Impact of Generative AI

Natural Language Processing and Cognitive Computing

May 2023, Chicago IL
Submitted by Kshitij Mittal
The University of Chicago

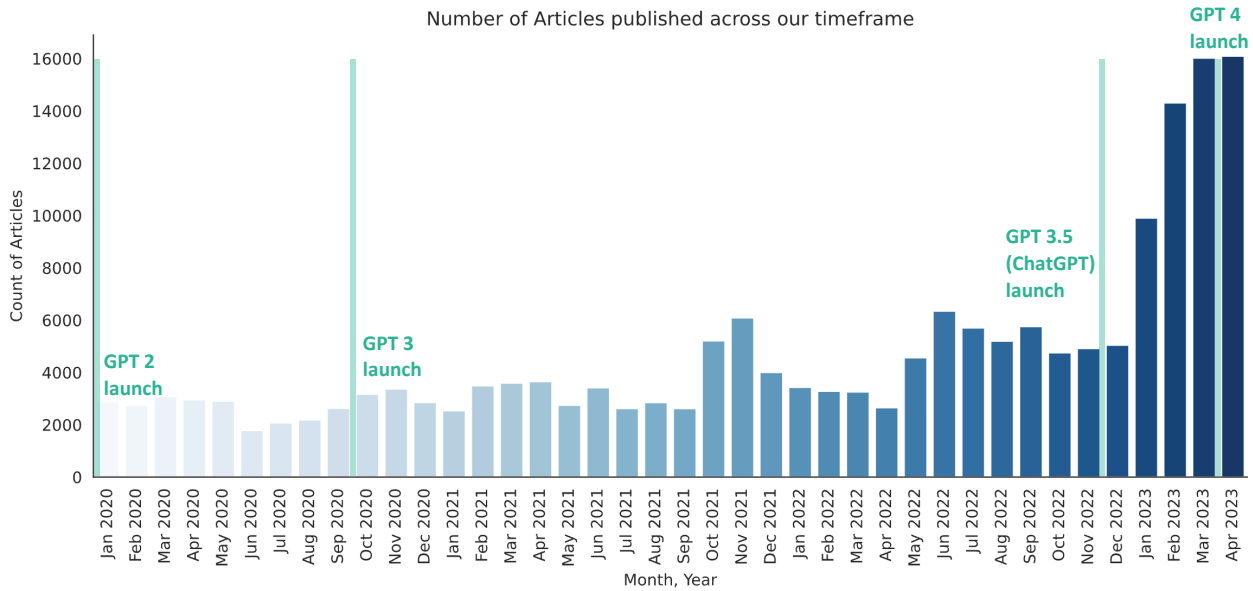
Executive Summary

AI as a field is has been going through tremendous developments in the past 5 years. New technologies and use cases are being rolled-out at an unparalleled speed. AI is rapidly embedding into every aspect of our lives, whether it is deciding one's stock portfolio, or the next AI enhanced picture someone wants to upload on Instagram.

Generative AI as a use-case has been in works ever since AI research emerged post the AI winter in the 1970s and 80s. Advancements in parallelized computing, deep learning and natural language processing have been propelling Generative AI throughout the last decade.

A watershed moment in the history of Generative AI was **November 30, 2022**. **OpenAI**, with its series of successful generative models like GPT1, GPT2 and GPT3, for the first time released a consumer facing product: **ChatGPT**
It marked as the starting point for the new **'AI Wars'.**

As with any new technological development, there is an associated **Hype Cycle**.
As a part of this analysis, we analyze if Generative AI can cross this initial hype and investigate **what types of tasks and jobs are likely to see the biggest impact**



200K news articles related to AI are collected between Jan 2020 and April 2023. While we see a positive trend across the first 3 years, there is a surreal increase in articles for 2023.

Executive Summary

Key Findings and insights

- Recent large language models are being fine-tuned and evolved for a plethora of use-cases:
  - Healthcare (can impact para-meds)
  - Financial Industry (can impact quant analysts)
  - Automotive sector (can impact CAD designers)
- New use cases which were previously un-touched by AI are also evolving:
  - AI in Legal systems (can impact paralegals)
  - AI for music generation (can impact composers)
- This generation of AI is faced with various challenges as well:
  - Regulatory concerns
  - Issues of Deepfakes in Image generation
  - AI for cryptocurrencies is facing pressure from the current downtrend
  - Layoffs and underwhelming Big Tech performance can dampen AI progress
- Industries like Agriculture and Ed-tech are averse to current Generative AI capabilities (as observed from the data)
- The current hype cycle is leading to a big market shake-up, especially among the MAANG companies:
  - **Microsoft**, for the first time in 2 decades, overtook **Google** and other companies owing to it's GenAI integrations
  - **OpenAI** propelled in the past 2 years, and **ChatGPT** has been one of the fastest adopted products in Tech history
  - **Google's** Bard received a huge negative connotation from the market, putting some shade on it's GenAI capabilities
  - **IBM, Samsung** are losing relevance in GenAI

Different fractions of the economy are reacting differently to Generative AI

Article Clean-up

200,322

News Articles related to AI are present in our corpus

- All articles were in English
- There were no Null rows in the dataset

TIME RANGE
January **2020**
March **2023**

Clean-up

1 Articles Text was cleaned to remove:
- Mentions and URLs
- New Line (/n), Tab Spaces (/t), Carriage Return (/r), Hashtags (#)
- Uncanny long words (>15 characters)

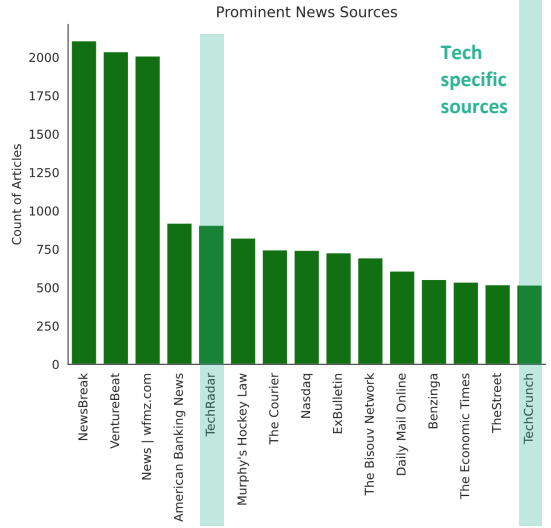
2 After manual inspection, remnants of Web Crawl were removed separately. Words like: AdBlock, Refresh, page, Search, Login, Register, Cookies, All Rights Reserved etc.

