The AI Effect on Industry

Yiling Ding

Contents

Executive summary: AI's Industry Impact

Top Impacted Industries: Finance, Retail, Healthcare, Legal, Media

Automate Jobs:

- **Finance**: Fraud detection, chatbots
- **Retail**: Checkout, inventory tracking
- **Healthcare**: Diagnostics, scheduling
- **Legal**: Contract review, compliance monitoring

Boost Productivity

- Use AI assistants for writing, coding, and summarization
- Apply NLP for summarizing emails and documents
- Optimize schedules with AI planners and routing tools

Adoption Strategy

- Start with low-risk, high-impact pilots
- Invest in high-quality data
- Upskill employees with AI literacy
- Build trust with transparency and fairness
- Focus on augmenting—not replacing—human work

Project overview

We analyzed ~200,000 AI-related news articles to uncover how different industries are being impacted by artificial intelligence. Using topic modeling, sentiment analysis, and entity extraction, we identified key trends, technologies, and organizations shaping AI adoption.

Goal:

- Reveal where AI is gaining momentum
- Who is leading it
- How industries are responding over time

Dataset Overview

	url	date	language	title	text
0	http://businessnewsthisweek.com/business/infog...	2023-05-20	en	Infogain AI Business Solutions Now Available i...	\n\nInfogain AI Business Solutions Now Availab...
1	https://allafrica.com/stories/202504250184.html	2025-04-25	en	Africa: AI Policies in Africa - Lessons From G...	\nAfrica: AI Policies in Africa - Lessons From...
2	https://asiatimes.com/2023/07/yang-lan-intervi...	2023-07-25	en	Yang Lan interviews academics on AI developmen...	\nYang Lan interviews academics on AI developm...
3	https://cdn.meritalk.com/articles/commerce-nom...	2025-02-04	en	Commerce Nominee Promises Increased Domestic A...	\nCommerce Nominee Promises Increased Domestic...
4	https://citylife.capetown/hmn/uncategorized/re...	2023-11-11	en	Revolutionizing the Manufacturing Industry: Th...	Revolutionizing the Manufacturing Industry:...

Source: AI/ML/DS news article dataset (Parquet format)

Total Articles: 200,083

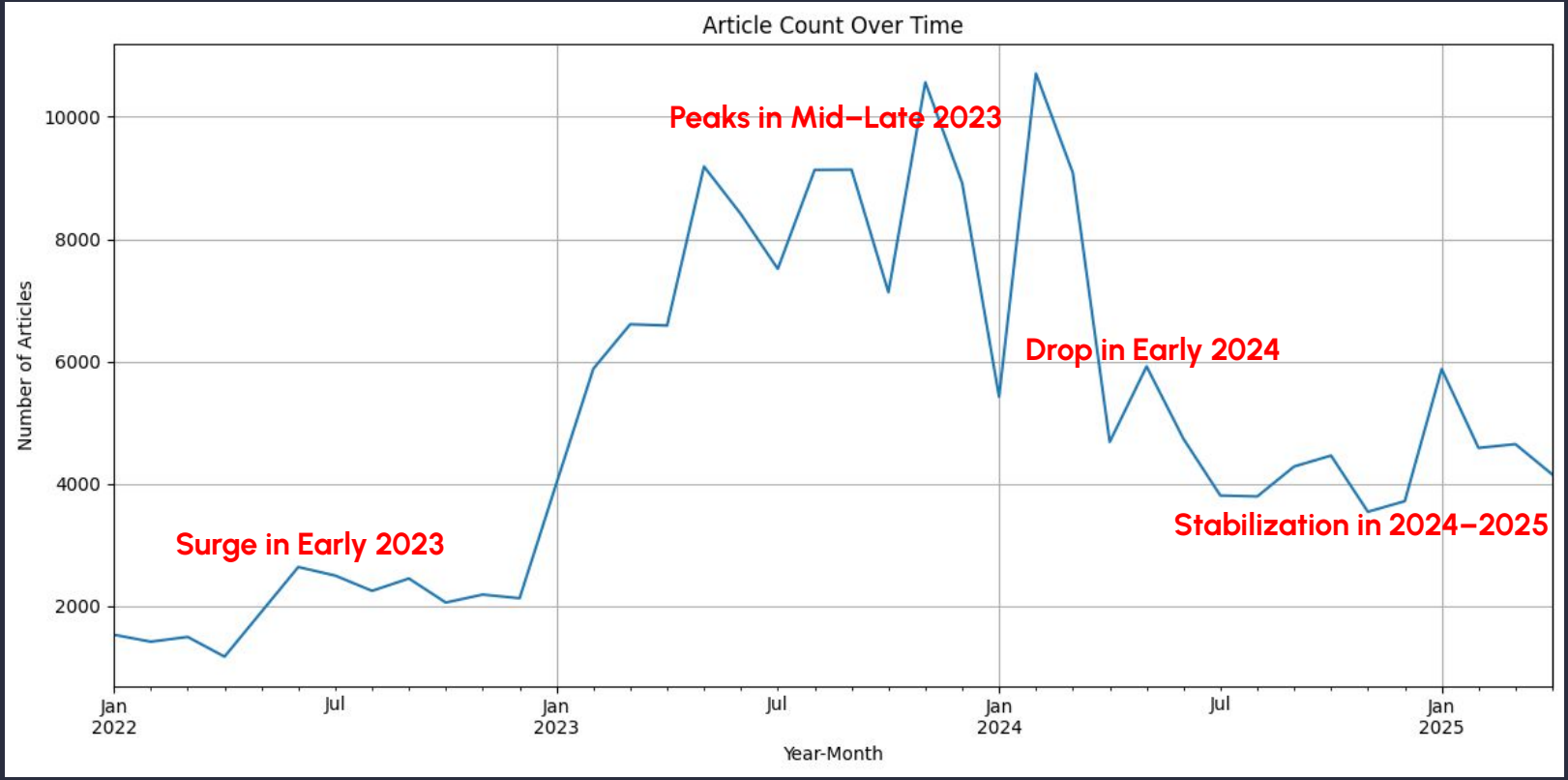
Language: 100% English (`language` = 'en')

