



Natural Language Processing and Cognitive Computing

Final Project

Sirinda Leesuravanich (Pang)

Executive Summary

This project leverages BERTopic, NER, and Sentiment Analysis to analyze 200,083 AI-related news articles. Key findings include:

- **Industries Most Impacted by AI**

- *Technology* is the most impacted industry, as AI is now central to innovation, with nearly all major tech companies investing in or integrating AI into their products.
- *Marketing and Healthcare* show successful AI adoption, with high article volume and mostly positive sentiment.
- *Legal Services, Media & Entertainment, and Employment & Workforce* face challenges with AI integration. The frequent negative sentiment reflects concerns about job loss, AI misuse, and creative disruption, despite clear potential.

- **Drivers of Successful AI Adoption**

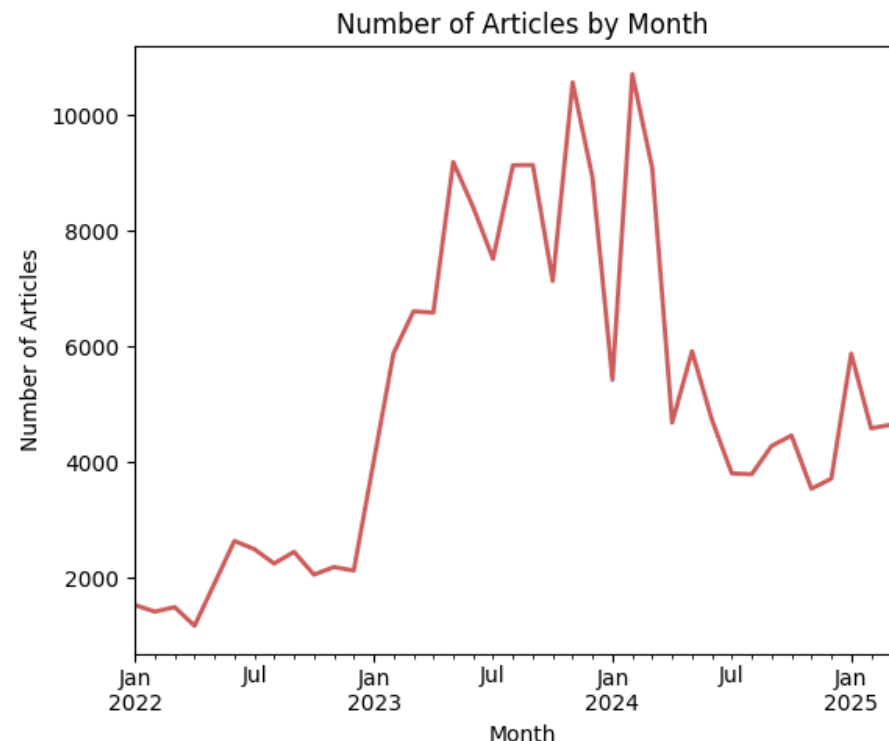
- **Generative AI Tools:** Tools like ChatGPT, Bard, and Gemini receive very high positive sentiment, indicating strong public acceptance. They are also widely applied across almost all industries, as they are referenced in nearly every sector.
- **Supportive Government Policy:** Public sector leadership (e.g., Biden, Trump) plays a critical role in guiding AI use and addressing concerns around ethics, regulation, and job displacement. If public concerns are addressed, AI could be applied more successfully in industries like Media & Entertainment.
- **Upskilling the Workforce:** Many articles emphasize the need for AI-skilled employees. Therefore, workers should learn to use AI as a tool, stay updated on new capabilities, and avoid roles that AI can fully automate.

Data Overview

This project analyzes *200,083 news articles* related to data science, machine learning, or artificial intelligence.

Data Summary Table

No	Column	Description
1	url	Link to the news article
2	date	Date the article was published
3	language	Language used in the article
4	title	Title of the news article
5	text	Main content of the article



- The data covers articles published from **January 1, 2022, to April 28, 2025**, with the majority concentrated between early 2023 and early 2024.
- All **200,083 URLs are unique**, representing distinct articles and sources. This column will serve as the primary key.

Data Preprocessing (1)

Most issues were found in the *title* and *text* columns. The following steps were taken to clean the data:

Data Quality Issues

1. Non-English content is included in some titles and articles.
2. Unwanted whitespace characters such as `\n`, `\t`, and `\xa0`.
3. Presence of headers and footers from the original websites are included in the article content.
4. Some articles are not related to AI

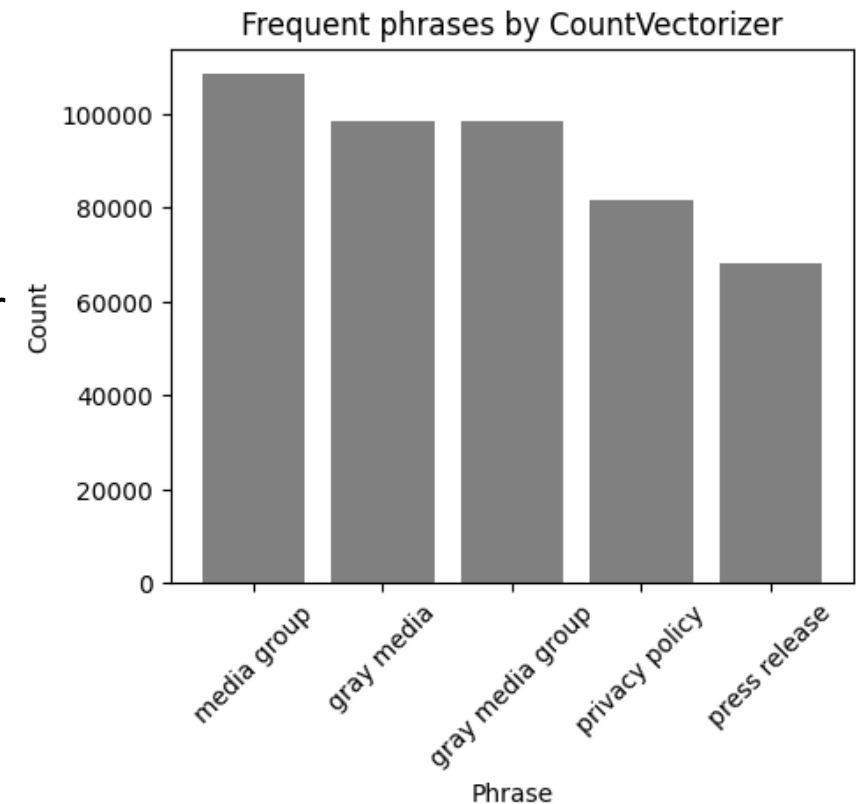
Title Cleaning Process

1. Remove non-English titles using the langdetect library.
2. Strip unwanted characters like `\xa0` and extra spaces.
3. Trim source names and dates (e.g., “May 10, 2023 –”) using regular expressions.
4. Remove trailing segments that are too short to be meaningful by splitting on common separators, then discard parts that are too short.
5. Normalize spacing.

Data Preprocessing (2)

Text Cleaning Process

1. Remove non-English articles using the langdetect library.
2. Clean whitespace characters like `\n`, `\t`, `\xa0`, and extra spaces using regex.
3. Identify header and footer patterns through manual inspection of sample articles.
4. Remove footers using frequent phrases detected by CountVectorizer with `ngram_range = (2,3)` and trim text at the matched keyword.
5. Remove headers by detecting common phrases (e.g., “about us”, “contact us”, “donate”) and trimming early content within a max distance threshold.
6. Keep only valid characters (English letters, numbers, and basic punctuation) using regex.
7. Remove unusually long words (over 25 characters)



Data Preprocessing (3) – Discarding Irrelevant Articles

Initial dataset size: 200,083 articles

Data filtering steps:

No	Steps	Remaining articles
1	Remove null and non-English articles	192,706
2	Filter articles truly related to AI by using regex to search for AI-related keywords such as "artificial intelligence", "machine learning" and related companies or products such as "ChatGPT" and "NVIDIA".	190,586
3	Remove non-news content by using BERTopic to detect topics unrelated to news, such as free AI-generated photos, and remove based on topic keywords and title patterns	181,299

Final dataset after filtering: **181,299 articles (9.4% removed)**

As most of the data remains, this dataset will be used for further analysis.