3 Post initial topic modeling, we were seeing a topic solely composed of country names and months. These were remnants of web crawl, and were removed from the text

4 Articles Titles were also cleaned with the above steps. News Source was stripped away from Titles

5 News Titles and Article texts were lower cased, stripped off their punctuations and non-alphabetical characters, and then tokenized using Gensim

6 However, both original text (cased, with punctuations) and tokens were retained for the entirety of proceeding analysis

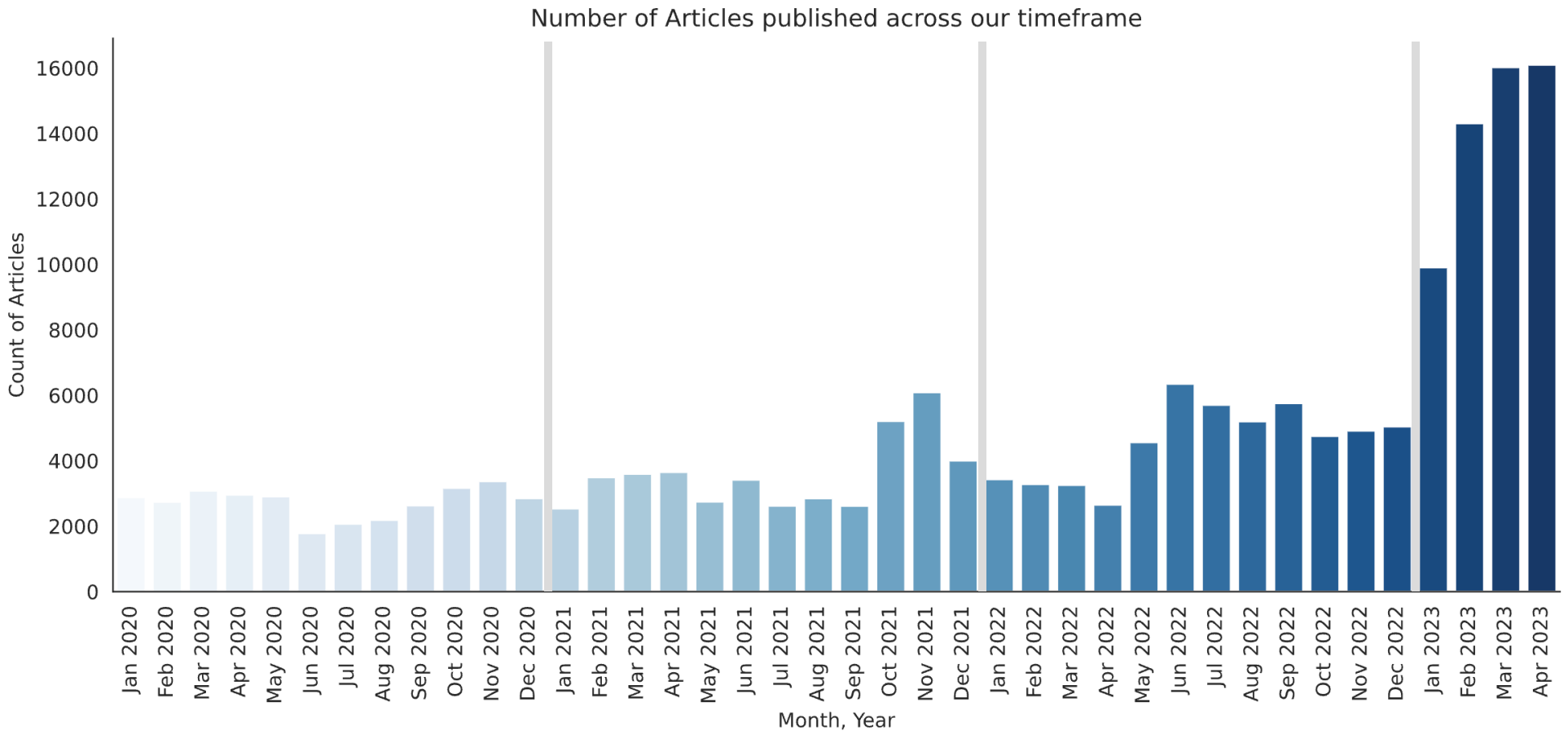


New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

Topic Modeling

Understanding the time dimension

The articles were collected between Jan 2020 and April 2023. While we see a positive trend across the first 3 years, there is a surreal increase in articles for 2023. This played a big role in understanding topics and sentiments prevalent in the market about AI.



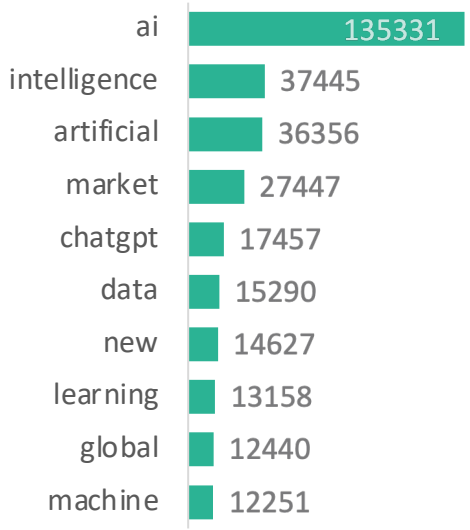
New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

Article Filtering

200,322 News Articles → **198,373** News Articles → **194,894** News Articles → **186,815** News Articles

On an initial view, most articles were revolving around 'ai', 'artificial intelligence', 'machine learning'.

Top 10 tokens from the entire corpus



DUPLICATION ANALYSIS

for URL, Titles, and Article Texts

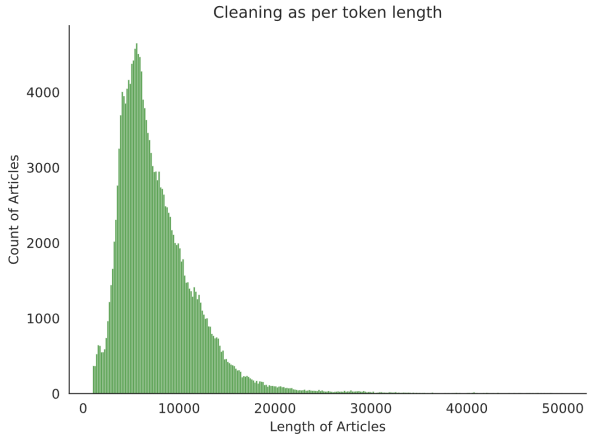
- 0 duplicates were identified in URL
- ~80K articles had duplicate titles – These could have been the same articles being reported by different news outlets
- ~2000 articles had the same text composition. These were flagged and removed from the analysis

RELEVANT TOKENS

Articles containing Tokens like artificial, intelligence, machine, ai, ml, data, analytics, gpt were filtered in

TOKEN LENGTH

Very small articles (<1000 tokens) and very big articles (>40000 tokens) were filtered out



KEYWORD ANALYSIS

RAKE was implemented to extract major keywords from each article.

