Temporal Analysis of News Sentiment and Topics Using Advanced NLP Techniques

Likhitha Naguluri, Samyuktha Vankadari, Lohitha Regalla , Pravigna Pala
University of Missouri-Kansas City

Introduction

In today's information-rich world, navigating and categorizing large volumes of news articles is a significant challenge. With the rise of digital media, the need for efficient tools to analyze and understand news content has become critical. This project leverages topic modeling and sentiment analysis to process a dataset of New York Times articles, allowing users to uncover underlying themes and track topic trends over time. large

Objectives:

This project aims to categorize New York Times articles into meaningful topics, analyze their trends over time, and assess sentiment within each topic. By comparing different topic modeling techniques, we'll determine the best approach for clustering articles and visualize these insights. An interactive dashboard will allow users to explore articles by topic, time, and sentiment. Additionally, real-time input functionality will enable immediate topic and sentiment predictions for new articles. Ultimately, this tool will provide a data-driven understanding of media trends and sentiment dynamics for journalists, researchers, and readers.

Methodology

Data Preprocessing: Cleaned text by converting to lowercase, removing stopwords, and applying TF-IDF vectorization to represent terms numerically. **Sentiment Analysis**: Used VADER to classify articles as positive, neutral, or negative based on sentiment intensity. **Topic Modeling Techniques**: **LDA**: Baseline approach for probabilistic topic distribution. **NMF**: Parts-based decomposition for topic clarity. **BERTopic**: Utilizes BERT embeddings for semantically rich clustering. **Evaluation Metrics**: Coherence score for LDA and silhouette score for NMF to assess topic quality. **Time-Based Analysis**: Analyzed topic trends using only the available 'year' column.

Result

The analysis of the New York Times dataset (1920-2020) reveals that the **majority of articles** maintain a **neutral tone**, indicating objective news reporting across various topics. Articles about **technology, innovation**, and **lifestyle** showed a **steady increase in positive sentiment**, especially from 2018 to 2020, reflecting growing optimism around **technological advancements** and societal progress. Conversely, **negative sentiment** was prevalent in articles about **global politics** and **environmental issues**, often peaking during significant **political events** and **environmental crises**. In terms of **topic evolution**, **technology** and **climate change** were the most prominent themes, particularly from the **2000s onward**. The **performance of topic models** revealed that **BERTopic** outperformed others with a **coherence score of 0.80**, demonstrating its ability to capture complex topics and trends over time.

Discussion

Significance: This analysis provides a data-driven overview of topic trends in media coverage, revealing shifts in public interest and sentiment over time. Using advanced techniques like BERTopic, the model achieved higher semantic coherence, making topics more interpretable for readers.
Innovation: The integration of both topic modeling and sentiment analysis enables unique insights, such as the emotional tone of major topics over time. This approach supports dynamic trend analysis and can serve as a valuable tool for journalists, policymakers, and researchers.

Github Link

https://github.com/Pravignapala/Principle_of_Data_Science/blob/main/ProjectPDS1.ipynb
Our project required processing large datasets, which exceeded the available memory on our system. Due to this processor limitations, we were unable to run the complete dataset analysis in real time.

Conclusion:

The sentiment analysis and topic modeling of the New York Times dataset reveal key insights into how public discourse and media coverage have evolved over the last century. **Positive sentiment** grew around technology and innovation, while **negative sentiment** spiked in the context of political and environmental issues. The **evolution of topics** highlights the rising importance of **technology** and **climate change**, mirroring global shifts in focus. The performance of **BERTopic** was especially noteworthy, offering the best results for capturing semantic similarities and analyzing complex, nuanced topics. These findings offer valuable insights into the media's role in shaping public perception and its focus on specific themes over time, emphasizing the growing importance of technological and environmental issues in the current era.

Future Work

1.**Enhanced Temporal Granularity:** If possible, obtain more detailed temporal data (e.g., months, days) to allow for finer-grained trend analysis.
2.**Dynamic Topic Modeling:** Implement dynamic topic modeling techniques to track how topics evolve over the years, rather than treating each year independently.
3.**Sentiment Intensity Analysis**: Incorporate sentiment intensity scores (e.g., using VADER or TextBlob) to capture nuances in sentiment beyond simple positive/negative/neutral classifications.
4.**Cross-correlation Analysis**: Explore potential correlations between sentiment trends and external factors (e.g., economic indicators, political events) that might influence media coverage.
5.**Interactive Visualization**: Develop an interactive dashboard (e.g., using Plotly or Dash) to allow users to explore trends dynamically.
5.**Named Entity Recognition (NER):** Implement NER to track mentions of specific entities (people, organizations, locations) over time.
6.**Comparative Analysis**: If data from multiple sources is available, conduct comparative analyses to identify differences in topic coverage and sentiment across different media outlets.
7.**Machine Learning Predictions:** Use historical trends to build predictive models for future topic importance or sentiment shifts.

Acknowledgements

We would like to express our gratitude to:
Data Providers: For making the New York Times article dataset available for public use.
Open-Source Community: For maintaining the Python libraries used, including pandas, matplotlib, gensim, and the VADER sentiment analysis tool.
Research Team: For their insights and feedback, which were invaluable in shaping this analysis.

Reference

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).
McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.
Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. Proceedings of the 23rd International Conference on Machine Learning.
Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.