**Final Project Report**

Xiao Zhou

This project focuses on performing sentiment analysis on climate-related news articles collected from two major international media outlets: BBC and CNN. The entire workflow includes web scraping, text cleaning, entity recognition, and sentiment scoring, with the aim of understanding how different countries or regions are represented in terms of sentiment within environmental journalism.

**BBC News Workflow:**

1. Scraping Article Metadata

Using bbc\_scraper.py, I collected titles, timestamps, and links from BBC's search results.

Because BBC's pagination is difficult to automate, I manually turned pages and scraped a total of 403 articles.

The scraped metadata was saved in BBC\_Climate\_Articles\_Limited.csv.

2. Scraping Article Content

Then, I used bbc\_content.py to extract the full content of each article and saved the results in bbc\_climate\_articles\_with\_content.csv.

3. Sentiment Analysis

I analyzed the textual sentiment using senti\_local\_bbc.py.

The final output, sentiment\_bbc.csv, contains country mentions along with average sentiment scores.

**CNN News Workflow**

1. Scraping Article Metadata

I used cnn.py to scrape article titles and links, which were saved to cnn.csv. I scraped a total of 573 news.

2. Scraping Article Content

I then ran cnn\_content.py to extract article content, storing it in cnn\_content.csv.

3. Sentiment Analysis

Sentiment analysis was performed using senti\_local\_cnn.py.

Text cleaning and sentiment scoring were handled within the same script.

**Tools & Techniques Used**

1. Selenium + webdriver\_manager: Automated Browser Control

- Why: CNN and BBC render search results dynamically using JavaScript, so standard requests scraping does not work.

- What it does: Automates interaction with search pages, simulating a real browser to collect links and headlines.

2. spaCy: Named Entity Recognition (NER)

- Used to identify geopolitical entities (GPEs) such as countries and cities from the article content.

- The lightweight English model en\_core\_web\_sm was used to detect location mentions.

- Limitations: spaCy sometimes misidentifies cities as countries, and occasionally includes non-location entities as GPEs.

3. VADER: Sentiment Analysis

- The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analyzer from the vaderSentiment library was used.

- VADER calculates a compound sentiment score ranging from -1 (very negative) to +1 (very positive).

* Each identified country (or location) inherits the sentiment score of the article in which it is mentioned.

Findings & Observations

Although I didn’t implement complex post-processing or filtering of the results, some interesting patterns emerged from the raw data.

**1. BBC vs CNN Tone**

- BBC tends to use more positive wording in climate reporting.

- CNN shows a more balanced or mixed tone across articles.

The following charts (see attached figures) show the sentiment distributions across countries.

图表, 直方图

AI 生成的内容可能不正确。图表

AI 生成的内容可能不正确。

**2. Sentiment Differences within the UK**

In analyzing different terms referring to the UK, I noticed striking sentiment differences: This suggests that different parts of the UK may be framed with slightly different emotional tones, even in climate-related reporting.

|  |  |  |
| --- | --- | --- |
| **Terminology** | **Sentiment** | **Mention Times** |
| **UK** | **0.5643** | **150** |
| **England** | **0.6839** | **53** |
| **Scotland** | **0.3769** | **43** |
| **Wales** | **0.7543** | **22** |
| **Britain** | **0.5786** | **21** |
| **North Ireland** | **0.3612** | **18** |

**3. Surprising Mentions**

One unexpected finding was the mention of “Soviet Union” in recent news coverage — a term seemingly outdated.

• It had a very negative sentiment score: -0.7288

• In a trial using only the most recent two months of news, it scored -0.9969, the most negative value in the dataset.

日历

AI 生成的内容可能不正确。

**4. Geographic Inequality in Media Attention**

Countries and regions in Africa and South America were mentioned significantly less frequently in the dataset.

This imbalance may partially reflect climate injustice in media representation — where regions most vulnerable to climate change receive disproportionately less coverage in mainstream environmental reporting.

There are more findings to find in the final csv files.