AI Impact Analysis: Unveiling the Future of Jobs and Tasks

For Nick Kadochnikov, Ignas Grabauskas and NLP Colleagues
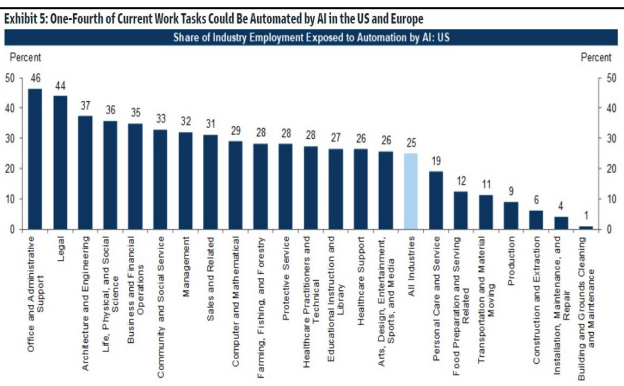
Parth Bansal

Agenda

- Executive Summary
- Actionable Recommendations
- Article Clean-up and Filtering
- Sentiment Analysis
- Sentiment Over Time Analysis
- Topic Detection
- Entity Identification
- Targeted (Entity) Sentiment Analysis

Executive Summary

Exhibit 5: One-Fourth of Current Work Tasks Could Be Automated By AI in the US and Europe

Share of Industry Employment Exposed to Automation by AI: US

1. Around 200K news articles on Data Science, Machine Learning and Artificial Intelligence were cleaned, filtered and analyzed to gain insights and identify the industries and job lines that are most likely to be impacted to impacted by AI. The articles were thoroughly preprocessed and tokenized resulting in a final dataset of ~170K news articles.

2. Using VADER and a custom Yelp Model, I performed sentiment analysis on these articles to the success or failure of AI applications. Successful data science initiatives were associated with terms with research, market analysis learning and skill while failed initiatives were associated with terms rights reserved, people, risk and privacy.

3. The sentiments were analyzed over time. All in all, the overall sentiment towards data science technologies remains highly positive despite the multiple privacy and legal concerns. Hence, as suggested in the study in question, innovations in fields of deep learning, natural language processing, computer vision, and reinforcement learning are reshaping the field and demanding a certain set of skills from the workforce (data science related.) Hence, relevant actions need to be taken by individuals to avoid being replaced by AI (discussed later).

Executive Summary cont.

4. Using NER SpaCy, entities were extracted from the news articles and analyzed for both positive and negative sentiments and plotted against a timeline (shown later in the ppt). Google and Microsoft had the highest mentioned which is due to their advancements in generative AI technologies like ChatGPT and BART. Although these companies have garnered positive associations for their innovations, they have also faced negative sentiments primarily driven by concerns surrounding privacy, data collection, and potential monopolistic practices.

5. The top five negative topics (limited mentions) in the analysis included AI's impact on fake image recognition and recent developments, emerging trends and developments in AI news, negative sentiment around AI and cancer research, general AI risks including ethical concerns, and AI's impact on job loss. These topics reflect the concerns and challenges associated with the widespread adoption and advancement of AI technologies. Even though these topics are very relevant and highlight the cons of AI, their mentions were very limited indicated the positive impact/ gravity of AI over everyday tasks.

Actionable Recommendations

1. Companies should invest in AI-related training and upskilling programs to enhance your knowledge and skills in AI concepts, algorithms, and best practices. This will help develop strong data literacy skills and embrace data-driven decision-making to remain valuable in an AI-driven workforce.

2. Governments and Organizations should prioritize human-AI collaboration, invest in reskilling initiatives, and create new roles that effectively utilize AI technology while retaining human workers.

3. Individuals should encourage a culture of continuous learning and collaboration to stay updated with the latest advancements in AI technology. Individuals needs to stay informed about emerging trends and opportunities in the AI industry, particularly in domains like ChatGPT and OpenAI, and consider leveraging them to enhance your professional prospects. Also, individuals should also keep an eye on evolving job market demands and explore opportunities to reskill or transition into roles that complement AI technology rather than being replaced by it.

4. Lastly, government should stay develop and update laws arounds on privacy and data laws related to AI, image generation, and AI usage in media, and advocate for the implementation of more specific and robust regulations in these areas.

