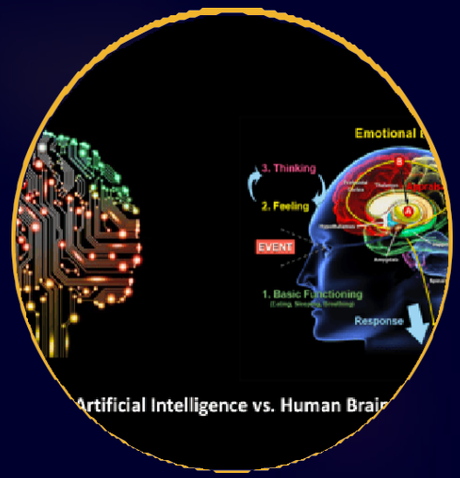
Exploring the Landscape of AI and Data Science Applications Through NLP



by Sanchit Narayan Kumar

Executive Summary

Artificial Intelligence has burgeoned in recent years, carving out an indispensable role within a host of sectors such as healthcare, law, and education. This analysis uses NLP techniques to decipher meaningful insights from news articles – spanning January 2020 to April 2023.

This project uses sentiment analysis, entity recognition, and topic modeling to identify successful data science and AI initiatives and reasons for that. In addition, it also identifies unsuccessful products or industries that have failed in its adoption of AI. Specifically, Bertopic and LDA was employed for topic modeling, Spacy Large for entity extraction and twitter-roberta-base-sentiment-latest from hugging face for sentiment.

After topics were extracted from positive and negative sentiment articles respectively, targeted entity extraction was implemented to identify sentiment towards these entities. In addition, Flan-t5-base was used for text summarization and targeted sentiment analysis.

This analysis promises to shed light on both the strides made in the AI industry and its potential areas of growth. This research stands to benefit stakeholders such as investors (venture capitalists), policymakers, companies looking to invest into data science in different industries, and business professionals.

Actionable Recommendations

Invest into AI

There's a high level of positive sentiment around businesses investing in data science. Sectors like law, education, and healthcare, which are increasingly leveraging AI, appear promising investments for VCs.

Competitive AI

Healthy competition is crucial – governments should ensure that the AI landscape isn't monopolized by a few. Incentives should be in place to stimulate a fair, competitive environment in AI development.

AI Hardware

Emphasize investments in areas like quantum computing and IoT – can ensure the high computational capability necessary for AI implementations.

Open-source AI

Strong support exists for further open-source contributions from major tech entities like Google and Microsoft. Arguably, increased open-sourcing is vital, as it encourages more collaborative innovation.

AI without bias

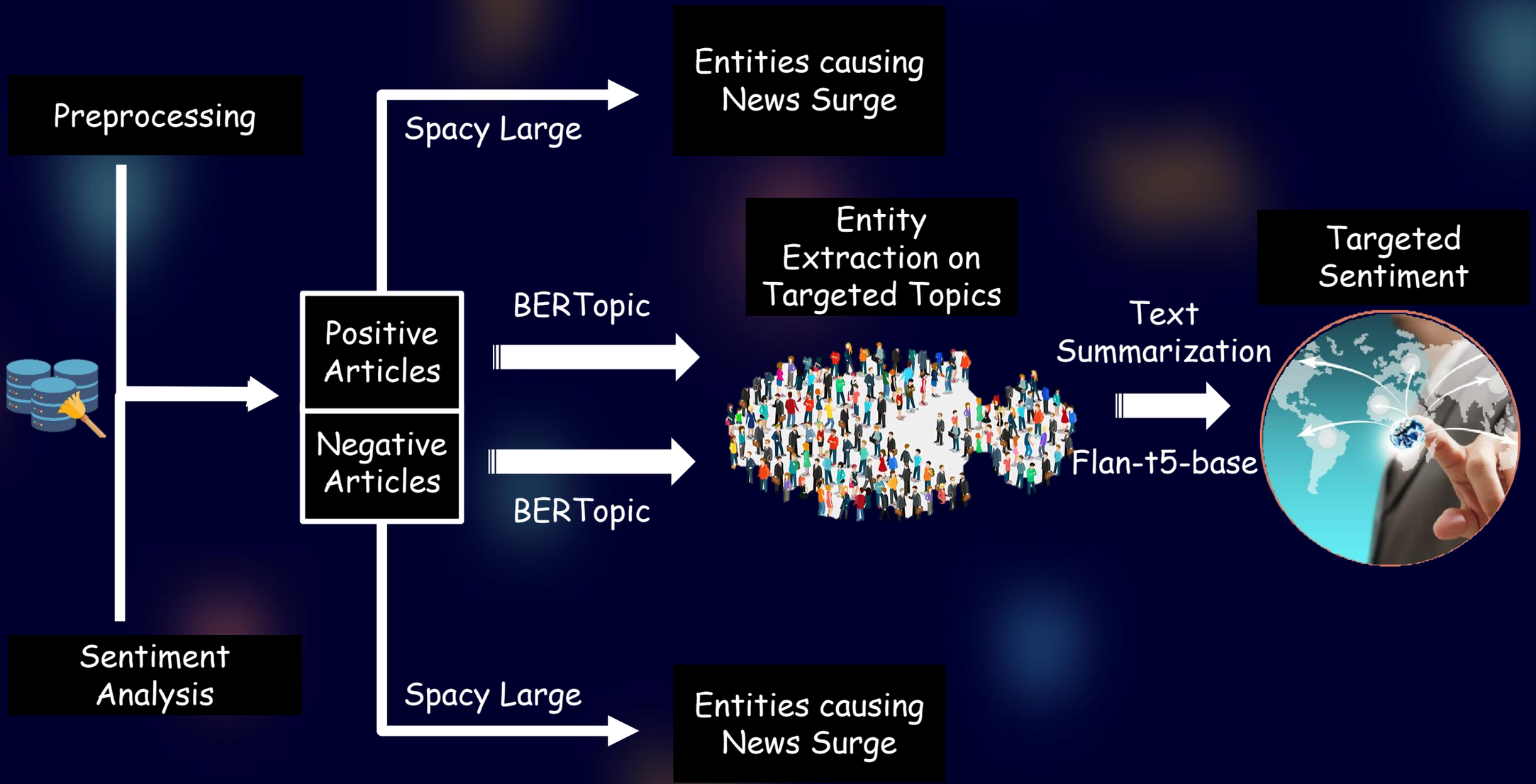
Ethical factors must be at the forefront when innovating and deploying Artificial Intelligence technologies. This entails a commitment to fairness, transparency, and impartiality in AI systems.

