Media Matters: Text-Based Analysis of How The Guardian and The New York Times Cover US News

Students: Mohan Krishna Askani and Naveen Korrapati

Course Name: Data Wrangling

Instructor: Stevenson Bolivar Atuesta

May 9, 2025

Executive Summary

With growing concerns around media bias, information bubbles, and the polarization of public discourse, understanding how news is framed by different publishers is more relevant than ever. This project compares national news coverage from *The New York Times* and *The Guardian*, uncovering how each frames stories thematically, emotionally, and geographically.

Data Sources

Articles were collected over the past month (April 2025) using official **APIs** provided by *The New York Times* and *The Guardian*. These sources were selected due to their high journalistic influence, structured access to full-text data, and established reputations for national and international coverage.

Main Objective

To answer: *What topics do both of these publishers emphasize? How emotional is their tone? Do they give equal attention to different regions in the U.S.?*. This helps identify **editorial imbalance** and **narrative skew** which is crucial for organizations making decisions based on public sentiment, reputation management, or regional outreach. It also reveals how **news content may influence perception** through emotional tone or topical omission.

Key Insights & Summary

- **Topic Divergence**: The Guardian highlighted themes around global politics (*e.g., "ceasefire"*, *"workforce"*), while NYT emphasized institutions and U.S.-centric terms (*e.g., "harvard"*, *"republican"*), as shown through TF-IDF analysis.
- **Sentiment Variation**: The Guardian exhibited more frequent and intense emotional language (e.g., spikes in *fear*, *anger*), while NYT maintained a steadier and more neutral tone.
- **Geographic Focus**: Both publishers covered major cities, but NYT showed tighter regional focus, while The Guardian had more dispersed geographic mentions across states.
- **Temporal Trends**: Sentiment and article volume varied across time, with The Guardian showing more volatility, possibly reflecting different editorial pacing or audience engagement strategy.

Brief Methodology

Cleaned and standardized article text for both sources. Used NLP methods (TF-IDF, sentiment analysis, NER) to extract structure. Visualized trends using charts and animated maps. Compared by publisher across topic, sentiment, emotion, and geography.

Introduction: Context & Project Relevance

Problem Statement

In today's fragmented media landscape, the same national events can be presented through vastly different editorial lenses depending on the news outlet. This variation in topic selection, emotional tone, and regional emphasis can influence public perception, create information silos, and shape national discourse. Understanding these differences is crucial for media consumers, analysts, and researchers interested in bias detection, media accountability, and information transparency.

Project Relevance

This project is designed to analyze and compare how two major news publishers ***The New York Times*** and ***The Guardian*** report on U.S. national news. The insights can benefit:

- **Media analysts** to benchmark editorial strategies and tone.
- **Journalism educators and researchers**, to illustrate framing, bias, and narrative construction.
- **Policy makers and advocacy groups**, to understand how regional or thematic issues are covered.
- **General news consumers**, to build media literacy and evaluate information critically.

Datasets Overview

The analysis draws on article data collected over the past month from:

- The Guardian (via Open Platform API): Articles from the U.S. section.
- The New York Times (via Article Search API): Articles tagged under U.S./National categories.

Each dataset contains metadata (title, date, URL) and full article body text. Initial observation revealed stylistic differences, variation in article length, and inconsistencies in how publishers tag or present geographic references.

Goal of the Analysis

The primary objective is to uncover ***how the two publishers differ*** in terms of:

- Topic emphasis (via unigram, bigram, and TF-IDF analysis)
- Sentiment and emotion framing (using Bing and NRC lexicons)
- Geographic distribution of coverage (through Named Entity Recognition and map visualizations)

The expected outcome is a clearer understanding of each publisher's editorial tendencies providing a foundation for media comparison, bias evaluation, and improved news literacy.

Data Wrangling & Cleaning

Initial State of the Data

We started with two datasets:

- **The Guardian** dataset from its U.S. section via the Open Platform API.
- **The New York Times** dataset via its Article Search API focused on national topics.

Each dataset contained basic metadata (e.g., title, publication date, URL) and full article bodies. However, both had issues:

- The Guardian text included HTML tags, boilerplate sections, and inconsistent punctuation.
- NYT articles often lacked full body content or returned JavaScript-rendered placeholders (e.g., "We are having trouble retrieving the article content").
- There were missing or malformed date entries, inconsistent casing, and extraneous characters in both datasets.

