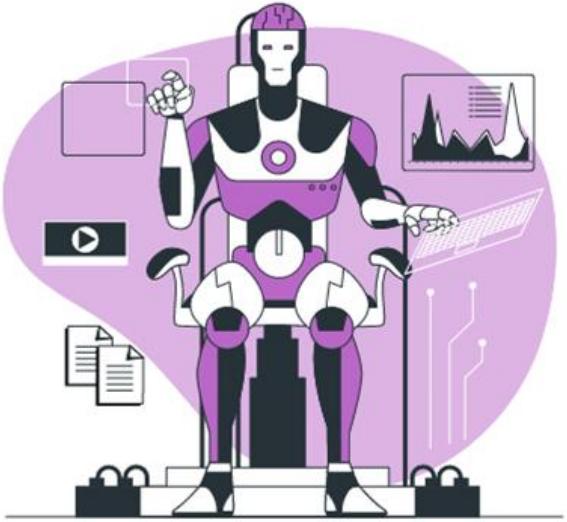


Will AI Take Your Job?

ADSP 32018 IP02 Natural Language Processing
Final Project
William DeForest

Agenda

01 Executive Summary

02 Data & Methodology

03 Article Cleaning & Filtering

04 Top Industries for AI Integration

05 Sentiment Analysis

06 Key Orgs, Products, & People

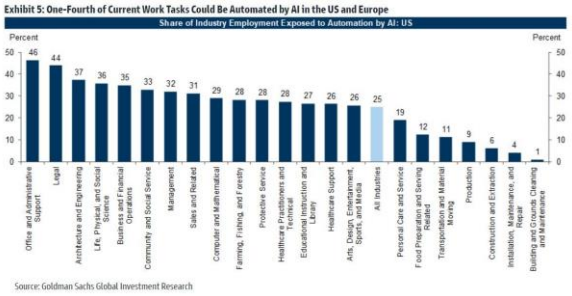
07 Recommendations - Work With AI, Not Against It

Executive Summary

Background & Context

The recent **renaissance** in **artificial intelligence (AI)** is set to transform industries by **automating** tasks and enhancing **efficiency**.

In fact, **Goldman Sachs** recently published a report in which they estimate that **25% of current work tasks** can be automated by AI.



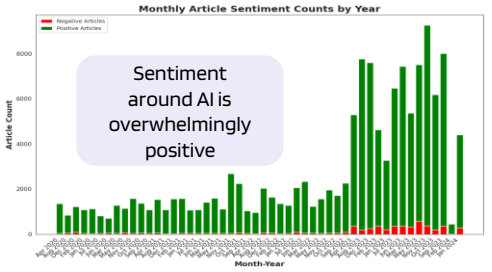
Problem & Methods

This project aims to identify what types of **tasks** and **jobs** are most likely to **feel the largest impact from the advent of AI.** These insights are invaluable as they can help people identify strategies to **make AI work for them** rather than have AI replace their work.

A collection of **~200k news articles** related to AI, machine learning (ML), and data science (DS) were made available for analysis

To achieve this objective, a variety of **natural language processing (NLP)** techniques such as topic modeling, sentiment analysis and named entity recognition are employed

Findings & Next Steps

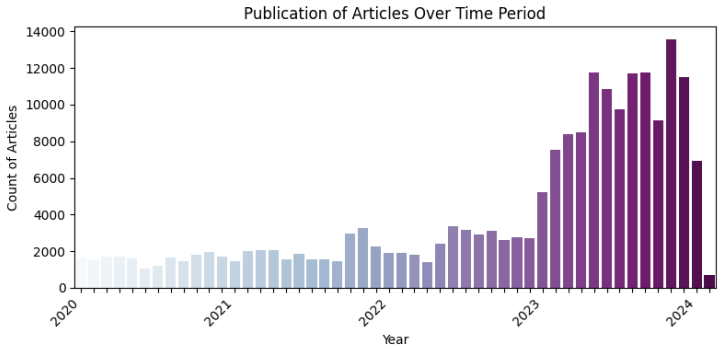


Some areas will be hit harder than others:

- The **semiconductor industry** will continue to explode as computing demand grows
- **Administrative** and **content creation** roles face significant automation
- ChatGPT is **losing hype**, but it's true value as a **tool** for building new applications will be realized

Paddling hard before the AI wave hits will be the key to riding it to a **brighter future**.