Article keywords which include words pertaining to Artificial Intelligence were filtered in

Article Filtering was an iterative process, and continued during Topic Modeling and Sentiment Analysis

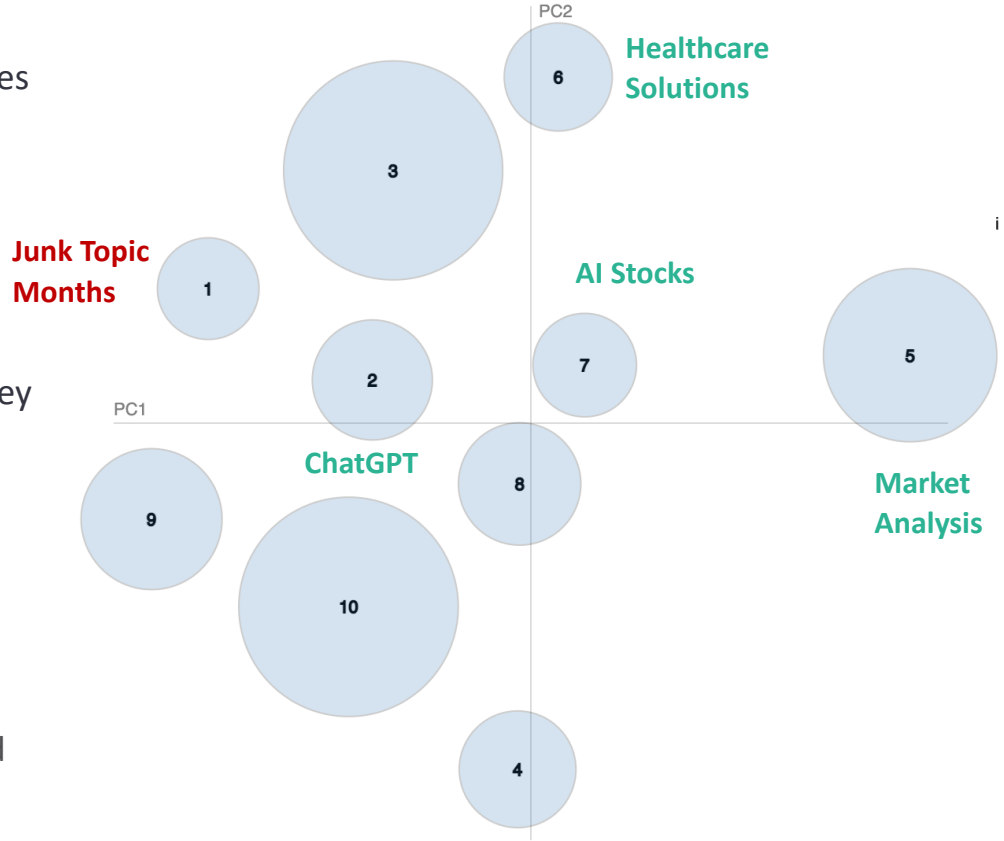
Topic Modeling

Topic Modeling was conducted in 3 phases to extract the most insights from our corpus:

1. **LDA on the entire corpus**
2. **KTrain to find year-wise Topics**
1. **POST Sentiment Analysis – BERTopics**

LDA

- Multicore LDA was conducted on a sample of 5000 articles
- After conducting tuning and manually inspecting the topics, following hyperparameters yielded the best coherence:
  - Topics: 10
  - Alpha: 0.51
  - Beta: 0.01
- More number of topics were experimented upon, but they were losing coherence beyond 18 topics
- LDA gave the following starter topics. However, these topics were not clean
  - Market Analysis
  - Healthcare
  - AI Stocks
  - ChatGPT
- Interestingly, it also pointed a topic which just comprised of months, and were adding no intrinsic value to the analysis. These words were cleaned out during text cleaning



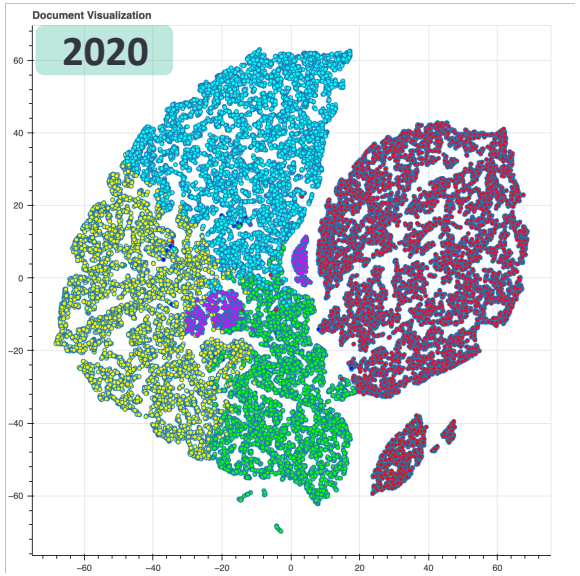
New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