Topic Modeling (1) – Methodology

1. Removed general AI-related keywords such as “AI”, “machine learning”, and “data” from titles to better capture industry-specific topics.
2. Applied **BERTopic** with the following configurations:
 1. Clustering: *HDBSCAN* to identify meaningful topic groups and ensure robust clustering by requiring a minimum number of articles per topic.
 2. Embedding Model: *paraphrase-MiniLM-L6-v2* for its balance of speed and semantic accuracy
 3. Vectorizer: *CountVectorizer with English stopwords* to filter out common terms to focus on industry-specific words.
 4. Increased *top_n_words* to ensure the inclusion of key industry-specific terms beyond generic or frequently used words
3. Post-processing to handle outliers was done using the *reduce_outliers()* function, which reassigned many outliers to the most appropriate existing topics. As a result, the number of **outlier articles was reduced from 77,704 to just 3, while retaining all 305 topics**

Top 5 topics from 305 identified by BERTopic

Name	Count
0_chatgpt_prompts_blocking_regrid	3311
1_nvidia_chips_trillion_jensen	2857
2_gmail_workspace_duet_maps	2694
3_stocks_wall_bubble_bull	2314
4_skills_workforce_americans_fear	2860

Topic Modeling (2) – Methodology

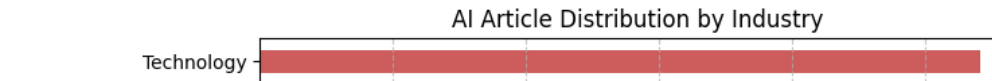
4. Manually reviewed the top topics generated by BERTopic and identified 17 potential industries, such as Technology, Healthcare, Finance & Stock Markets and Education.
5. Zero-Shot classification for topic labeling
 - Utilized the *facebook/bart-large-mnli* zero-shot model to classify each topic into the most likely industry.
 - Used both `representative_docs_` and `topic_keywords` as input prompts
 - Voting mechanism applied across outputs to assign the most probable industry to each topic.
6. Mapped Topics to Industries and Articles

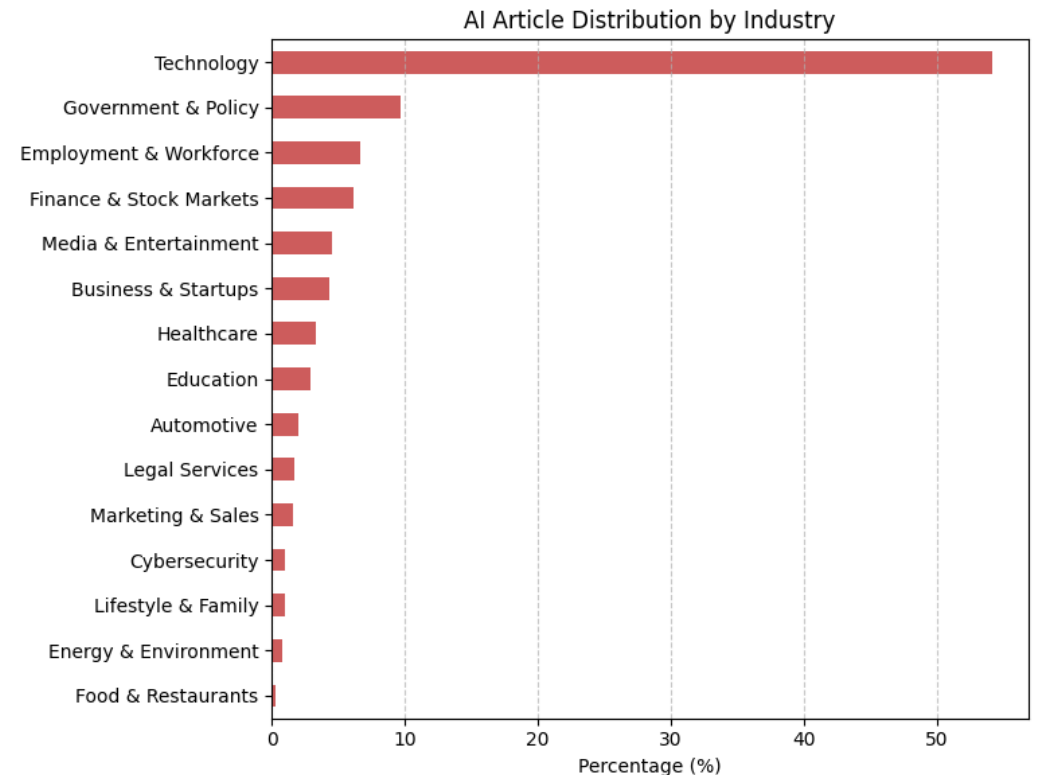
Top 5 topics matched with industries by zero-shot modeling

Name	Industry	Count
0_chatgpt_prompts_blocking_regrid	Technology	3311
1_nvidia_chips_trillion_jensen	Technology	2857
2_gmail_workspace_duet_maps	Technology	2694
3_stocks_wall_bubble_bull	Finance & Stock Markets	2314
4_skills_workforce_americans_fear	Employment & Workforce	2860

Topic Modeling (3) – AI in Different Industries

Most articles (54.2%) are associated with the *Technology industry*, primarily for two reasons:

- AI Integration in Tech Companies
 - AI is easily integrated into technology companies like Apple, especially in smartphone products. The frequent mention of “smartphone” in topic keywords (as shown in the word cloud) highlights this trend.
 - Focus on AI Development
 - Many articles discuss new AI features and conversational tools like ChatGPT, rather than practical applications in other industries.
 - Prominent terms such as “chatgpt”, “launch”, and “feature” reflect the industry’s emphasis on innovation.
- 
- | Industry | Count |
|-------------------------|-------|
| Technology | 10 |
| Government & Policy | 4 |
| Employment & Workforce | 3 |
| Finance & Stock Markets | 3 |
| Media & Entertainment | 2 |
| Business & Startups | 2 |
| Healthcare | 1 |