Cleaning Process

To address these issues:

- We removed all HTML tags, special characters, and non-informative phrases from the article bodies using a custom text cleaning function.
- Articles with body text under 300 characters were dropped to ensure only complete content was analyzed.
- We standardized punctuation, stripped possessives and contractions (e.g., “Trump’s” became “Trump”), and removed all numeric or very short tokens.
- Stop words were removed, and all remaining tokens were lemmatized (reduced to their root form).
- Part-of-speech filtering was applied to retain only meaningful words: nouns, verbs, adjectives, and adverbs.

Data Merging

Once cleaned, both datasets were combined into a unified structure with three key columns:

- publisher: Source of the article
- pub_date: Date of publication
- cleaned_body: Final processed body text

Merging was straightforward after ensuring consistent column naming and formatting. Some challenges included aligning date formats and ensuring no duplicated articles were present.

Final Cleaned Dataset

The resulting dataset contains **1,405 articles** across both publishers, each with:

- Cleaned and standardized article text
- Consistent publisher and date metadata
- Ready-to-analyze structure for tokenization, sentiment analysis, and geographic extraction

Additionally, new features such as token counts, sentiment scores, and geographic mentions were derived from this cleaned text for downstream analysis.

Before-and-After Summary

Aspect	Before Cleaning	After Cleaning
HTML tags/noise	Present in Guardian content	Fully removed
Missing / Short bodies	~8% of NYT articles incomplete	Removed articles < 300 words
Duplicate entries	Not present	Verified and removed if any
Non-standard punctuation	Present (quotes, apostrophes, symbols)	Normalized
Columns retained	title, date, url, body	publisher, pub_date, cleaned_body
New features created	None	tokenized text, sentiment, locations

Exploratory Data Analysis (EDA)

Overview of the data after cleaning

After cleaning, we retained **1,405 high-quality articles** published over the past month approximately half from *The Guardian* and half from *The New York Times*. Each article includes metadata (publisher, pub_date) and a cleaned full-text body ready for textual analysis.

Some descriptive stats:

- **Average article length**: ~250–300 words after cleaning.
- **Date range**: Uniform distribution over the most recent 30 days.

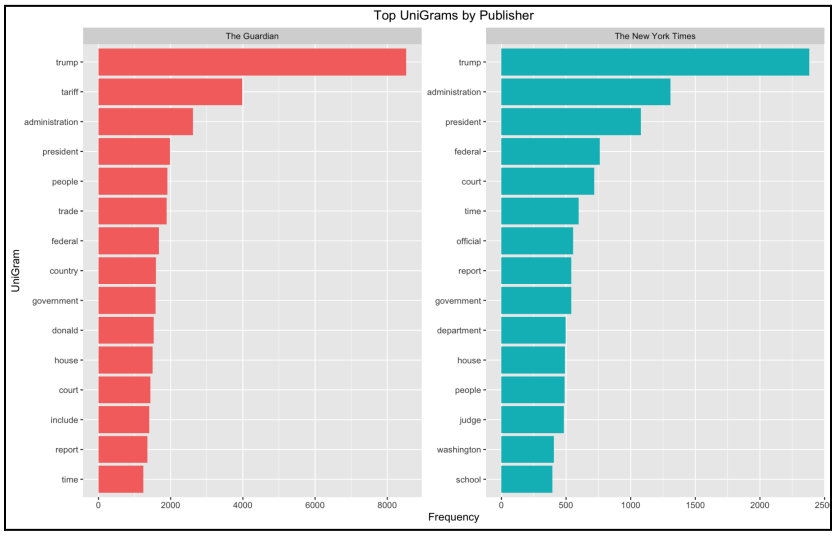
- **Balance**: Roughly equal number of articles from both publishers, ensuring fair comparative analysis.