AI as a tool to augment

AI ought to be viewed as an instrument that bolsters. Striving to maintain equilibrium between potentials and constraints of AI is crucial, and we should be wary of undue dependency on this technology.

*Through text summarization

Solution Design - Methodology



Preprocessing

Clean-up tokens

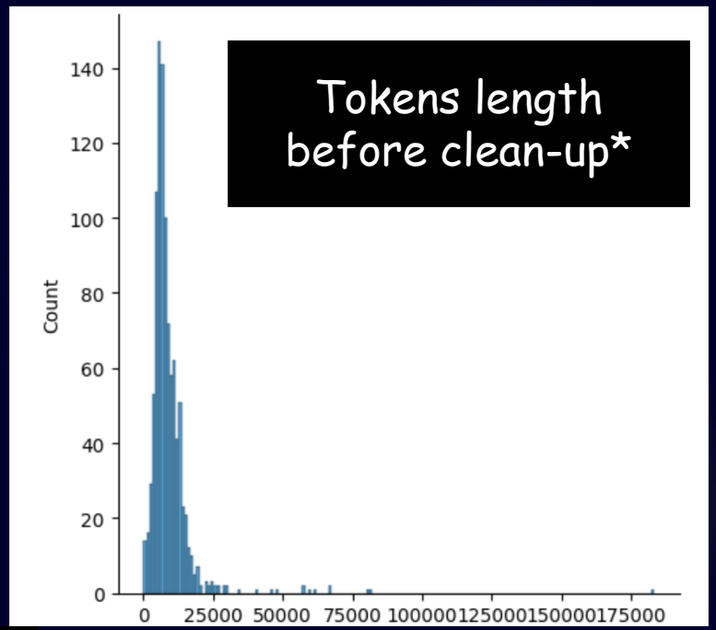
Removed noise, by eliminating newlines, tabs, whitespaces, special characters, URLs, links, remnants of web crawls, sentences with word length greater than 100 and other irrelevant text

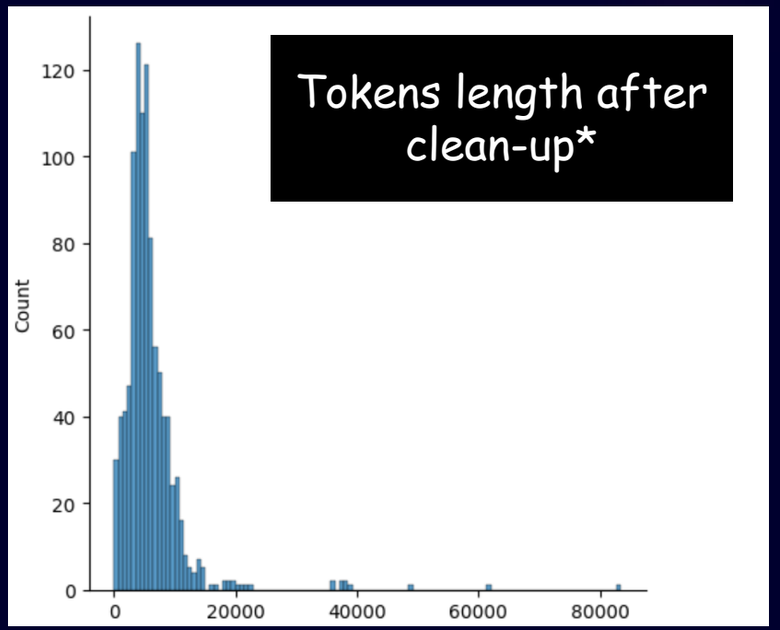
Keyword Extraction

Use selected keywords that are related to artificial intelligence and data science to filter out irrelevant articles - used bigrams, trigrams and fourgrams to find out other tokens most similar to AI and data science that were added to keywords list

Filtering

Dropped duplicate text only and not both text and title because it is important to capture similar titles - suggests that multiple news agencies are reporting on the topic - this will suggest news surge and topic importance for further analysis