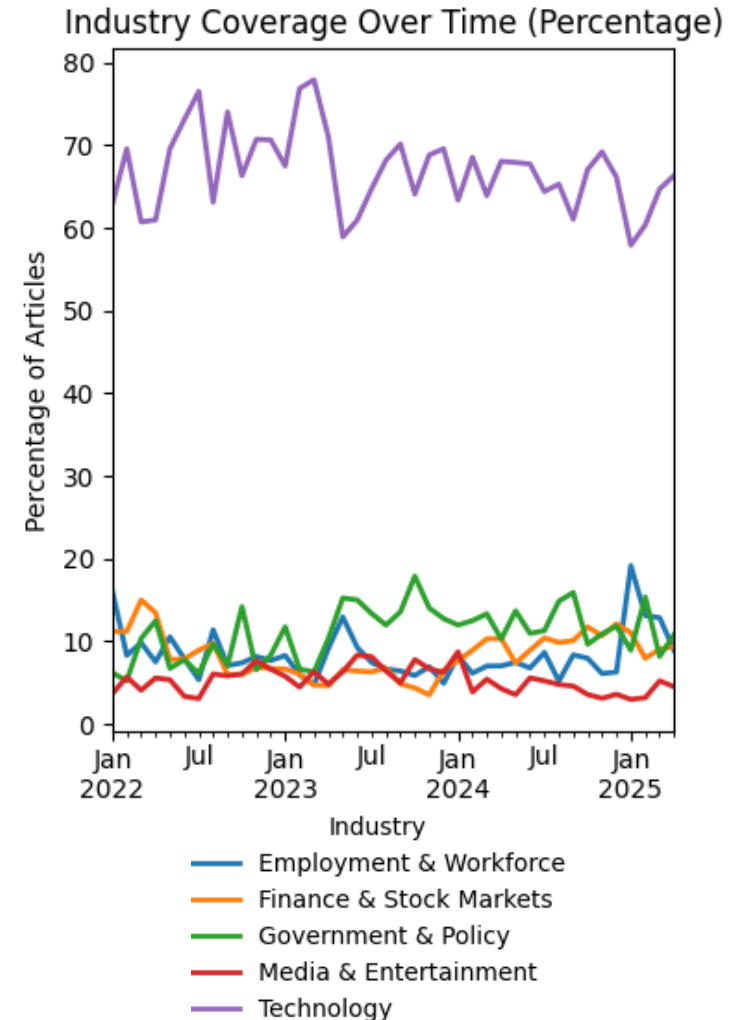


Topic Modeling (4) – AI in Different Industries

While Technology dominates consistently, maintaining 60–75% of article coverage throughout the period, other industries also show significant connections to AI:

- *Government & Policy* (9.6%): Highlights steady attention on how governments are regulating or responding to AI developments across countries
- *Employment & Workforce* (6.7%): Reflects AI's impact on the job market, including salary shifts, the integration of AI tools in the workplace, and visible spikes during key periods
- *Finance & Stock Markets* (6.1%): Often covers the performance of AI-related companies
- *Media & Entertainment* (4.5%): Indicates moderate discussion on AI's use in creative fields such as content generation

On the other hand, the industries with the lowest proportion of AI-related articles are *Lifestyle & Family* (0.94%), *Energy & Environment* (0.76%), and *Food & Restaurants* (0.29%). This suggests that AI has had limited visibility or adoption in these areas compared to more tech-driven sectors.



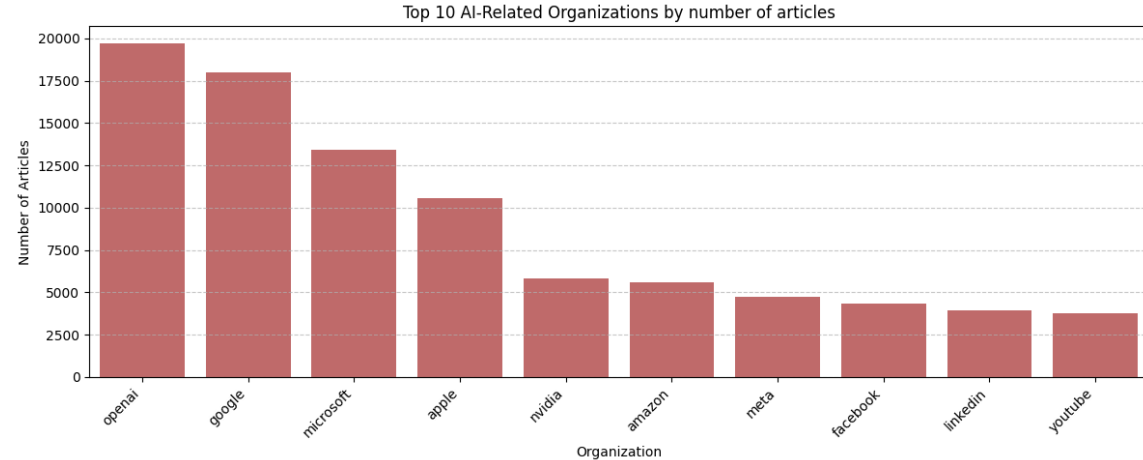
Entity Extraction (1)

Methodology

To better understand the technologies and factors driving AI integration across industries, **Named Entity Recognition (NER)** was used to identify organizations, products, people, and locations related to AI.

1. Created specific patterns to detect AI-related entities (e.g., OpenAI, ChatGPT, Sam Altman).
2. Used spaCy's *"en_core_web_trf"* model to perform NER and extract entities from article content in batches.
3. Normalized extracted entities by converting text to lowercase and removing punctuation.

Top Organizations Driving AI Integration

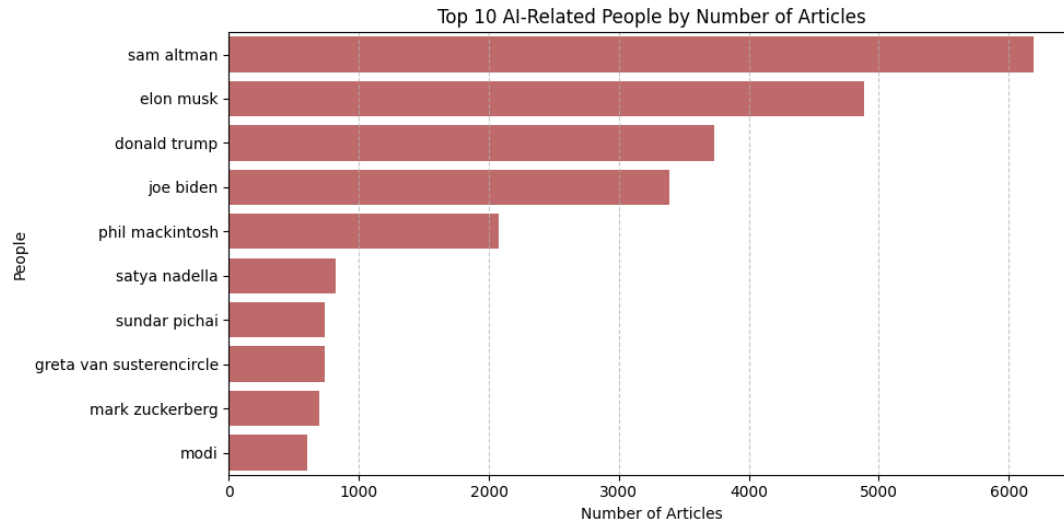


While the top-mentioned organizations all come from the technology sector, their roles in AI integration differ significantly:

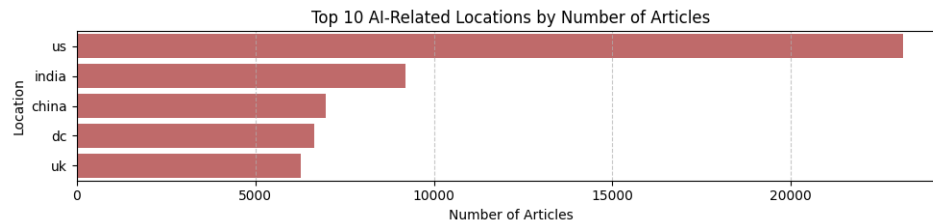
- *OpenAI and Google* lead as developers of foundational models like ChatGPT and Gemini, often mentioned alongside their products.
- *Microsoft* plays a dual role as an investor in OpenAI and a key integrator, embedding AI into products like GitHub and Office.
- *Apple* focuses on applying AI to enhance user experiences, rather than developing its own foundation models.

