NLP TECHNIQUES

To examine NYTimes user comments on articles about Ozempic

Problem Statement

The goal of this analysis is to examine NYTimes user comments on articles about Ozempic using various Natural Language Processing (NLP) techniques. This study aims to uncover patterns in reader discussions, including sentiment, key themes, and linguistic trends.

Key Objectives:

- Extract and preprocess NYTimes user comments related to Ozempic.
- Apply multiple NLP techniques, such as:
  - Sentiment Analysis (positive, negative, neutral classifications).
  - Topic Modeling (identifying key discussion themes).
  - Named Entity Recognition (NER) (detecting important entities like brands, drugs, or side effects).
  - Word Frequency & N-gram Analysis (understanding common words and phrases).
- Compare sentiment patterns and discussion trends across different articles.

Topic - Ozempic

What is Ozempic?

Ozempic is a prescription medication primarily used to treat type 2 diabetes, but it has gained significant attention for its off-label use in weight loss due to its ability to regulate appetite and blood sugar levels.

What We Are Trying to Find with NYT Comments on Ozempic

By analyzing comments from NYT articles related to Ozempic, we aim to identify public sentiment, key concerns, and emerging themes surrounding its use. This includes understanding opinions on its effectiveness, side effects, accessibility, ethical considerations, and societal perceptions of weight-loss drugs.