Source Data



~200k news articles

- Published between **2020-2024**
- Collected using a web crawler
- Focused on **data science, machine learning and artificial intelligence** topics
- Features include article title, text, url, publish date, and language (all in English)
- As a result of the web scraping, the raw article text is **noisy**
- Website headings, links and other errors irrelevant to the article are removed

Methodology

1. Article titles and text are cleaned using **regular expression pattern matching**.
2. Extremely short/long articles and articles lacking data science, machine learning and AI keywords are filtered out.
3. Key topics and industries are identified using **Latent Dirichlet Allocation (LDA)** topic modeling from **GenSim** and **ktrain**.
4. An **SVM sentiment analysis model** is trained on an open-source corpus of ~4.6k news articles.
5. This model is applied to each sentence of each article and aggregated to determine **article sentiment.**
6. **Named Entity Recognition (NER)** is performed at a sentence level using **spaCy's** large English pipeline trained on written web text (**en_core_web_lg**)
7. **Targeted sentiment analysis** is conducted by combining article topics, entities and sentiments and analyzing their **changes over time**

Article Cleaning & Filtering

1

Media outlet name was removed from the article title if present

2

URLs, links, special characters and exceptionally long word (15+ characters) were removed from article text

3

Common **web crawl remnants** unrelated to articles ('Cookies', 'Contact Us', 'Subscribe', etc.) were removed

4

A common structure for **website headings** was several capitalized words in a row. When five or more were in a row, they were removed.

5

Tokenized articles and filtered out **extremely short/long** articles.
Too short < (1/50)*mean token count
Too long > 5000 tokens

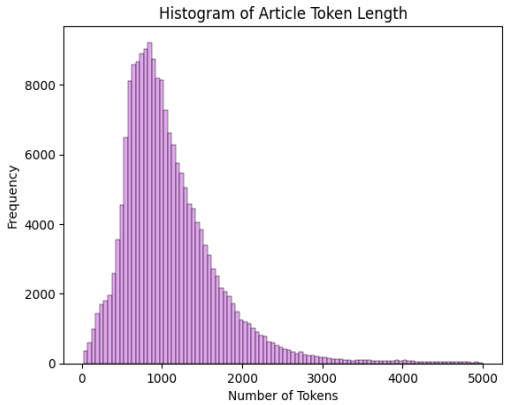
6

Combined clean title and text tokens and used **nltk's FreqDist** to find the top 100 most common tokens. Chose **top 10 AI, ML, Data Science keywords**, enriched with related ChatGPT generated keywords, and filtered out articles missing these tokens in the **first 20%** of the article.

Top 10 Keywords

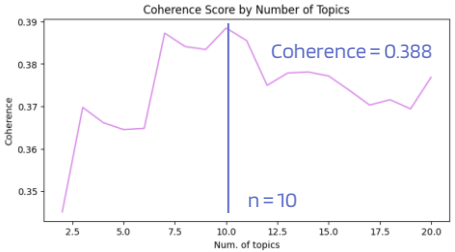
1. Ai
2. Data
3. Intelligence
4. Artificial
5. Technology
6. ChatGPT
7. Learning
8. Generative
9. Machine
10. Analytics

~193k Cleaned & Filtered Articles



Topic Modeling to I.D. AI-Impacted Industries

LDA Topic Modeling (Gensim)

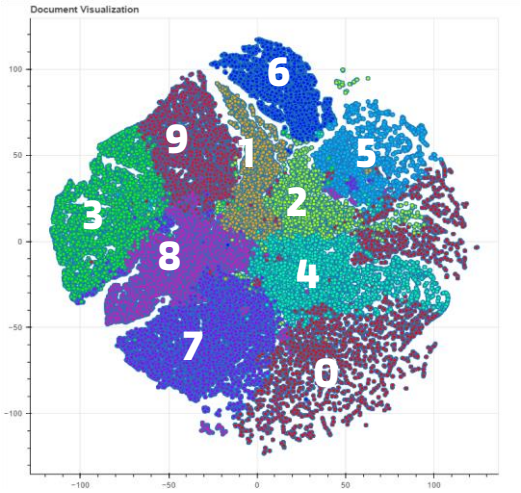


Gensim LDA Topic Modeling was performed on a sample of 10,000 articles. **Coherence score** was calculated for **n = 2-20 topics**.