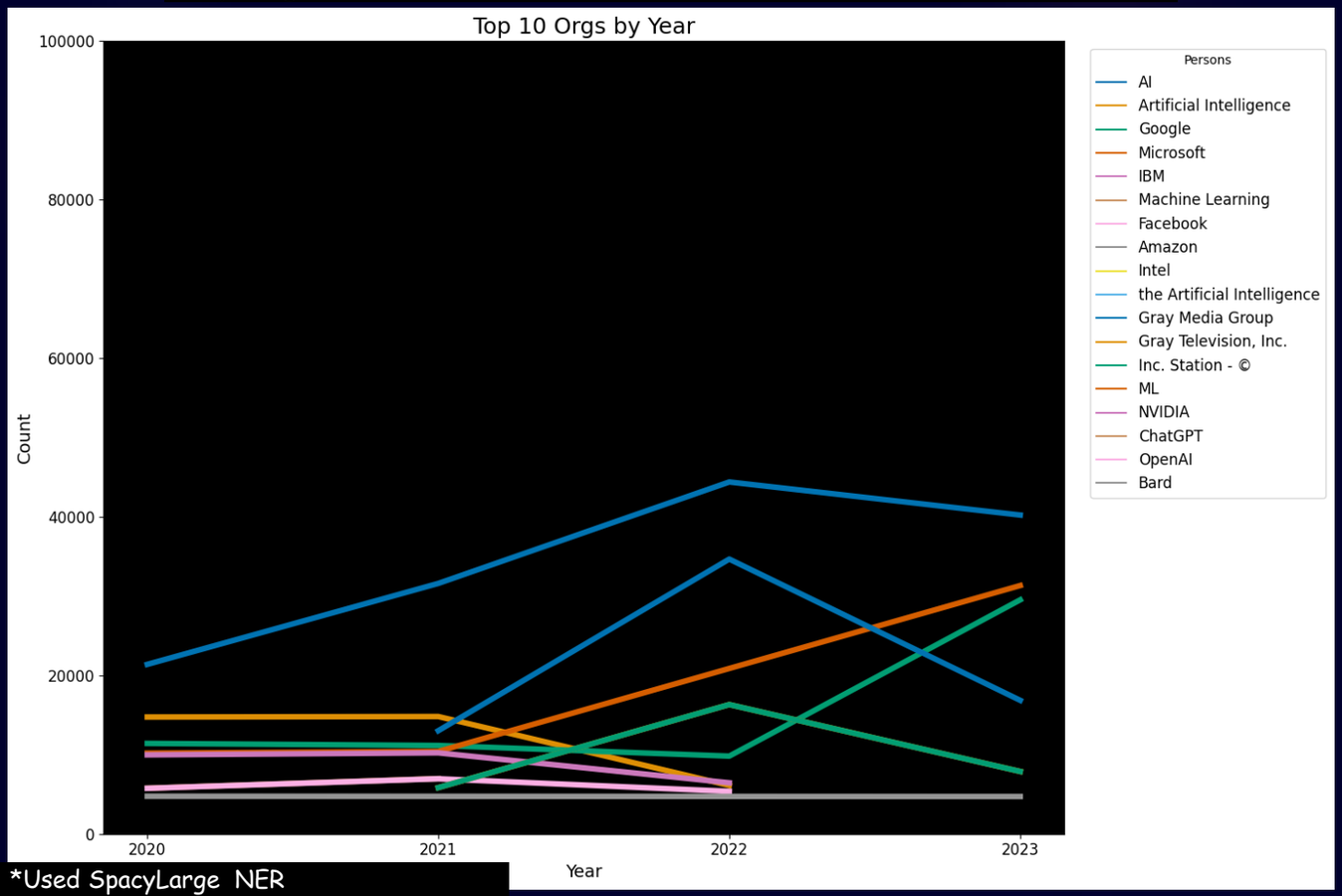


*For 1000 samples

ChatGPT Drives Positive Sentiment Trends

- Goal is to separate positive and negative sentiment from the corpus of news articles
- Implemented and fine-tuned model 'twitter-roberta-base-sentiment-latest' (Hugging Face) – chosen as it distinguishes between positive and negative sentiments effectively
- Experimented with different hugging face models for sentiment analysis – FinancialBert, Sentiment-Roberta-Large-English, Roberta-base-sentiment, news-sentiment-analysis

Top Organizations in Positive Sentiment News Articles Over Time



*Used SpacyLarge NER

Strong positive sentiment caused by major tech companies like Google, IBM, Microsoft, Nvidia, OpenAI and products like ChatGPT and hardware improvements (GPUs)

ChatGPT Also Drives Negative Sentiment Trends

Negative Sentiment around ChatGPT

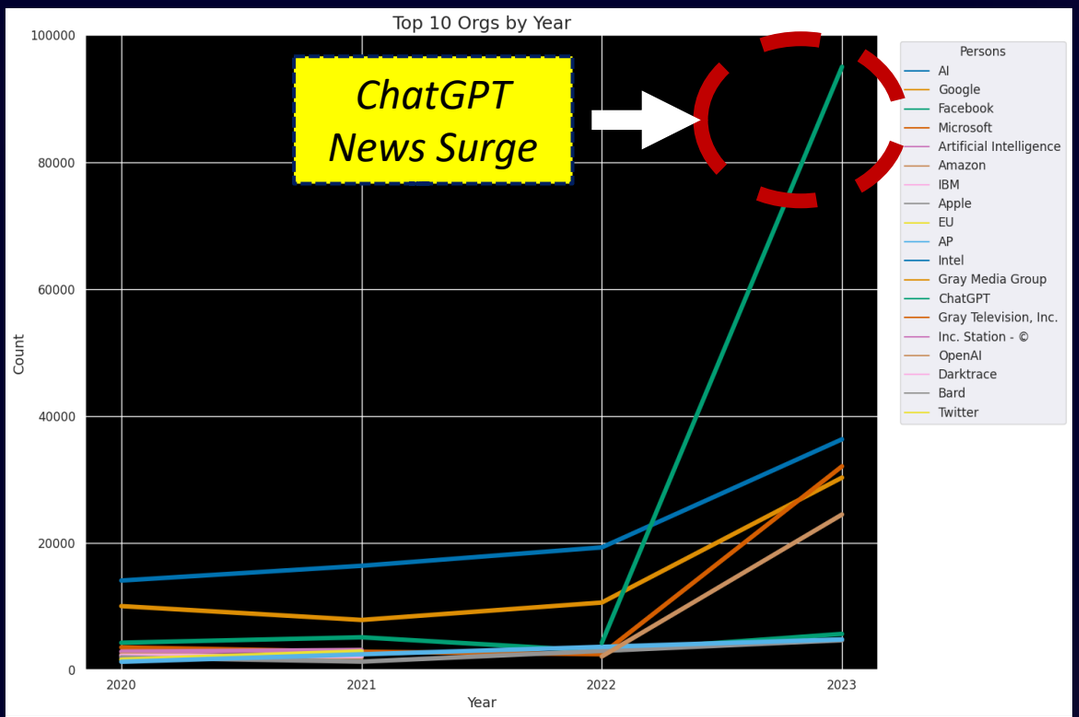
Strong negative sentiment also caused by major tech companies like Google, IBM, Microsoft and products like ChatGPT and Bard in particular