Time Range: 2022–2025 (based on `date` column)

Fields:

- **title** – Article headline
- **text** – Full article body
- **date** – Publication date
- **url** – Article source
- **language** – Detected language

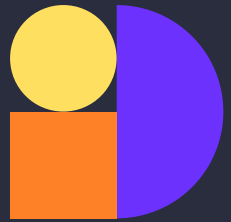
EDA–Article Count Over Time



Data Preprocessing

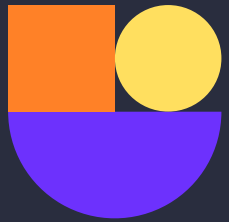
Removed Noise

Filtered out HTML tags, excessive whitespace, and boilerplate phrases



Filtered Short Articles

Removed entries with less than 25 characters in the body text



Normalization

Applied lowercasing, cleaned formatting inconsistencies

Result: 199,484 rows remain

Modeling Roadmap

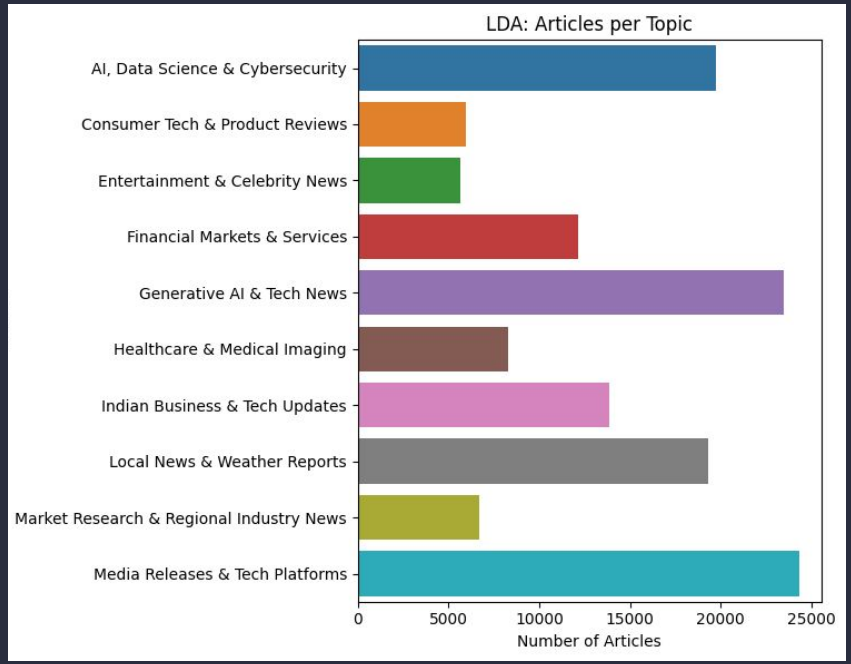
Topic Detection

LDA

We initially applied **LDA (Latent Dirichlet Allocation)** to discover topics in the dataset.

While it provided a basic segmentation, the results were not as useful as expected:

- Topics were **numerically labeled** with unclear real-world meaning, and it is hard to assign main industry to each topic.
- LDA's **bag-of-words approach** missed semantic context and nuance.



Problem:

- There are some topics that are extremely similar (AI & Data Science VS Generative AI News)

Topic Detection

NMF

After that, I tried NMF, but the result is still not too satisfied.

```
Topic 0: ai, data, gray, generative, platform, technology, group, customer, learning, cloud
Topic 1: rawpixel, jpeg, px, generated, image, rawpixelrawpixelelementsdesignsdesign, topicselement, topicsboards, usimagine, comkeywords
Topic 2: nasdaq, symbols, quotes, add, watchlist, symbol, markets, stocks, fool, market
Topic 3: overviewview, cision, products, consumer, services, entertainment, general, transportation, telecommunications, resources
Topic 4: ago, hours, weather, video, stories, news, bestreviews, file, app, days
Topic 5: newswires, presswire, ein, south, releases, distribution, republic, north, dakota, virginia
Topic 6: best, google, apple, chatgpt, ai, iphone, pro, samsung, galaxy, reviews
Topic 7: price, stocks, market, share, stock, trading, shares, growth, indices, ai
Topic 8: news, openai, said, chatgpt, ai, new, altman, sports, facebook, icon
Topic 9: salary, compensation, scientist, data, salaries, total, paying, pay, levels, fyi
```

Problem:

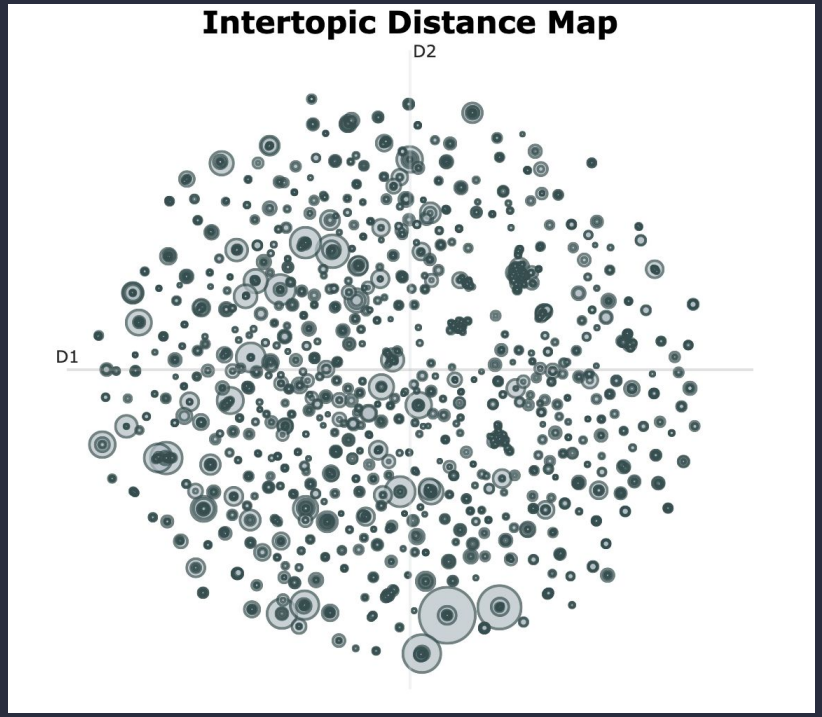
- There are some topics that are extremely similar (topic 5 and 8 are both news)

Topic Detection
BERTopic

It initially generated **3,090 fine-grained topics**, which was **too many** for effective interpretation.