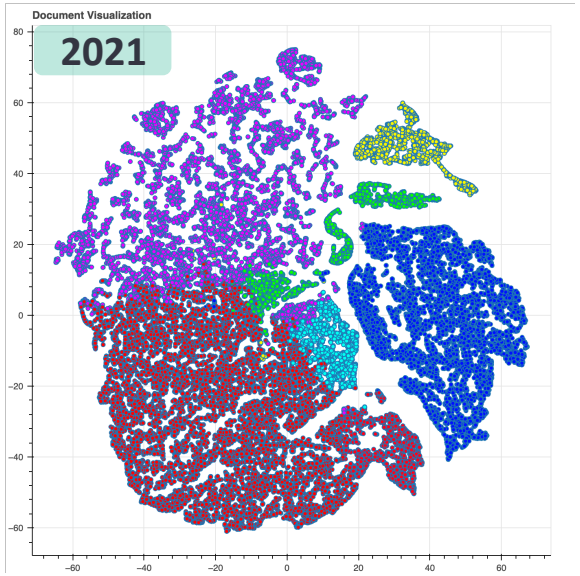
Topic Modeling

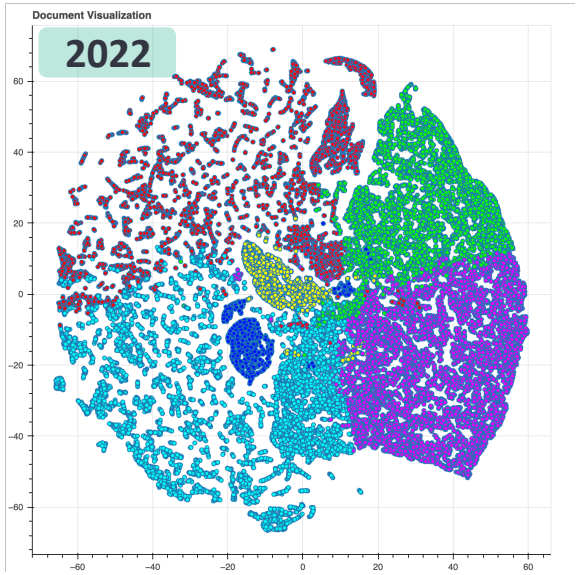
K-Train

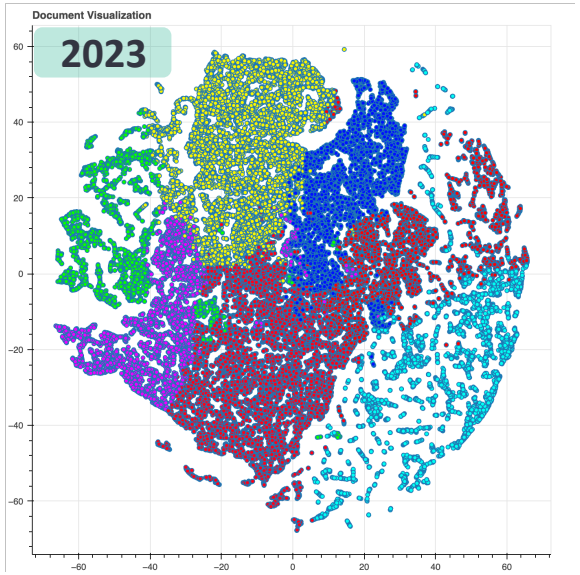
- Seeing unsatisfactory results from LDA, the dataset was divided as per years to understand the implications of timeline on AI topics
- 7-10 topics were investigated for each year separately. The following topics were common to most years:
  - Industry Automation
  - AI Research
  - Market Research and analysis
  - Consumer Applications
  - Financial Markets
- For the year of 2023, we were seeing two extra topics around Conversational AI
- Running on a similar LDA core, K-Train was unable to identify highly nuanced topics and had a lot of noise
- However, it chunked all the topics irrelevant to AI, but containing similar words together:
  - Air India (An Indian airlines division) uses 'AI XXX' as their flight codes.
  - Ai Weiwei is a Chinese contemporary artist, who was present in our data but without relevant articles

These insights were critical for identifying noisy articles, which were later flagged









Sentiment Analysis

Multiple approaches were utilized for conducting Sentiment Analysis for News Article Texts

1. A custom binary classifier was trained on Yelp reviews dataset.

SVM, Naïve Bayes, and Logistic Regression classifiers were implemented, and then extrapolated on news article texts.

RESULTS

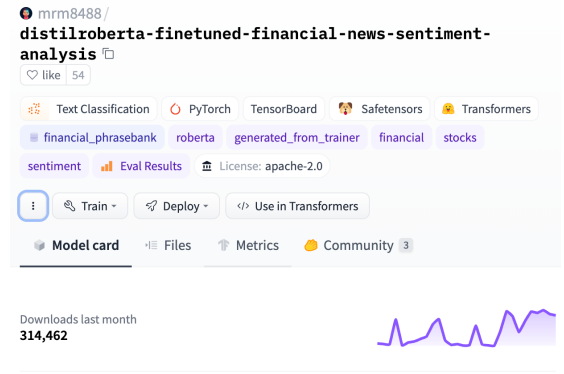
- Binary Classifiers were trained only to identify positive sentiments. When concerned with news, we have a big majority of articles which are just factful, and do not pertain to a particular sentiment. Binary classifiers are unable to tease out such neutral articles
- 95% articles were being put in '0' – Non-positive category

2. A Hugging Face Transformer fine-tuned on news sentiments was employed

distilRoberta-financial-sentiment is a fine-tuned version of distilroberta-base on the financial_phrasebank dataset.

RESULTS

- A pipeline was created to provide individual Positive, Negative, and Neutral sentiment probabilities for each news article
- The transformer was able to identify positive sentiments accurately, but was extremely strict for negative sentiments. Capture for negative sentiments was increased by reducing the probability threshold of negative percentage to 10%



New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

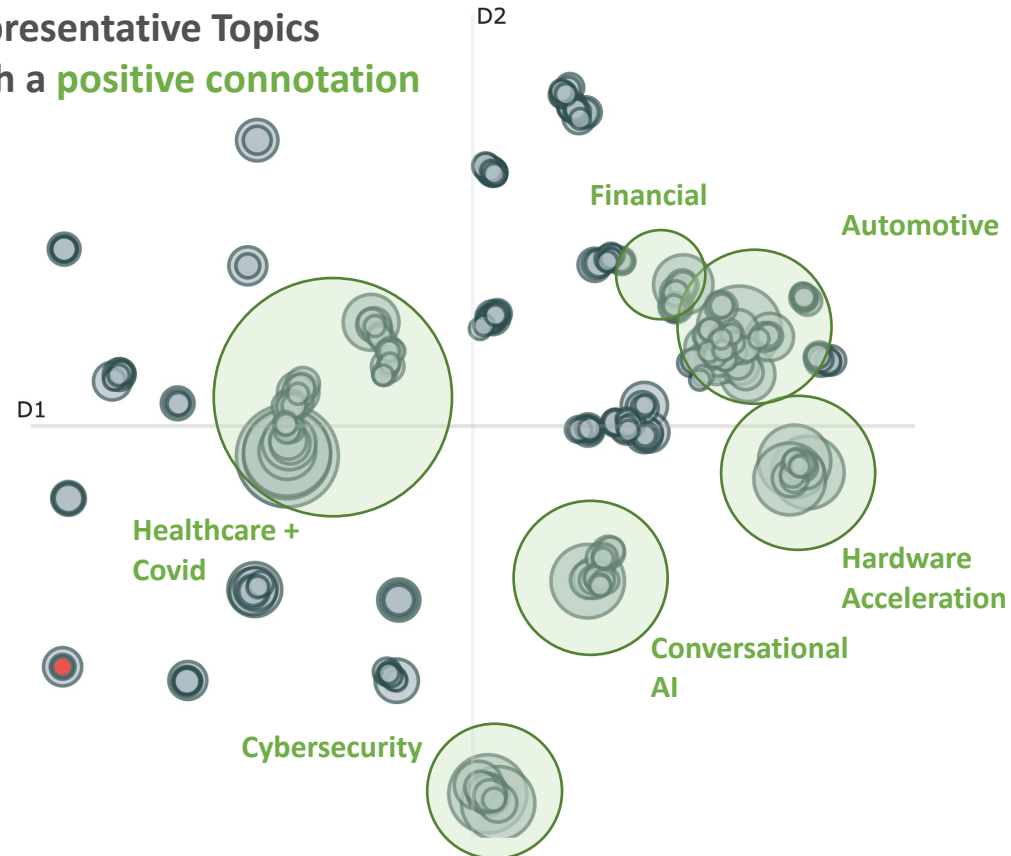
Topic Modeling (post Sentiment Analysis)