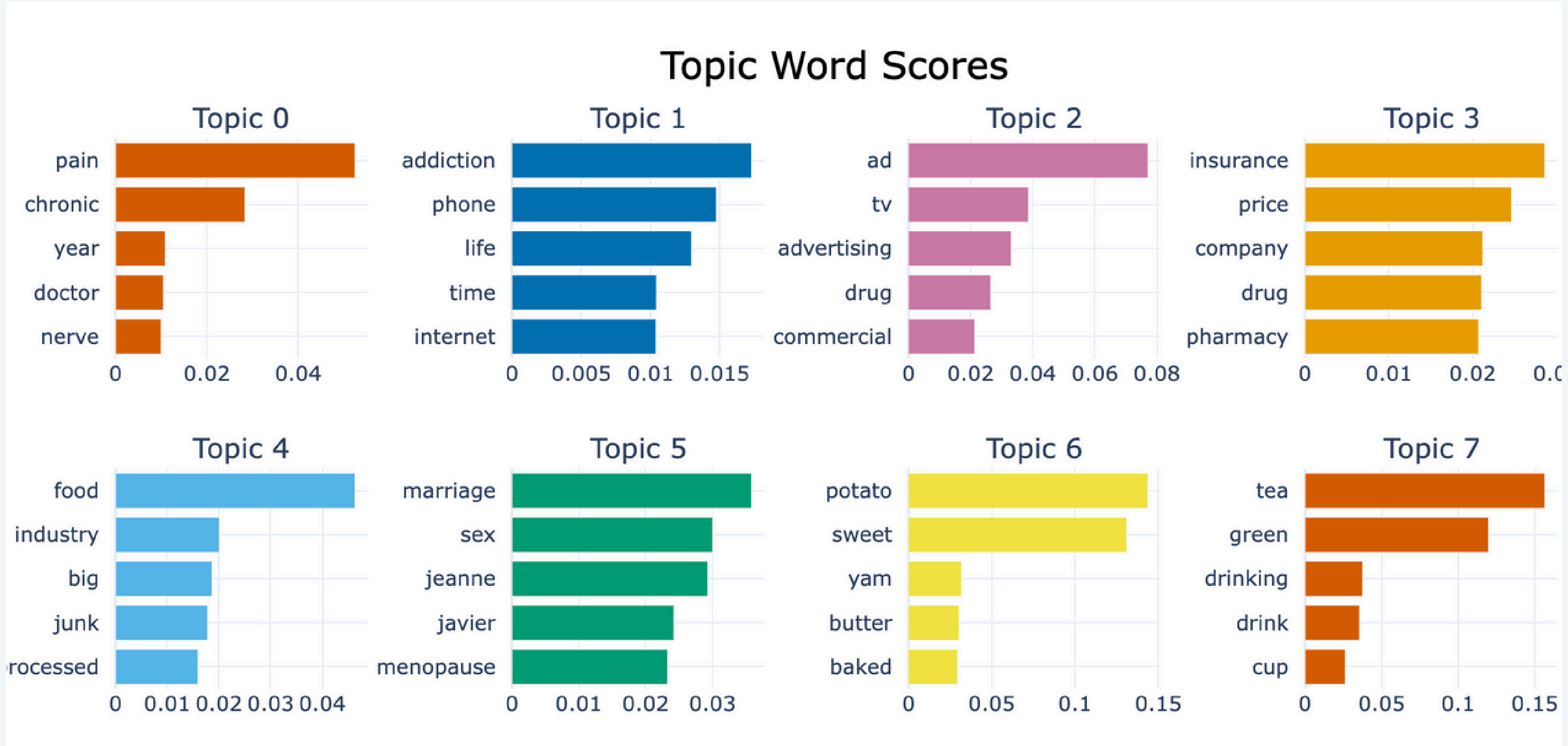
Topic Modeling

Topic modeling is an unsupervised Natural Language Processing (NLP) technique used to discover hidden topics in a large collection of text. It helps in organizing, summarizing, and structuring textual data by identifying groups of words that frequently occur together.

Technique Used: BERTopic

BERTopic leverages BERT embeddings along with UMAP for dimensionality reduction and HDBSCAN for clustering. This approach has helped me extract more meaningful and coherent topics from unstructured text data, making it particularly effective for analyzing short texts like customer feedback, social media posts, and business documents.

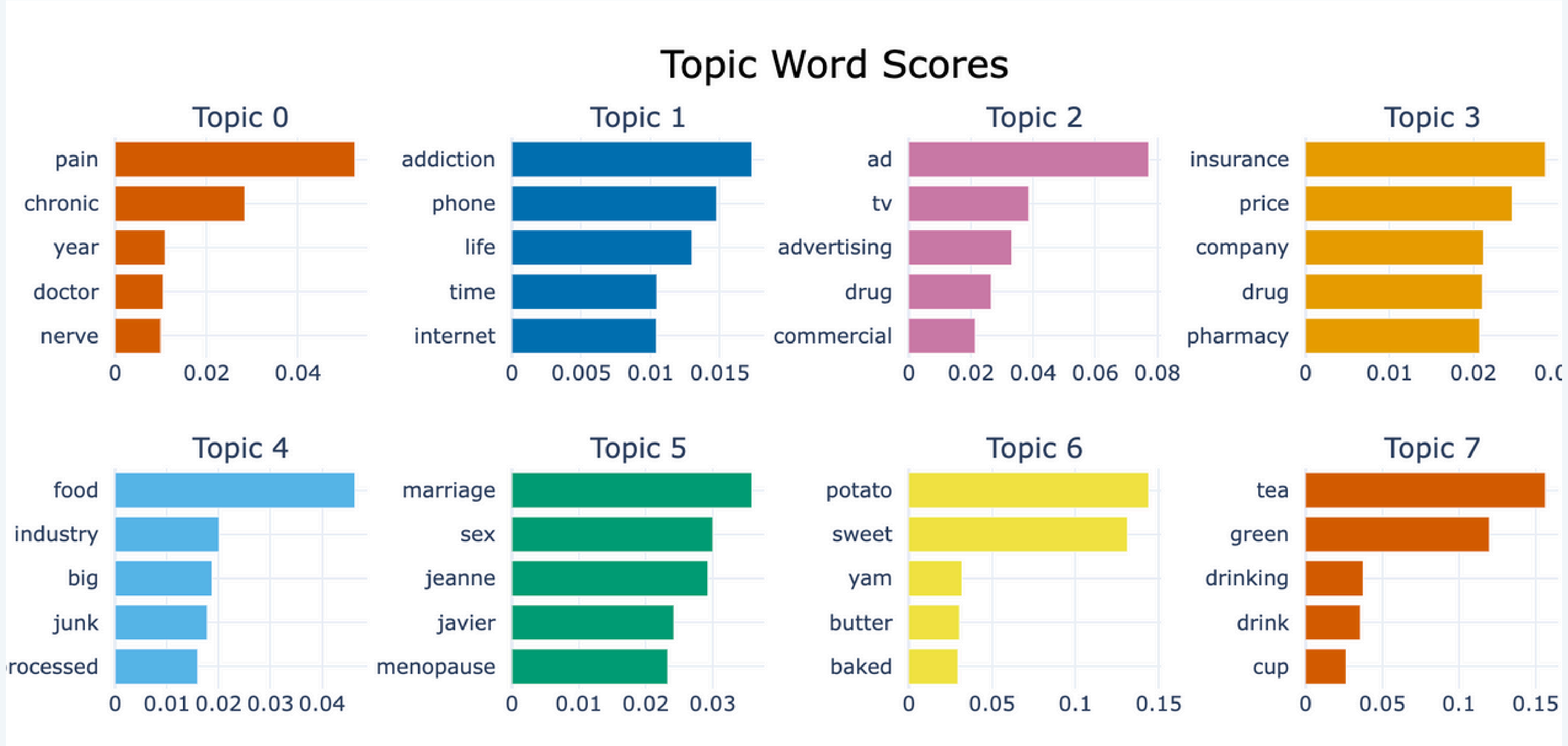
Topic Modeling



What these Topic represents?