caused by

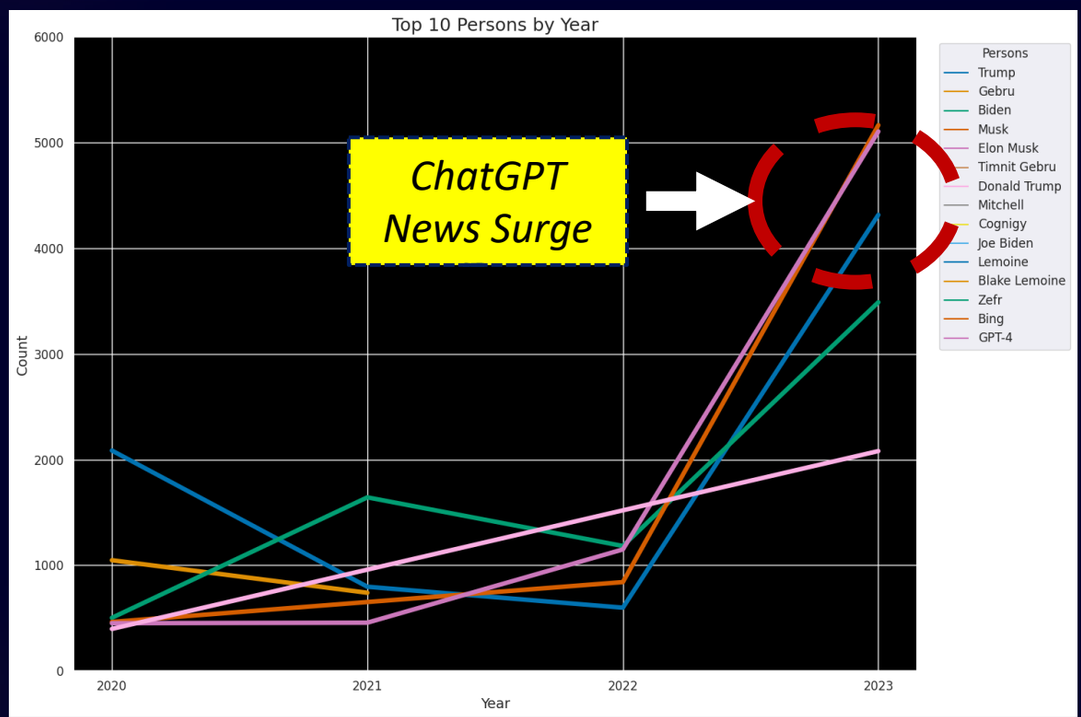
Centered around Regulation, Government, and Rivals

Driven by government officials like Trump / Biden, Elon Musk (rival ChatGPT) and AI ethics researchers like Gebru

Top Organizations in Negative Sentiment News Articles Over Time



Top Persons in Negative Sentiment News Articles Over Time



*Used SpacyLarge NER

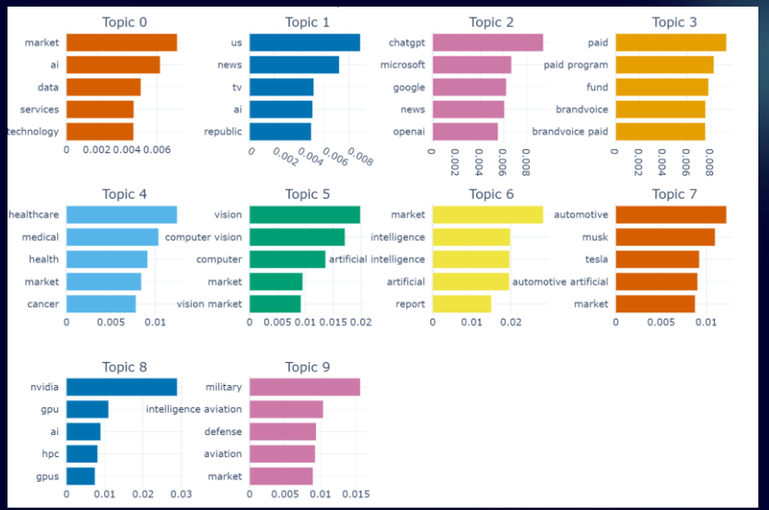
Overview of Industries with Successful and Failed AI Adoption

1. Successes
   - Technology
   - Legal
   - Retail
   - Fashion
   - Military
   - Healthcare
   - Quantum Computing
   - Food
   - Education

2. Failure
   - Agriculture
   - Oil & Gas
   - Insurance
   - Airlines
   - Automotive
   - Shipping

Identified through the major topics (LDA and BERTopic) on positive and negative sentiment articles respectively. Number of topics were determined by visualizing topic clusters (reduced topics to 100 and 60 respectively). Entity extraction on specific topics applied for targeted sentiment using Spacy Large.