To address this, we considered two solutions:

- **Define 10 industry categories** and map BERTopic topics using keyword matching
- Select the **top 10** most frequent topics from BERTopic, then assign them to industries based on top keywords



Topic Detection

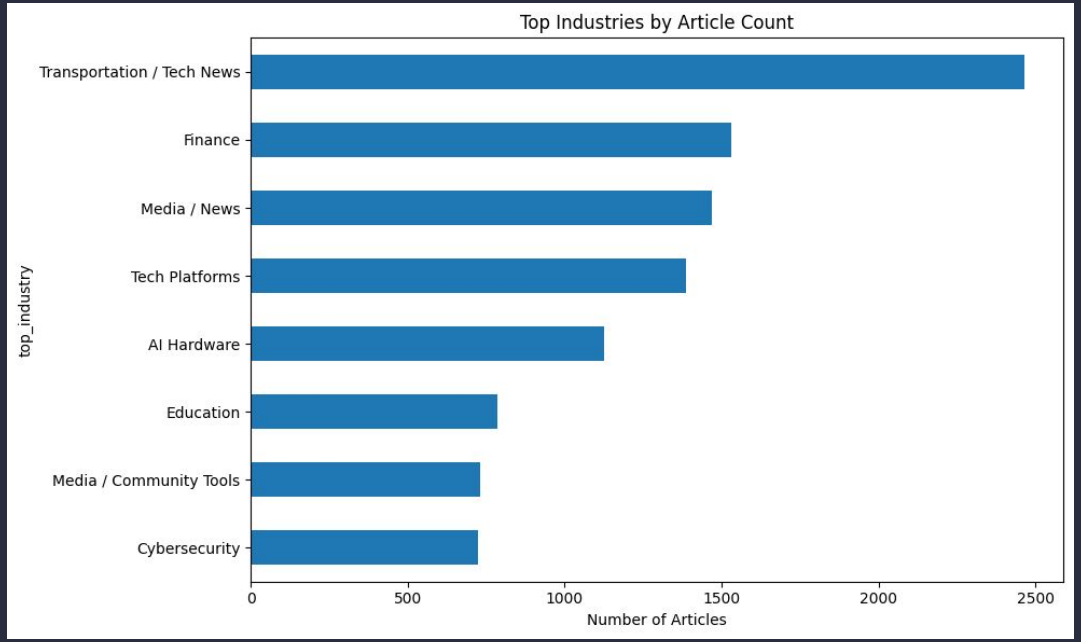
BERTopic-Method 1

To create meaningful, executive-level insights, we **selected the 10 most frequent topics based on article count**.

- For each, we reviewed the top keywords and manually assigned a descriptive industry label (e.g., "Media / News", "Finance", "Education").

Problem:

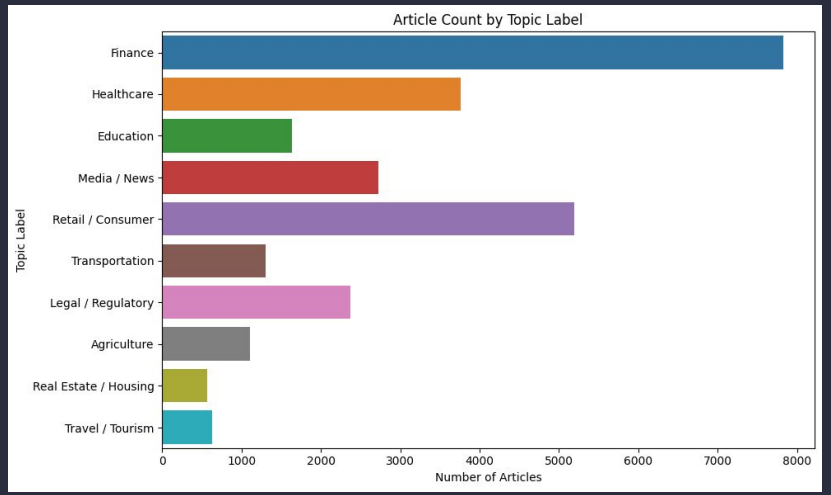
- Still have the problem with unclear boundaries between topics (News VS Tech News)



Topic Detection

BERTopic-Method 2

Right now, we can clearly divide the topics. Thus, we decided to choose this method.



As an alternative to choose top 10 industries, we manually:

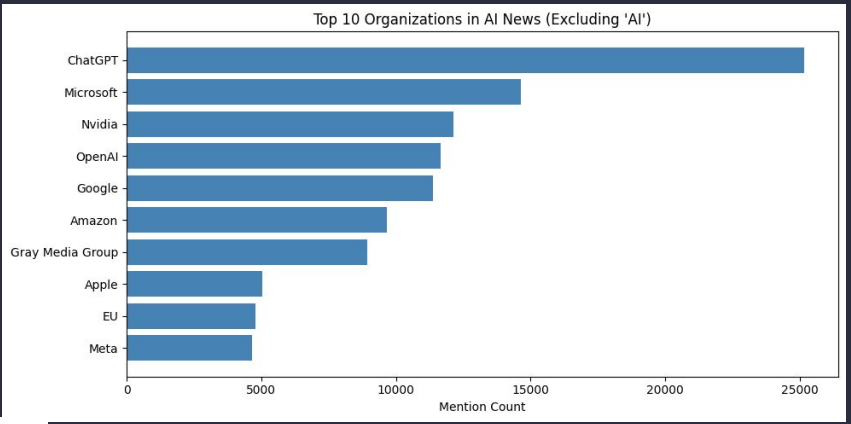
- Defined **10 high-level industry categories** (e.g., Finance, Tech Platforms, Healthcare)
- **Created keyword sets** for each category based on domain knowledge
- **Mapped** BERTopic topics to industries by matching top 10 keywords from each topic to our predefined keyword sets