Entity Extraction (2)

Top People Driving AI Integration



Top Locations Driving AI Integration



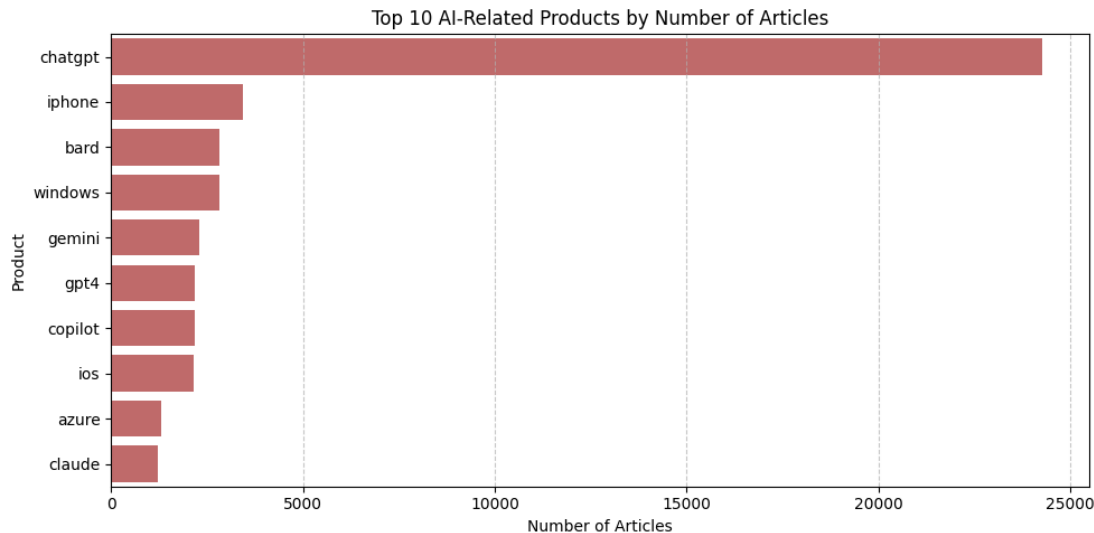
Individuals frequently mentioned in AI-related articles can be grouped into two main categories:

- **Technology Leaders:** Sam Altman, Elon Musk, Satya Nadella, Sundar Pichai, and Mark Zuckerberg are CEOs of major tech companies. They play a central role in either developing foundational AI technologies or integrating AI into widely used products and platforms.
- **Government Figures:** Donald Trump and Joe Biden are the most frequently mentioned political leaders, reflecting the significant influence of government policy on the direction of AI.

-
- *The United States* is the most frequently mentioned location in AI-related articles, reflecting its position as home to many leading tech companies and its global leadership in AI research and innovation.
 - It is followed by *India and China*, which are also actively investing in AI development and integration

Entity Extraction (3)

Top Products Driving AI Integration



ChatGPT leads all AI products, with significantly more mentions than others. It is driving change across many industries.

Other popular products mentioned include:

- Generative AI tools such as *Bard, Gemini, and Claude*
- Products that use AI, like *iPhone, Windows, Copilot, iOS, and Azure*

This highlights that the **technology industry is the most impacted by AI.**

Summary

From the NER analysis, the most influential organization driving AI integration is *OpenAI*, the developer of *ChatGPT*, followed by other major tech companies. This **highlights the significant impact of generative AI across industries**, with the technology sector being the most affected, due to its ability to quickly integrate AI into widely used products like iPhone and Azure.

Sentiment Analysis (1) – Methodology

To investigate public opinion toward AI in different industries, sentiment analysis was conducted using the following steps:

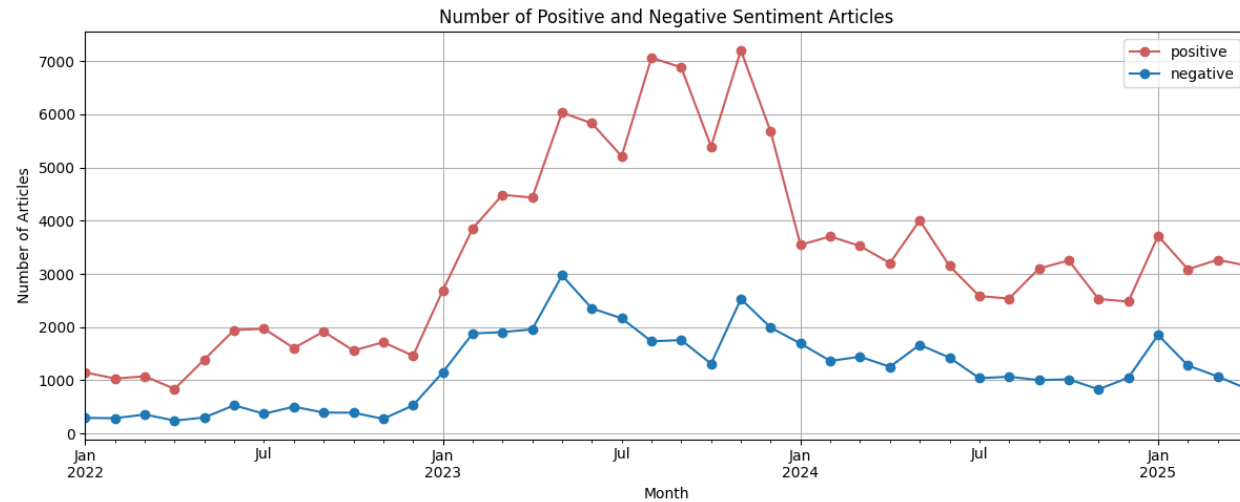
1. Labeled 320 sample articles using *GPT-3.5* to classify sentiment as positive or negative based on article content.
2. Split the labeled data into training and testing sets for model development.
3. Fine-tuned a pre-trained transformer model (*distilbert-base-uncased*) on the labeled data, as it offers a good balance between speed and accuracy for text classification tasks like sentiment analysis.
4. Evaluated model performance on the test set to ensure reliability.
5. Predicted sentiment for the entire dataset of articles using the fine-tuned model.

Classification Report

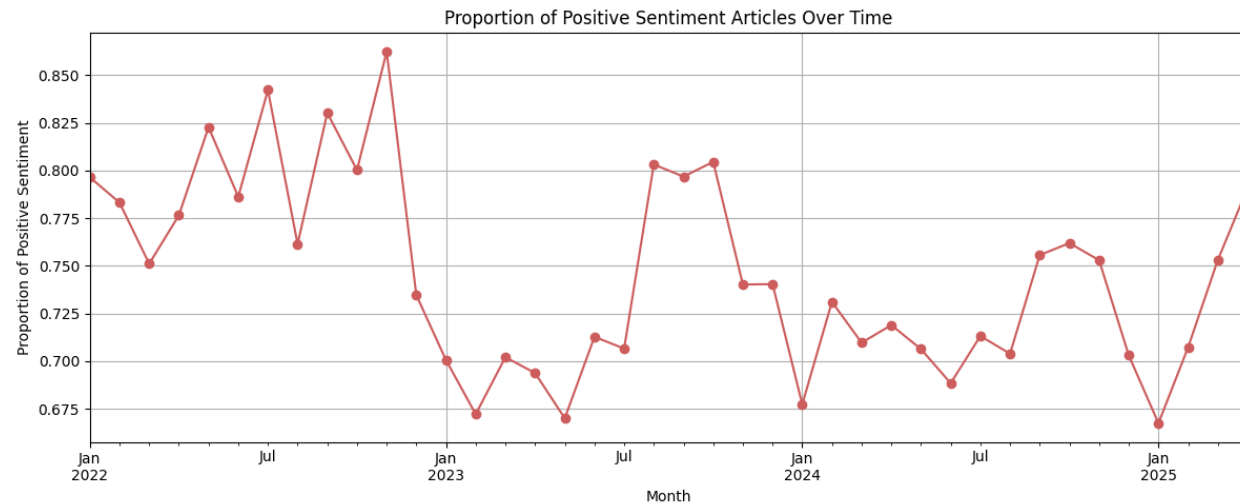
negative	precision	recall	f1-score	support
negative	0.71	0.63	0.67	19
positive	0.85	0.89	0.87	45
accuracy			0.81	64
macro avg	0.78	0.76	0.77	64
weighted avg	0.81	0.81	0.81	64

Classification report shows **81% accuracy**. The model performs strongly on positive sentiment and reliably detects negative sentiment despite class imbalance, making it suitable for predicting sentiment across the full dataset.

