Analysis of AI Trends in News Articles Using NLP

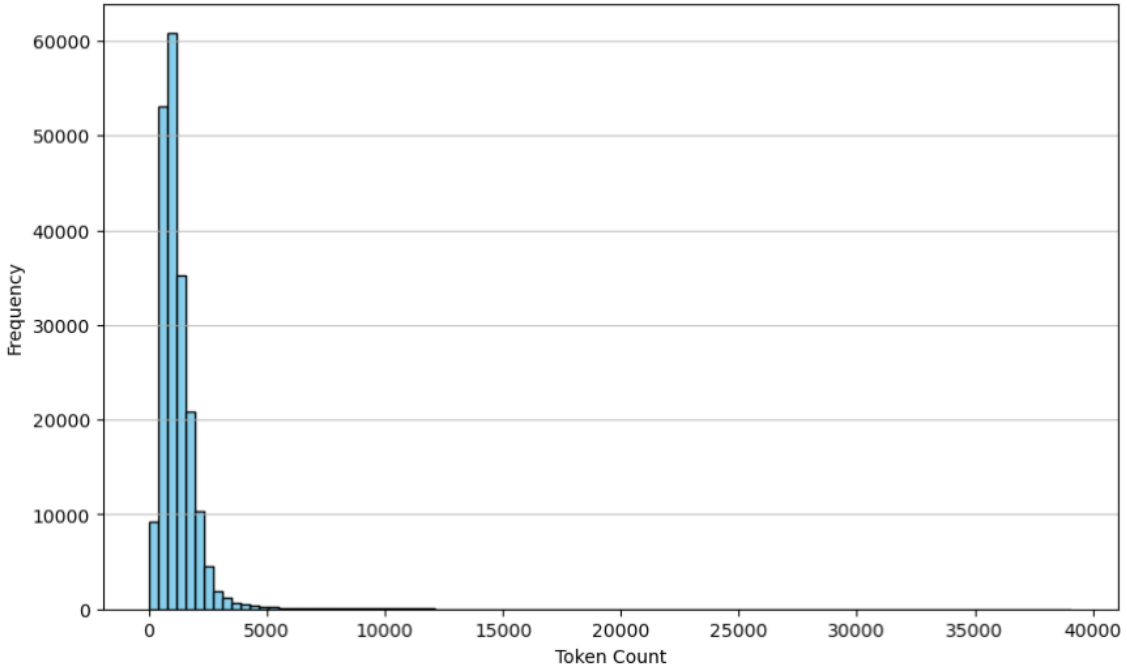
Daichi Ishikawa

March 2024

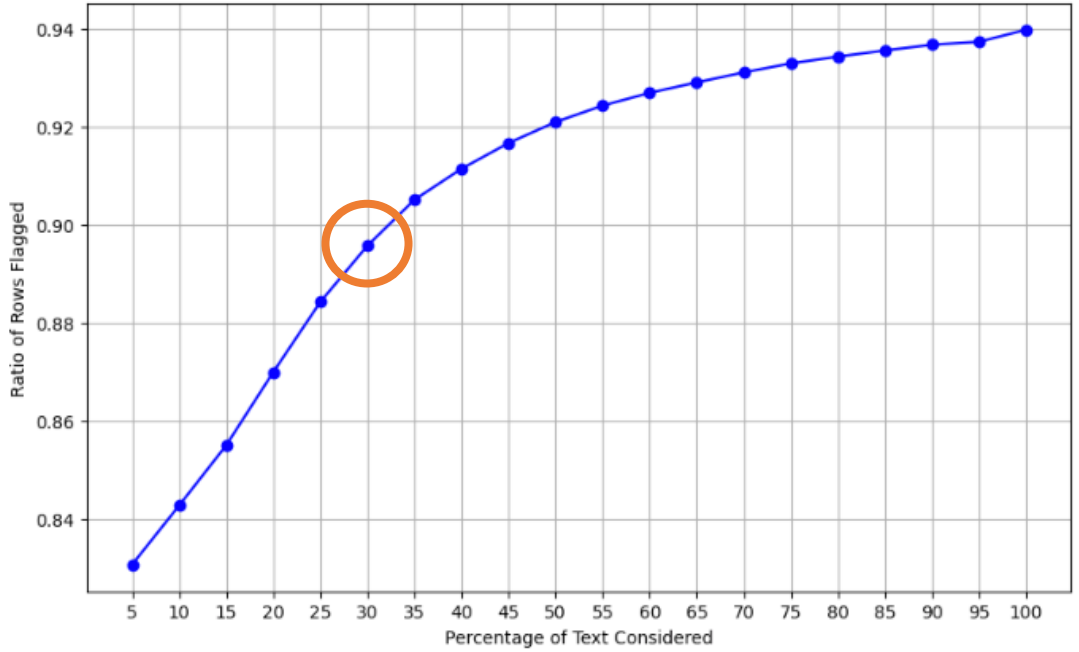
Dataset Overview and Initial Cleaning

- The original dataset comprised **200k news articles**, with an average token count of about 1,200 per article, spanning from January 2020 to February 2024. Following a process, it was **reduced to 157k** by:
- **1. Exclude Outlier:** The distribution of Token Count is right-tailed, meaning that articles with a very high Token Count take a long time to analyze and should be excluded. Similarly, articles with a very low Token Count should also be excluded due to their lack of information. Therefore, the top and bottom 1% of articles by Token Count are excluded.
- **2. Removing News Weakly Related to AI**: Articles that did not mention 'Artificial Intelligence' or 'AI' within the first 30% were excluded, by adopting the point where the curve becomes more gradual (refer to the right figure). Notably, using 'Artificial Intelligence' alone would exclude too many articles, whereas including related terms like 'Machine Learning' and 'Data Science' did not significantly alter the outcome.
- **3. Removing Non-English News:** Articles not in English were eliminated.
- During data cleaning, **URLs and special characters were removed, and whitespace was normalized using regular expressions.**

Histgram of Token Count



Ratio of News with Relevent Words(AI, Artificial Intelligence) by Text Percentage



Topic Modeling

- Conducted topic modeling using **LDA through Ktrain**, initially setting the number of topics to 10. After this initial run, three topics related to AI were identified as being too general. Therefore, performed topic modeling again for the three AI-Related groups with the number of topics set to 10.
- Additionally, topics with little relevance were grouped as noise(Other/Not Assigned).
- As a result, as shown in the table on the right (Final Modeling), the news were divided into **nine topics: Finance, Market Research, HealthCare, AIproduct_Cloud_Enterprise, ConversationalAI_LLM, Digital_Innovation, Google_Services, Chip_Computing, and CEO_Event_Regulation.**
- In determining the content of the topics, representative news articles were also referenced.
- Before executing the model, data cleaning steps were taken, including converting to lowercase, removing stopwords, and lemmatizing.