Article Clean-up and Filtering
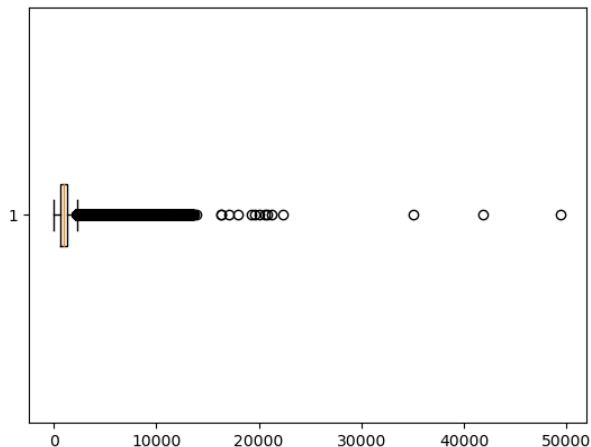
Visualization of Text Length Distribution:
Understanding Optimal Lengths for Articles

1. To ensure consistency and relevance, an initial filtering process was implemented to include only articles written in English.

2. Subsequently, an article cleanup procedure was performed, aiming to identify and eliminate any URLs, remnants of web crawls, tabs, newlines, non-printable characters, non-ASCII characters, and email addresses.

3. Then, they were transformed into word tokens, allowing for more in-depth analysis. In order to eliminate irrelevant words and numerical values, a filtering step was applied to remove stop words and numbers. Additionally, words were further processed using lemmatization, which reduced them to their base form, enabling more accurate analysis and comprehension.

4. Articles that were excessively long or short were identified and subsequently removed from the dataset. The filtering was based on the 5th to 95th percentile values, allowing for the retention of articles that fell within the optimal range.

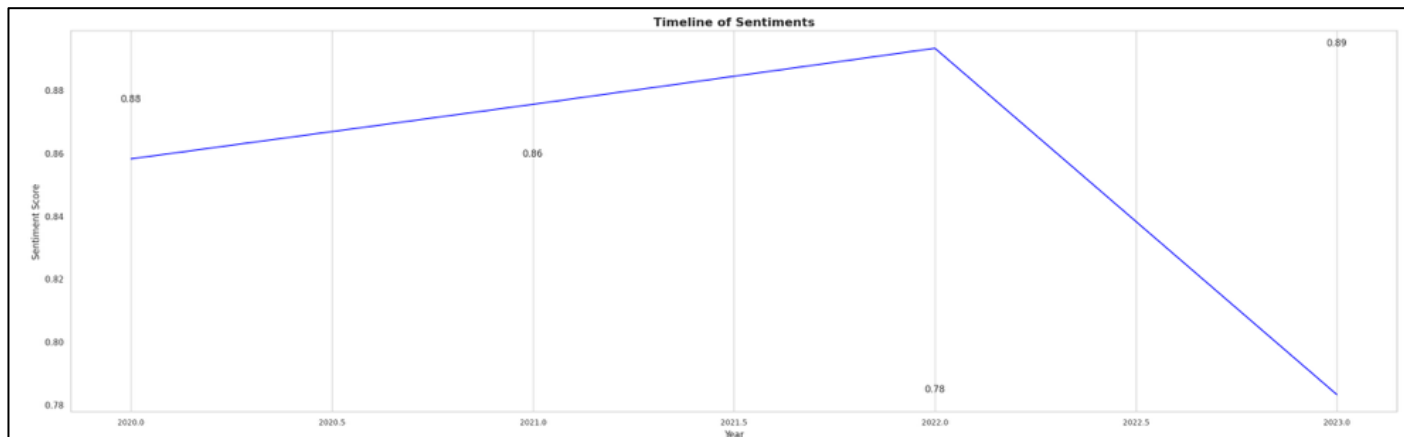5. In the end, **there were 170,984 articles remaining.**

Sentiment Analysis

1. To identify the top reasons for successful data science / AI initiatives, I predicted the sentiments of the articles and analyzed the sentiment brackets separately using 2 different approaches.

2. The first approach utilized a dictionary-based method using VADER (Valence Aware Dictionary and Sentiment Reasoner) for sentimental analysis. VADER assigns sentiment scores to lexical features, enabling the bracketing of sentiments into different categories.

3. The second approach involved training a custom sentiment prediction model using pre-existing labeled Yelp data and classifier.

4. The models were then analyzed separately by predicting the sentiments of our dataset. The labelled articles were then analyzed by reading some of them manually. **VADER was chosen as the preferred approach to move forward with sentiment classification.** Its accuracy and effectiveness in analyzing sentiment scores outweighed the performance of the custom Yelp model.