BERTopic

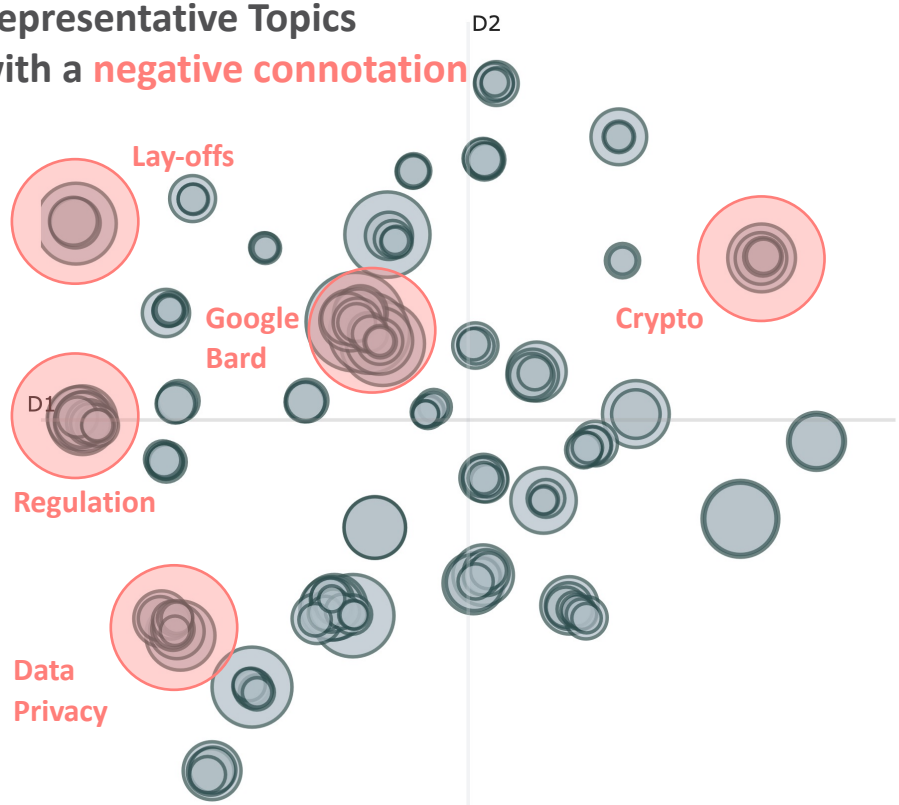
BERTopic was used to analyze key topics relevant to positive and negative sentiments

- It embedded each document in the input text corpus using the pre-trained BERT model
- It then applied HDBSCAN to group similar news articles based on their vector representations
- It identified representative keywords for each cluster by analyzing TF-IDF
- With tuned hyperparameters, it identified ~300 positive and ~150 negative topics
- After manual inspection, these topic counts were reduced to 60 and 20 respectively

Representative Topics with a positive connotation



Representative Topics with a negative connotation



New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

Why are some data science applications succeeding?

Text summarization was applied on topics anticipating new developments in the Generative AI space

Healthcare

The global AI in drug discovery market, valued at USD 260 million in 2019, is projected to experience a robust growth rate of over 40.7% during the forecast period of 2019-2026. This market offers significant opportunities driven by the increasing demand, business strategies, and advancements in AI technology, while also facing upcoming challenges in its implementation and ethical considerations.

Hardware Acceleration

Nvidia and Microsoft are joining forces to develop a powerful AI supercomputer with extensive capabilities. This collaboration aims to create an infrastructure that combines Nvidia's expertise in AI hardware and software with Microsoft's cloud computing capabilities. The partnership holds the potential to drive advancements in AI research and development, enabling breakthroughs in various fields.

Automotive

Invisible AI, a computer vision platform provider, has successfully secured $15 million in a Series A funding round to accelerate the expansion of its innovative technology across car manufacturing facilities. The funding will enable Invisible AI to scale its computer vision platform, which offers advanced capabilities for enhancing efficiency and productivity in manufacturing processes. The investment reflects the growing recognition of computer vision's potential to revolutionize industrial operations and drive further advancements in the manufacturing sector.

Financial

FundMore, a prominent AI innovator, has formed a partnership with M3 Financial Group, a major player in Canada's mortgage industry, to expedite mortgage approval processes for Canadians. This collaboration aims to leverage AI technology to streamline and enhance the mortgage approval experience, making it faster and more accessible to all individuals. The partnership signifies the increasing adoption of AI in the financial sector, specifically in mortgage origination, to improve efficiency and customer satisfaction.