Coherence score (a measure of how meaningful and interpretable the identified topics are) is **maximized at n = 10 topics**.

LDA Topic Modeling (ktrain)

Using the optimized n = 10 topics from Gensim LDA, **ktrain topic modeling** was performed due to ktrain's superior topic analysis capabilities.



Analyzing the **word distributions** and the **summaries** of top-ranked articles for each topic, **six** clear areas being **impacted by AI** emerged.

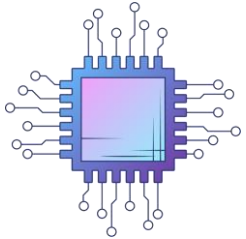
Word Distributions of Top 6 Topics

2: machine, model, learning, network, computer, application, human, algorithm
4: customer, service, experience, generative, platform, process, enterprise
5: patient, health, care, medical, study, clinical, disease, cancer, improve
7: risk, human, student, need, concern, job, government, world, tech
8: user, model, chatgpt, chatbot, tool, email, image, language, search
9: open, video, game, create, voice, feature, people, icon, good

Interpreting ktrain Topics: Text Summarization

The TransformerSummarizer from ktrain was used to summarize the 50 top-ranked articles for each topic. These summaries provided additional insight and aided in the identification of the following industries/areas.

2 Semiconductors



Summary Excerpt: "chip chip artificial intelligence chip specialized processor design "

4 Conversational AI

Summary Excerpt: "vast amount datum improve service enhance customer experience"

5 Healthcare

Summary Excerpt: "recent clinical trial combination immunotherapy nasopharyngeal cancer"

7 AI Regulation

Summary Excerpt: "concern regard artificial intelligence repeatedly warn potential danger"

8 ChatGPT & LLM's

Summary Excerpt: "restrain request rate thus limit collection datum high traffic site analitici"

9 Entertainment

Summary Excerpt: "light big rays design design design light big ray design designdesign"