Initial Topic Modeling

Topic	Keywords	Number of news
Finance	stock market, share price, nasdaq, fund, investment, investor	9869
Market Research	market, global analysism, growth, research forecast, key trend, player size	8919
HealthCare	health, healthcare, patient, science, medical, machine	22408
AI-Related	product, entertainment, consumer, resource, general,computer	3470
	solution, customer, platform, generative, cloud	41858
	google, tech, image, video, chatgpt, user, model	30024
Other	email, facebook, twitter, search, schedule, community, music, program, radio	8160
	newswires, south africa, united, north, country, international, island	4461
	weather, sport, local	21928
	september,experience, august, october, marketing,website	4870
Not Assigned	Probability < threshold = 0.25	405

Final Modeling

Topic	Keywords	Number of news
Finance	stock market, share price, nasdaq, fund, investment, investor	9869
Market Research	market, global analysism, growth, research forecast, key trend, player size	8919
HealthCare	health, healthcare, patient, science, medical, machine	22408
AIproduct_Cloud_Enterprise	customer, generative, cloud, enterprise, experience, organization, capability, digital, application	10799
ConversationalAI_LLM	model, language, chatgpt, tool, gpt, text	3684
Digital_Innovation	digital, innovation, market, development, partner, team	6106
Google_Services	google, mobile, apple, user, tech, search, video (Top relevant article: Google duet AI, gemini ai bard google pixel pro)	6057
Chip_Computing	nvidia, edge, computing, chip, device, performance	3484
CEO_Event_Regulation	tech, event, Microsoft, ceo, openai, privacy, generative, policy	6957
Other/ Not Assigned	Wide range of general news articles	78771

NER(Named Entity Recognition)

- After chunking the article into sentences, Named Entity Recognition (NER) was performed using **spaCy**.
- Additionally, for Job titles (JOB), industries (IND), and technology names (TECH) that are difficult to recognize with default spaCy, **lists were manually created utilizing websites and GPT-4, and recognition was simplified using regular expressions.** Although not appearing at the top of the table, TECH includes new technology terms, and flagging new technologies enriched the subsequent analysis with more insights.
- The recognized entities are listed below. Some category misclassifications are observed, but overall, **the recognition quality looks high.**
- A detailed analysis of each entity will be presented later. During the following analysis, entities with incorrect categories or those serving as noise (e.g., the news media name 'Gray Media Group' or the too common 'AI') were manually excluded. Additionally, variations in the representation of the United States, such as US and U.S., were standardized.

Top 15 Entiries by Category(spaCy)

ORG	PRODUCT	GPE	PERSON
AI (163769)	AI (375934)	US (41972)	GPT-4 (4057)
Google (46702)	Google Cloud (2698)	India (31765)	Elon Musk (3962)
Microsoft (41188)	UsMeet (2582)	PRNewswire (28700)	Musk (3930)
ChatGPT (34820)	Galaxy (2375)	China (25463)	CaptioningAudio DescriptionAt (3399)
Gray Media Group Inc. (32587)	Windows (2364)	U.S. (22862)	Sam Altman (3112)
Gray Media Group (22491)	YouTube (2300)	UK (20429)	CaptioningAudio (2713)
Gray Television Inc. (21534)	Windows 11 (2226)	Japan (12749)	Bing (2362)
IBM (20299)	SAP (2202)	Canada (10733)	Biden (2264)
Nvidia (17196)	HPC (2153)	France (8698)	AdvertisingAt Gray (2158)
Artificial Intelligence (13726)	JavaScript (2139)	Germany (8314)	Altman (2061)
NVIDIA (13503)	Bing (1827)	Australia (7933)	Greta Van SusterenCircle - Country (1544)
Apple (13424)	CRM (1819)	the United States (6778)	Vicky Stavropoulou (1311)
Amazon (13416)	Core (1218)	Italy (5973)	Sundar Pichai (1239)
OpenAI (12583)	Pixel (1018)	Russia (5757)	Phil Mackintosh TradeTalks (1229)
Facebook (10785)	Generative AI (1009)	Taiwan (5026)	Satya Nadella (1213)

Top 15 Entiries by Category(Regular expressions)

IND	JOB	TECH
Software (88018)	Analyst (14105)	Cloud (88007)
Financial (56239)	Professor (8344)	Generative AI (86145)
Healthcare (51707)	Editor (7973)	Machine Learning (47919)
Energy (41140)	Scientist (4548)	ChatGPT (44138)
Education (31711)	Writer (4336)	OpenAI (26638)
Government (30855)	Engineer (3212)	Cybersecurity (19014)
Finance (26946)	Athlete (2336)	Chatbot (18308)
Manufacturing (25890)	Designer (1812)	ML (14740)
University (25307)	Teacher (1404)	GPT (14668)
Consumer (22120)	Artist (1004)	IoT (12930)
Legal (21254)	Architect (998)	Conversational AI (12876)
Game (20688)	Data Scientist (973)	Deep Learning (11495)
Retail (20578)	Product Manager (719)	Computer Vision (11100)
Gaming (20119)	AI Developer (521)	Blockchain (11046)
Hardware (19808)	Software Engineer (416)	Bard (10783)

*The numbers in parentheses indicate the count.

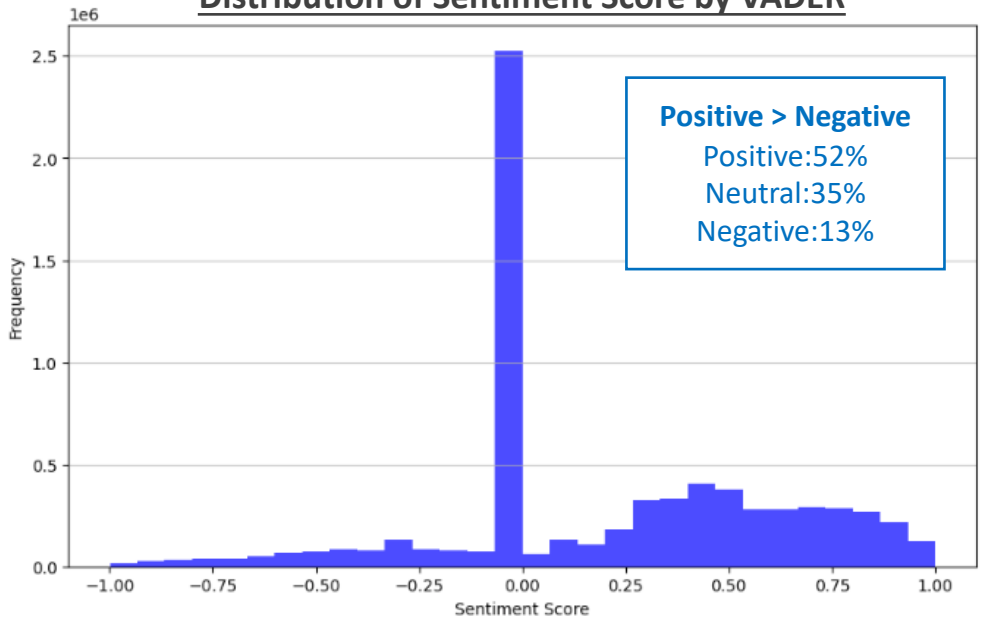
Sentiment Analysis

- In comparing DistilBERT with **VADER (Valence Aware Dictionary for Sentiment Reasoning)**, VADER achieved higher accuracy, leading to the decision to utilize VADER. DistilBERT often assigns excessive evaluations to sentences that VADER categorizes as neutral.
- The reason for VADER's higher accuracy may lie in its analysis at the level of short sentences, where its simpler design potentially worked more effectively.
- Furthermore, given the short length of the texts and the already satisfactory results, it was determined that significant improvements in accuracy through customization were unlikely, and therefore, no customization has been undertaken.
- The decision to conduct sentiment analysis on a **sentence-by-sentence basis**, rather than on the entire article, was driven by the **changes in content and sentiment within news articles and the high potential for improved accuracy in targeting sentiments towards entities.**

DistilBERT vs VADER

sentence	DistilBERT	VADER
The autonomous AI does not take the place of a doctor rather it is a tool to better streamline the clinical process.	-0.99	0.44
A Gray Media Group Inc. Station - 2002-2023 Gray Television Inc.	-1.00	0.00
Find refine and apply your ICP across the go-to-market motion with Patri's AI-Generated ICP Engine.	-0.95	0.00
AdvertisementAdvertise with NZME.LaMDA not only communicates via language Lemoine points out but has 'eyes' too capable of interpreting images as well as words	-1.00	0.72
Samsung Foundry's commitment to advancing semiconductor technology aligns with our vision for advancing RISC-V and AI and makes them an ideal partner to bring our AI chiplets to market	1.00	0.79