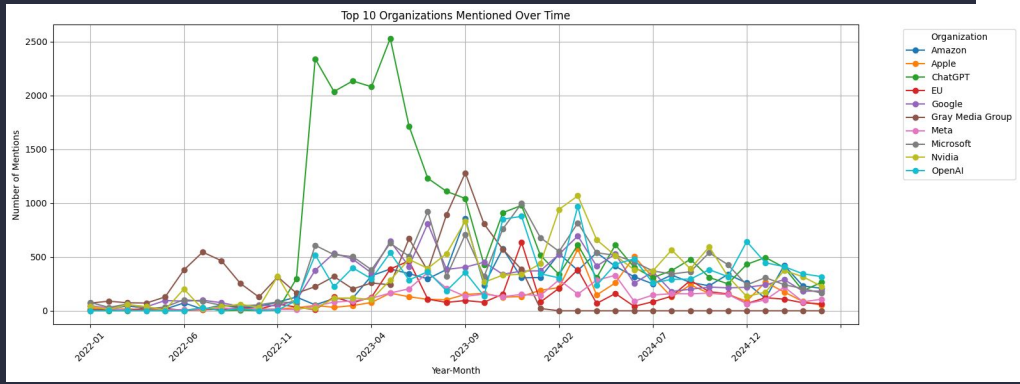
Results:

- **Finance** is the most discussed topic, suggesting it is highly
- impacted by or relevant to AI applications, especially in areas like algorithmic trading, fraud detection, and robo-advisors.
- **Retail / Consumer** and **Healthcare** follow closely, reflecting strong AI influence in e-commerce personalization, customer behavior prediction, diagnostics, and patient management.
- **Education, Media / News,** and **Legal / Regulatory** show moderate AI relevance, likely due to AI's role in content generation, sentiment analysis, and compliance automation.
- **Agriculture** and **Travel / Tourism** have the least mentions, possibly indicating they are at earlier stages of AI integration.

Entity Extraction: Organizations

- **ChatGPT dominates** the conversation, peaking sharply in early 2023—likely linked to its public release and rapid adoption.
- **Microsoft** and **OpenAI** maintain high and consistent visibility, suggesting strong engagement in AI deployment and investment.
- **Nvidia** and **Google** also sustain notable presence, driven by hardware (GPUs) and foundational models respectively.
- **Other tech giants like Amazon, Apple,** and **Meta** are mentioned, but at relatively lower levels.

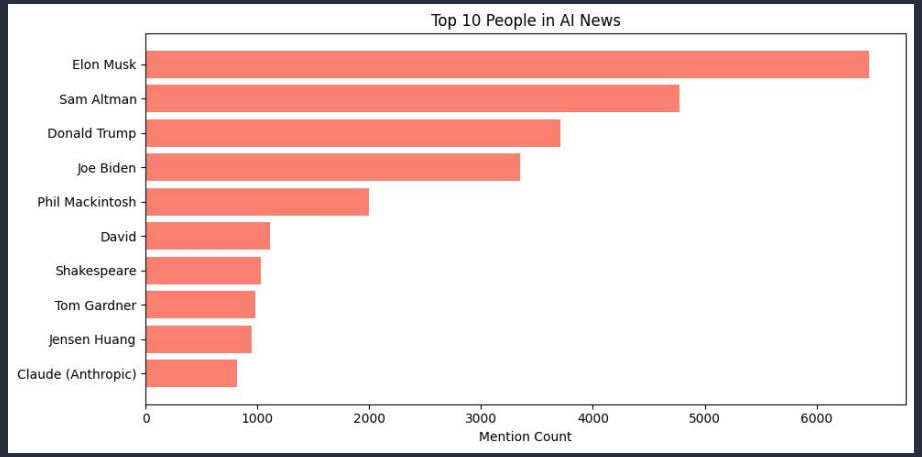


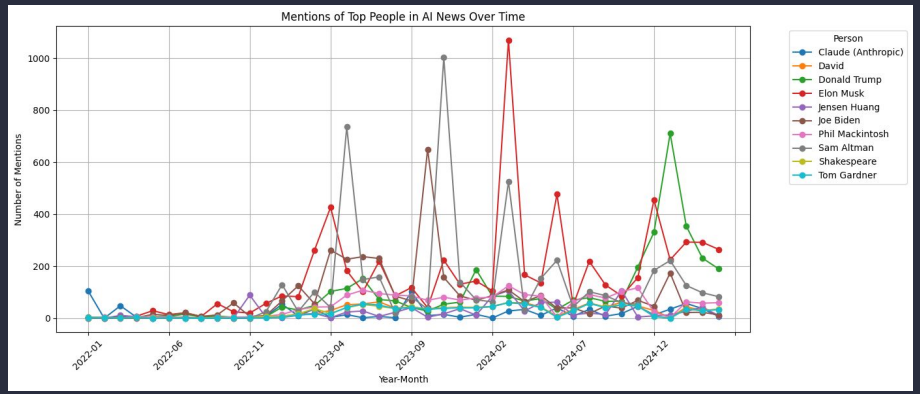


- The **surge in ChatGPT** mentions reflects its impact on public discourse.
- **Steady trends for Microsoft** and **Nvidia** show strategic, sustained AI engagement.
- **Lower but consistent mentions of regulators (EU)** suggest ongoing policy coverage.

Entity Extraction:
People

- **Elon Musk** leads by a wide margin, reflecting his dominant presence in AI-related discussions — likely due to his roles at Tesla, X (formerly Twitter), and xAI.
- **Sam Altman**, CEO of OpenAI, is a close second — expected given ChatGPT's prominence.
- **Donald Trump** and **Joe Biden** follow, suggesting their involvement in public discourse or policymaking around AI.