- **Topic 0** - Chronic Pain & Healthcare → Some discussions around Ozempic include side effects like nausea and potential long-term health impacts.
- **Topic 1** - Technology & Addiction → Ozempic has gained viral attention on social media, with discussions about people becoming reliant on it for weight loss.
- **Topic 2** - Advertising & Media Influence → Ozempic is widely advertised, influencing consumer perceptions, which is reflected in the ad, TV, advertising, drug, and commercial keywords.
- **Topic 3** - Pharmaceutical Industry & Pricing → The high price and insurance coverage of Ozempic are major topics of debate, shown by the insurance, price, company, drug, and pharmacy keywords.
- **Topic 4 -** Food Industry & Junk Food → Ozempic is often discussed in the context of reducing cravings for processed and junk food, which aligns with this topic.
- **Topic 5** - Marriage & Relationships → Weight loss from Ozempic has reportedly impacted relationships and body image discussions.
- **Topic 6 -** Sweet Potatoes & Cooking → Diet changes with Ozempic include altered food preferences, such as decreased cravings for carbs.
- **Topic 7** - Tea & Beverages → Ozempic users often discuss changes in eating and drinking habits, including increased consumption of tea and lighter foods.

Topic Modeling



Why "drug" Appears Twice

- **Topic 2 (Advertising & Media)** → How Ozempic is marketed and promoted.
- **Topic 3 (Pharmaceutical Industry & Pricing)** → How Ozempic is priced and regulated.

This reinforces the dual nature of Ozempic's popularity—its media hype and its cost/availability challenges.