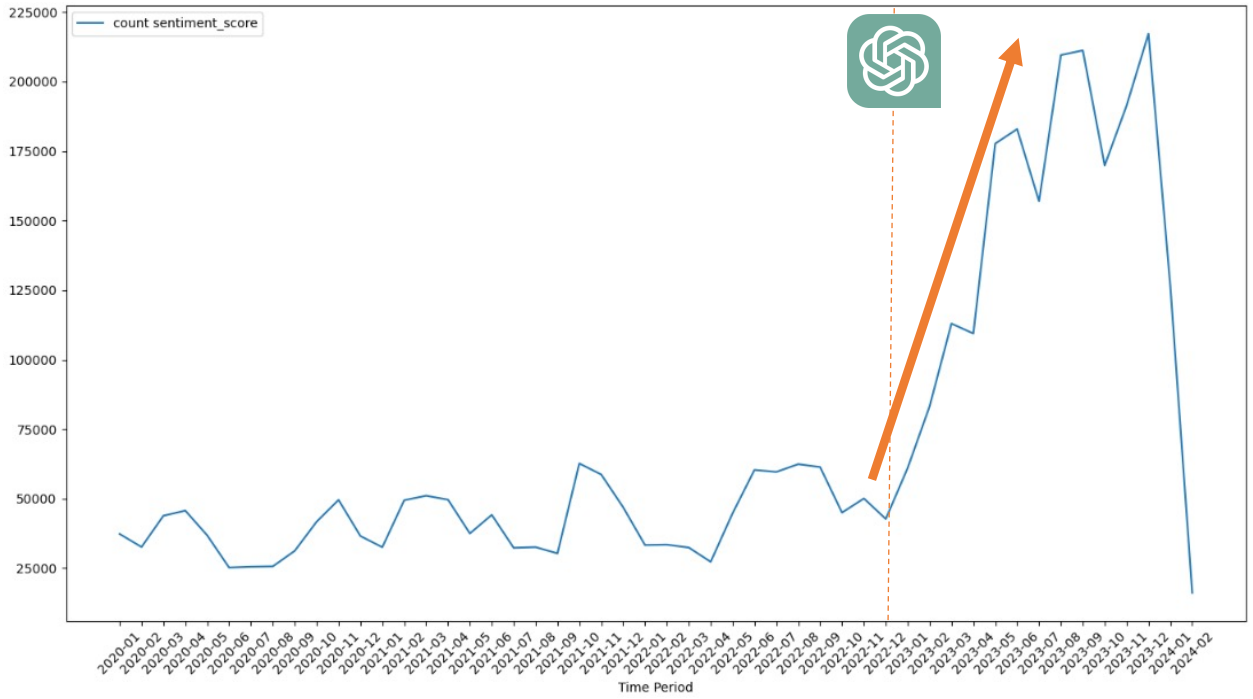
Distribution of Sentiment Score by VADER



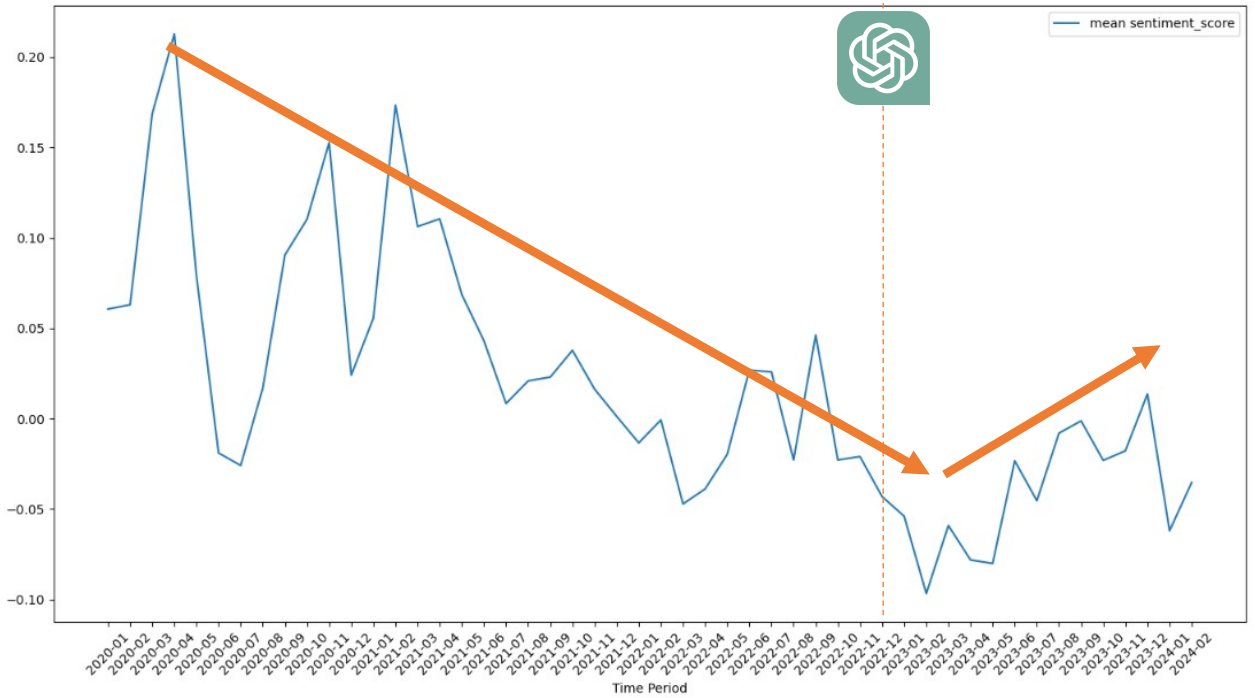
Overall Sentiment Trend

- The news has been on a sharp increase since the end of 2022 (the drop in February 2024 is simply due to fewer data being collected).
- The average sentiment score had been declining, but **there was a rebound starting from 2023**.
- These trends are expected to be closely related to **the release of ChatGPT** at the end of November 2022.