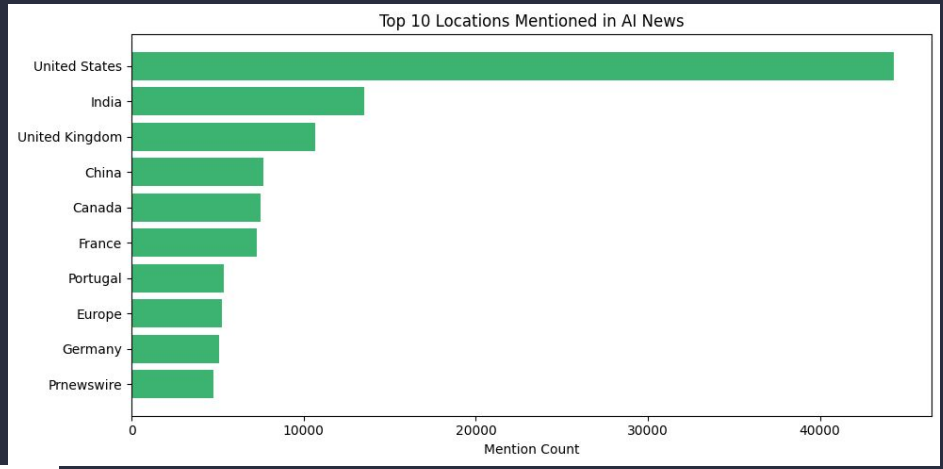


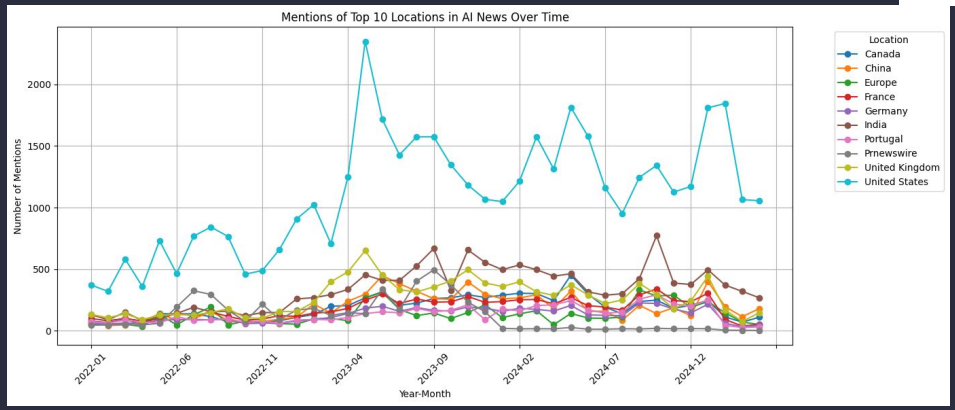


- **Musk** and **Altman** show **spikes** tied to major AI events (e.g., model launches, board changes).
- **Trump** and **Biden** have **periodic spikes** around regulatory moments.
- Some names (e.g., "Claude") show **steady** mentions, reflecting consistent relevance.

Entity Extraction: Location

- The **United States leads** AI news coverage by a large margin.
- **India, United Kingdom**, and **China** follow with strong representation.
- Other frequently mentioned regions include **Canada, France, Portugal, Germany, and Europe.**





- The **US** shows **consistent** and **high mention** volume, with several peaks indicating major AI-related developments.
- **India** has seen **rising** mentions, especially from mid-2023 onward.
- **UK, China, and Canada** show **steady** coverage, reflecting sustained AI engagement.
- Overall, global AI attention is concentrated in North America, South Asia, and parts of Europe.

Entity Extraction:
Organization, People, Location

We used **spaCy's Named Entity Recognition (NER)** to extract three key entity types from article text:

- ➔ Organizations (ORG)
- ➔ People (PERSON)
- ➔ Locations (GPE)

To improve accuracy and consistency:

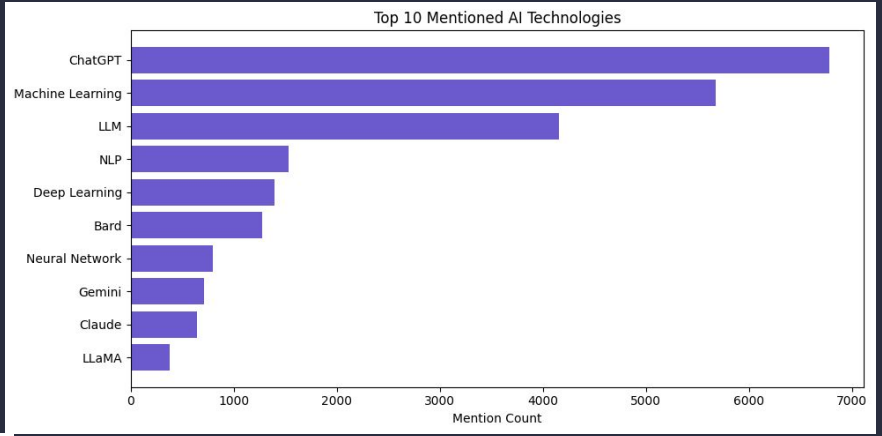
- We **normalized** entities with common aliases (e.g., "U.S." and "US" → "United States", "Elon" → "Elon Musk").
- We **excluded ambiguous terms** (e.g., removing "AI" as an organization).

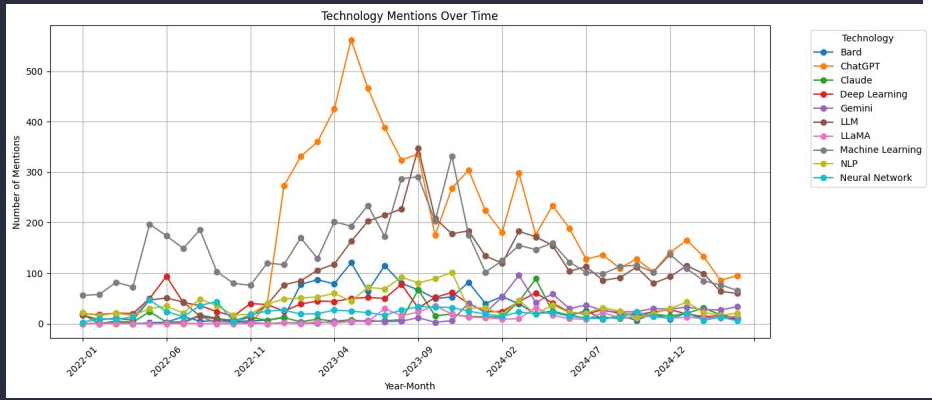
This approach was selected because:

- spaCy's pre-trained NER is fast, scalable, and effective for general-purpose entity detection.
- It automates extraction without the need for hardcoded keywords.
- The results support temporal trend analysis and sentiment attribution, offering valuable insights into who, where, and what is influencing AI discourse.