Sentiment Analysis (2) – Trend over time



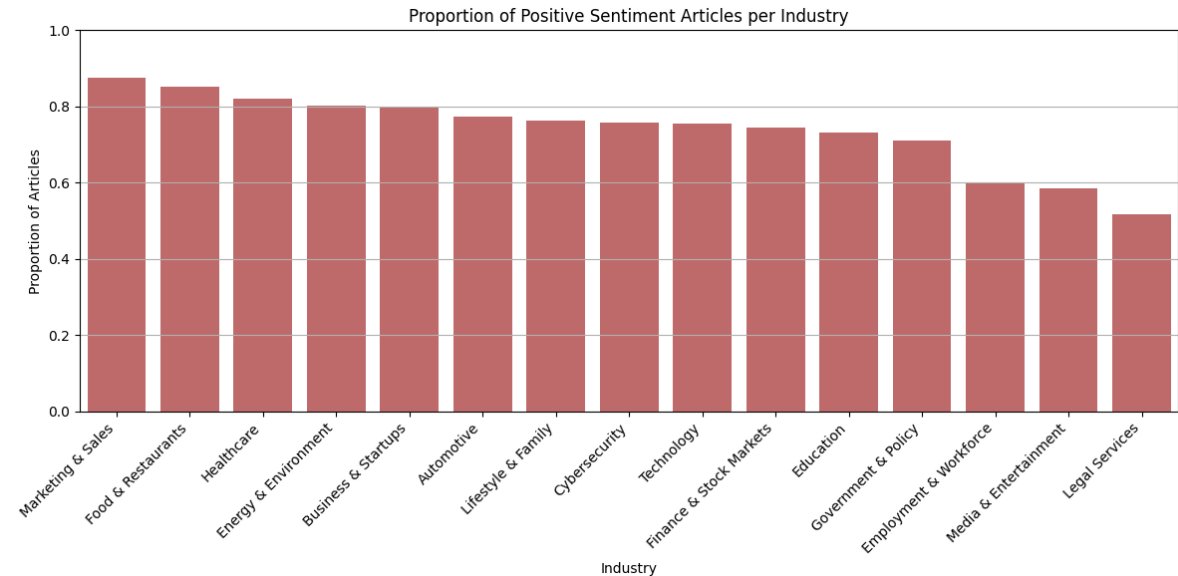
The majority of articles express a positive opinion toward AI (73.5% of all articles). **From 2022 to 2025, positive sentiment consistently exceeds negative sentiment.** Besides, beginning in January 2023, after the release of ChatGPT-3.5 and followed by other major AI models such as Llama and Claude, there was a noticeable spike in the number of AI-related articles.



At the same time, the proportion of positive sentiment declines, suggesting a rise in public concern about AI. Since then, the percentage of positive articles has remained lower than it was prior to 2023, when article volume was lower. **This trend reflects growing public interest in AI, along with increasing caution and critical attention.**

Sentiment Analysis (3) – AI Integration across Industries

Different industries have been impacted by AI in distinct ways. Based on sentiment analysis, industries with a **higher proportion of positive sentiment articles may indicate more successful AI integration.**



Successful AI integration

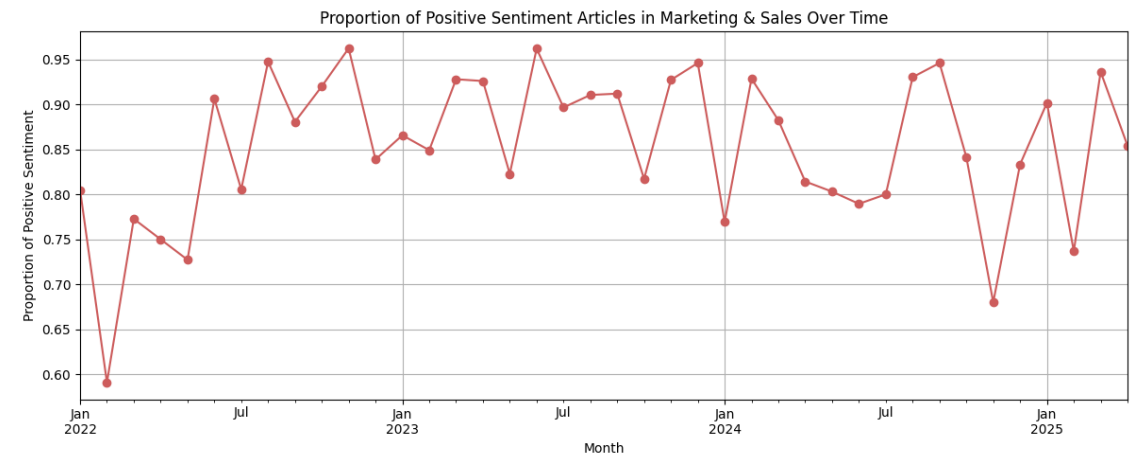
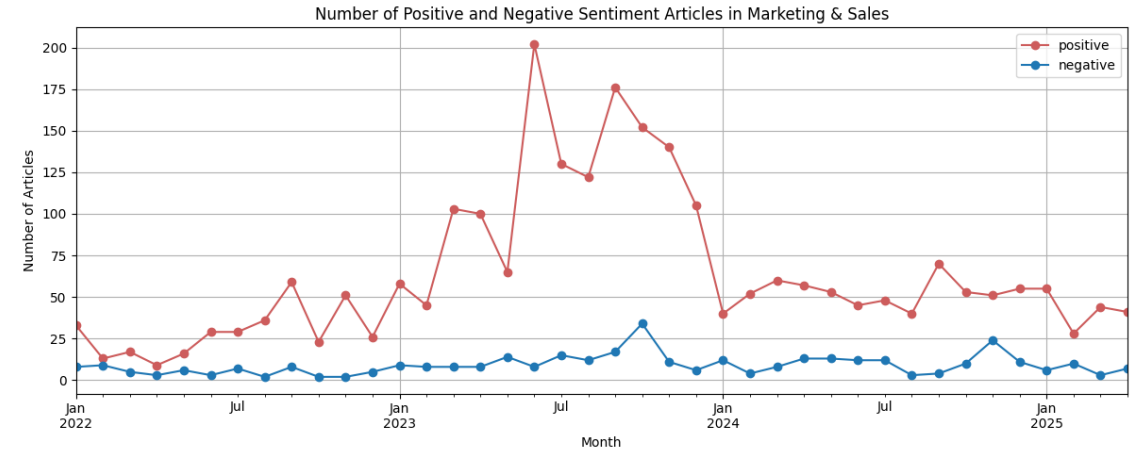
Marketing & Sales and *Healthcare* show strong positive sentiment, suggesting AI is making a valuable contribution. While Food & Restaurants also shows a high proportion of positive sentiment, the overall number of articles is too small to draw firm conclusions. This suggests that while AI may be beneficial in this sector, its impact is still limited.