News(Sentence) Count



Average Sentiment Score

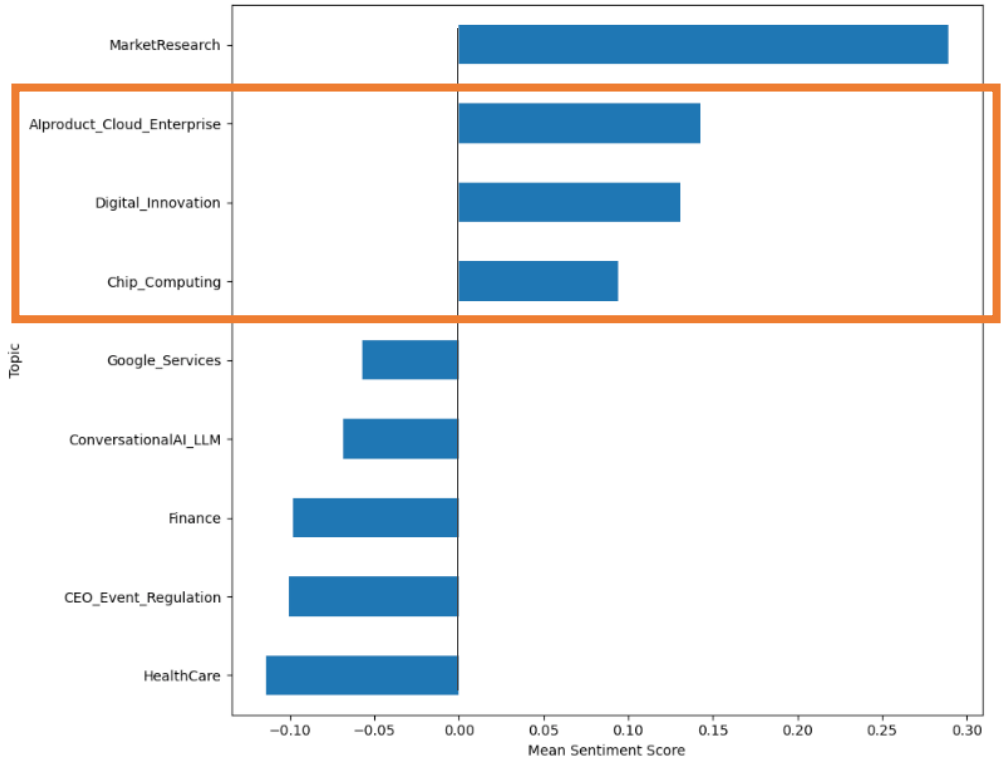


*When aggregating sentiment, the 'sum' is generally used. This approach is taken because the extent to which a topic is discussed in the news is considered important for measuring the overall market sentiment. However, relying solely on the sum may overlook entities or words with a low frequency of mention. In cases where insights from entities or words with fewer mentions are crucial, the 'average' is also sometimes employed. Unless specifically stated otherwise, the analysis presented in this slide is based on the 'sum'. Furthermore, due to the relatively higher number of positive scores overall, standardization of sentiment scores is conducted to make relative comparisons more straightforward.

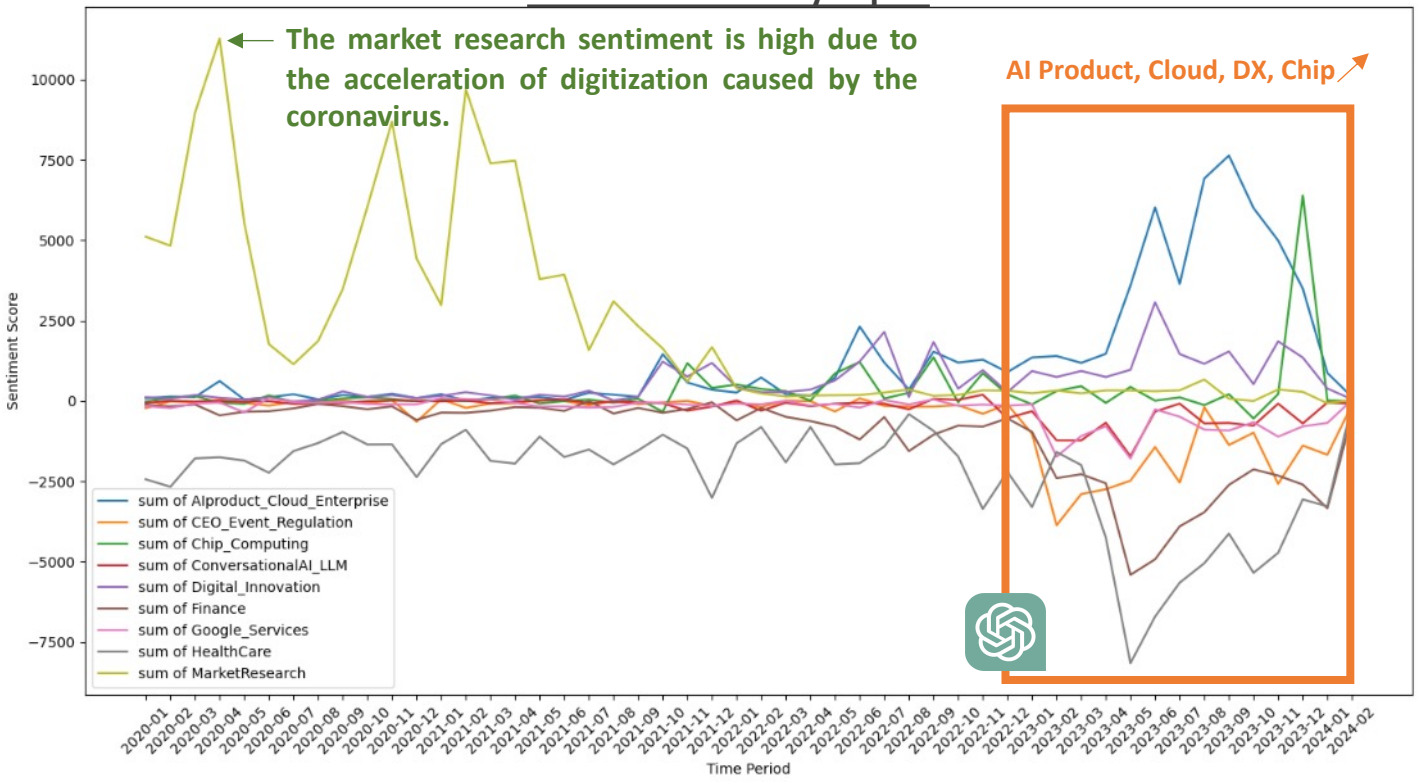
Sentiment Analysis by Topics

- **After appearance of ChatGPT, sentiments for topics such as AI Product_Cloud_Enterprise, Digital_Innovation, and Chip_Computing have surged dramatically.**
- On the other hand, the scores for **HealthCare and CEO_Event_Regulation have declined**. In healthcare-related contexts, the presence of words such as 'Cancer' or 'Kill' may lead to a tendency for the sentiment to become negative. Therefore, it is necessary to verify whether the sentiment towards healthcare themes themselves is negative.
- It is evident that **discussions related to regulation are prevalent in AI after the introduction of ChatGPT**.
- Regarding **Google_Services**, while there is a high regard for the cloud, the low evaluation of Bard could potentially be a headwind for sentiment (details to follow).