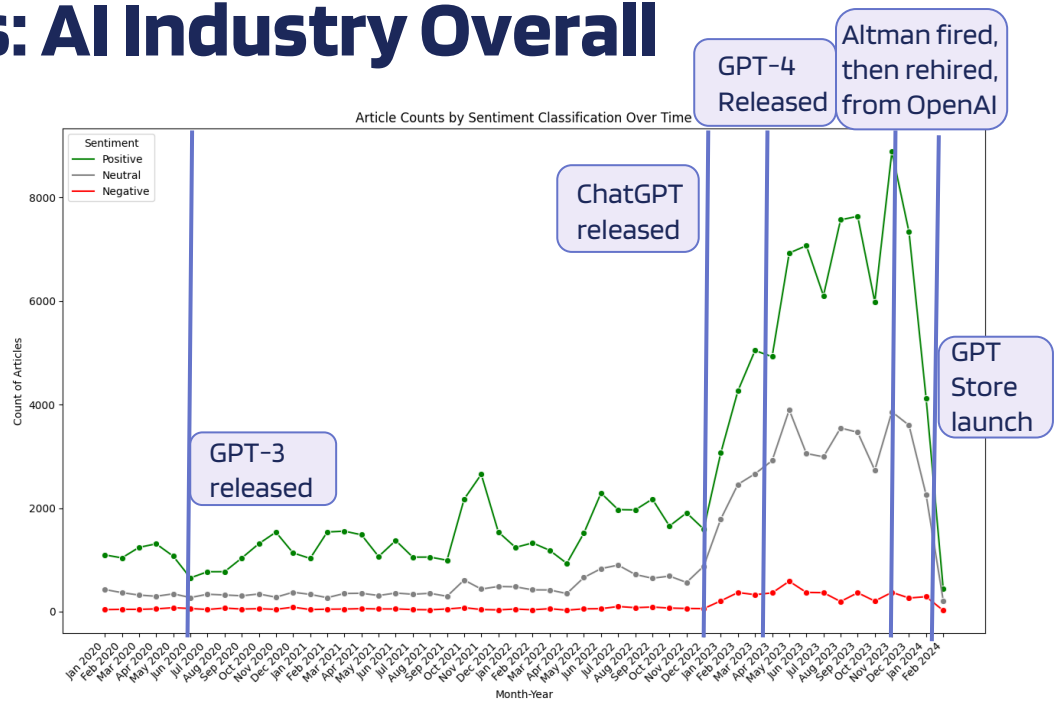
Sentiment Analysis: AI Industry Overall

Custom SVM Classifier

A **custom SVM classifier** was trained on the '**ML-news-sentiment**' dataset from Hugging Face.

- Dataset contains 4.6k news articles
- Labels: positive, neutral, or negative.

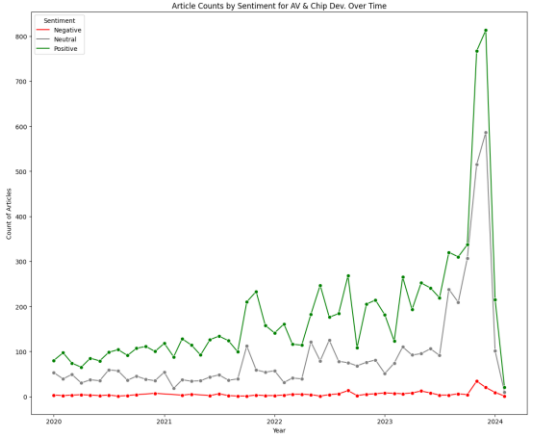
Chosen because it is **trained on news articles** (rather than other forms of text such as reviews) and because it provides a **neutral** classification. Often, the factual or descriptive statements in news articles have no clear sentiment. Rather than inaccurately classifying these statements as positive/negative, this model labels them as neutral.



Articles were chunked into **sentences** using **spaCy's** sentence recognizer **'senter'** and the customized sentiment model was deployed at the sentence level. As can be seen above, sentiment surrounding AI, ML and DS was **overwhelmingly positive** from the start of 2020 through 2023, though a waning count of articles written in early 2024 could warn of a potential **"AI winter"**

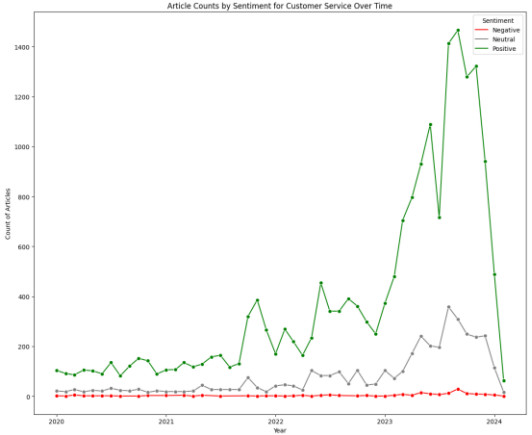
Sentiment Trends Provide Insight Into Burgeoning Or Contracting Industries (1/2)

Semiconductors



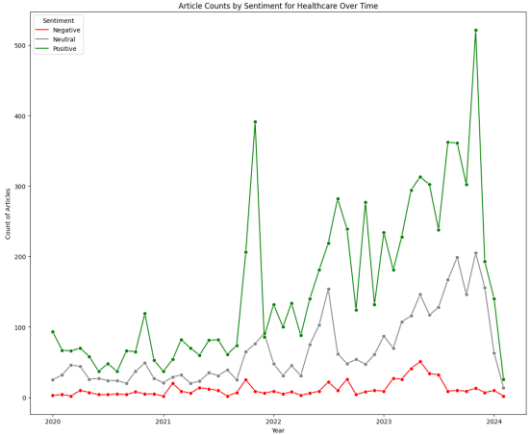
Investigating the sharp increase in positive sentiment in late 2023 reveals that **NVIDIA** has seen a massive increase in stock price as demand for its AI-powering chips reaches **fever pitch**