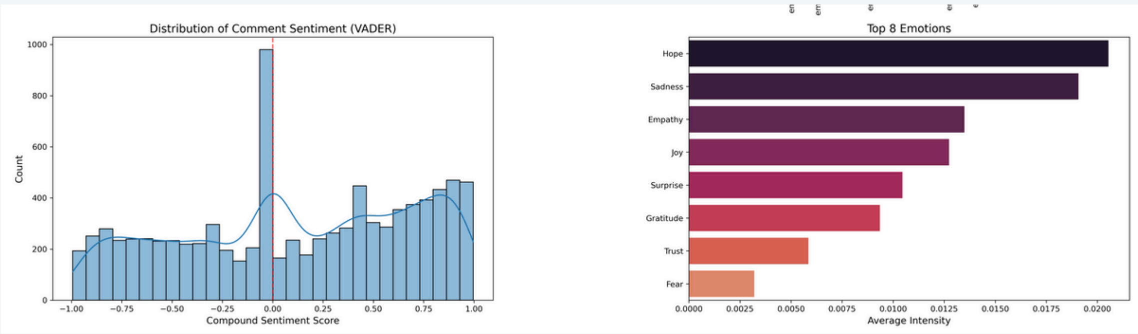
UNVEILING READER EMOTIONS

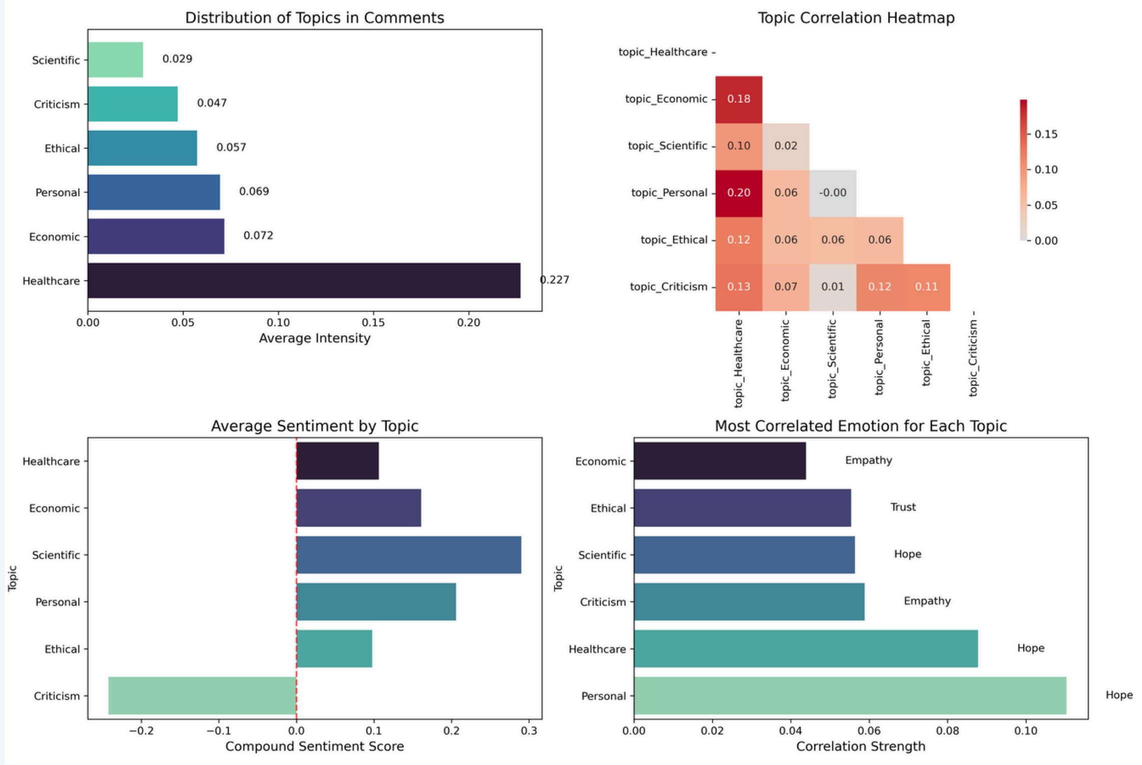
Emotion Insights

Hope (0.021) and Sadness (0.019) are the dominant emotions expressed in comments. Empathy and Joy follow closely, suggesting readers engage emotionally with content. Negative emotions like Anger and Disgust are least prevalent, indicating mostly constructive discourse.

Emotional Distribution

VADER Analysis reveals a bimodal distribution with peaks at neutral (0) and positive (0.75). Scientific topics show the most positive sentiment (0.25), while Criticism is slightly negative (-0.05).
Most topics trend positive, suggesting constructive rather than critical commentary.





Topic-Emotion Relationships

Hope correlates strongly with Healthcare, Personal and Scientific topics.
Empathy correlates with Economic and Criticism topics, showing reader connection.
Trust correlates with Ethical topics, reflecting discussions about moral dimensions of healthcare.

Topic Analysis

Healthcare (0.227) is the dominant topic by a significant margin. Economic issues (0.072) and Personal experiences (0.069) are secondary topics. Scientific content received less attention despite being health-related, suggesting commenters focus more on healthcare access/delivery than research.

Sentiment Analysis

Sentiment analysis is a Natural Language Processing (NLP) technique used to determine the emotional tone behind a body of text. It helps identify whether the expressed opinion is positive, negative, or neutral, providing quick insights into the overall attitude or mood in the text.