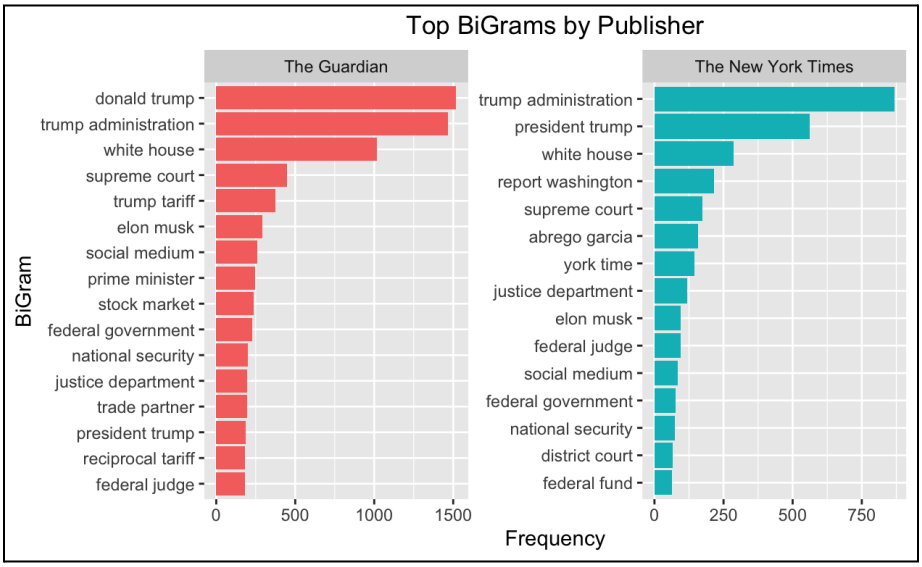
Visualizations & Patterns Identified

1. Top Unigrams & Bigrams

Using frequency counts, we visualized the most common individual words and two-word phrases for each publisher.

- *The Guardian*: Frequently used terms like “trump tarrif”, “trade”, and “elon musk”.
- *The New York Times*: Focused on names and institutions like “report washington”, “supreme court”, and “white house”.

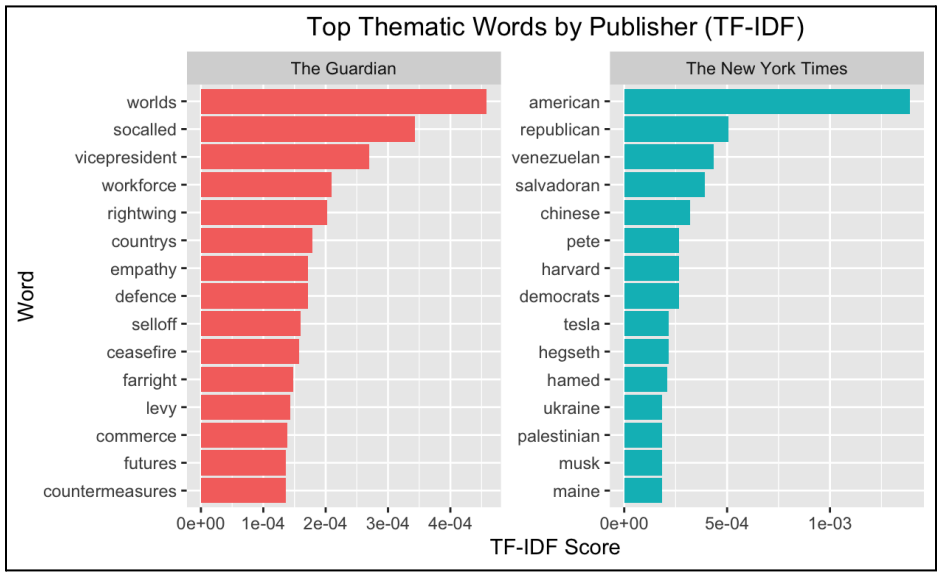




2. TF-IDF analysis

TF-IDF scores helped us uncover **distinctive vocabulary** for each publisher words that were not just frequent but also uniquely emphasized.

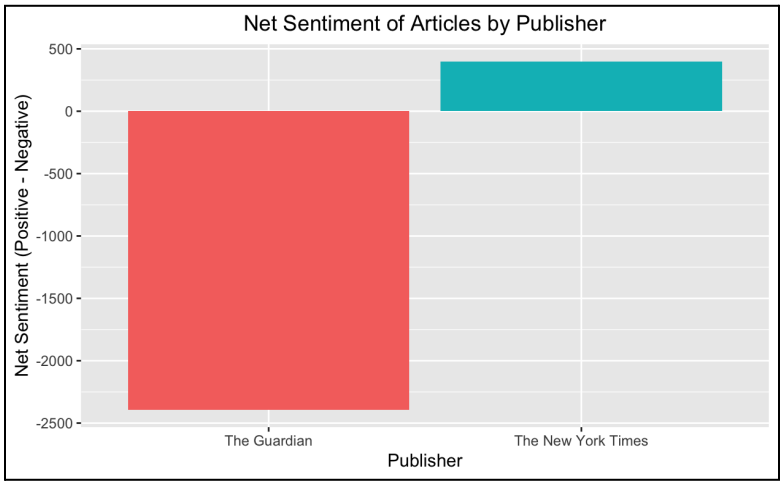
- The Guardian used more ideological and international terms.
- NYT emphasized more event-specific and U.S.-institutional language.