Unsuccessful AI integration

Employment & Workforce, *Media & Entertainment*, and *Legal Services* have a higher proportion of negative sentiment, reflecting public concern and potential challenges in integrating AI effectively into these fields

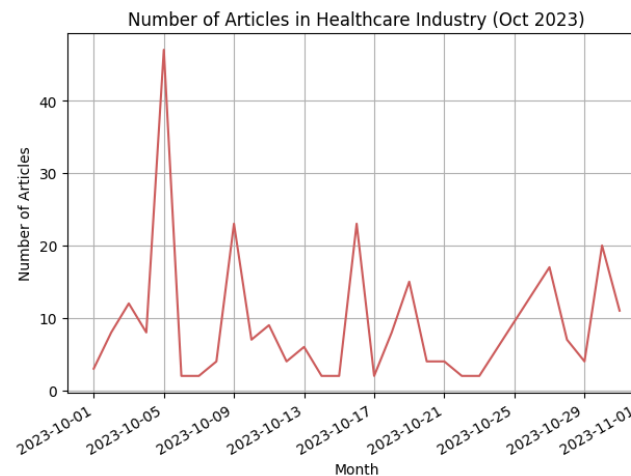
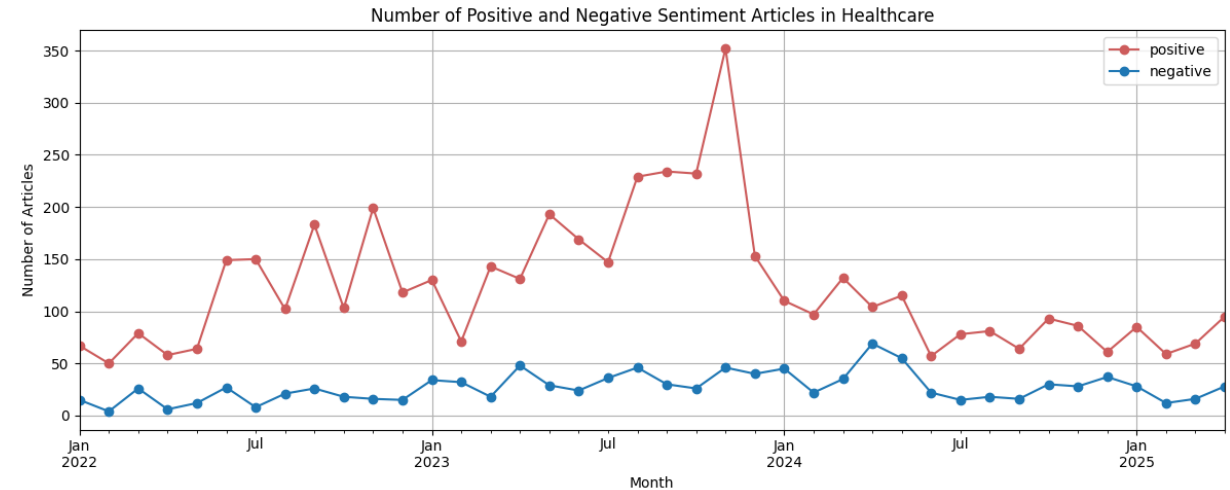
Successful AI Integration (1) – Marketing & Sales

- Marketing & Sales has the highest proportion of positive articles (87.5%), indicating strong sentiment toward AI adoption in this sector.
- From 2022 to mid-2023, there was a clear upward trend in both the number of articles and the proportion of positive sentiment, peaking at over 95% in several months.
- This high level of positivity has remained steady, reflecting how AI has quickly established a strong and stable presence in marketing workflows.
- **These patterns suggest that AI integration in Marketing & Sales is both effective and widely accepted, marking it as a clear success story.**



Successful AI Integration (2) – Healthcare

- Healthcare ranks in the top 3 industries with the highest proportion of positive articles (82%).
- The number of positive articles shows a steady uptrend from 2022 through 2024, reflecting growing positive opinion about AI's role in the sector.
- A spike in October–November 2023 corresponds to the launch of *RhythmX AI* (Oct 5) and *ThinkAndor's* expansion in Nov, highlighting interest in new AI-driven health technologies.
- However, since 2024, the proportion of positive sentiment has slightly declined, indicating emerging concerns about AI integration in healthcare.
- **AI is widely used in healthcare with many new tools, and while some concerns have grown recently, most articles are still positive, showing healthcare as another successful AI integration industry.**

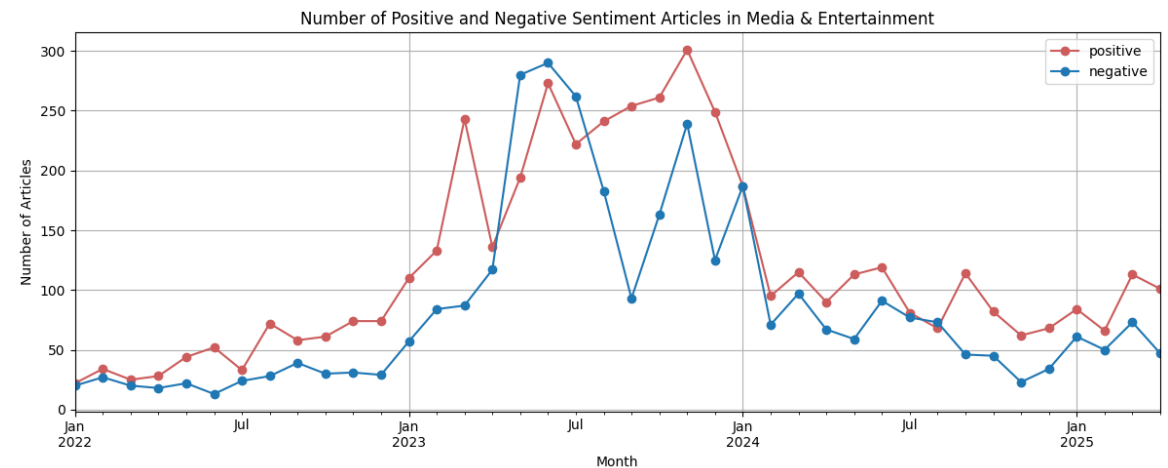
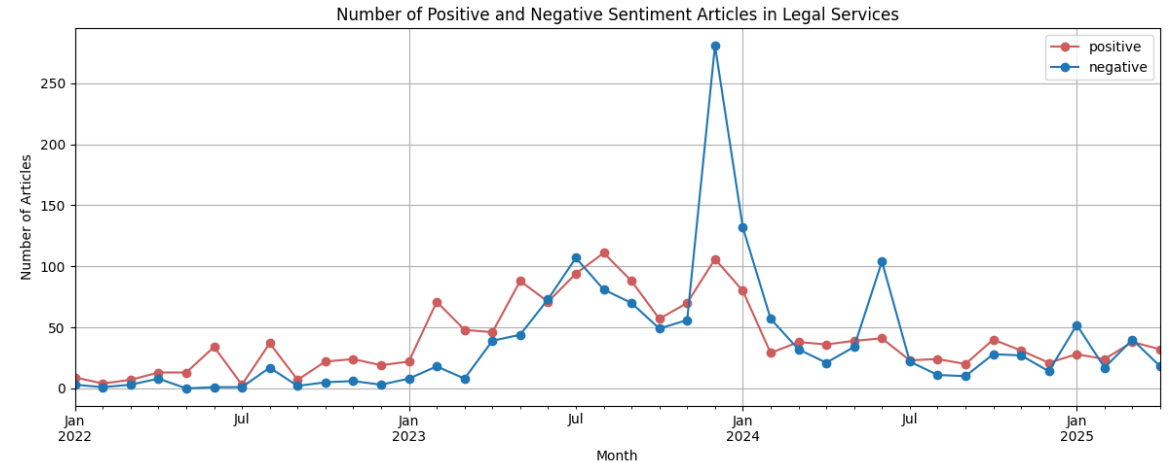


Word Cloud: Products mentioned in Healthcare (Oct-Nov 2023)