Conversational AI



Positive sentiment around **conversational AI** has risen hand-in-hand with LLMs as ChatGPT, PaLM 2, and Bard showcase strong **question answering** capabilities - a key for conversational AI applications

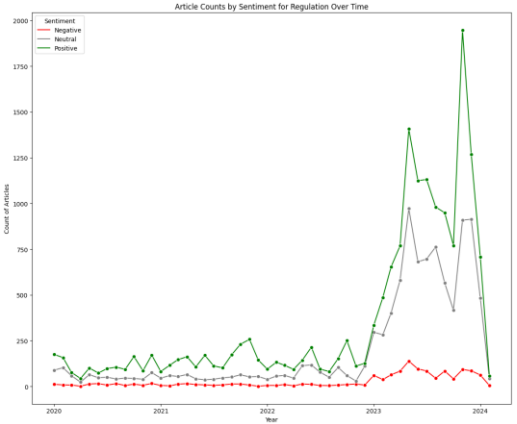
Healthcare



Relative to other industries, **healthcare** has seen gradual growth in positive sentiment and has had more variation in negative sentiment. There is massive potential in the automation of **admin. work**, but data confidentiality concerns and **regulation** are slowing progress

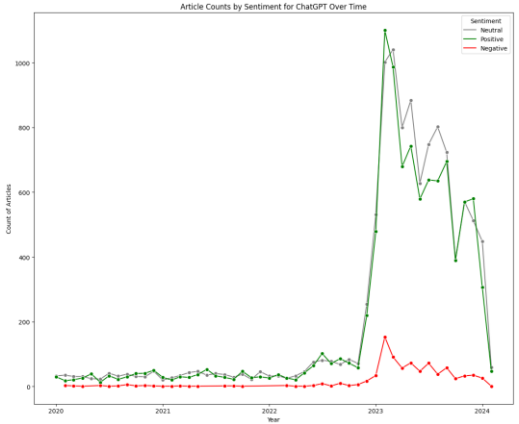
Sentiment Trends Provide Insight Into Burgeoning Or Contracting Industries (2/2)

AI Regulation



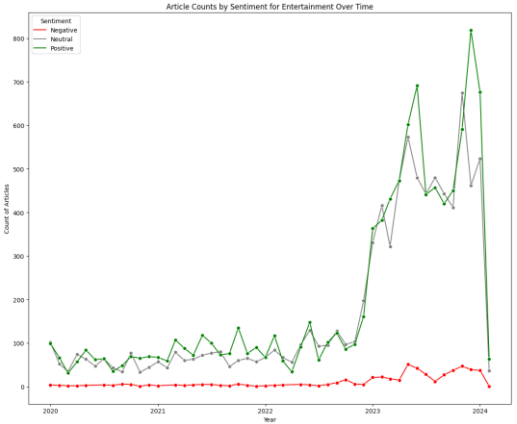
AI regulation peaks in positive sentiment simultaneously with major OpenAI news. The first rise begins in late 2022 with the **release of ChatGPT**. The highest peak occurs in conjunction with the **firing of Sam Altman** which set off the **AI acceleration vs. alignment** debate

ChatGPT



ChatGPT is such an integral feature to the AI landscape that it's an entire topic. However, after its initial spike, we see **decreasing positive sentiment** and high levels of neutral sentiment, This trend suggests that ChatGPT is **losing its novelty** and becoming a more mature constituent of the space

Entertainment



Positive sentiment regarding AI in the entertainment industry has largely followed breakthroughs in generative AI - releases of **ChatGPT** and the multimodal **GPT-4**. While breakthroughs, they raise thorny questions around intellectual property and **threaten many creative jobs**

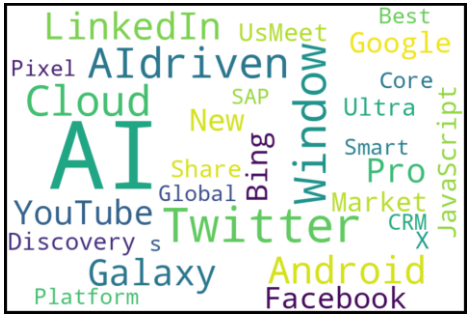
Named Entity Recognition (NER) Identifies the Organizations, Products, and People Driving AI

Named Entity Recognition (NER) was performed using spaCy's "en_core_web_lg" model. A pipe with only "ner" enabled was used to perform NER at the sentence level.