Average Sentiment Score by Topics



Sentiment Score by Topics



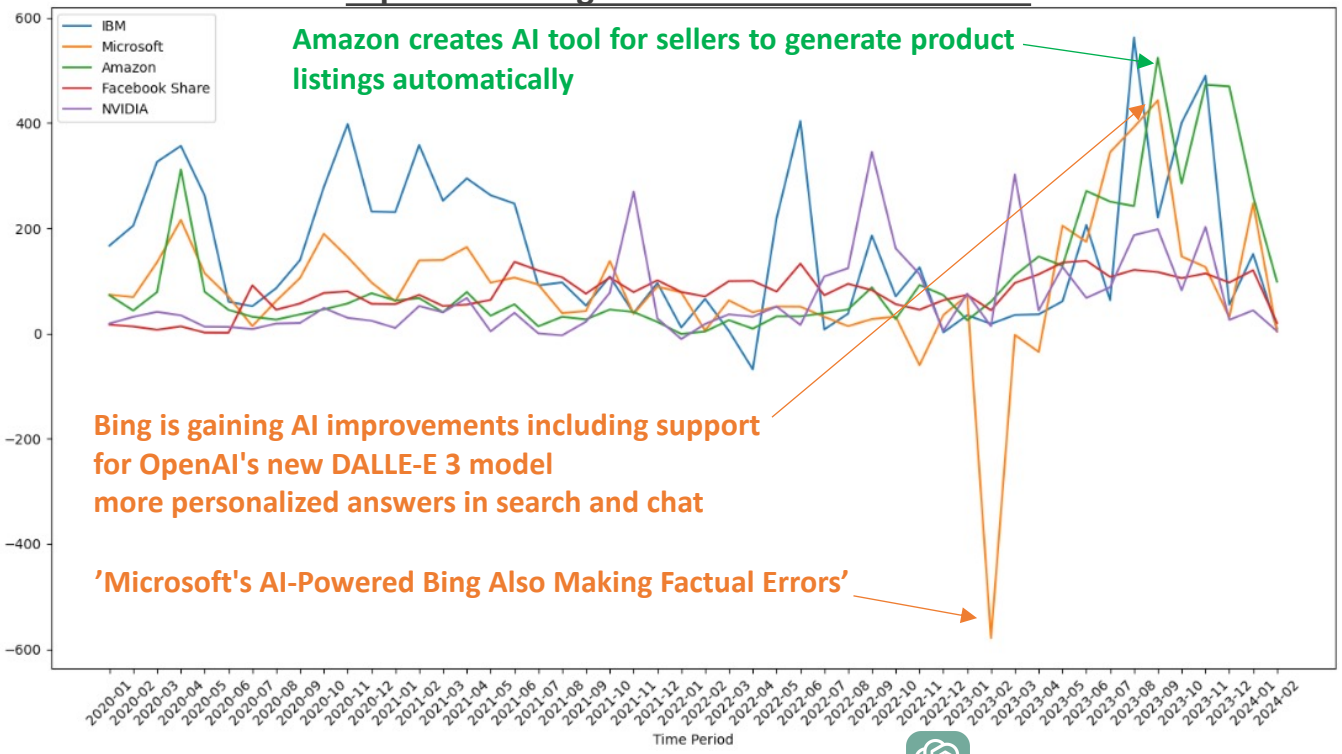
Sentiment Analysis by Organizations

- The top positive sentiment organizations include **IBM, Microsoft, Amazon, Facebook, and NVIDIA**.
- Looking at the scores of the top 5 over time, it is evident that concerns about **Microsoft's Bing** increased after the introduction of ChatGPT. However, the sentiment improved following news about Bing's enhancements.
- **Amazon** saw a rise in sentiment due to **the utilization of AI tools in its e-commerce services**.
- Regarding **Facebook**, there has been **no significant change in sentiment**, suggesting the possibility that it has not been able to ride the wave of GenAI.

Top Positive Sentiment Organizations

*The size of the text is related to the strength of the sentiment.

Top Positive Organizations - Sentiment Score



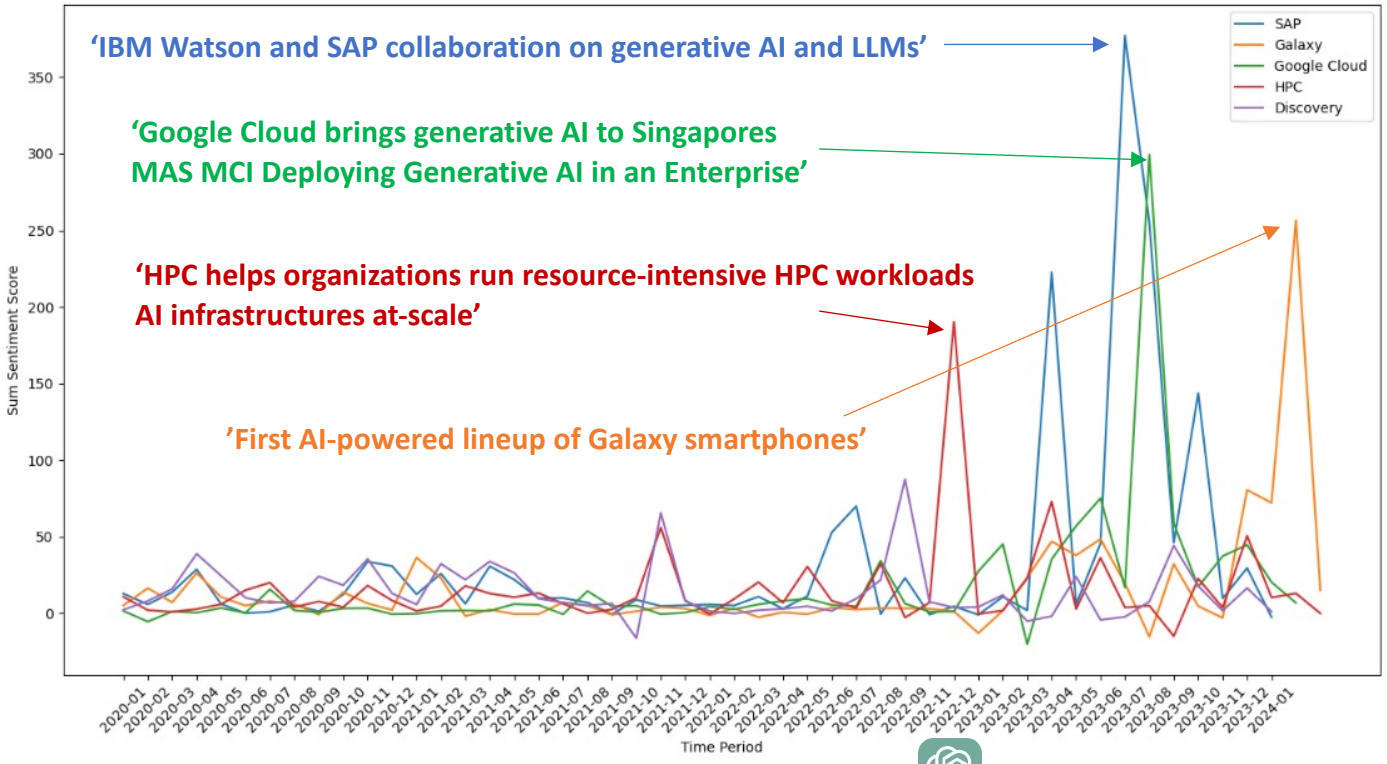
Sentiment Analysis by Products

- The top sentiments by product are **SAP, Galaxy, Google Cloud, HPC (High-performance computing),** and **Discovery (Vertex AI Search by Google)**. **Microsoft and Windows-related products** also ranked highly.
- **SAP** saw a rise in sentiment due to its **collaboration with IBM on GenAI and LLMs**.
- There was increased focus on **Galaxy AI** and other smartphone and tablet technologies.
- The sentiment towards **GCP** rose due to news of providing **GenAI-related solutions** to Singapore's MAS, MCI and other companies.
- **HPC gained attention as an essential computing resource for AI** around the time of ChatGPT's release.

