Seattle News Sentiment Analysis
Automated Scraping and Political Candidate Analysis

Data Science Project

August 11, 2025

1 Overview

This project performs automated sentiment analysis on Seattle news articles, focusing on political candidates and municipal issues. The system scrapes articles from multiple news sources, analyzes sentiment using natural language processing, and generates visualizations showing sentiment patterns across sources, candidates, and themes.

2 Project Structure

2.1 Core Components

- `newspaper_sentiment_scraper.py` - Python web scraper for collecting news articles
- `sentiment_analysis.r` - R script for sentiment analysis and visualization
- `sentiment_analysis_clean.r` - Simplified version with enhanced error handling

2.2 Data Sources

The scraper targets the following Seattle news outlets:

- Seattle Times
- KING 5 News
- KIRO 7 News
- KOMO News
- KUOW Public Radio
- The Stranger
- Capitol Hill Seattle

3 Installation and Setup

3.1 Python Dependencies

```
pip install requests beautifulsoup4 feedparser pandas
pip install urllib3 dataclasses logging functools
```

3.2 R Dependencies

```
install.packages(c("tidyverse", "tidytext", "textdata",
                   "ggplot2", "scales"))
```

4 Usage

4.1 Data Collection

Run the Python scraper to collect articles:

```
python newspaper_sentiment_scraper.py
```

This generates timestamped CSV files: `seattle_news_YYYYMMDD_HHMMSS.csv`

4.2 Sentiment Analysis

Execute the R analysis script:

```
source("sentiment_analysis_clean.r")
```

5 Methodology

5.1 Web Scraping

The scraper employs multiple strategies:

- RSS feed parsing for structured data
- Direct web scraping with CSS selectors
- Wayback Machine fallback for unavailable content
- Weighted keyword relevance scoring

5.2 Keyword Detection

Articles are filtered using weighted keyword categories:

High Priority (Weight = 3):

- 2025 mayoral candidates: Harrell, Wilson, Armstrong, Bliss, Mallahan, Molloy, Whelan, Willoughby, Savage
- Political terms: mayor, election, candidate, city council

Medium Priority (Weight = 2):

- Urban issues: housing, homeless, transportation, budget
- Labor topics: union, organizing, workers

Low Priority (Weight = 1):

- Policy terms: ordinance, legislation, referendum

5.3 Sentiment Analysis

The analysis uses the AFINN lexicon with the following process:

1. Text preprocessing and tokenization
2. Stop word removal
3. Sentiment scoring: $S_{article} = \frac{\sum_{i=1}^{n} w_i}{n}$
   where w_i is the AFINN score for word i and n is the word count
4. Aggregation by source, candidate, and theme

6 Output Files

6.1 Data Files

- `article_sentiments.csv` - Individual article sentiment scores
- `sentiment_by_source.csv` - Aggregated sentiment by news source
- `candidate_sentiment_by_source.csv` - Candidate sentiment by source
- `sentiment_by_theme_and_source.csv` - Thematic sentiment analysis
- `candidate_timeline_data.csv` - Timeline data for candidates

6.2 Visualizations

- `source_sentiment_plot.png` - Bar chart of sentiment by news source
- `candidate_sentiment_heatmap.png` - Heatmap of candidate sentiment across sources
- `candidate_sentiment_timeline.png` - Timeline of candidate sentiment over article sequence
- `thematic_sentiment_plot.png` - Faceted plot of sentiment by theme and source

7 Key Features

7.1 Robust Data Collection

- Multiple fallback mechanisms for failed requests
- Rate limiting and respectful scraping practices
- Duplicate detection and removal
- Date format normalization

7.2 Comprehensive Analysis

- Source-level sentiment comparison
- Candidate-specific sentiment tracking
- Thematic categorization (Politics, Urban Issues, Public Safety)
- Timeline analysis using article sequence

7.3 Error Handling

- Graceful handling of network failures
- Alternative PNG saving when ggsave() fails
- Comprehensive logging and debugging output

8 Configuration

The scraper supports configuration via `scraping_config.json`:

```
{
    "max_articles_per_source": 10,
    "request_timeout": 10,
    "delay_between_requests": 1.0,
    "min_content_length": 100,
    "use_wayback_fallback": true
}
```

9 Limitations and Considerations

- Limited to English-language sentiment analysis
- AFINN lexicon may not capture political context nuances
- Web scraping subject to site structure changes
- Timeline analysis uses article sequence rather than publication dates due to inconsistent date formats

10 Future Enhancements

- Integration of transformer-based sentiment models
- Real-time monitoring and alerts
- Expanded keyword detection using NER
- Interactive dashboard development
- Cross-validation with human-annotated sentiment

11 Technical Notes

11.1 Date Handling

The system handles multiple date formats:

- ISO 8601: `2025-01-30T14:30:00`
- RFC 2822: `Wed, 30 Jan 2025 14:30:00 +0000`
- US format: `January 30, 2025 at 2:30 pm PDT`

11.2 Visualization Details

All plots use consistent color schemes:

- Positive sentiment: Blue
- Negative sentiment: Red
- Neutral reference: Dashed line at y=0

12 Contact and Support

For questions or issues, refer to the project documentation or examine the comprehensive logging output generated during execution.