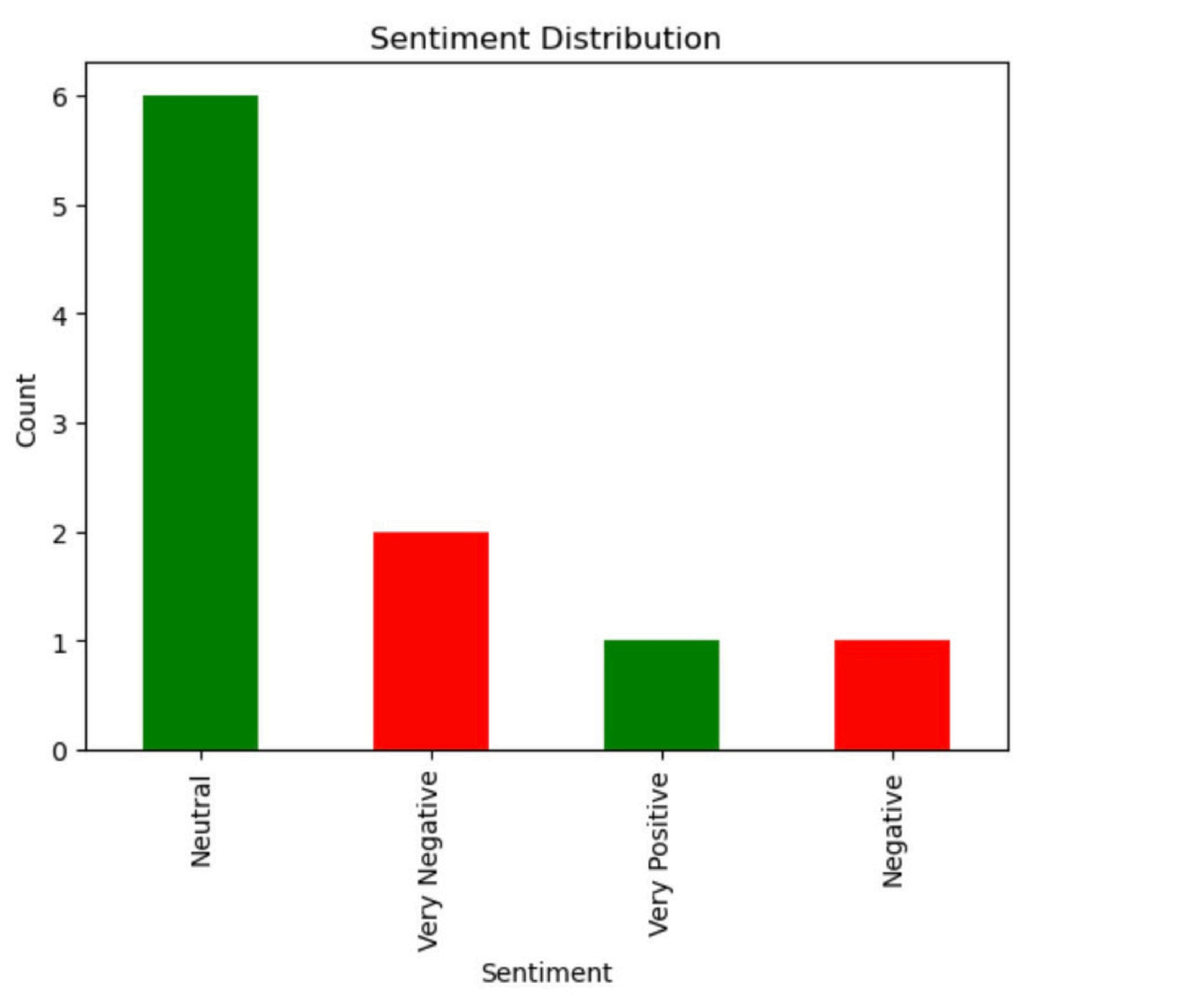
Technique Used: Hugging Face Transformers

We utilized a pretrained Hugging Face model (DistilBERT fine-tuned on SST-2) for sentiment classification. This model reads each article's text and outputs a sentiment label (Positive, Negative, or sometimes Neutral) along with confidence score.

Why It's Useful:

- **Quickly Gauges Tone**: We can see if coverage around Ozempic skews positive or negative.
- **Scalable**: Works on multiple articles at once, letting us track sentiment trends across a dataset.
- **Easy to Implement**: Hugging Face pipeline allow us to integrate powerful language models with just a few lines of code.

Sentiment Analysis



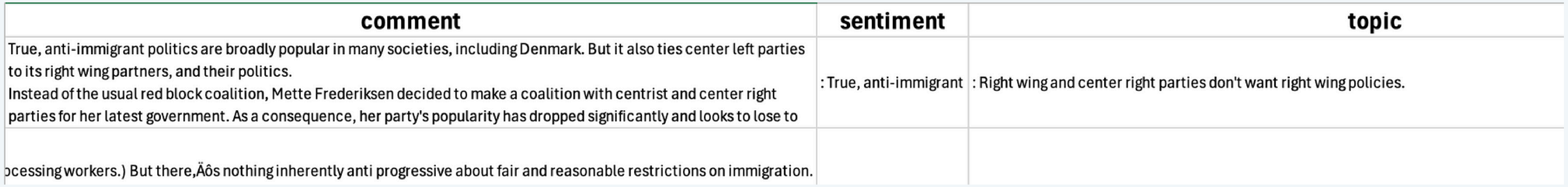
Main Observations

- **Majority are Neutral**: Most articles present a balanced or informational tone rather than a strongly positive or negative stance.
- **Noticeable Negative Proportion**: A significant share of articles lean negative, possibly discussing concerns (e.g., side effects, cost, controversies).
- **Smaller Positive Portion**: Fewer articles highlight positive aspects (e.g., effectiveness for treating conditions, success stories).