Entity Extraction: Technologies

- **Dominant Technologies**: **ChatGPT, Machine Learning,** and **LLM** dominate the AI conversation, with ChatGPT receiving the highest mentions overall.
- **Trend Spikes**: Mentions of **ChatGPT surged sharply** in early 2023 and gradually declined but r**emained high**—likely reflecting public launch and ongoing product evolution.





- **Sustained Presence: Machine Learning** shows consistent interest over time, indicating its foundational role in AI.
- **Rising Terms**: Terms like **LLM** and **NLP** gained traction **post-2022**, suggesting growing awareness of model types and natural language processing applications.

Entity Extraction:
Technologies

Approach Used:

- **Keyword mapping** + **frequency** counting over time
- Mapped **multiple aliases** (e.g., "LLM" = "Large Language Model") to ensure accuracy and avoid fragmented results.
- Used **monthly aggregation** to balance granularity and trend visibility, avoiding noise from daily variations and lag in publication dates.

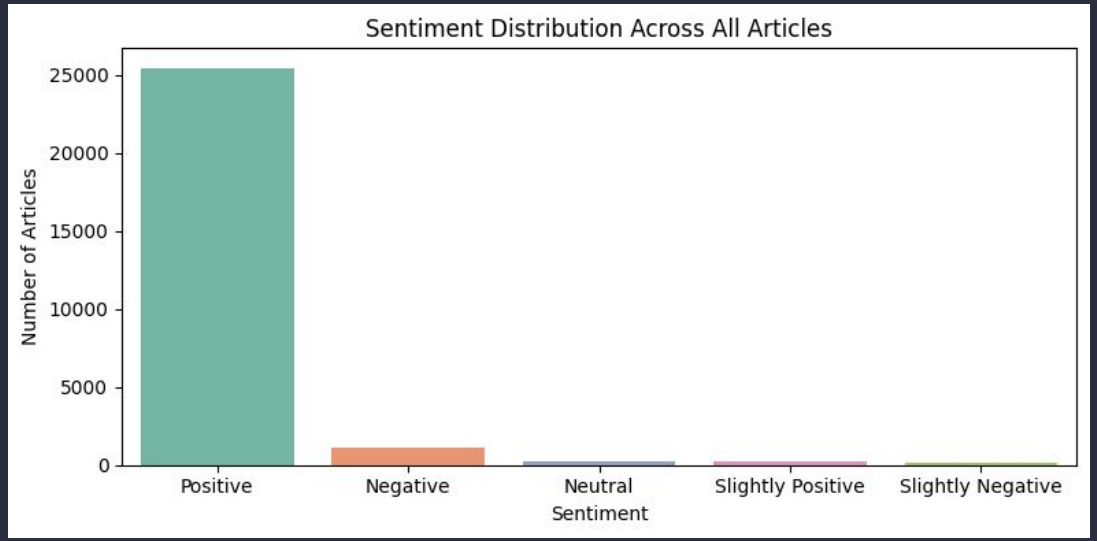
Why It's Right:

- Transparent & interpretable: Easy to explain to executives and stakeholders.
- Scalable & flexible: Can incorporate new technologies or synonyms easily.
- Data-driven insights: Reveals actual shifts in media attention over time, not just one-off mentions.

Topic-level Sentiment Analysis

NLTK

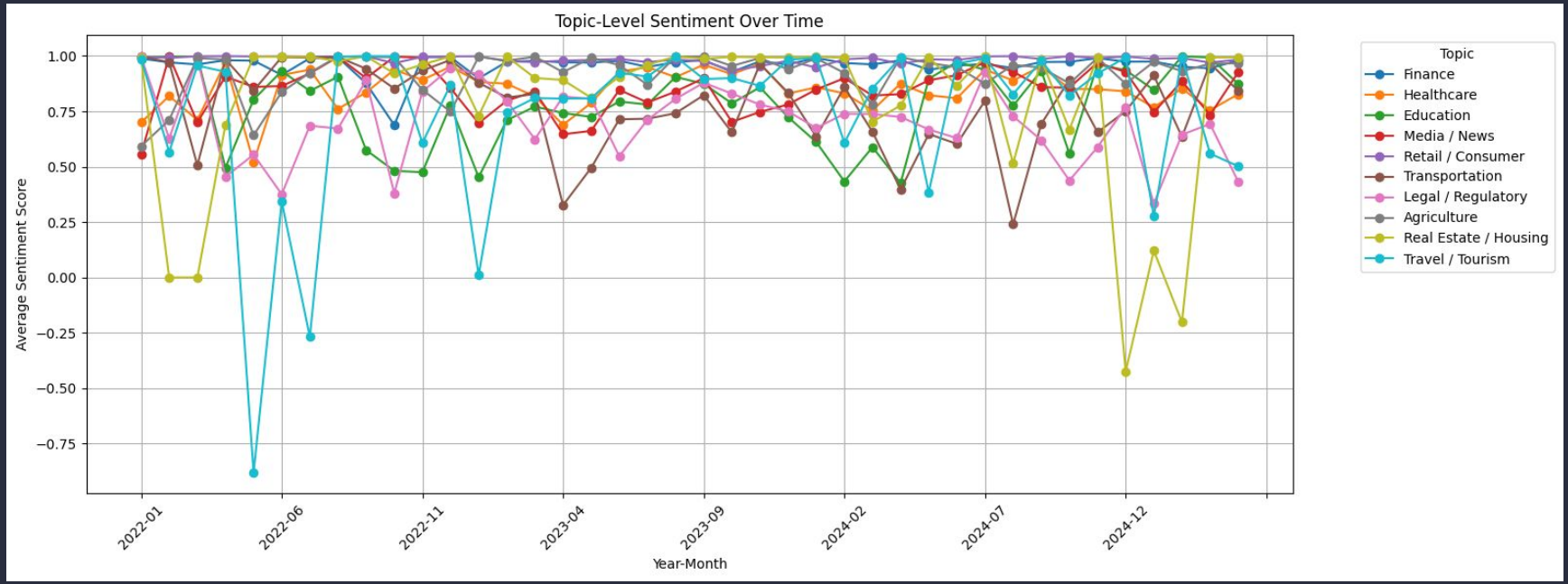
We used **NLTK's VADER** for topic-level sentiment analysis due to its speed, interpretability, and strong performance on media-style texts. It's ideal for capturing public tone across AI-related topics without the overhead of transformer models.