Top Positive Sentiment Products

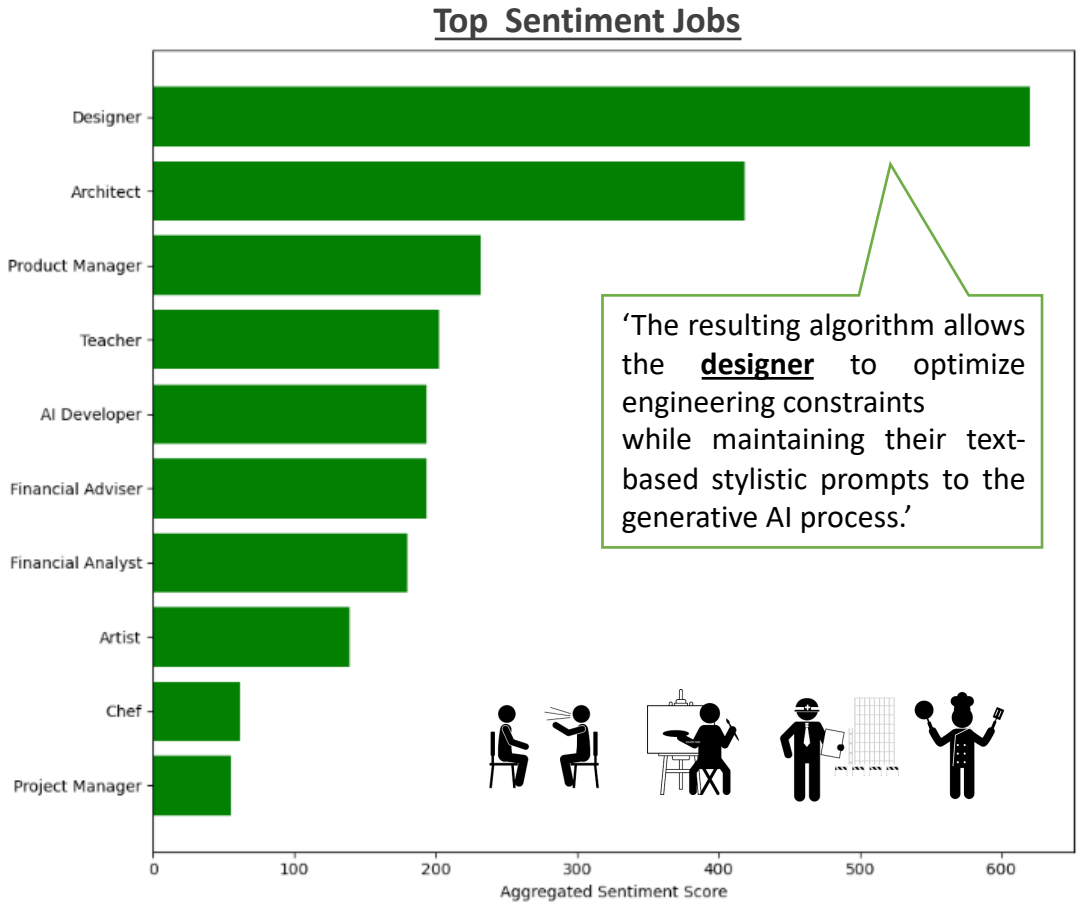
*The size of the text is related to the strength of the sentiment.

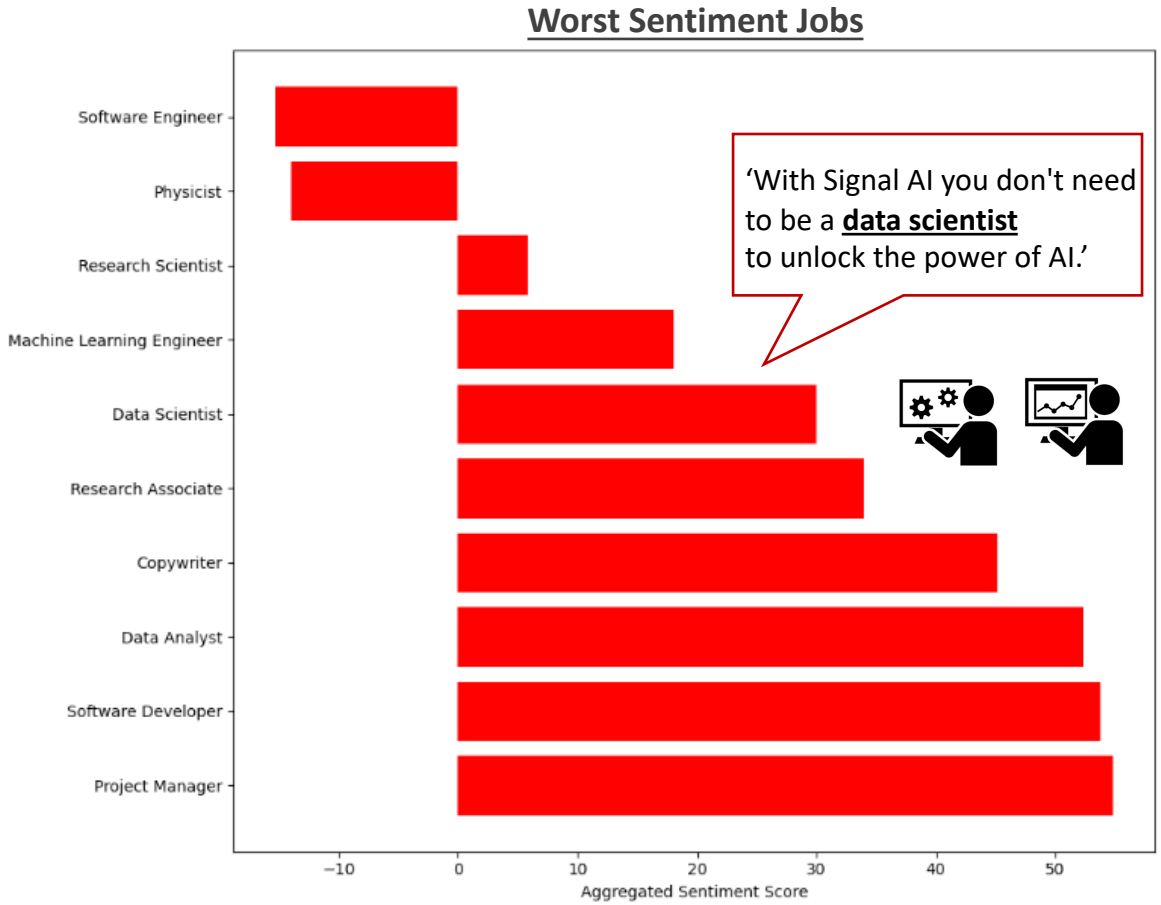
Top Positive Products - Sentiment Score



Sentiment Analysis by Jobs

- Jobs with the highest sentiment include roles that require **creativity, such as Designers, Artists and Chef** as well as positions involving **interaction with people, like Teachers, Financial Advisors, and Product Managers**. However, the development of Image GenAI could impact creative jobs in the future.
- Conversely, jobs with the worst sentiment are often **technical or code-based, such as Software Engineers, MLEs, and Data Scientists**. Interestingly, **AI Developer ranks high in sentiment, suggesting those who create AI could thrive.**
- As indicated by news examples, efficiently utilizing AI as a tool to enhance work and add value may turn individuals into assets **who can leverage AI, rather than being replaced by it.**