New Chat

Executive Summary

Actionable Recommendations

Article Clean-up and Filtering

Topic Detection

Sentiment Analysis

Sentiment Over Time

Entity Identification

Targeted Entity Sentiment

Upgrade to Plus NEW

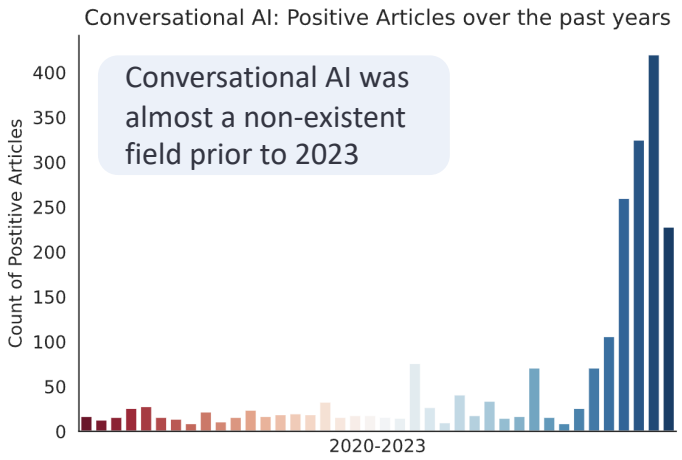
Kshitij Mittal

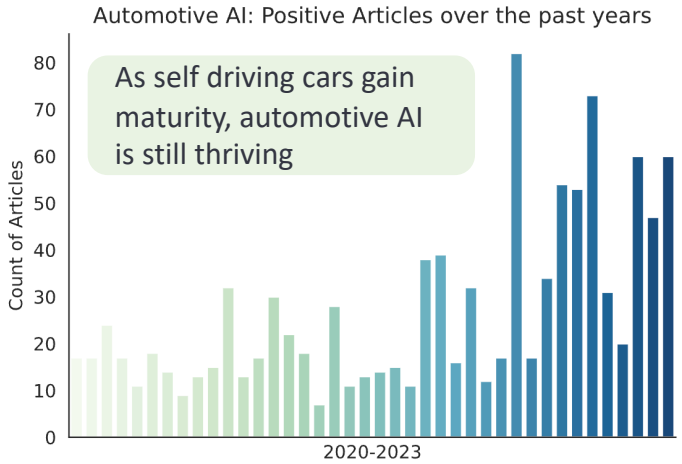
Sentiment Analysis over time

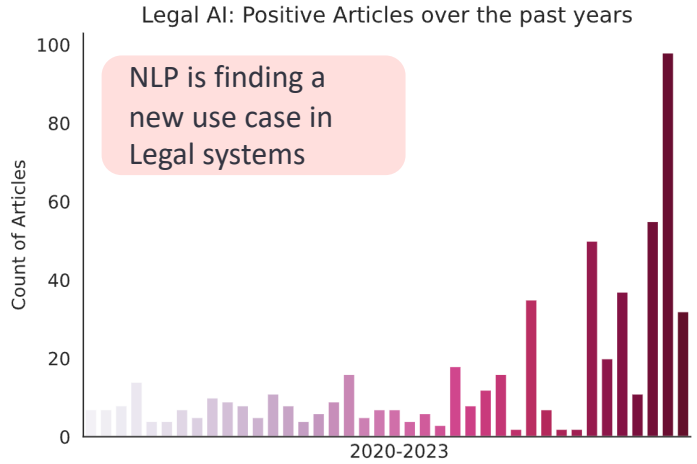
Positive Sentiments for the space far outweigh the negative sentiments. We see consistently high number for sentiments supporting overall space

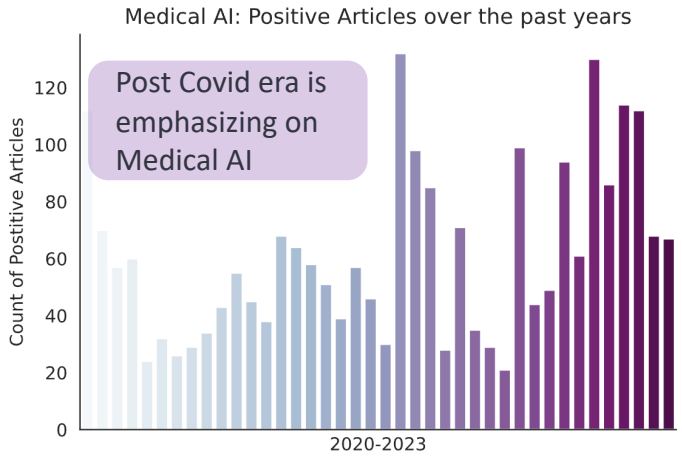
What new AI solutions will be affecting Data Science?

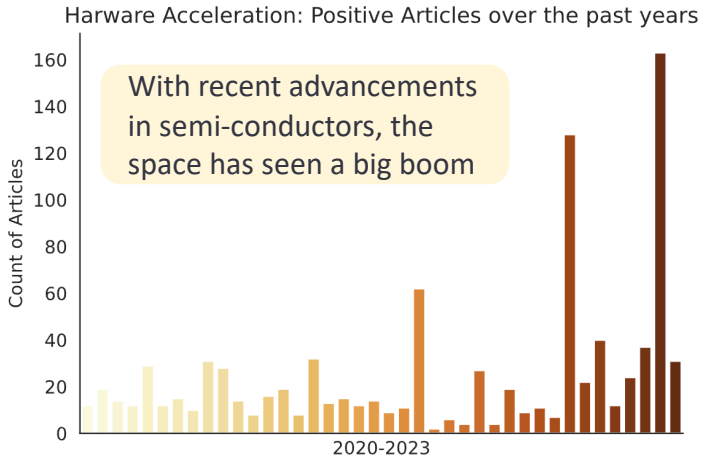
We see a lot of excitement for some data science applications:

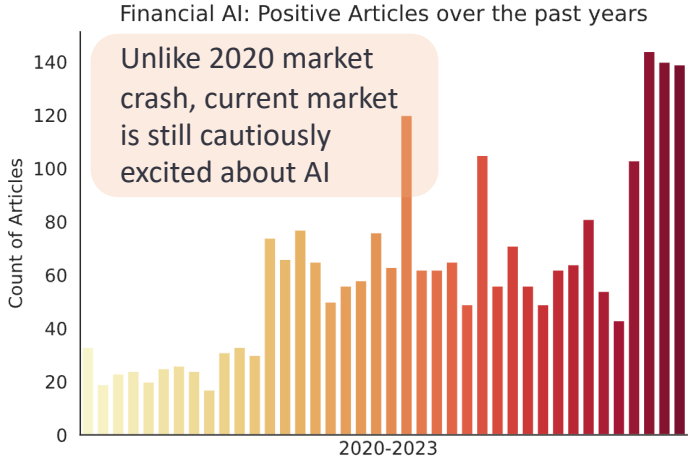








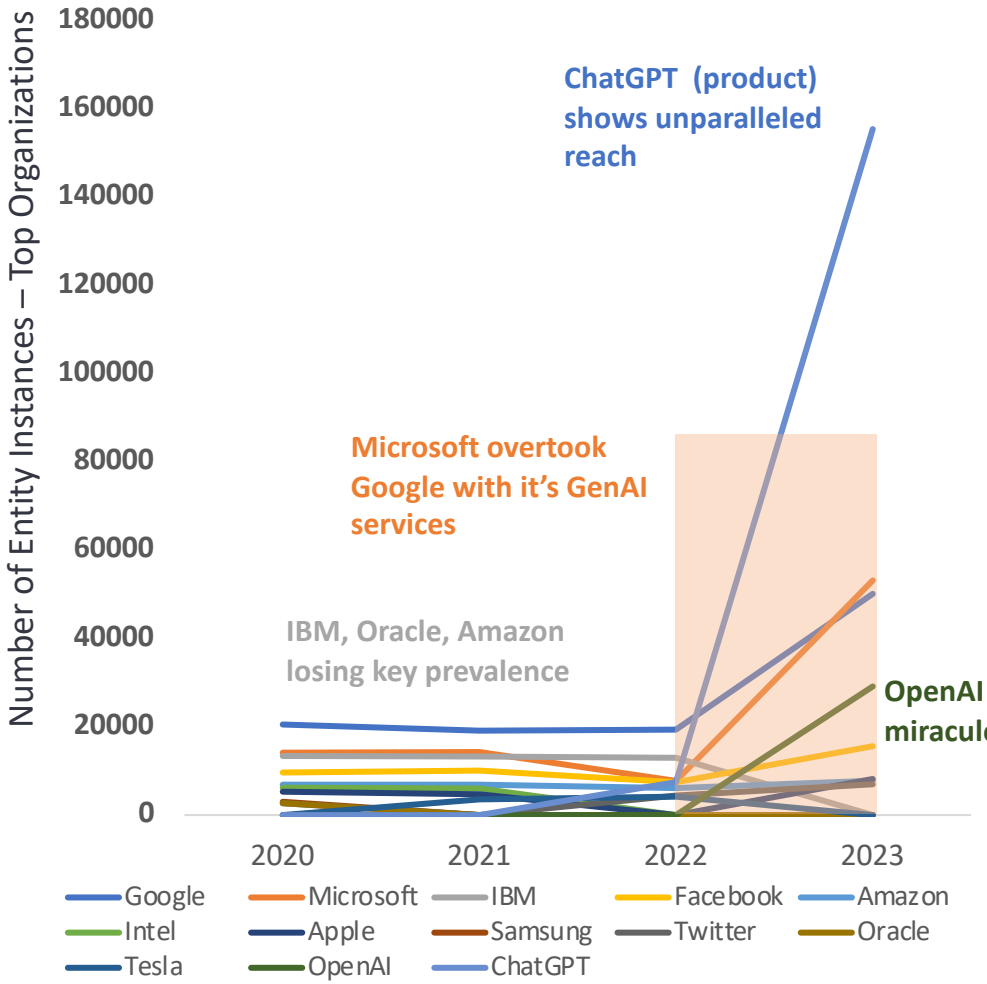


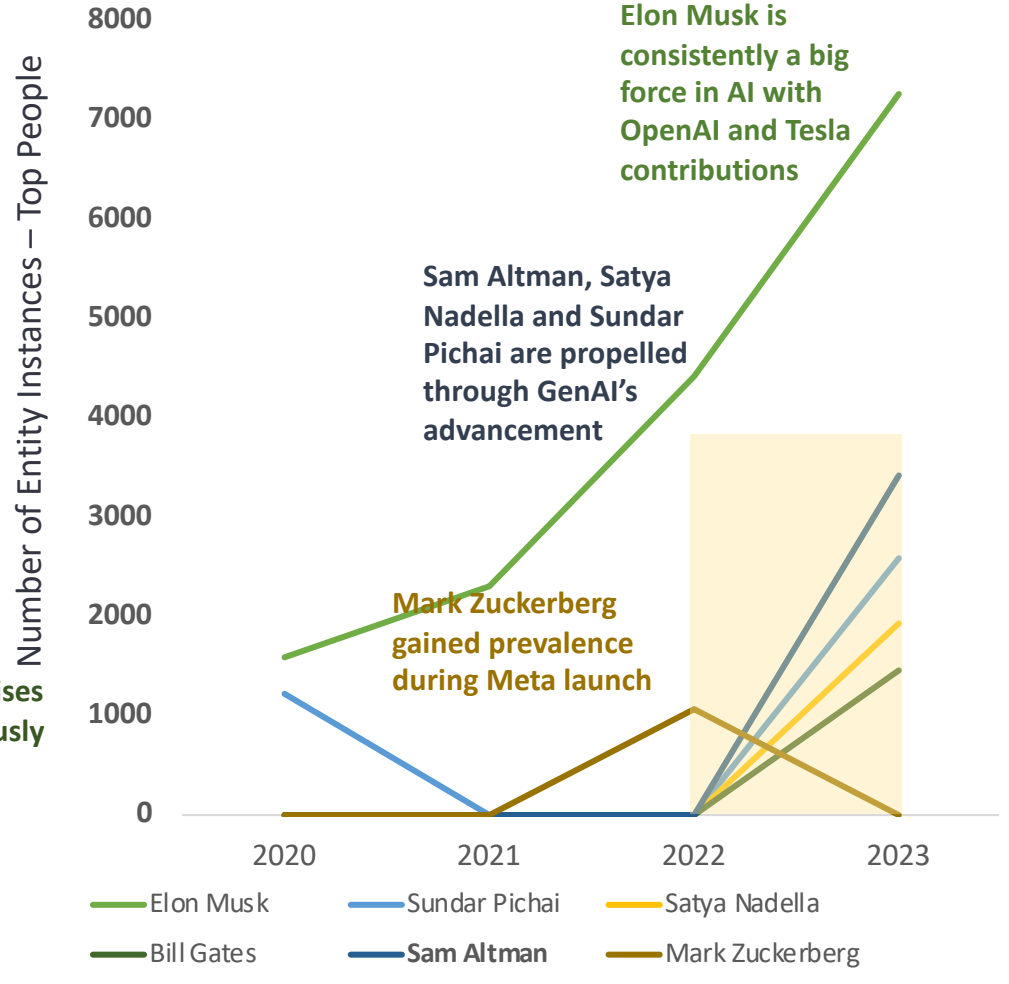


New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

Organizations, products and people popular in the space are changing every year

Spacy NER was applied to extract the following entities for each year: Organization, Persons, Products, National Groups

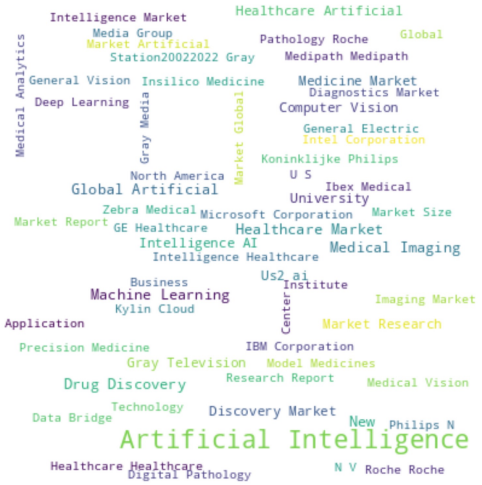




New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

The following lines of businesses should invest more in Data Science initiatives to propel their current success

Healthcare



- Companies like **Roche, Bayer, IBM** are doing significantly work to accelerate medical AI
- **Drug Discovery** and **Diagnostics** are critical avenues
- **Europe** can enhance its ecosystems for Medical AI research

Hardware

- **Nvidia** leads the chipset market, with its **CEO Jensen Huang** being a key voice
- Nvidia products like **DGX, HGX and V100** are becoming the new industry norms for AI applications
- Companies like **Intel, Oracle and Red Hat** and trailing in the space

Legal

- Startups like **CaseText**, with their product **CoCounsel** have been creating a lot of buzzword
- **Deloitte**, with its extensive consultancy background, can also pushing forward in the Legal AI space