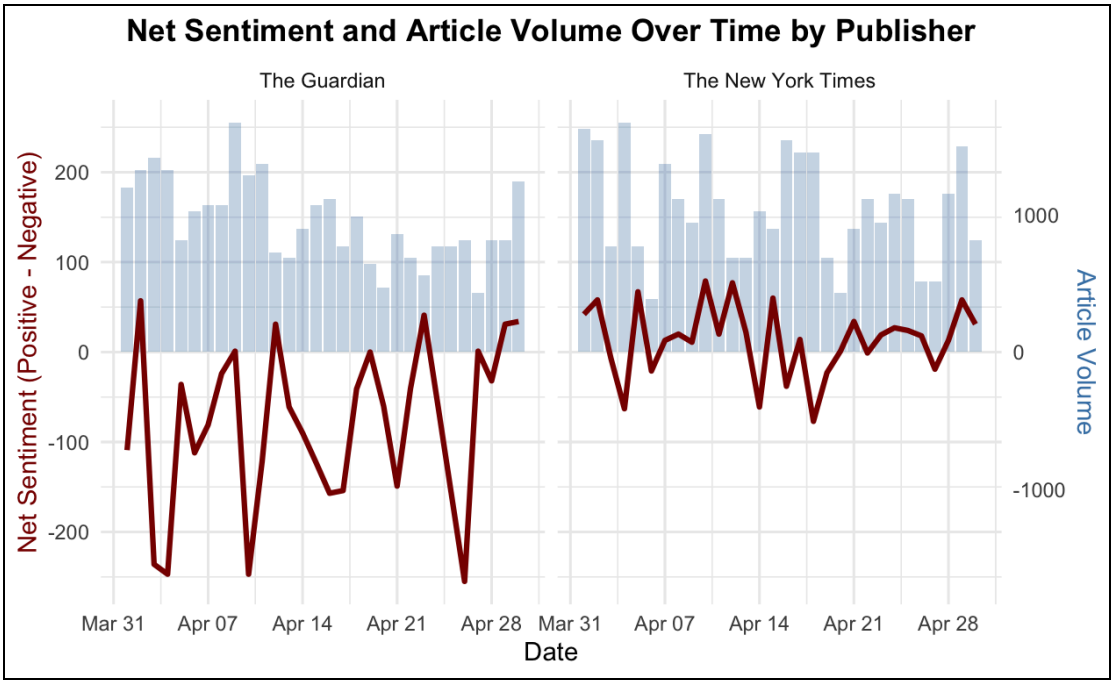


3. Sentiment Analysis

By applying the Bing sentiment lexicon, we tracked **net sentiment** (positive – negative words) across publishers and over time.

- The Guardian showed higher emotional variance, with more strongly negative tone on some days.
- NYT maintained a more **consistent and balanced tone**, rarely showing strong emotional spikes.