Organizations

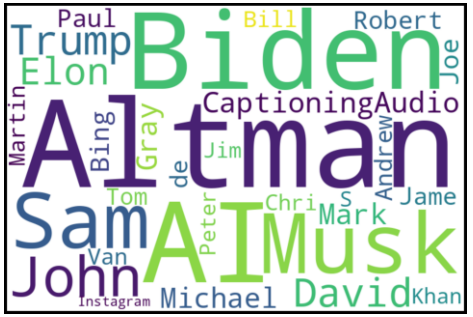
The organizations mentioned most frequently include **Google, Microsoft, Facebook** and **OpenAI**. Unsurprisingly, these organizations are the world's premier tech companies and are leading the way in AI R&D. ChatGPT is technically a product, but is so mainstream that it's often referred to like a company

Products



Top AI, ML, DS products include **YouTube, Galaxy, Bing** and **LinkedIn**. YouTube is a major source for learning about AI, Bing was integrated with the Copilot chatbot in early 2023, and Galaxy phones by Samsung are one of the first to integrate AI on-device. LinkedIn is an important connector of top AI talent

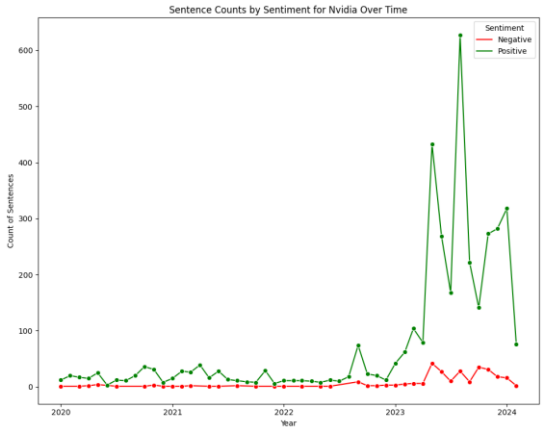
People



OpenAI CEO **Sam Altman**, President **Joe Biden**, and Tesla and xAI founder **Elon Musk** are the largest figures in the AI space. Altman for his role mainstreaming Generative AI, Biden in his position of balancing AI acceleration and regulation at the national-level, and Musk as a general, and very outspoken, leader in tech

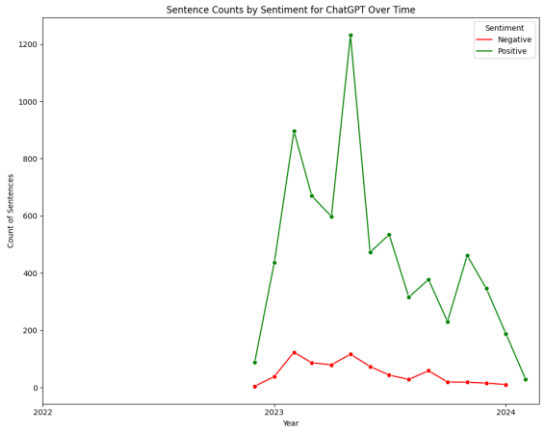
Targeted Sentiment Analysis of Important Entities In Their Respective Industries

Semiconductors



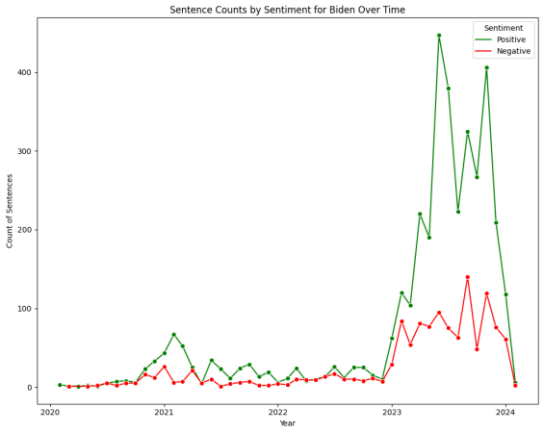
The positive sentiment around **Nvidia** jumped significantly in 2023. This reflects investors' excitement around Nvidia's positioning in the **booming semiconductor manufacturing market.**