- **Positive sentiment dominates** overwhelmingly, suggesting AI is mostly portrayed in a favorable light across articles.
- **Negative and slightly negative mentions are minimal**, indicating limited concern or criticism.

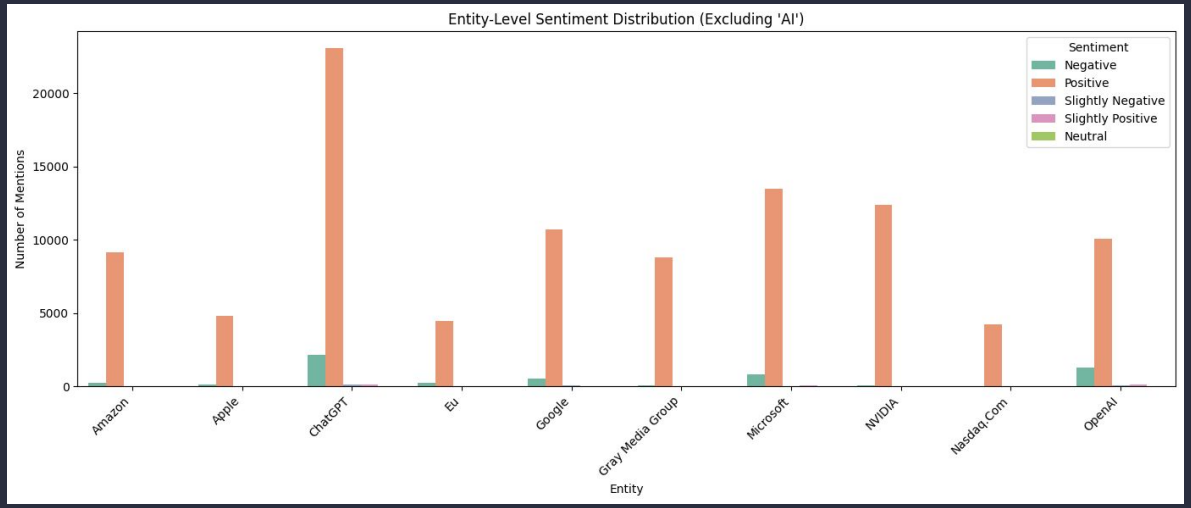
Topic-level Sentiment Analysis

Sentiment across AI topics remains **largely positive** over time, with Retail/ Consumer leading in consistency. Occasional dips in areas like Travel/Tourism and Real Estate/Housing may reflect event-driven backlash or low-article counts.



Entity-level Sentiment Analysis

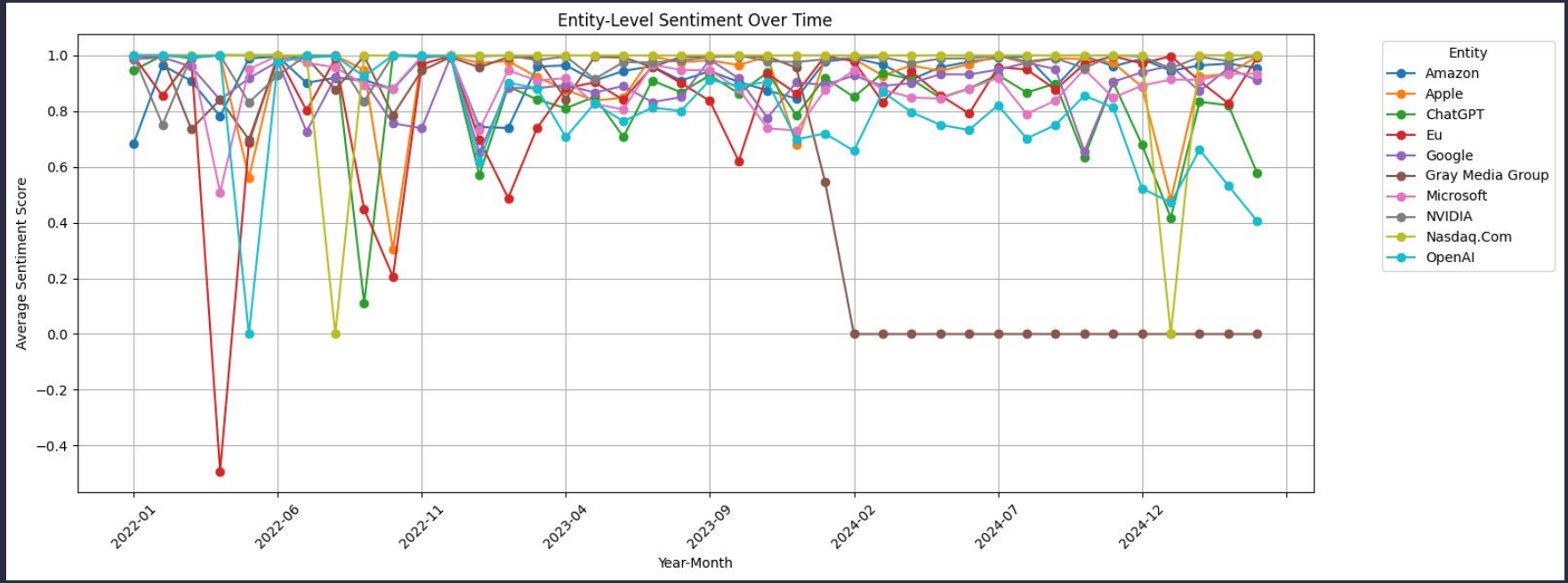
We focus on **organizations** in entity-level sentiment analysis because they are central to business impact, easier to normalize, and offer the clearest insights into how major AI players are perceived in the media.



