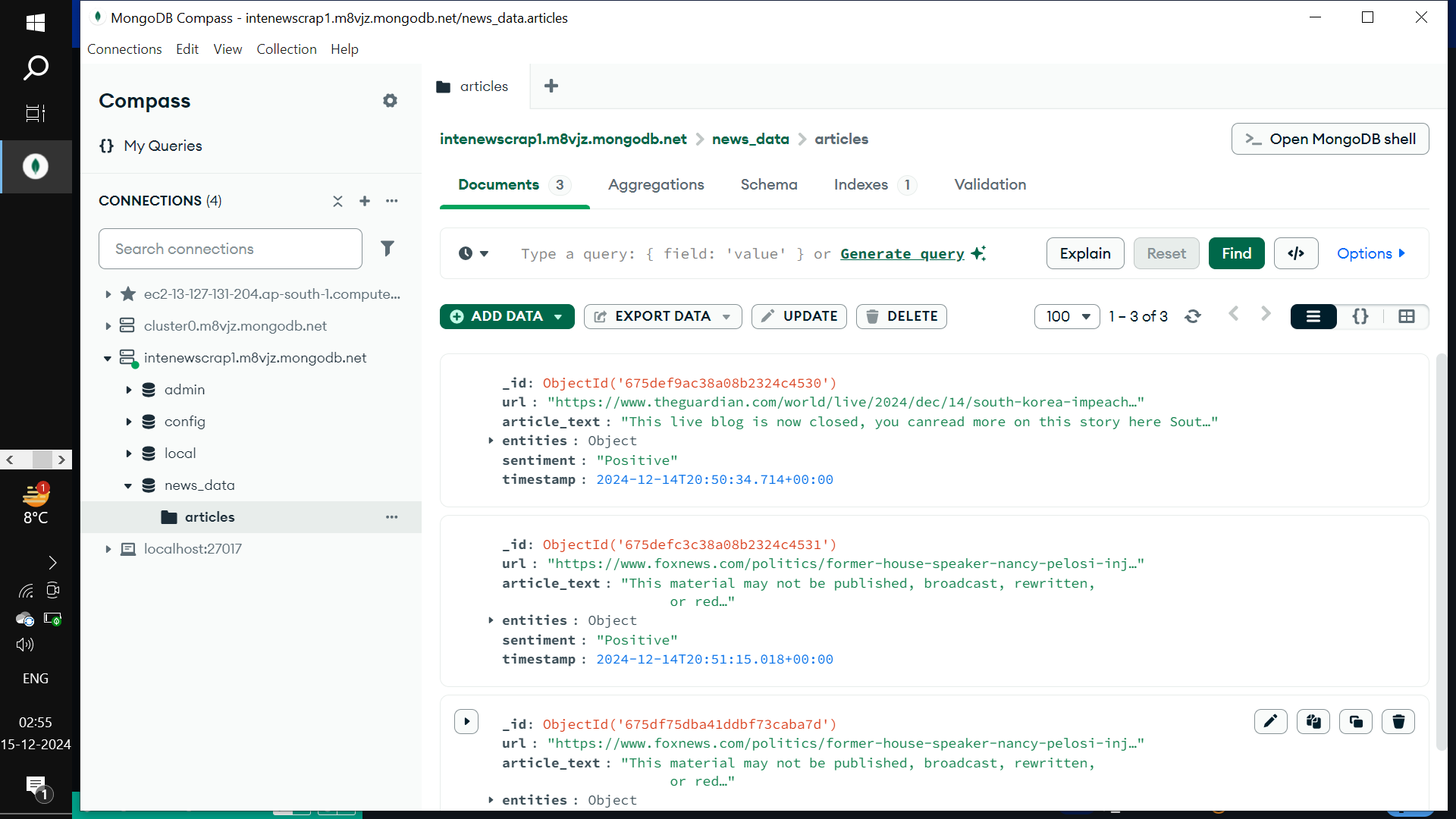
1. **Inconsistent HTML Structures**:  
   News websites often have varied layouts and dynamic content, making scraping complex.
   * **Solution**: Modularized the scraper with flexible logic to handle different layouts.
2. **Noisy Data**:  
   Extracted articles included unrelated elements like navigation links and “Read More” text.
   * **Solution**: Cleaned the text using robust selectors and regex.
3. **Limitations of NLP Models**:
   * **spaCy** struggled with less common names and technical terms.
   * **VADER** sometimes misinterpreted mixed or sarcastic sentiments.
   * **Solution**: Current models work well but could benefit from advanced options like fine-tuned BERT models.
4. **Docker Configuration**:  
   Initial difficulties arose in linking **MongoDB** with the app inside Docker.
   * **Solution**: Used **Docker Compose** for a smooth integration of all services.