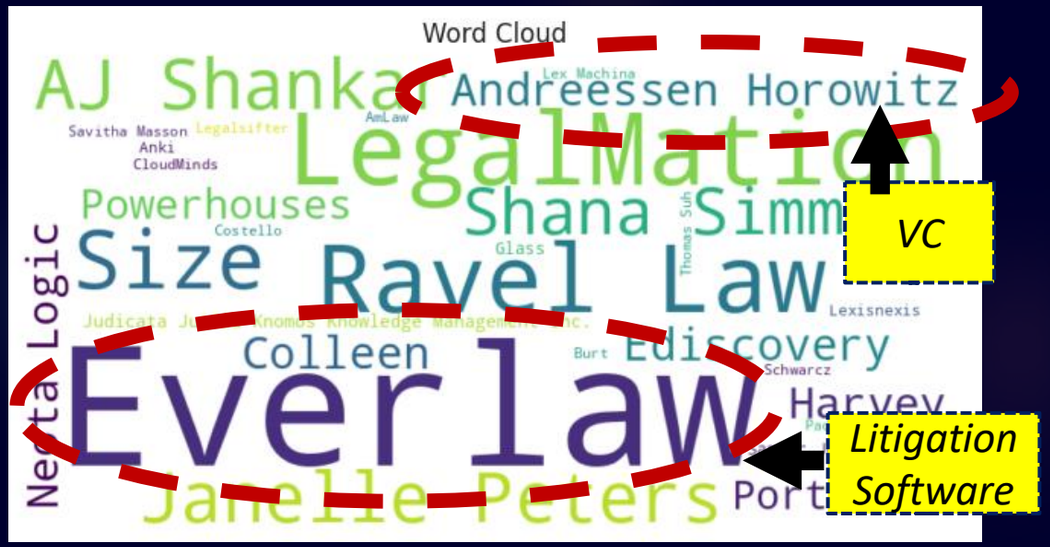
Positive Topics Tokens



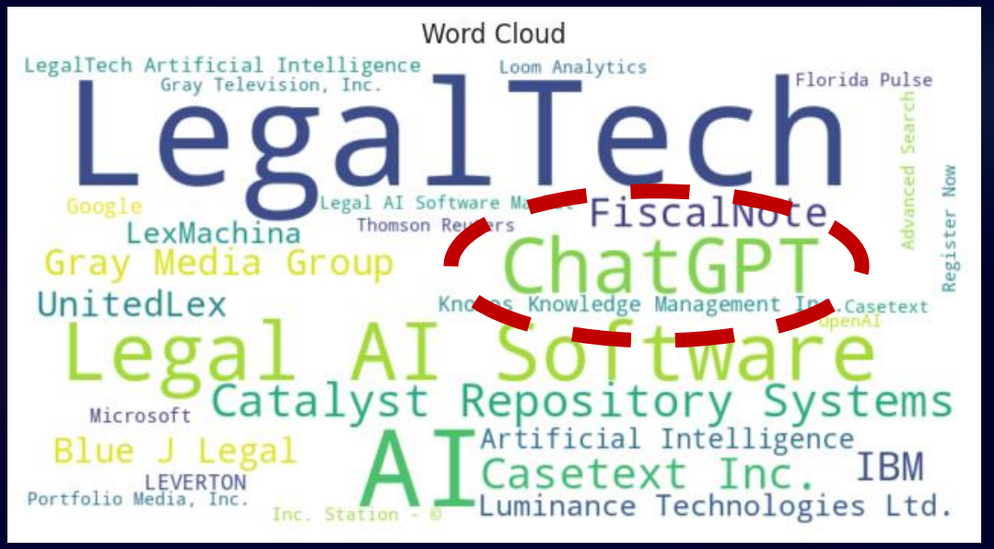
Negative Topics Tokens

Positive Sentiment Case Study - Law

Word Cloud of Top Persons in Positive Sentiment Articles in Law



Word Cloud of Top Organizations in Positive Sentiment Articles in Law

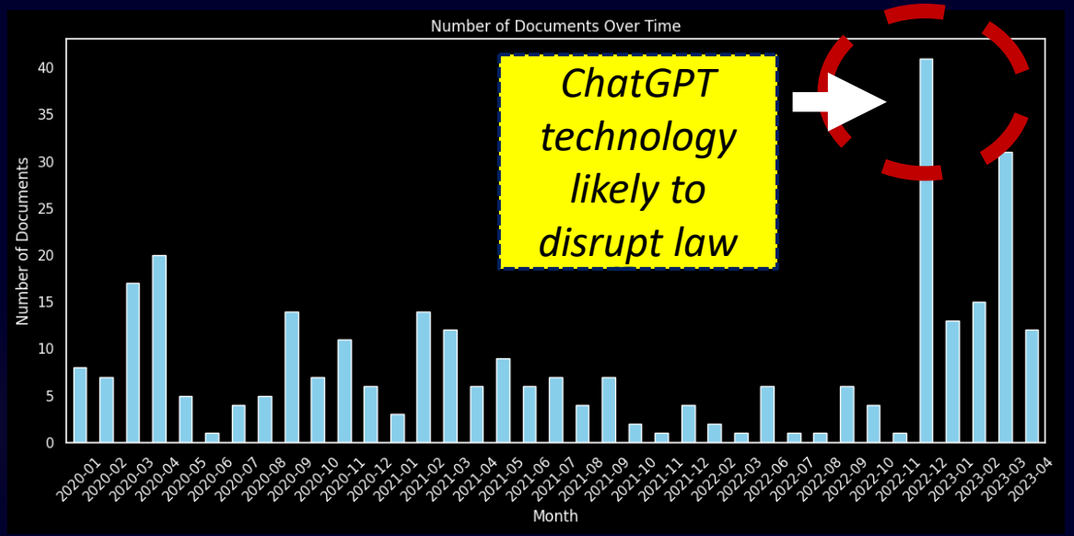


1

Successful Law Products

Focuses on Ediscovery, early case assessment, legal holds, trial preparations – key areas in data science involve document, image and audio processing and analysis