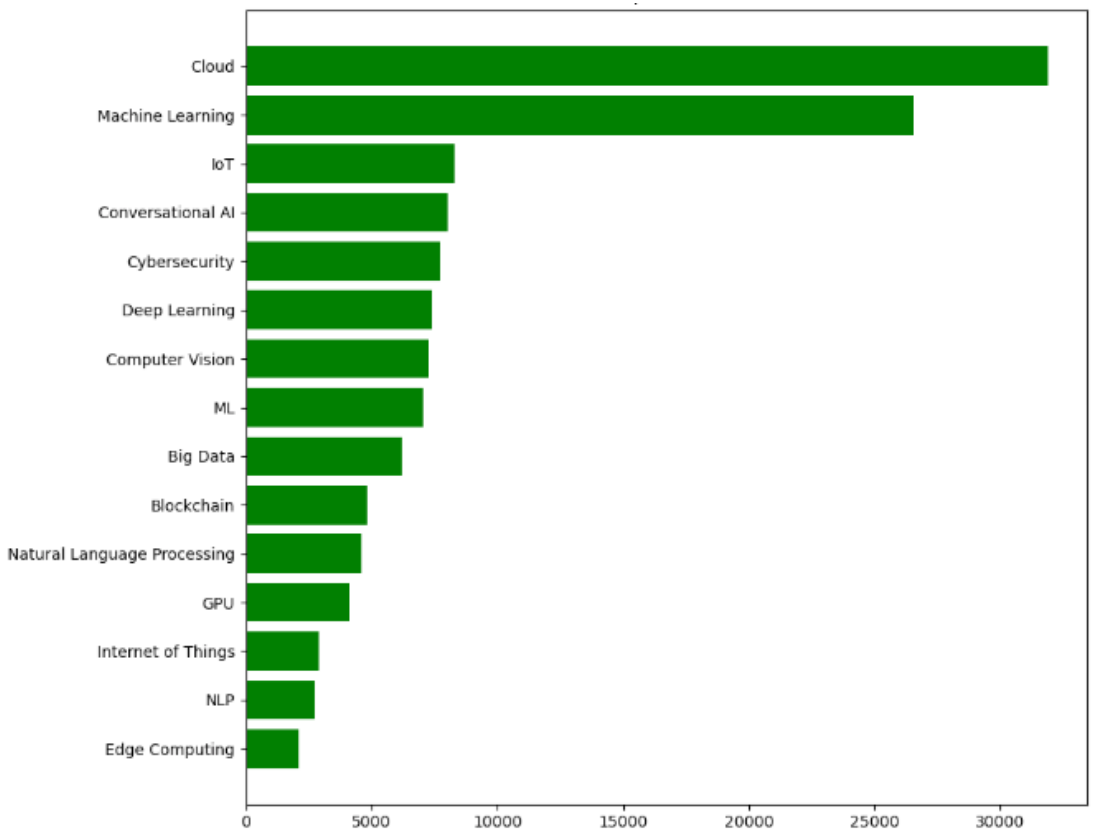




New Technologies

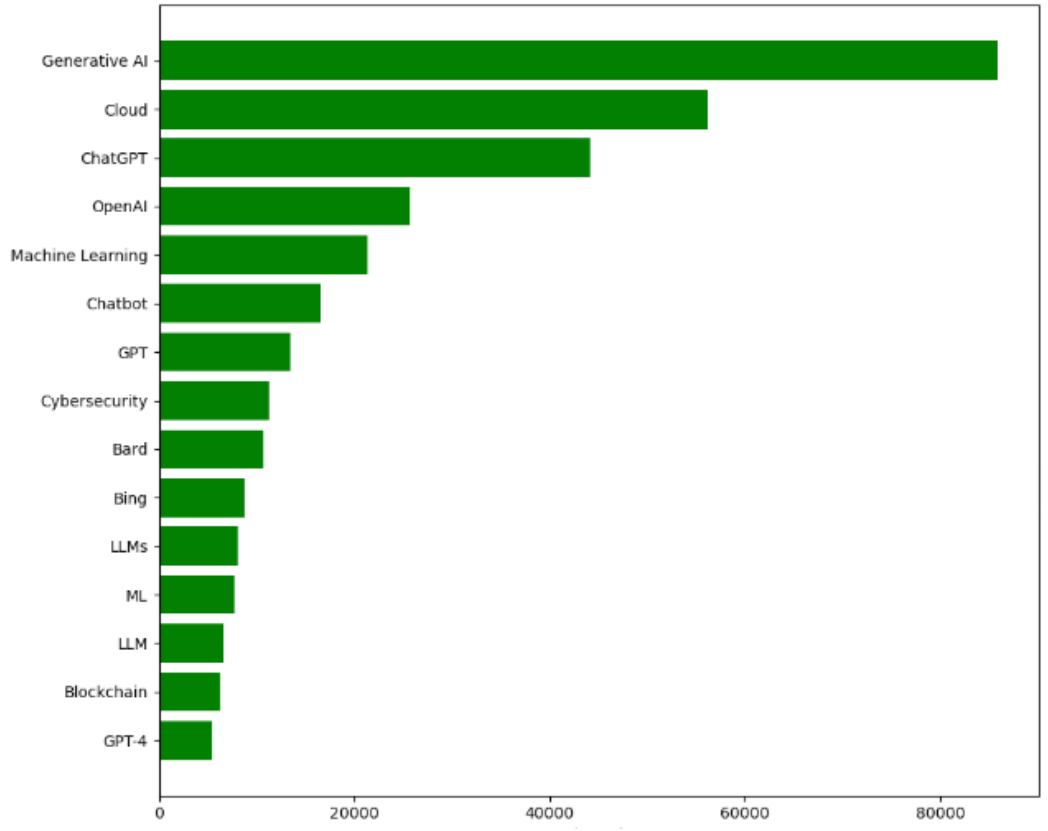
- Before the introduction of ChatGPT, topics such as **Cloud, IoT, Cybersecurity, Machine Learning, and Deep Learning** were at the forefront. There were mentions of Conversational AI, but not as many in terms of count.
- After the introduction of ChatGPT, there has been a significant increase in mentions of topics related to **Generative AI**, including **ChatGPT, Bard, Bing, and LLMs**, in the news. The fact that specific service names are ranking at the top is interesting.
- **Cloud and Cybersecurity** continues to receive high attention, suggesting that it remains an important technology.

Top Word Counts(Before ChatGPT)



Nov 2022

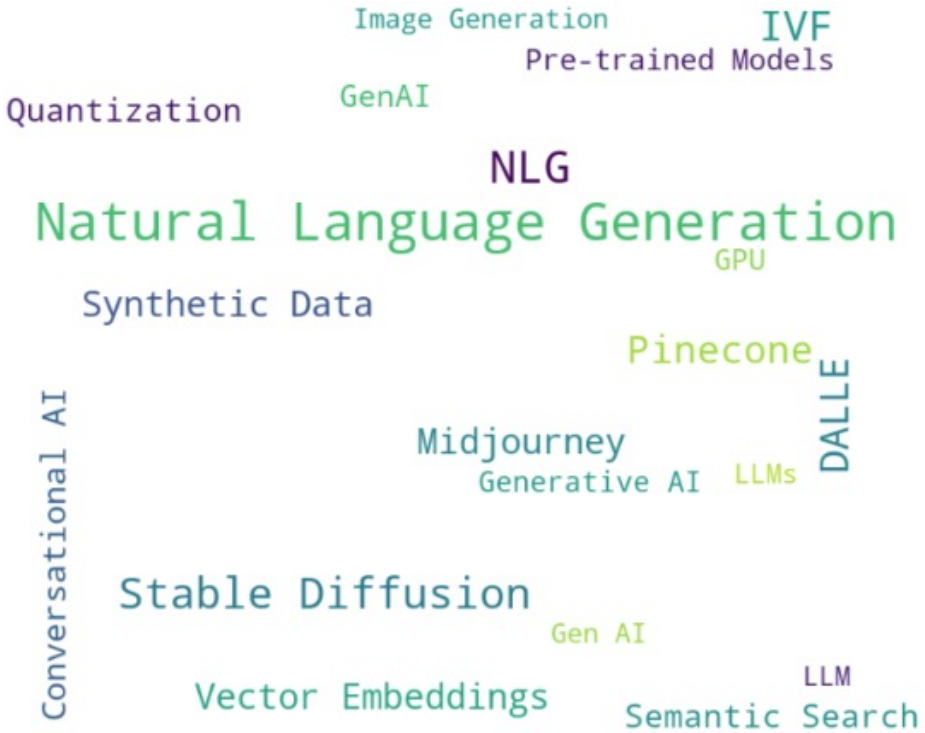
Top Word Counts(After ChatGPT)



Sentiment Analysis by New Technologies

- There are categories that showed a positive sentiment include **1. Text Generation** (NLG, LLMs, Conversational AI), **2. Image Generation** (DALL·E, Stable Diffusion, Midjourney), **3. Chip** (GPU), and **4. Search and Data Processing for GenAI** (Vector Embeddings, IVF, Semantic Search, Pinecone (Vector Database Service), Quantization).
- On the other hand, **Bard and Bing showed notable negative sentiments** due to its factual errors.