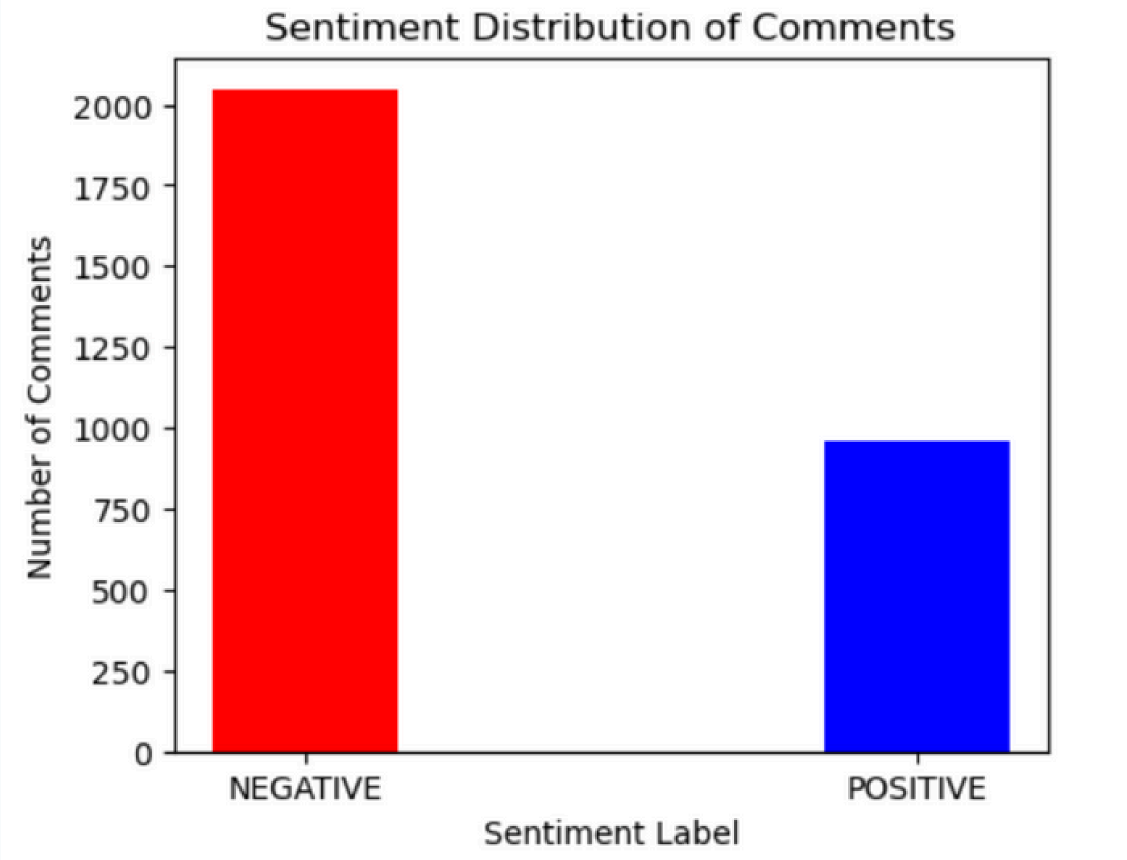
Insights & Limitations

- **Context Matters**: Medical/health-related topics often mix facts and opinions. A “neutral” label may indicate straightforward reporting.
- **Model Limitations**: Sentiment Analysis can oversimplify nuanced topics—especially in medical or policy-related articles.
- **Further Analysis**: We could dive deeper into each sentiment category to see which themes (e.g., cost, accessibility, side effects) drive these sentiments.

DistilBERT vs. T5: A Comparative Analysis



comment	sentiment	topic
True, anti-immigrant politics are broadly popular in many societies, including Denmark. But it also ties center left parties to its right wing partners, and their politics. Instead of the usual red block coalition, Mette Frederiksen decided to make a coalition with centrist and center right parties for her latest government. As a consequence, her party's popularity has dropped significantly and looks to lose to	: True, anti-immigrant	: Right wing and center right parties don't want right wing policies.
ocessing workers.) But there,Äôs nothing inherently anti progressive about fair and reasonable restrictions on immigration.		