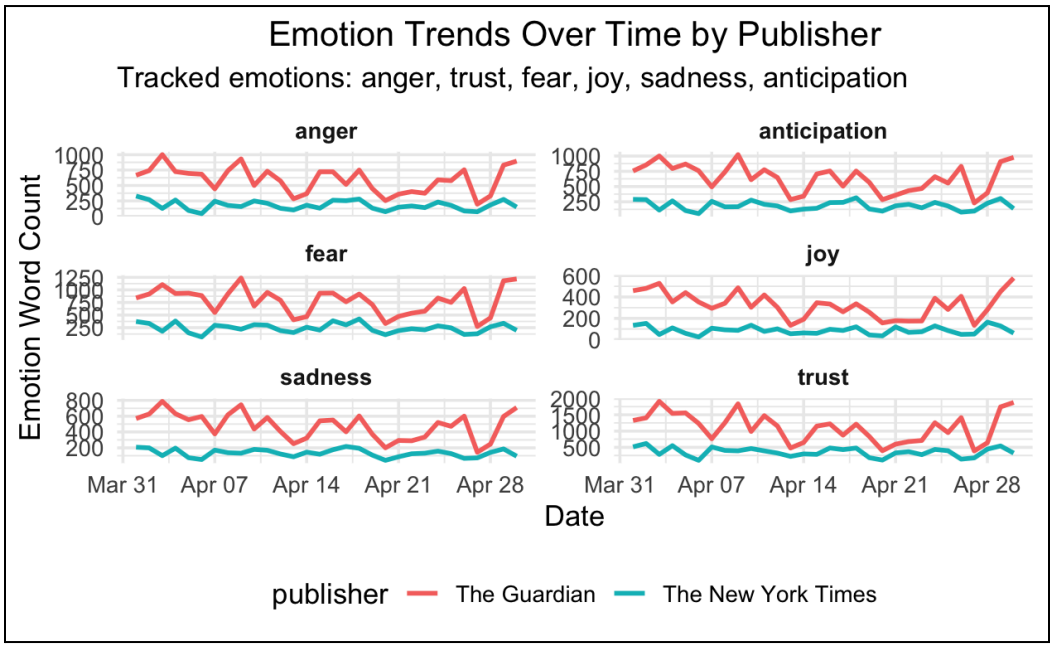




4. Emotion Trends

With the NRC lexicon, we tracked **emotion categories** like fear, anger, trust, and sadness.

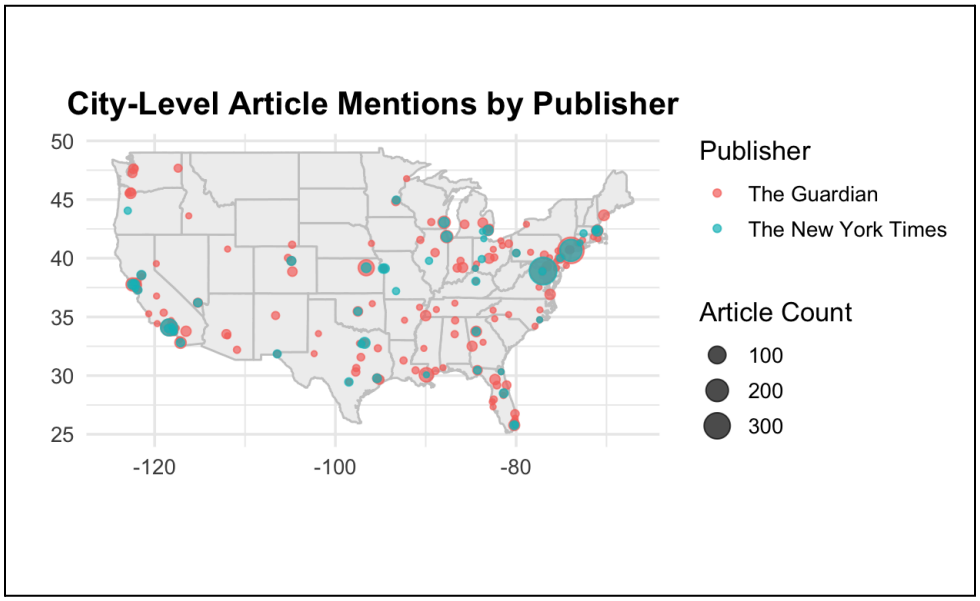
- The Guardian leaned more heavily into fear and trust language.
- NYT had lower emotion word counts overall, suggesting a more neutral framing.



5. **Geographic coverage**

Using Named Entity Recognition (NER), we extracted and mapped **mentions of U.S. cities and states**.

- NYT concentrated coverage around major U.S. cities and government hubs.
- The Guardian showed a **broader geographic spread**, with mentions from a wider set of states.



Key Insights for Decision-Makers

- **Editorial Tone**: Stakeholders interested in media framing should note The Guardian's more emotionally expressive tone vs. NYT's factual style.
- **Topical Focus**: Organizations monitoring narrative dominance (e.g., on global issues, protests, or education) will benefit from TF-IDF results identifying thematic emphasis.
- **Regional Impact**: State-level coverage insights can help advocacy groups or policymakers assess which regions are over- or under-represented in national discourse.
- **Media Monitoring**: Newsrooms and analysts can use these findings to benchmark tone, narrative style, and regional inclusion against peer outlets.