5. 158,890 were classified as positive, while 12,085 were classified as negative.

*refer to nlp04 and nlp05 notebooks for code reference

Sentiment Analysis

Associated with Positive Sentiment

Associated with Negative Sentiment

*refer to nlp04 and nlp05 notebooks for code reference

Sentiment Analysis over Time (Year)

The plot above depicts the overall sentiment over the years for data science and machine learning technologies reveals a consistent upward trend until 2022. However, a drop in sentiment is observed from 2022 to 2023. One possible reason for this is 'the initial hype circle'. Data science technologies experienced a period of hype and excitement, leading to an initial surge in sentiment. As the technologies matured and became more widely adopted, the initial enthusiasm might have waned, resulting in a decline in sentiment. All in all, as you can see, **the overall sentiment is still mainly positive (0.78-0.89) for data science technologies.**

*refer to nlp04 and nlp05 notebooks for code reference

Sentiment Analysis over Time (Month, Day)
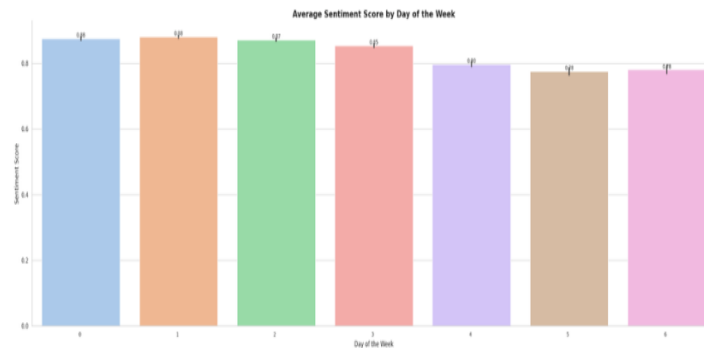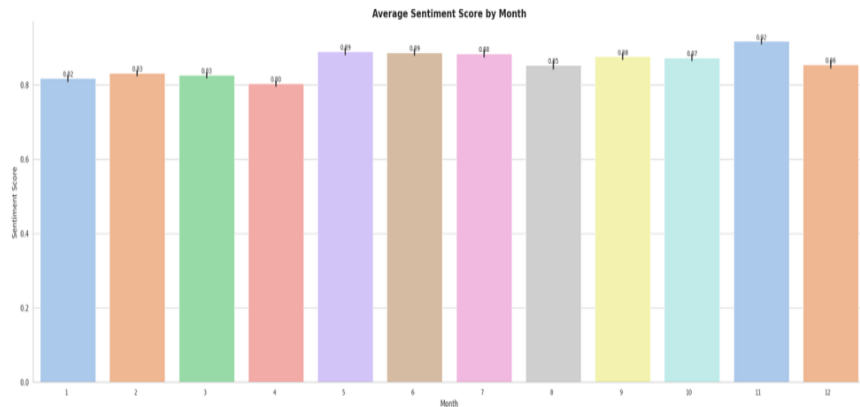
The displayed plots illustrate the average sentiments of Data Science and AI technologies over the years, indicating predominantly positive sentiments. There is no noticeable decline in positive sentiments. The relatively lower bars could solely be a result of publishing dates of the articles or the release dates of the AI technologies.