Insights & Limitations

- Upon conducting sentiment analysis with both models, DistilBERT proved to be fast and reliable, assigning clear Positive/Negative labels that align well with structured data visualization. However, it lacks depth in explaining sentiment beyond the classification.
- T5, on the other hand, generates contextualized sentiment explanations, capturing nuances in the text. While this makes it more human-like, it introduces variability and is computationally heavier, making structured analysis more challenging .
- Key Takeaway: DistilBERT is ideal for efficient, large-scale classification, while T5 offers richer sentiment insights but requires more refinement for structured use.

NER(Name Entity Recognition) Analysis

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique that identifies key entities such as people, organizations, locations, and products in a given text. It helps categorize and analyze which subjects are frequently mentioned.

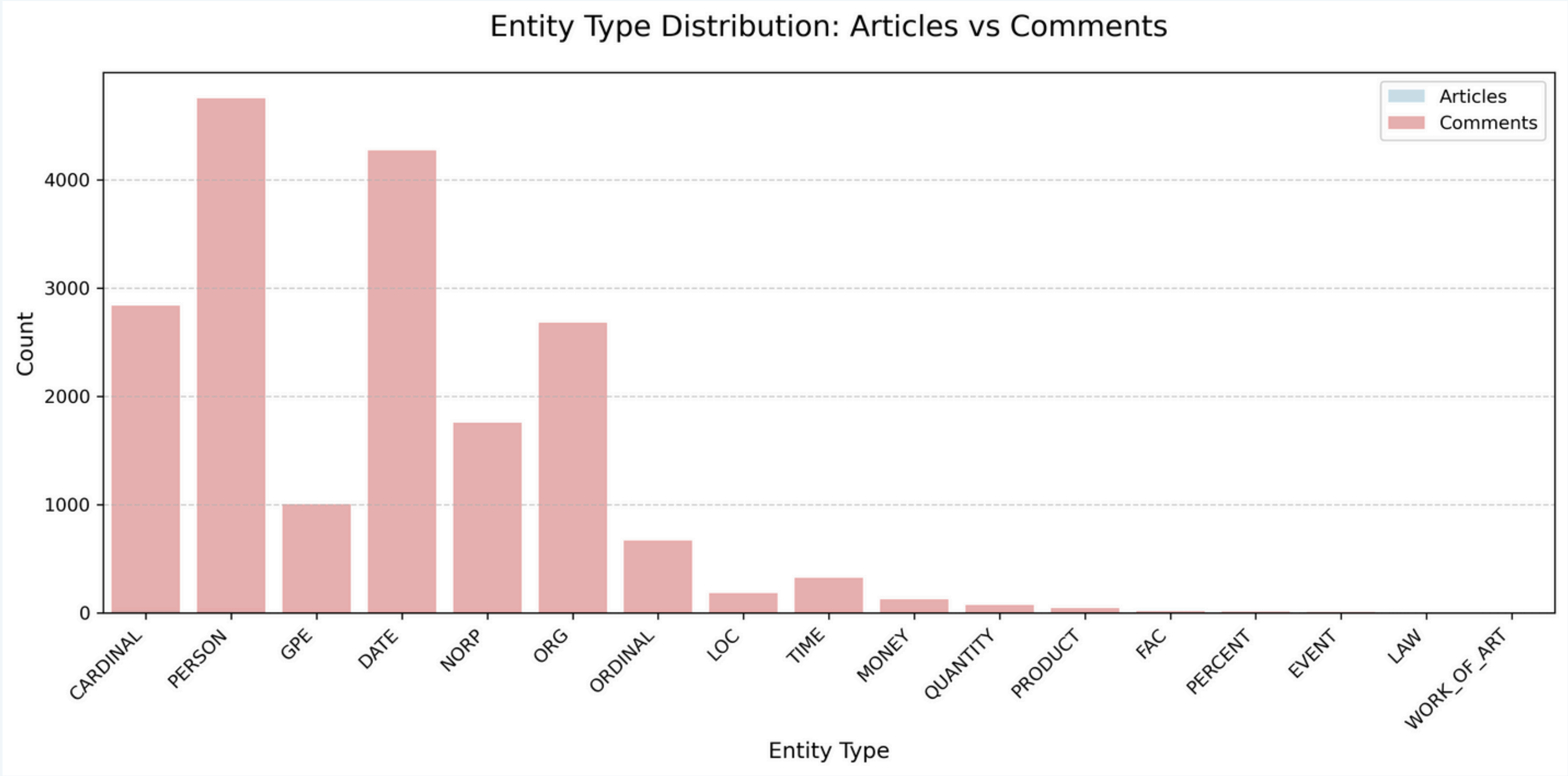
Technique Used: spaCy

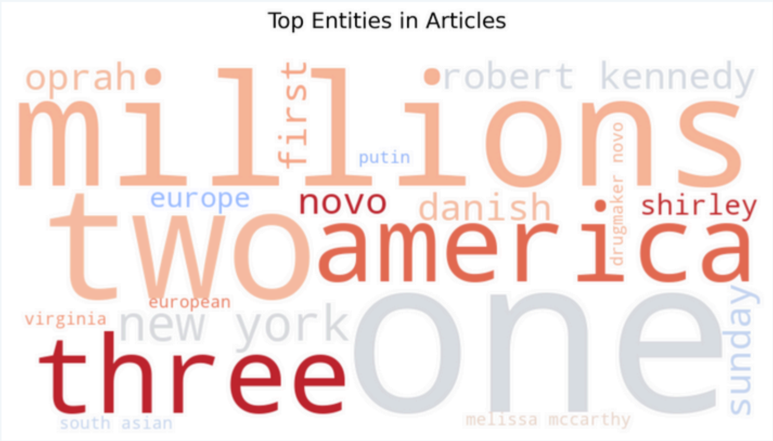
We used spaCy, a state-of-the-art NLP library, to extract named entities from both NYT articles and user comments. Additionally, a custom rule was added to ensure Ozempic is classified as a PRODUCT.

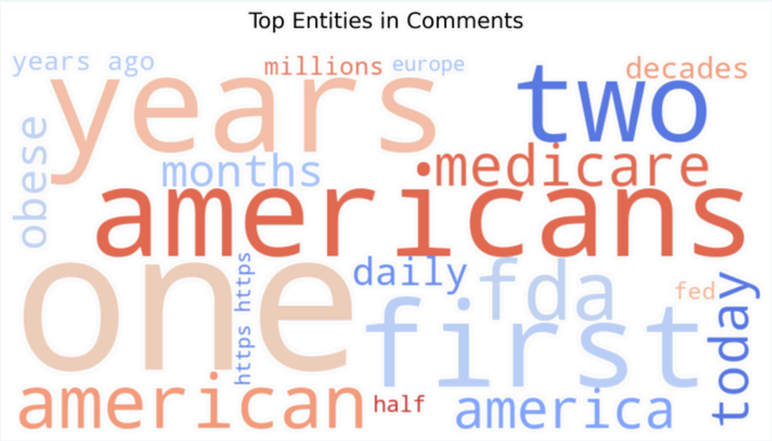
Why It's Useful:

- **Identifies Key Subjects**: Analyzes entities linked to Ozempic (e.g., companies, public figures, locations).
- **Compares Sources**: Reveals differences between media coverage and public discussions.
- **Enables Deeper Analysis**: Supports sentiment and topic modeling for further insights.