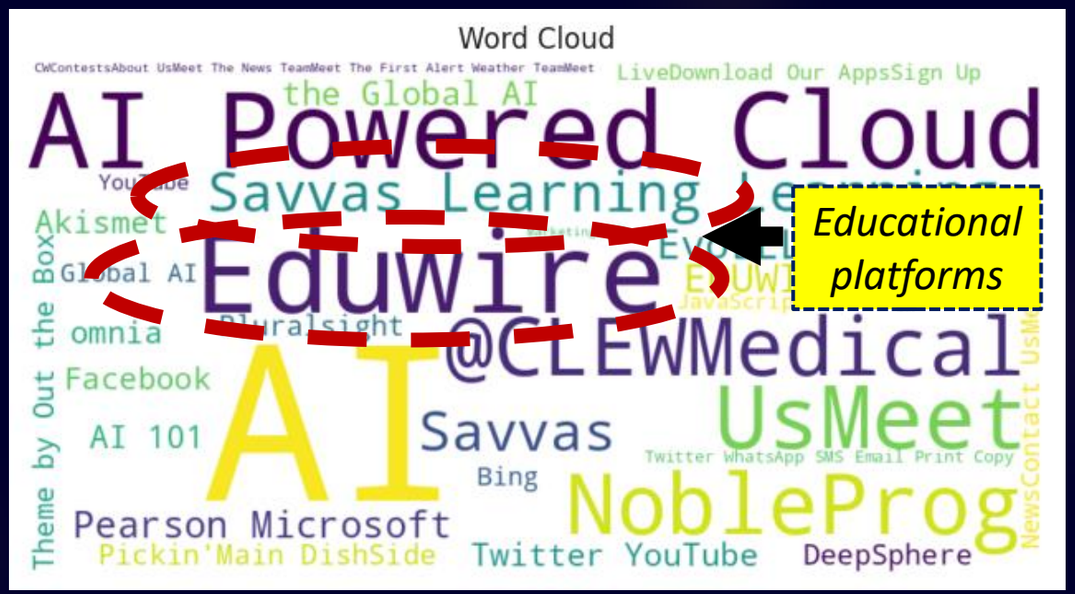
Positive Sentiment Articles in Law Over Time



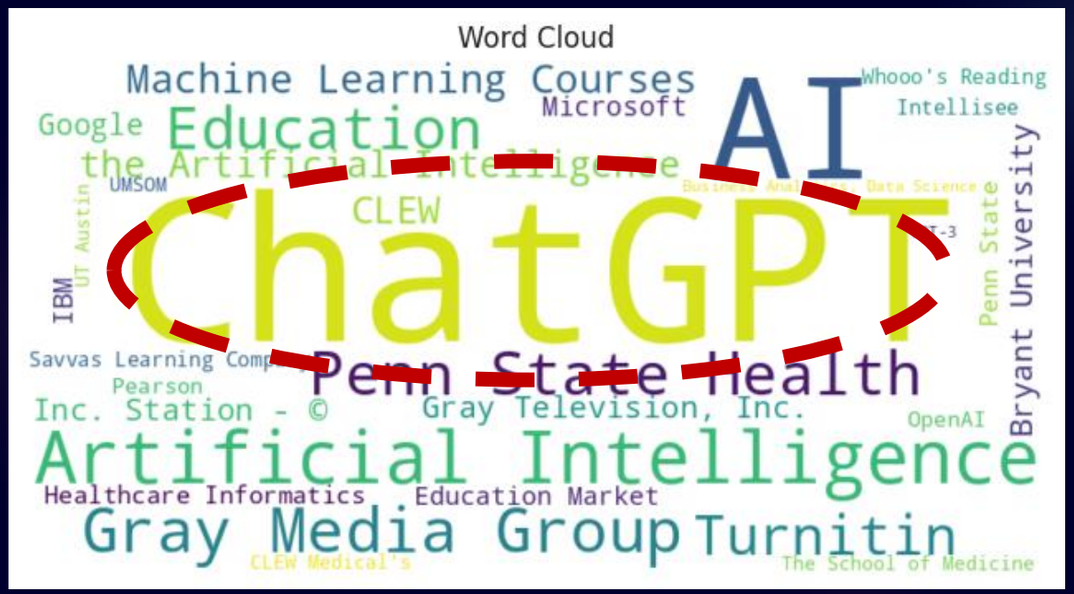
*Word Cloud based on SpacyLarge NER

Positive Sentiment Case Study - Education

Word Cloud of Top Persons in Positive Sentiment Articles in Education



Word Cloud of Top Organizations in Positive Sentiment Articles in Education

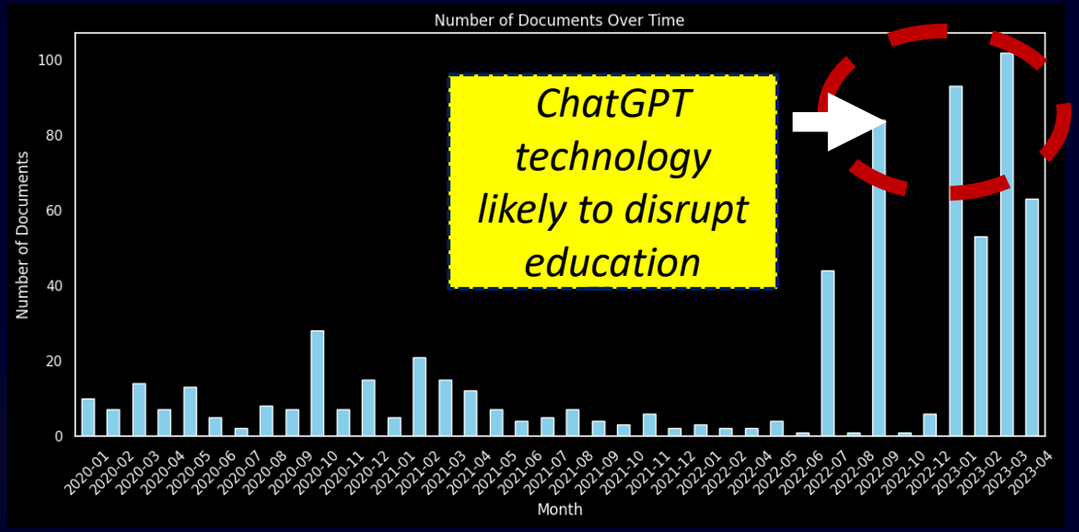


1

Successful Education Products

Focuses on leveraging conversational AI technology to provide digital learning solutions to create interactive classrooms, training programs, and digital assessments