*refer to nlp04 and nlp05 notebooks for code reference

Sentiment Analysis over Time (Date)

```
df_pos_max.groupby(["date"])["sentiment_score"].mean().plot()
plt.title("Timeline of highly positive sentiments")
plt.xlabel("Date")
plt.ylabel("Sentiment_Score")
```
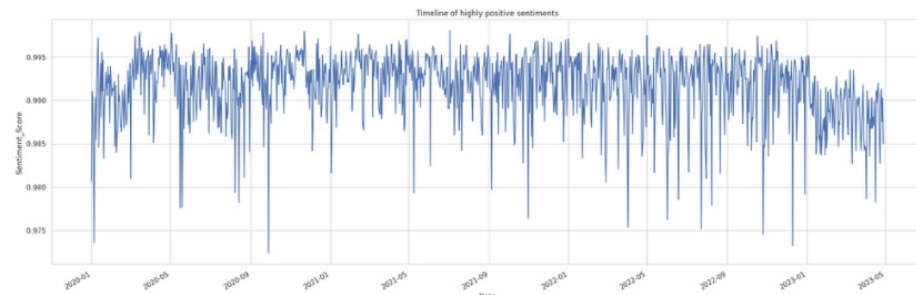
Text(0, 0.5, 'Sentiment_Score')
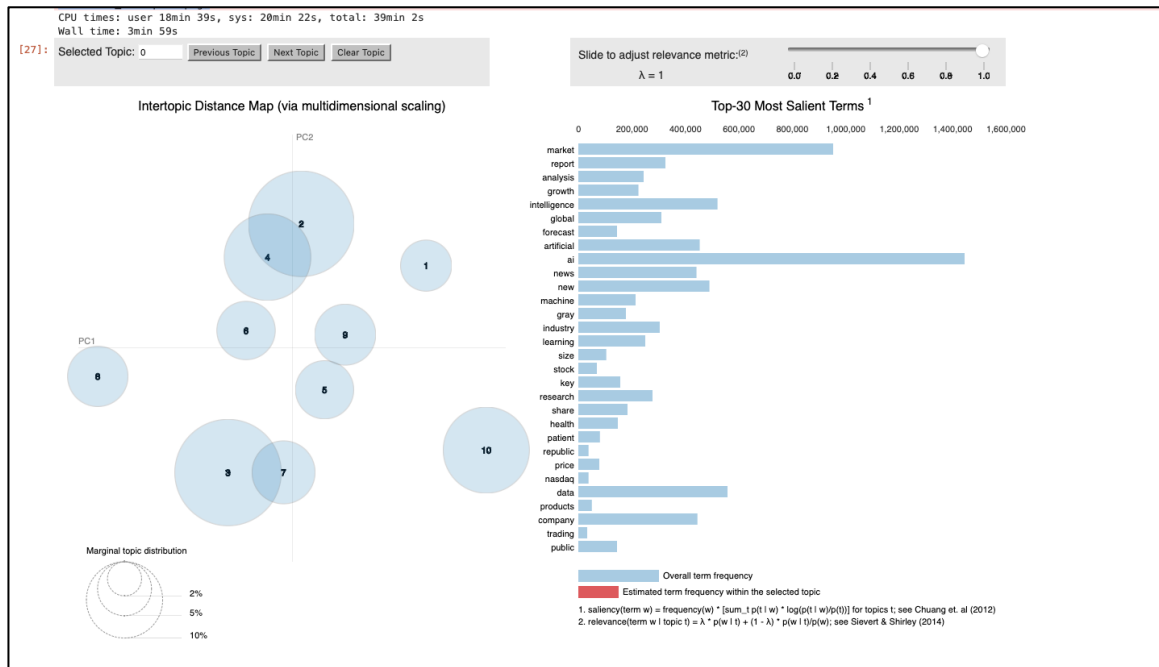

Timeline of highly positive sentiments

```
df_neg_max.groupby(["date"])["sentiment_score"].mean().plot()
plt.title("Timeline of highly negative sentiments")
plt.xlabel("Date")
plt.ylabel("Sentiment_Score")
```

Text(0, 0.5, 'Sentiment_Score')

Timeline of highly negative sentiments

*refer to nlp04 and nlp05 notebooks for code reference
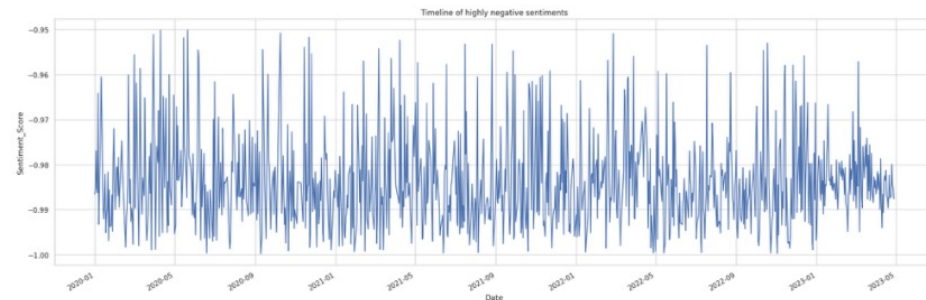
# Topic Detection (using LDA) – Positive Sentiments

Top 5 positive topics:

1. AI's Impact on the Stock Market and Company News

2. Advancements in AI Technology

3. AI Solutions for Customer Engagement and Media Platforms

4. The Role of AI in Artificial Intelligence News and Technology Updates

5. AI's Influence on Healthcare and Patient Data Management

Topic Detection (using LDA) – Negative Sentiments
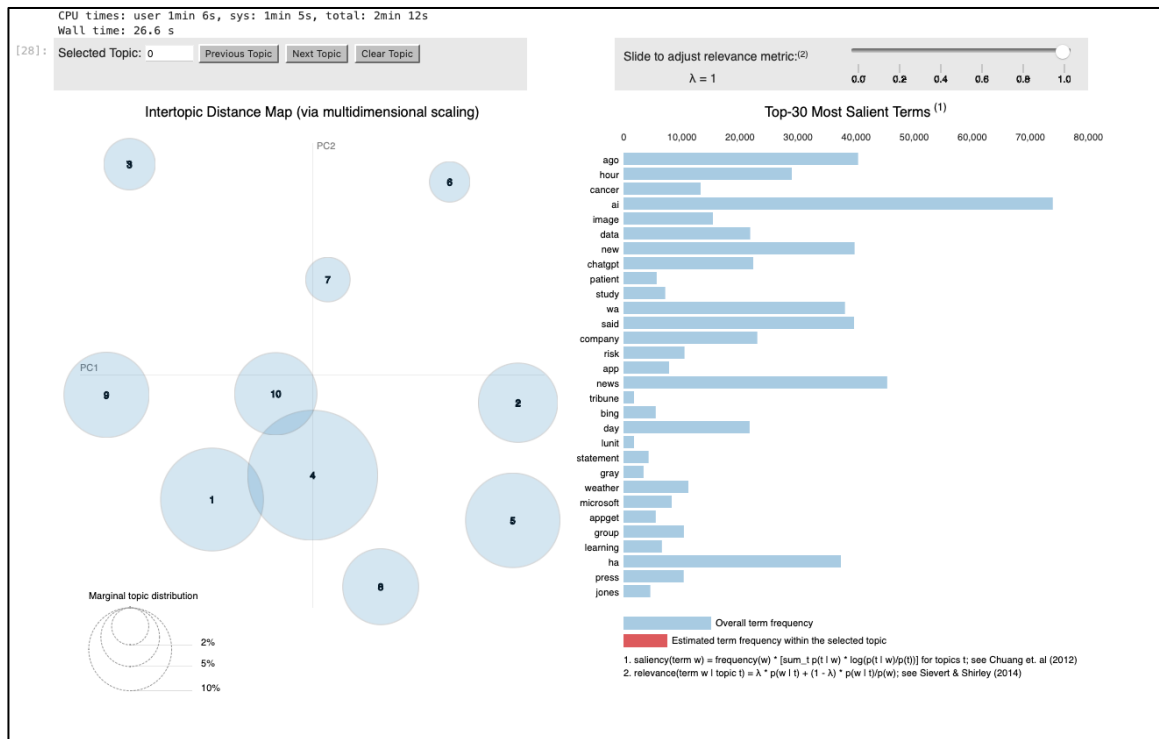
Top 5 negative topics:

1. AI's Impact on Fake Image Recognition and Recent Developments

2. Emerging Trends and Developments in AI News

3. Negative Sentiment around AI and Cancer Research

4. General AI Risks (+ Ethical concerns)

5. AI Impact on Job Loss

*refer to nlp06 notebook for code reference

Entity Identification

The **NER SpaCy model** was utilized to extract entities such as people, organizations, and locations from the sentiment analysis. Both Google and Microsoft emerged as prominent entities in the analysis, as they have made significant advancements in Generative AI. However, it is noteworthy that only articles associated with negative sentiment featured mentions of individuals among the top 10 named entities.
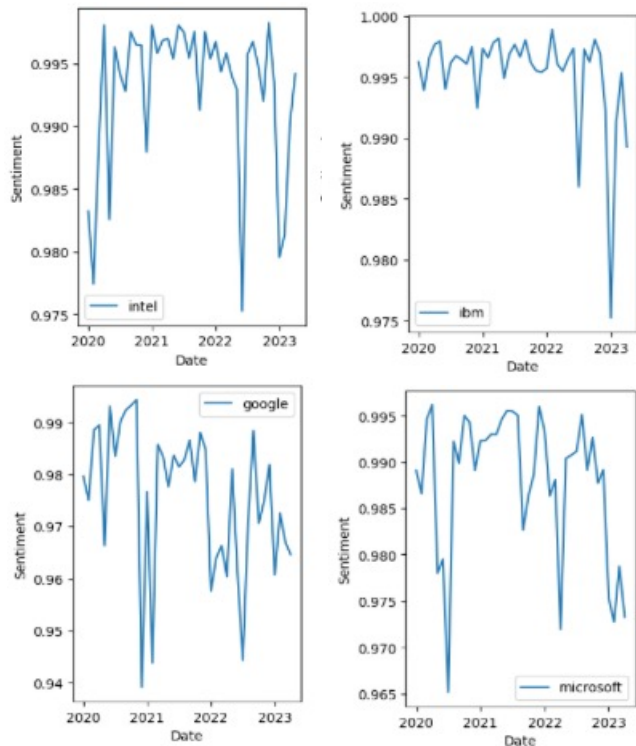
Associated with Positive Sentiment