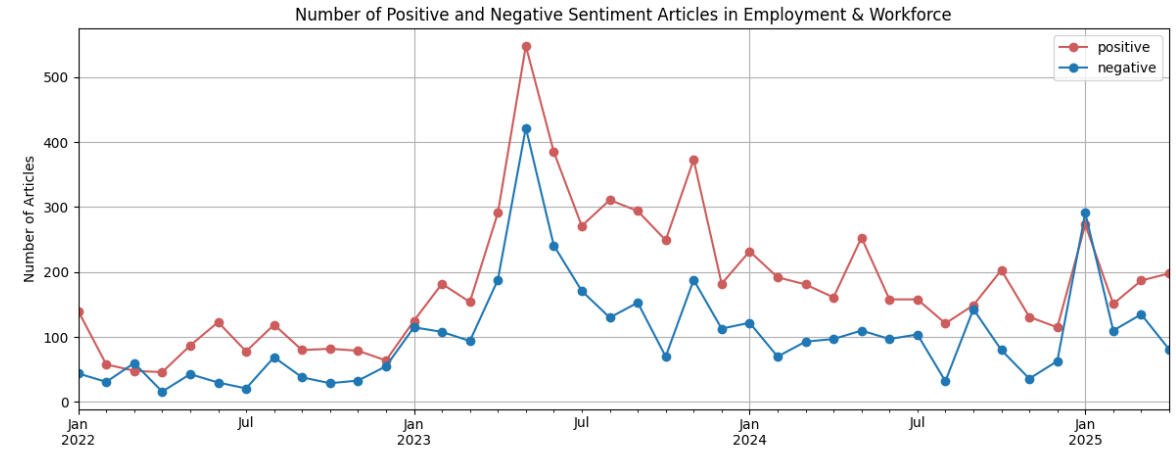
Unsuccessful AI Integration (1) – Legal Services & Media

- Legal Services and Media & Entertainment have the highest proportion of negative sentiment (over 40%), reflecting significant public concern about AI adoption in these sectors.
- In Legal Services, concerns center on generative AI producing unlawful content. A spike in negative sentiment occurred in December 2023 following *The New York Times* lawsuit against OpenAI over copyright infringement.
- In Media & Entertainment, AI is already widely used, but ongoing controversy around comparisons between human-made and AI-generated content highlights continued public resistance to fully accepting AI in creative work.
- **Overall, while both industries have strong potential for AI adoption, widespread acceptance remains limited. If key concerns are addressed, Media & Entertainment could become a successful case for AI integration.**

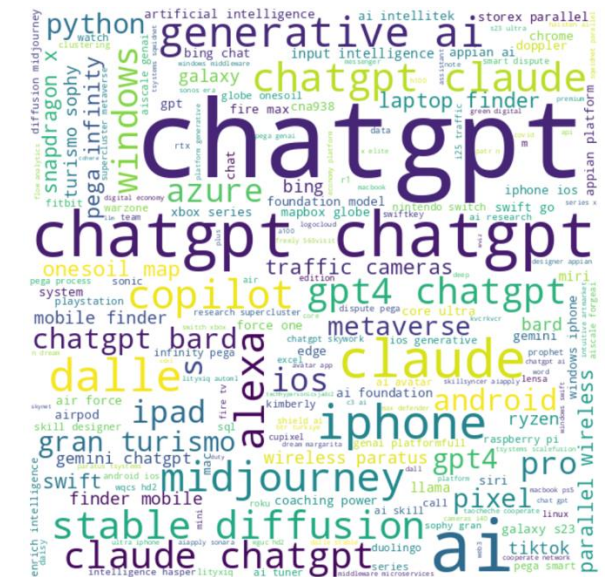


Unsuccessful AI Integration (2) – Employment

- Employment & Workforce is one of the top 3 industries with the highest negative sentiment (40%), showing strong public concern.
- **Generative AI has the biggest impact**, as shown in the word cloud. Also, the spike in May 2023 matches the launch of ChatGPT-4. This directly affects the job market in both positive and negative ways:
 - Positive: AI is used to improve working efficiency → companies need more people with AI skills.
 - Negative: Jobs that AI can do without human input are more likely to be replaced.
- Other key events also caused spikes:
 - May 2023: Vice President met with tech CEOs to talk about AI and its effect on jobs, showing the government's role in managing AI.
 - January 2025: Tech layoffs and the IMF report warning that AI could impact jobs added to public concern.

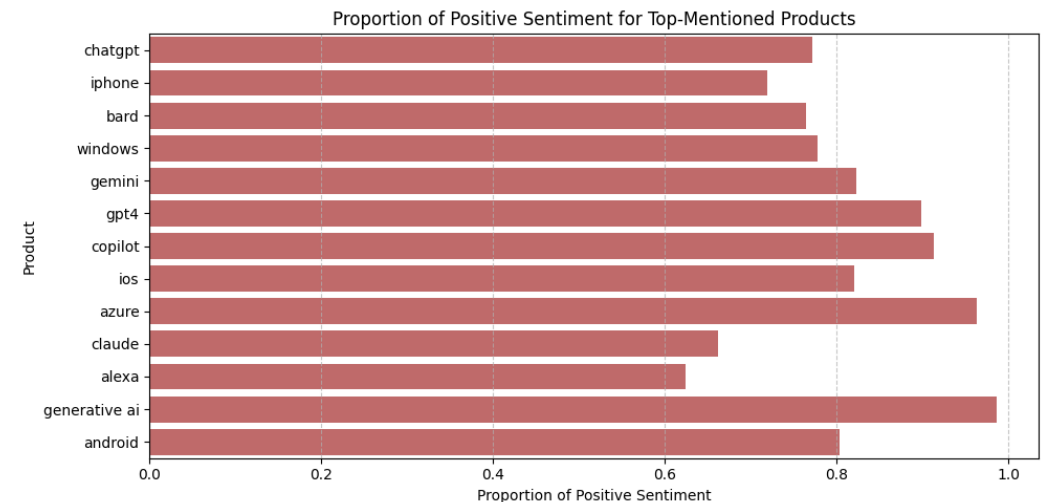
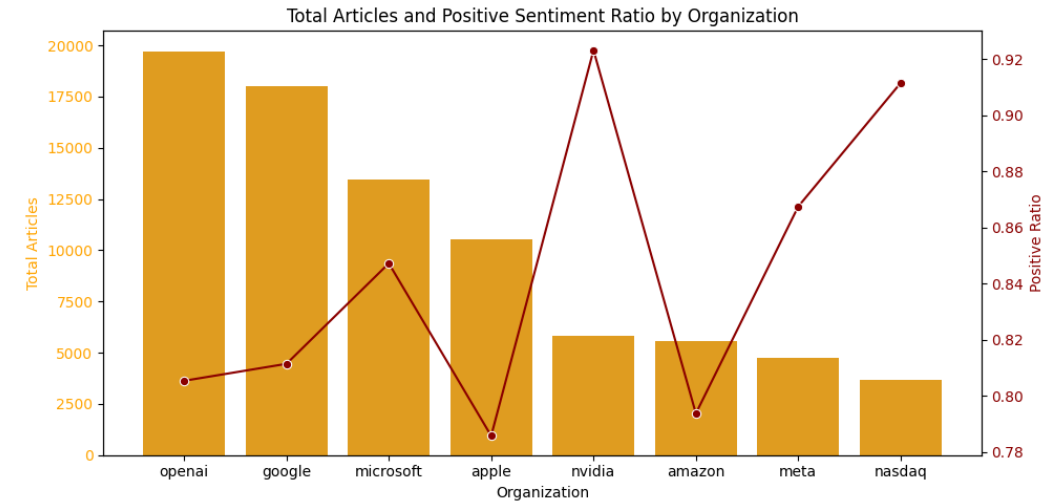