Key Insights & Discussions

Business Implications, Challenges, & Next Steps

The insights derived from this analysis offer valuable implications for a range of stakeholders:

- **News organizations** can use this framework to benchmark their editorial style, sentiment tone, and geographic reach against peer publications.
- **Media monitoring firms** can incorporate this methodology into real-time dashboards to detect bias, emotional framing, or underreported regions.
- **Policy analysts and advocacy groups** can evaluate which national topics are being emphasized or neglected, helping tailor their outreach strategies.
- **Researchers and educators** gain a reproducible pipeline for teaching and studying media bias, topic framing, and regional disparities in coverage.
- **News consumers** become more aware of editorial framing and narrative selection, improving critical media literacy.

This type of analysis moves beyond anecdotal bias claims and offers quantifiable patterns that can inform decisions in media strategy, public engagement, or policy research.

Challenges Faced

Several technical and data-related challenges emerged:

- **Incomplete Article Text**: Some New York Times articles lacked full body content, requiring filtering based on article length to maintain quality.
- **HTML and Formatting Noise**: The Guardian's text included heavy markup and redundant metadata, which required extensive regex cleaning.
- **Inconsistent Location Extraction**: City and state references were often embedded in context, leading to partial mismatches during geographic entity recognition.
- **Publisher Differences in Style**: Different narrative styles (e.g., The Guardian's frequent use of global references) introduced complexity in creating comparable topic models.

Despite these, systematic cleaning and standardization methods enabled meaningful cross-publisher comparisons.

Limitations & Next Steps

- **Temporal Scope**: The analysis was limited to the most recent 30 days. Expanding to multiple months could reveal seasonal or event-driven shifts in coverage.
- **Depth of Articles**: We relied on textual mentions rather than journalist intent or framing analysis. Future work could include manual annotations or discourse-level NLP.
- **NER Limitations**: Some U.S. cities with ambiguous names (e.g., “Springfield”) may have led to incorrect or missed geographic matches.
- **Sentiment Context**: Lexicon-based sentiment analysis does not always capture nuance (e.g., sarcasm, negation). A future extension could involve using BERT-based models for contextual sentiment classification.

In future work, incorporating more publishers, extending the timeline, or using user-engagement data (like social shares or comments) could significantly enhance the findings’ applicability to real-world decisions.

Conclusion

This project set out to compare how *The Guardian* and *The New York Times* report on U.S. national news using a structured data science approach. The workflow began with **data extraction** via official publisher APIs, followed by extensive **data cleaning** to standardize and prepare text for analysis. We then **merged** the datasets into a unified format, retaining key attributes like publication date, publisher name, and cleaned article body.

Subsequent text processing and natural language analysis enabled:

- **Topic modeling** using unigrams, bigrams, and TF-IDF to identify the most frequently and uniquely used terms per publisher.
- **Sentiment and emotion tracking** using lexicons like Bing and NRC to capture differences in emotional framing and tone.
- **Geographic mapping** to analyze the spread of coverage across U.S. states and cities.

Key Takeaways

- *The Guardian* leaned more toward emotionally expressive and internationally contextualized reporting, frequently using terms associated with global politics and social issues.

- *The New York Times* maintained a steadier tone and focused more on U.S. institutions, named figures, and localized events.
- Geographically, The Guardian showed broader state-level coverage, while the NYT concentrated its attention around key urban centers.
- Sentiment and emotion analysis confirmed that The Guardian used stronger expressions of fear, trust, and anger, whereas NYT displayed relatively neutral sentiment across time.

Together, these insights offer a powerful lens through which editorial focus, framing, and potential biases can be understood benefitting researchers, media strategists, educators, and informed readers alike.

References

- **The Guardian Open Platform API**
  Data retrieved from: https://open-platform.theguardian.com
  Used for collecting full-text U.S. national news articles from The Guardian.

- **The New York Times Article Search API**
  Data retrieved from: https://developer.nytimes.com
  Used for extracting recent national news articles tagged under U.S. topics.

- **NRC and Bing Sentiment Lexicons**
  Hu, M., & Liu, B. (2004). *Mining and Summarizing Customer Reviews*.
  Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a Word-Emotion Association Lexicon*.

- **TF-IDF Methodology**
  Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.