ChatGPT



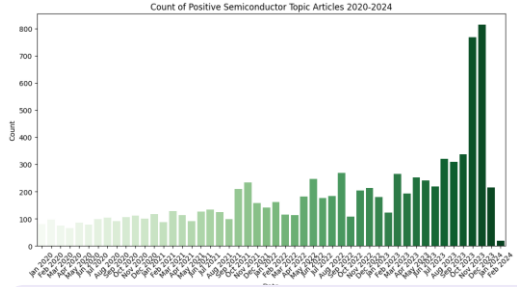
Positive sentiment around ChatGPT (entity) **declined** during 2023, despite the release of industry-leading GPT-4. This shows that ChatGPT is losing hype as **competitors** have produced similarly performing models.

AI Regulation



Sentiment regarding **President Biden** is significantly mixed. While partially due to **political polarization** in the U.S., we do see a spike in negative sentiment in Oct. 2023 when Biden issued an executive order establishing **AI safety and security standards**.

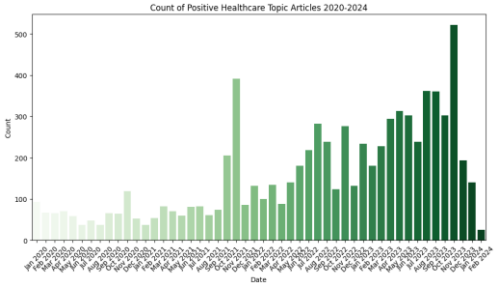
Findings & Next Steps → **Recommendations**



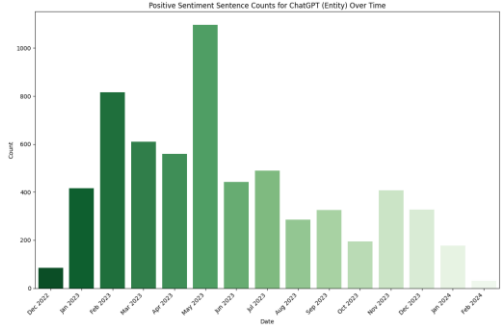
1. Invest in semiconductor/chip manufacturing. Furthermore, learn how to use a GPU!

While the semiconductor market has been red-hot, it is well warranted. Only a handful of companies like **Nvidia**, **AMD** and **Intel** have the infrastructure to build these high performance chips. There is strong positive sentiment and demand is likely to increase as more AI products are built. Learn how to **leverage** this powerful tech before it replaces you.

2. Administrative and content generation roles are at particular risk of being automated by AI.



The repetitiveness of **admin. tasks** and the sheer amount of training data available for **content generation** have people excited about AI automation in these areas. If you are in such a role, **embrace AI** and use it to improve your work. You are already an expert and have a step ahead of everybody else. **Don't let them catch up** by refusing to accept AI.



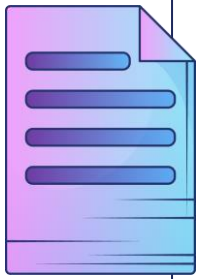
3. ChatGPT is losing its novelty and hype. Take this opportunity to separate yourself from others.

While the **hype** of ChatGPT is dying down, it is still **the most powerful tool** the world has ever seen. Industries like **healthcare** and **entertainment** will experience major overhauls as new GPT iterations augment existing, or invent new, processes. **Educate yourself** about LLMs, learn the basics of **integrating** them into existing structures and then **get creative**. Use ChatGPT to **create a new job** for yourself, rather than waiting around for it to take yours.

Thanks!

Do you have any questions?

wdeforest@uchicago.edu
+1 206 678 6215
https://github.com/wdeforest23



References

Goldman Sachs Article: https://www.aei.org/articles/why-goldman-sachs-thinks-generative-ai-could-have-a-huge-impact-on-economic-growth-and-productivity/

Sentiment Analysis Training Data:
https://huggingface.co/datasets/sara-nabhani/ML-news-sentiment?row=3

Animated Images:
https://storyset.com/search