Entity	Count
Microsoft	74,279
Google	61,410
IBM	36,324
Gray Television, inc	34,960
Intel	18,624

Associated with Negative Sentiment

Entity	Count
Google	8,093
Microsoft	7,223
GPT	1,059
Donald Trump	1,231
Media Inc.	1,923

*refer to nlp06 notebook for code reference

Targeted Entity Sentiment Analysis – Positive

The growing popularity of products from both Microsoft and Google was accompanied by a mixture of negative and positive sentiments. However, it is important to note that the positive sentiments significantly outweighed the negative ones. In the case of Microsoft, the partnership with OpenAI and the subsequent acquisition brought about a surge of positive sentiments, highlighting the potential benefits and advancements in the field of AI. Similarly, Google's continuous innovation and development of new technologies like BARD garnered widespread positive sentiments, reflecting the positive impact they have had on various industries. Despite the occasional negative sentiments that may arise alongside increased popularity, the overall sentiment towards both Microsoft and Google remained largely positive, highlighting their success and influence in the AI landscape.

*refer to nlp06 notebook for code reference

Targeted Entity Sentiment Analysis – Negative
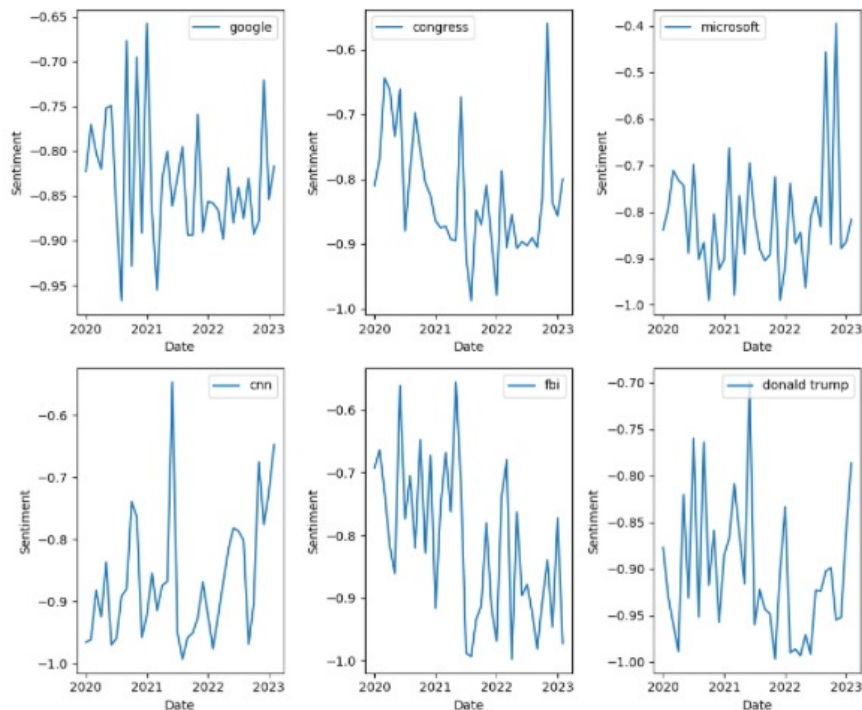
The presented plots depict the timeline of negative sentiments associated with the top entities over the years. As previously discussed, both Google and Microsoft have faced negative sentiments primarily stemming from concerns regarding privacy, data collection, and monopolistic practices. It is noteworthy that CNN, being a prominent news organization, also exhibits a significant number of articles associated with negative sentiments. This observation can be linked to the insights derived from the Negative Topic Detection slide, which highlights emerging trends and developments in AI news.

*refer to nlp06 notebook for code reference

Thank you

Have a good summer :)