NER(Name Entity Recognition) Analysis







Main Observations

- **Articles vs. Comments:** Articles focus on organizations (Novo Nordisk, FDA) and geographic locations (America, Europe), reflecting a policy-driven narrative.
  - Comments highlight personal concerns, with frequent mentions of Medicare, cost, and obesity, showing a focus on accessibility and impact.
- **Word Cloud Insights**: Articles emphasize corporate and regulatory entities.
  - Comments reflect public discourse on health, affordability, and policy.

Insights & Limitations

- **Different Perspectives:** Media focuses on policy and companies, while comments emphasize personal stories and healthcare concerns.
- **NER Challenges**: Some detected words (e.g., "one," "two") lack context—filtering needed for better insights.
- **Further Analysis**: Co-occurrence analysis could reveal how entities relate to concerns like cost or regulation.

Result

The analysis shows a clear contrast between media coverage and public discussion on Ozempic. Articles focus on corporate and policy aspects, frequently mentioning Novo Nordisk and the FDA, while comments highlight personal concerns like Medicare, cost, and accessibility. Sentiment trends reveal mostly neutral coverage, but comments lean more negative, reflecting affordability and healthcare frustrations.

Conclusion

Media presents a structured, policy-driven narrative, while public discussions focus on personal impact and healthcare costs. This contrast highlights the gap between reporting and real-life concerns. Future research could explore entity relationships and sentiment trends to better understand public concerns over time.

THANK YOU