Word Cloud:
Products
mentioned in
Employment &
Workforce



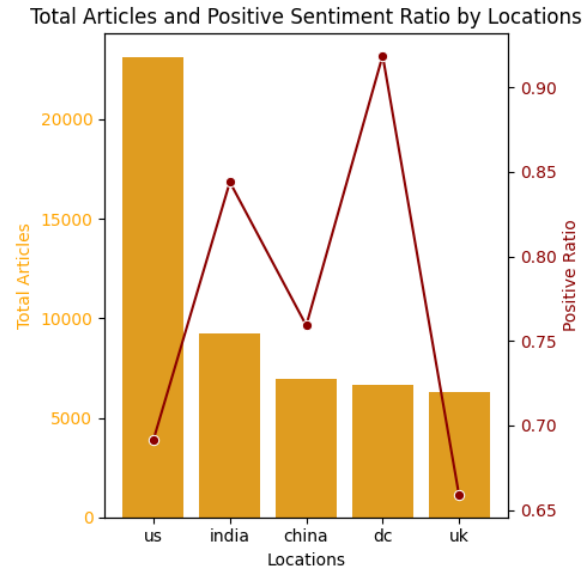
Entity-level sentiment analysis (1) – Success in Tech Industry

- Top organizations and products mostly come from the technology industry, and these articles are also largely positive in sentiment, especially *generative AI*, which receives the highest positive sentiment at 98.5%.
- Major tech companies like *Google, Microsoft, and Apple* are actively investing in or integrating AI into their products, further driving industry-wide adoption.
- Generative AI tools such as *ChatGPT, Bard, and Gemini* are frequently mentioned with strong sentiment, showing broad company use. In contrast, *Claude and Alexa* receive lower sentiment, highlighting the field's competitiveness.
- Customers also respond positively, with high positive sentiment toward AI-integrated products like *Azure, Copilot, iPhone, Android, Windows, and iOS*.
- **In conclusion, tech companies are heavily focused on developing and applying generative AI. With growing competition and continued positive customer feedback, this trend is likely to accelerate in the near future.**

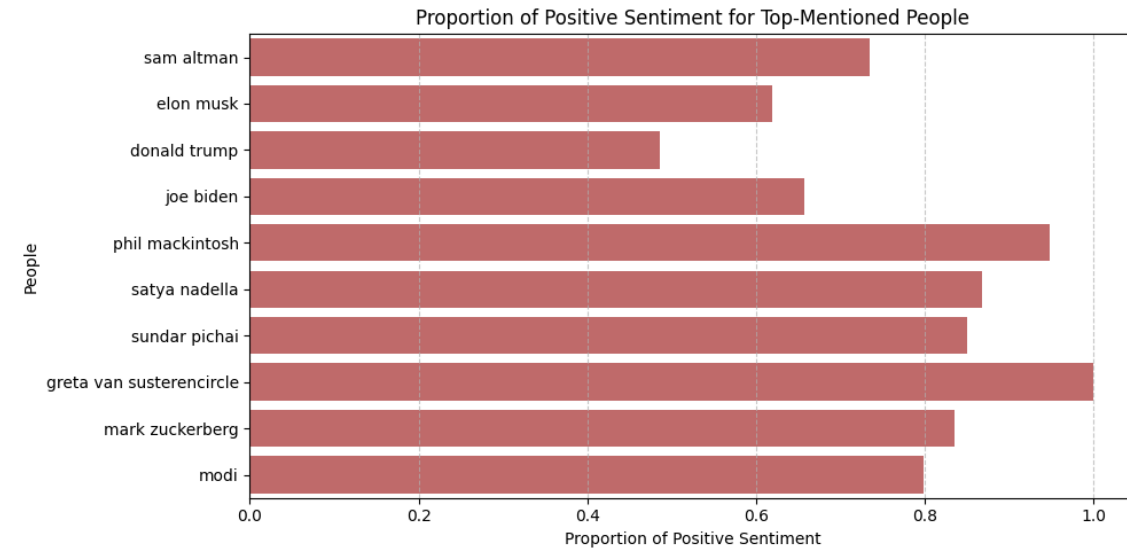


Entity-level sentiment analysis (2) – Government Impact

- Countries with a tech focus like the *US, India, China, and the UK* have a high number of AI-related articles but show differences in sentiment. India has the highest proportion of positive sentiment, while the US and UK show lower positivity, suggesting some public concerns about how AI is being applied in these countries.



- In terms of people, tech leaders such as Phil Mackintosh, Satya Nadella, Sundar Pichai, and Mark Zuckerberg receive significantly more favorable sentiment than political figures like Donald Trump and Joe Biden.
- This indicates that while the public generally supports AI and its potential, there are still concerns around government roles and policies. **It highlights the critical role of the public sector in accelerating the responsible development of AI.**

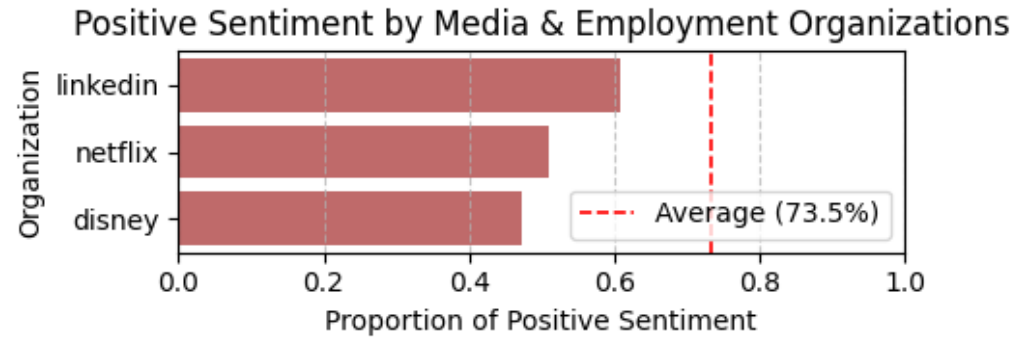


Example government-related articles:

- Biden to meet with experts on AI risks and opportunities.
- White House hosts AI forum with tech leaders in San Francisco.
- Major tech firms agree to AI safeguards set by the U.S. government.

Entity-level sentiment analysis (3) – Unsuccessful industries

- *Media & Entertainment* and *Employment & Workforce* are two industries with high AI adoption but also high negative sentiment.
 - Major organizations in these industries, such as *LinkedIn*, *Netflix*, and *Disney*, receive below-average positive sentiment.
 - Sample articles highlight concerns about AI replacing creative and professional roles, leading to increased public skepticism.
- **In conclusion, even though these industries have strong AI potential, the public still doesn't fully accept its impact.**



Examples of employment-related articles

- Double majors may lower risk of AI-related job loss.
- Governments partner with tech firms to boost AI skills.
- LinkedIn co-founder predicts the end of 9-to-5 jobs due to AI.

Examples of media-related articles

- Marvel used AI for *Secret Invasion*'s opening — received negative reactions.
- Hollywood strikes grow over fears AI could replace writers.
- *Black Mirror* highlights actors' anxiety about AI in entertainment.

Conclusion

AI has become the center of innovation, driving the technology industry and influencing almost every sector. This trend accelerated with the release of popular Conversational AI tools like ChatGPT, Llama, and Claude. However, alongside this rapid growth, public concern and critical attention toward AI are also rising. To ensure successful and responsible AI integration, the following actions are recommended:

- **Invest in Generative AI Tools:** Encourage the adoption of widely accepted tools like ChatGPT, Bard, and Gemini, which have high public approval and broad industry relevance. Tech-focused countries such as the U.S., China, and India are already investing heavily in these technologies, demonstrating that generative AI can be a key driver of national economic growth.
- **Establish Supportive Government Policies:** Develop clear guidelines and regulations that address ethics, job displacement, and data usage to promote responsible AI adoption, especially in sensitive sectors.
- **Prioritize Workforce Upskilling:** Launch training and reskilling programs to equip employees with AI-related skills, helping them adapt to evolving job roles.



Thank you