- **ChatGPT**, **Microsoft**, **Google**, and **NVIDIA** dominate in positive sentiment volume.
- **ChatGPT** and **OpenAI** has a notably higher count of negative mentions compared to peers, suggesting more public scrutiny.
- Other entities such as Apple and Amazon have **low polarity**, meaning coverage is less emotionally charged.

Entity-level Sentiment Analysis

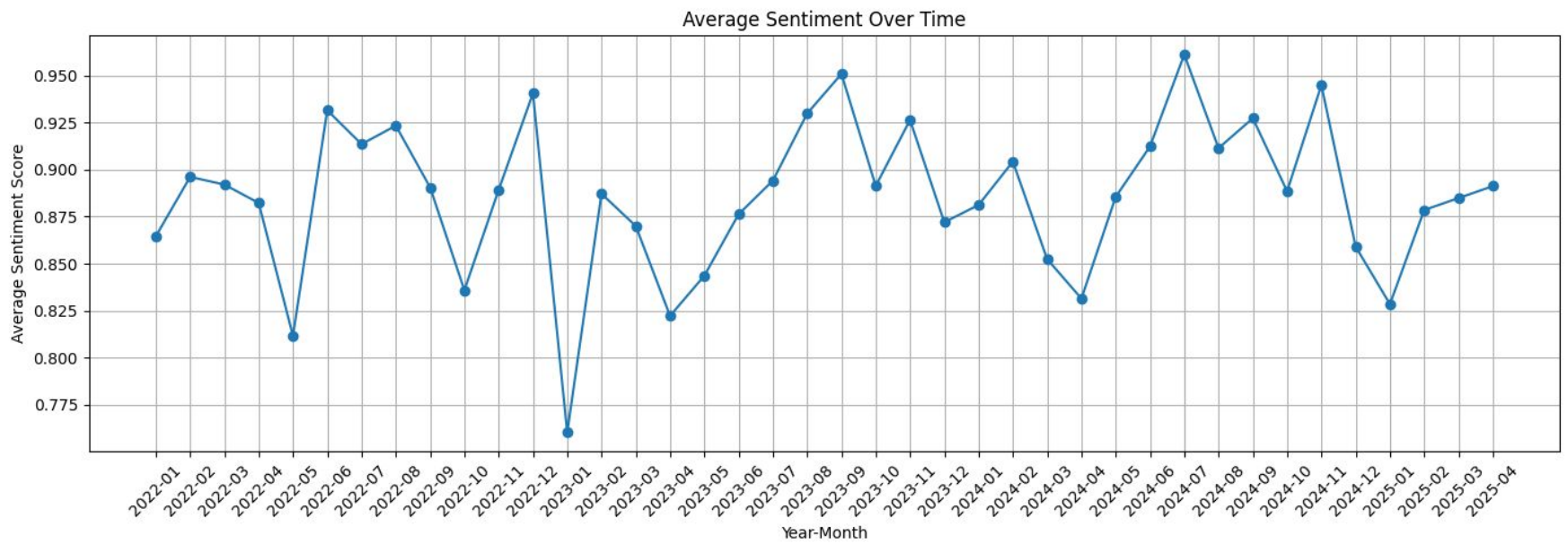
Most entities show **consistently high sentiment scores** near 1.0 across time. OpenAI, ChatGPT, and EU show sharp drops in some months, possibly tied to controversies or policy debates.



Visualization

Why Monthly Aggregation Was Used

- **Balanced granularity**: Captures key trends and turning points without being noisy.
- **Media cycles**: Monthly intervals align with how news sentiment and public narratives evolve.
- **Data density**: The dataset contains thousands of articles per month, making monthly aggregation statistically meaningful and stable.



Visualization

Major Turning Points, Trends & Outliers

- **2022-12: Significant drop in sentiment (~0.76),** possibly linked to concerns or criticism as AI technologies like ChatGPT gained public attention.
- **2023-09: Marked increase to ~0.95 sentiment,** reflecting positive reception—likely tied to improvements or responsible AI efforts.
- Overall sentiment is **consistently positive** (mostly >0.85).
- There's a **cyclical pattern** with dips followed by recovery, indicating alternating waves of hype and scrutiny in AI news.

Impact of New Technologies on Sentiment

- **Spikes** in sentiment often follow **the introduction or major update of technologies** (e.g., ChatGPT, LLMs).
- **Mid to late 2023** shows consistently high sentiment, aligning with **broader adoption and normalization of generative AI tools.**

Recommendation

What industries are going to be most impacted by AI over the next several years?

Using topic modeling (BERTopic), sentiment analysis (NLTK VADER), and entity extraction, we find that the most impacted industries will be:

- ➔ **Finance** – Driven by automation, fraud detection, algorithmic trading, and customer service AI.
- ➔ **Retail / Consumer** – Personalization, inventory optimization, and AI-driven marketing are key trends.
- ➔ **Healthcare** – AI in diagnostics, drug discovery, patient monitoring, and operational efficiency.
- ➔ **Legal / Regulatory** – Regulatory tech (RegTech), contract analysis, and AI for compliance tasks.
- ➔ **Media / News** – Content generation, curation, and audience targeting via AI models.

Recommendation

Automate Jobs

- ➔ **Finance:** Use AI for fraud detection, data entry automation, and customer chatbots.
- ➔ **Retail:** Automate checkout, inventory tracking, and demand forecasting.
- ➔ **Healthcare:** Automate diagnostics, transcription, and appointment scheduling.
- ➔ **Legal:** Use AI for contract review and compliance monitoring.

Improve Productivity

- ➔ Deploy **AI assistants (e.g., Copilot)** for writing, coding, and summarizing.
- ➔ Use **NLP tools** to summarize documents, emails, and meetings.
- ➔ Optimize scheduling and logistics using AI-powered planners and routing.

Ensure Successful Adoption

- ➔ **Start small:** Pilot low-risk, high-impact AI projects.
- ➔ **Invest in data:** High-quality data is essential.
- ➔ **Train staff:** Offer basic AI literacy and upskilling.
- ➔ **Build trust:** Ensure transparency and fairness in AI tools.
- ➔ **Focus on augmentation:** Use AI to enhance, not replace, human work.

Thank you

Appendix: Topic Detection