Music

- **Verbit** and **Stability AI** are leading the generative music space
- **TikTok (ByteDance)** can enhance it's product by generating personalized soundtracks
- However, there is a major concern around **Pro Music Rights**

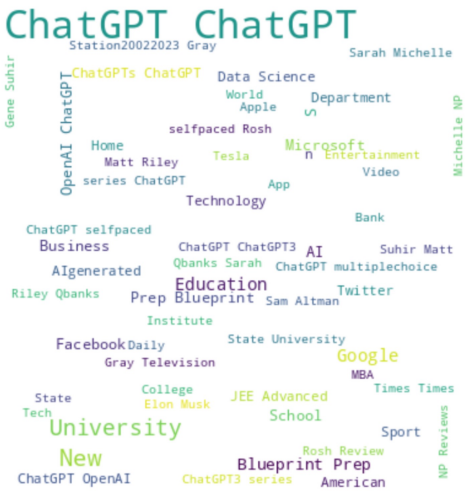
New Chat
Executive Summary
Actionable Recommendations
Article Clean-up and Filtering
Topic Detection
Sentiment Analysis
Sentiment Over Time
Entity Identification
Targeted Entity Sentiment
Upgrade to Plus NEW
Kshitij Mittal

Some sectors are not highly anticipating new AI advancements

Agriculture

- Few companies like **Rockwell Automation, GREEFA** and Honeywell are working towards Agriculture automation
- However, AI advancements have not found optimum use case in Agriculture

EdTech



- There is a lot of hesitancy in the market for **Generative AI being used in qualifying exams (LSATs, JEEs, GREs, GMATs)**
- Edtech companies like **Blueprint Prep** and **Posh Review** have not met with recent advancements well

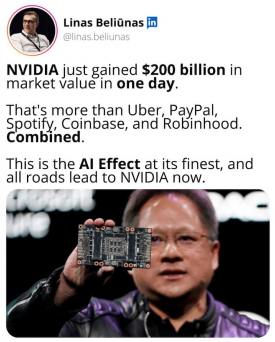
Image Generation

- While there has been a lot of excitement about companies like **Stability AI and Midjourney**, the issue of **Deepfakes** needs to be addressed before further work

Actionable Recommendations

Invest time and money in Hardware acceleration

- Advancements in AI are being led by many silent players working on the hardware aspect. Biggest of them right now is **Nvidia**, which is leading the chip development market.
- **GPUs** and **High-Performance Computing** are the back-bone of the current AI wars

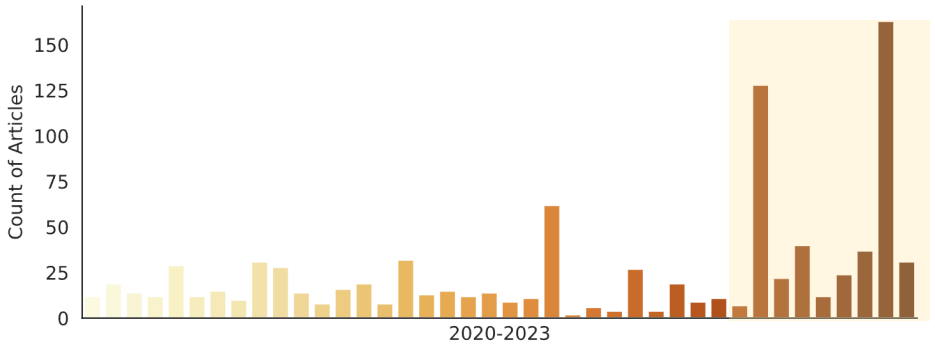


NVIDIA Brings Generative AI to World's Enterprises With Cloud Services for Creating Large Language and Visual Models

Adobe to Build Models for Next-Generation Creative Workflows; Getty Images, Morningstar, Quantiphi, Shutterstock Using NVIDIA AI Foundations Cloud Services to Customize Models for AI-Powered Applications

March 21, 2023

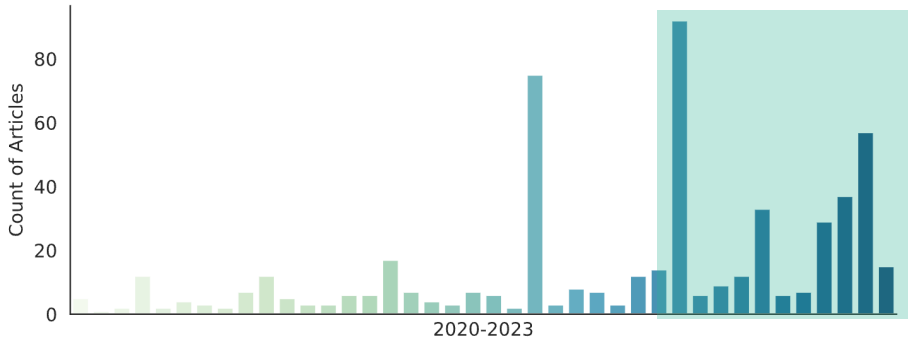
Harware Acceleration: Positive Articles over the past years



Invest in companies which aim to integrate current research with their product

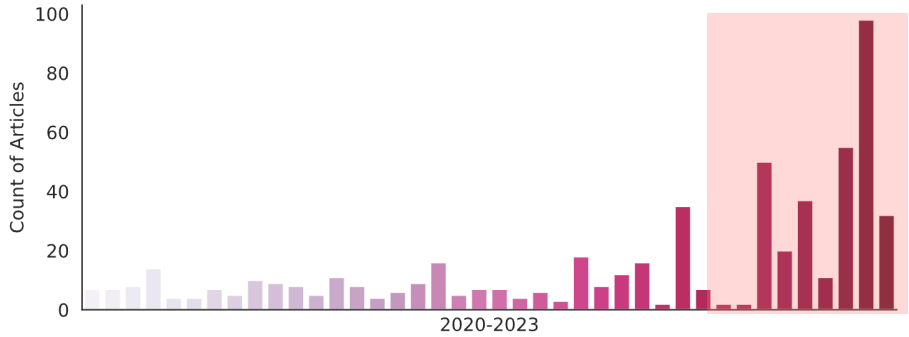
- Previously untouched sectors like **Legal** and **Music** generation provide novel opportunities for Generative AI integration
- Music generation capabilities are being offered by companies like **Verbit, TikTok and Stability AI**

Music AI: Positive Articles over the past years



- Legal companies like **CaseText** are making products that can make paralegal workforce more efficient

Legal AI: Positive Articles over the past years



Thank You!

Follow this project on GitHub

https://github.com/kshitij-mittal/Impact_Analysis_GenAI