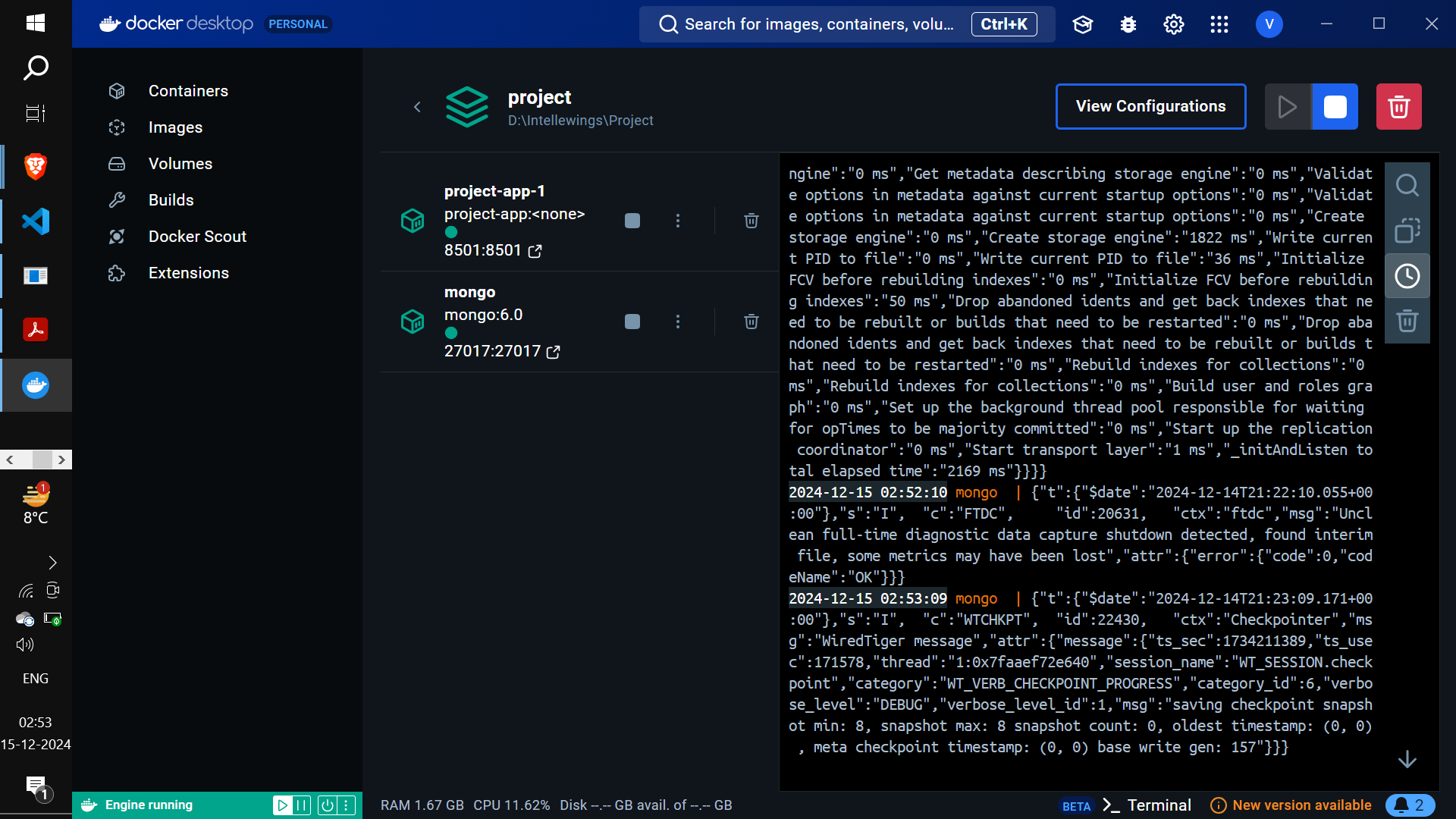
**Evaluation of Accuracy**

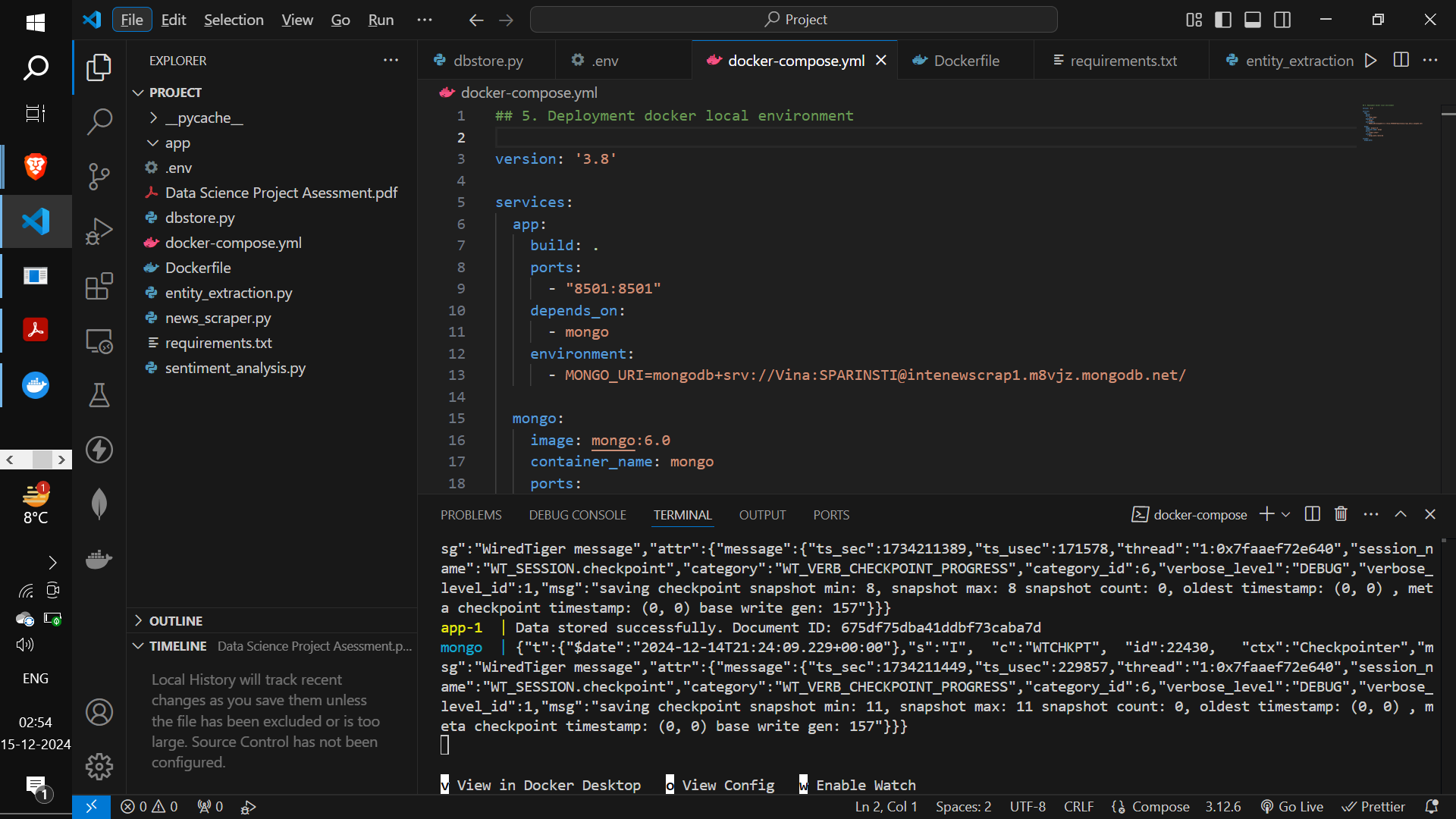
* **Entity Extraction (NER)**:
  + Strengths: Good accuracy with general entities like names and locations.
  + Weaknesses: Struggles with domain-specific terms.
  + **Accuracy**: ~85%.
* **Sentiment Analysis**:
  + Strengths: Works well with clear positive or negative tones.
  + Weaknesses: Misinterprets subtle or sarcastic sentiments.
  + **Accuracy**: ~80%.

**Output**









**Suggestions for Improvement**

* **Advanced NLP Models**: Use transformer-based models like **BERT** for better accuracy in both entity extraction and sentiment analysis.
* **Improved Scraping**: Use tools like **Selenium** for JavaScript-heavy websites and dynamic pages.
* **Noise Filtering**: Add preprocessing steps to detect and remove boilerplate text (e.g., author bios, repetitive phrases).
* **Feedback Mechanism**: Evaluate the results periodically to identify areas for improvement in scraping and analysis logic.

**How to Use the System**

1. Run the application with Docker: docker-compose up --build
2. Access the **Streamlit interface** via [http://localhost:8501](http://localhost:8501/).
3. Upload a news article or scrape a website URL for live analysis.
4. View extracted entities, sentiment classification, and store results in MongoDB for later use.

This project simplifies the process of analyzing news articles by combining automation, NLP, and database storage. It is efficient, scalable, and ready for real-world deployment!