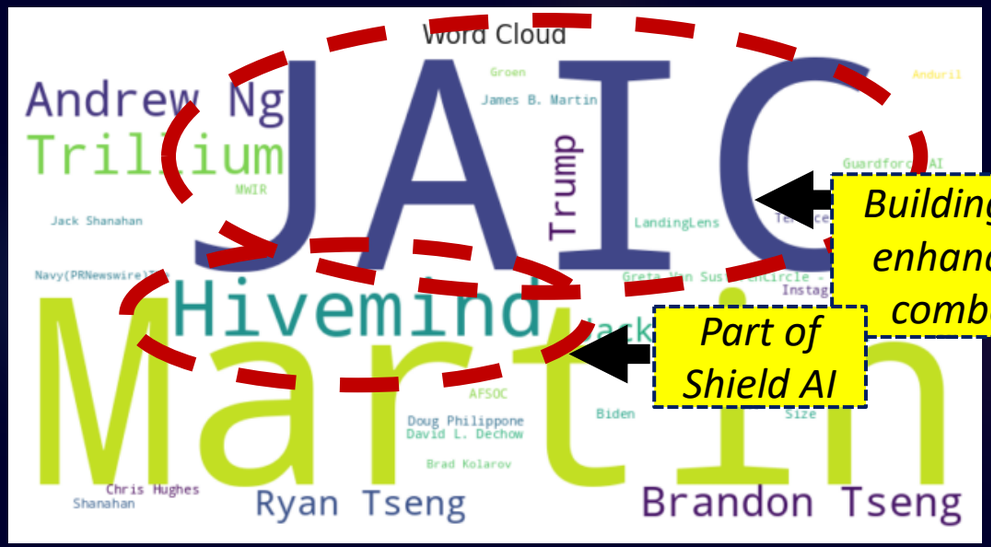
Positive Sentiment Articles in Education Over Time



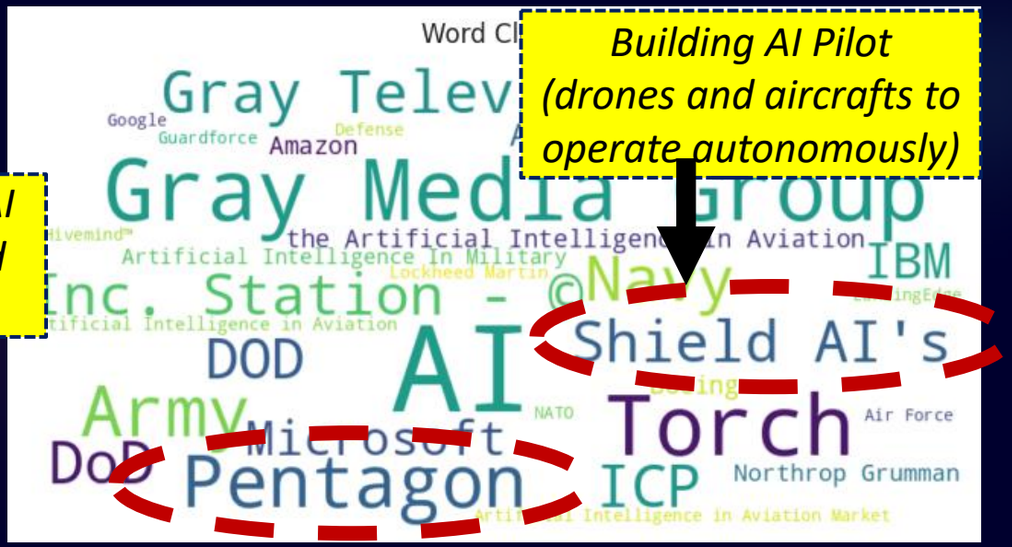
*Word Cloud based on SpacyLarge NER

Positive Sentiment Case Study - Military

Word Cloud of Top Persons in Positive Sentiment Articles in Military



Word Cloud of Top Organizations in Positive Sentiment Articles in Military

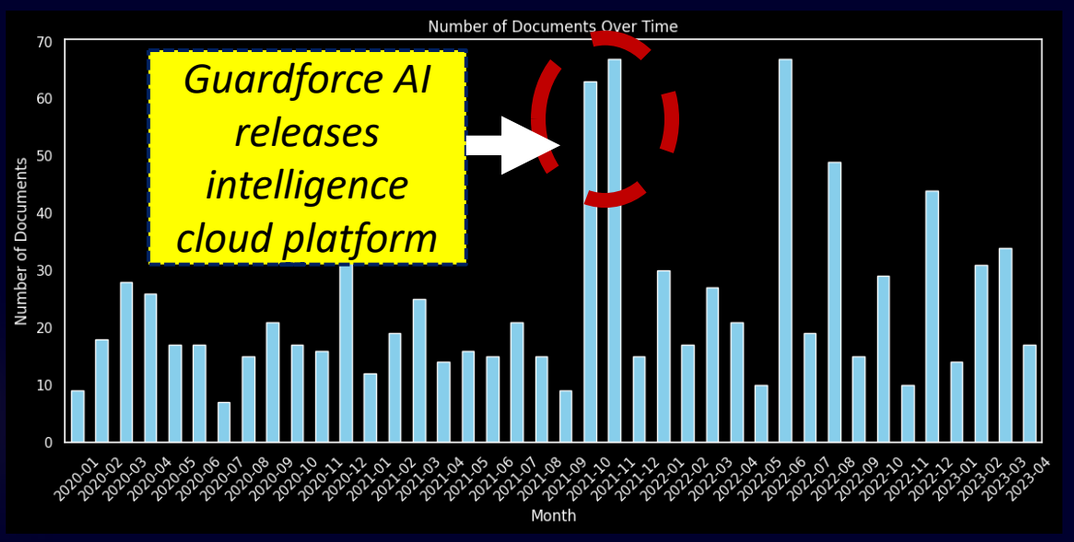


1

Successful Military Products

Focuses on drone and robotics technology that enables AI enhanced combat that can be used to evade defense systems and also can be used for surveillance purposes. AI seen as key strategy to beat rivals in combat