Top Positive Sentiment New Technologies (Average)



*The size of the text is related to the strength of the average sentiment.

Top Negative Sentiment New Technologies

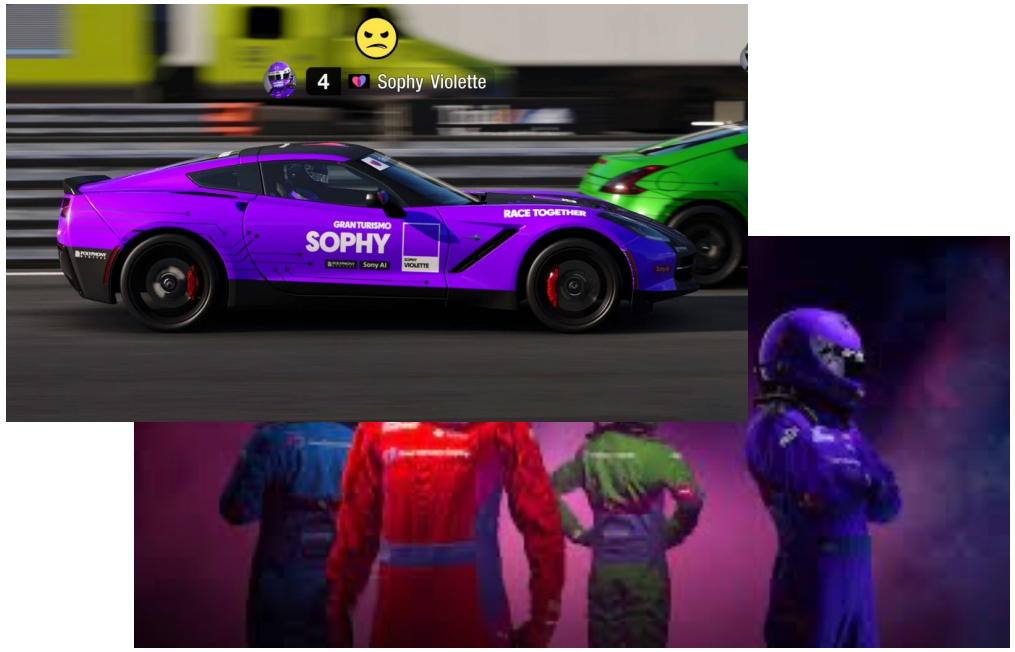
'After Google's Bard AI Microsoft's AI-Powered Bing Also Making Factual Errors.'

*The size of the text is related to the strength of the sentiment.

Other Insights from Sentiment Analysis

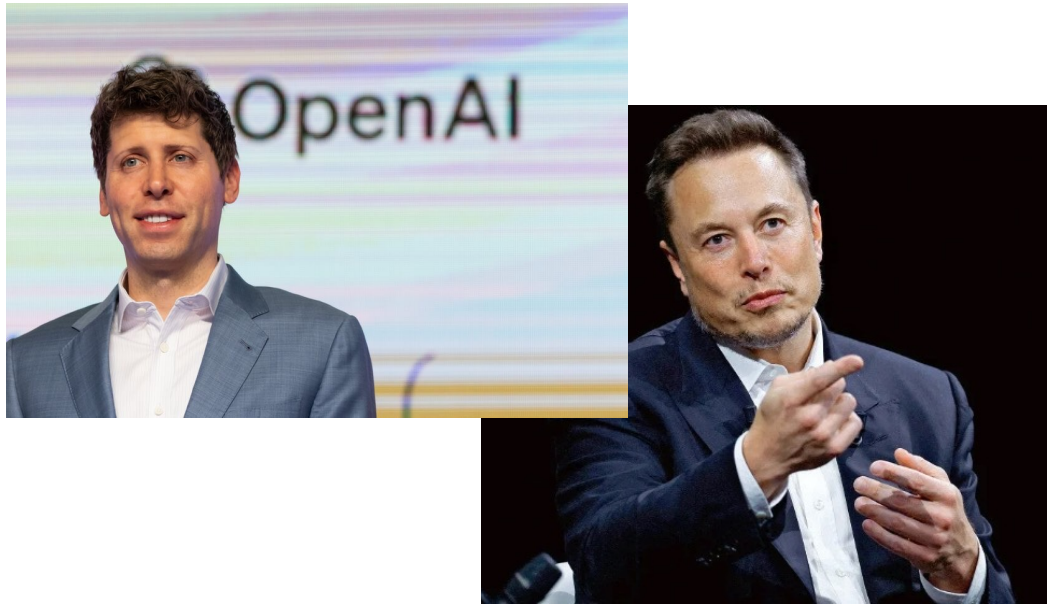
- The sentiment was high for Sony's **'Sophy', the Racing AI Agent** in PlayStation's 'Gran Turismo'. This indicates a growing interest in **AI within the gaming sector**.
- As for key figures in AI, **Sam Altman(OpenAI)** and **Elon Musk** are notable. The sentiment towards these individuals is **negative**, requiring close attention to discussions related to regulation and their statements.

Gran Turismo Sophy (Gaming AI)



Sophy: Sentiment Score approximately +800

Key Figures in AI



Sam Altman: Sentiment Score approximately -850
Elon Musk: Sentiment Score approximately -2,600

Summary

- With the emergence of ChatGPT, the spotlight on companies, technologies, and sentiments shifted dramatically, necessitating a catch-up to the 'new trends.' It is necessary to proceed with appropriate information gathering, capital investment, and technology implementation as described below.

Company	• **OpenAI, Microsoft, IBM, Google, Amazon and NVIDIA** have a significant influence on AI trends, and it's essential to pay attention to these entities.
Technology	• **Infrastructure (Keywords: Cloud, HPC, GPU, Cybersecurity):** These technologies is necessary for the GenAI era to process large and complex data efficiently. Cloud technologies also include GenAI functionalities. Besides, given the ongoing strong focus on personal information and cybersecurity, investments in security are also worth considering. • **Text Generation (Keywords: ChatGPT, LLMs, Conversational AI, NLG):** Capable of significantly reducing costs through high-quality customer service chatbot or data analysis via natural language prompts. • **Image Generation (Keywords: DALL·E, Stable Diffusion, Midjourney):** Utilizing these for creating marketing materials and advertisements could significantly improve efficiency. • **Search and Data Processing (Keywords: Vector Embeddings, IVF, Semantic Search)**: Useful for searching specific internal documents and deploying domain-specific chatbots, significantly improving operational efficiency.
Job	• Occupations that require **creativity (Designers, Artists, and Chefs)** and **interpersonal communication (Teachers, Financial Advisors, and Product Managers)** are less likely to be replaced by AI. However, the development of Image GenAI could impact creative jobs. • On the other hand, **technical or code-based jobs (Software Engineers, MLEs, and Data Scientists)** could be replaced, and from a worker's perspective, it's crucial to become an AI Developer or leverage AI for additional value creation. From a management perspective, replacing these high-cost talents with AI technology should be considered.
Industry/ Person	• The progression of AI affects a wide range of industries, with increasing attention in **gaming** recently. It's crucial to continually gather the latest information and consider how it can enhance existing businesses. • Attention should be paid to individuals like **Sam Altman** and **Elon Musk**, and particularly to events related to **regulation**.