Positive Sentiment Articles in Military Over Time



*Word Cloud based on SpacyLarge NER

#EU proposed ban to use of AI
eu_df_filtered_apr2021 = eu_df_filtered[eu_df_filtered['month_year'] == '2021-04']
#eu_df_filtered_apr2021.head()

Other Positive Sentiment Case Studies

Retail and Fashion

- Autonomous checkout, automation - warehouse and store operations; computer vision, customization of choices in retail through platforms

Healthcare

- Advancements in disease detection and discovery (early discovery of cancer) with FDA approval of new technologies

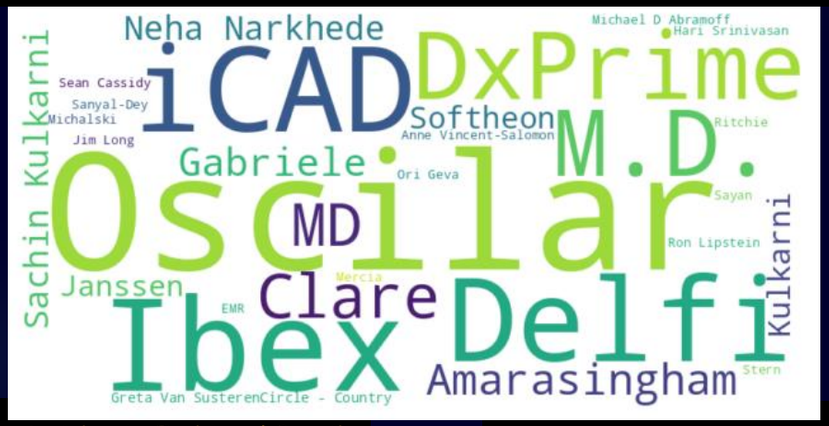
Hardware improvements

- Increased computing power – GPUs through NVIDIA and Intel

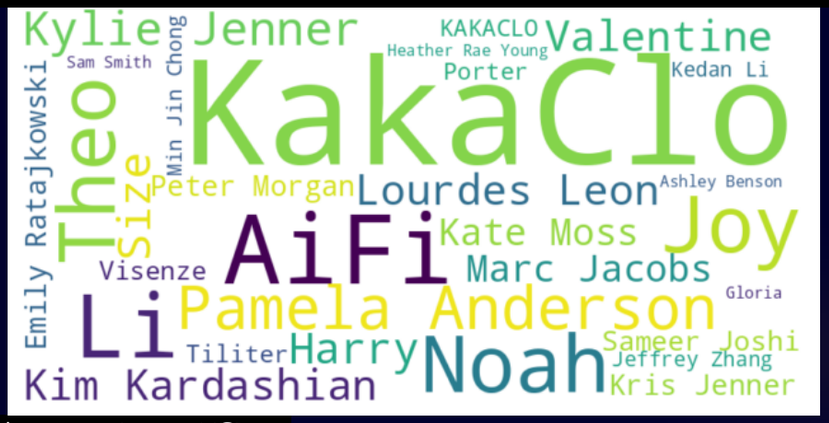
Cybersecurity

- Leading ai software detects, investigates, and responds to advanced threats

Word Cloud of Top Persons in Positive Sentiment Articles in Healthcare



Word Cloud of Top Persons in Positive Sentiment Articles in Fashion / Retail

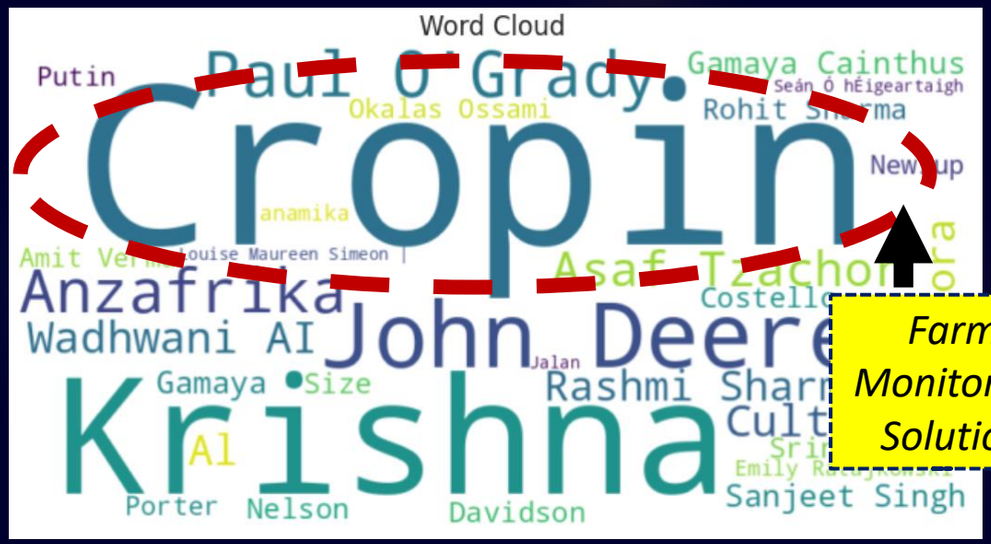


Word Cloud of Top Persons in Positive Sentiment Articles in Cybersecurity

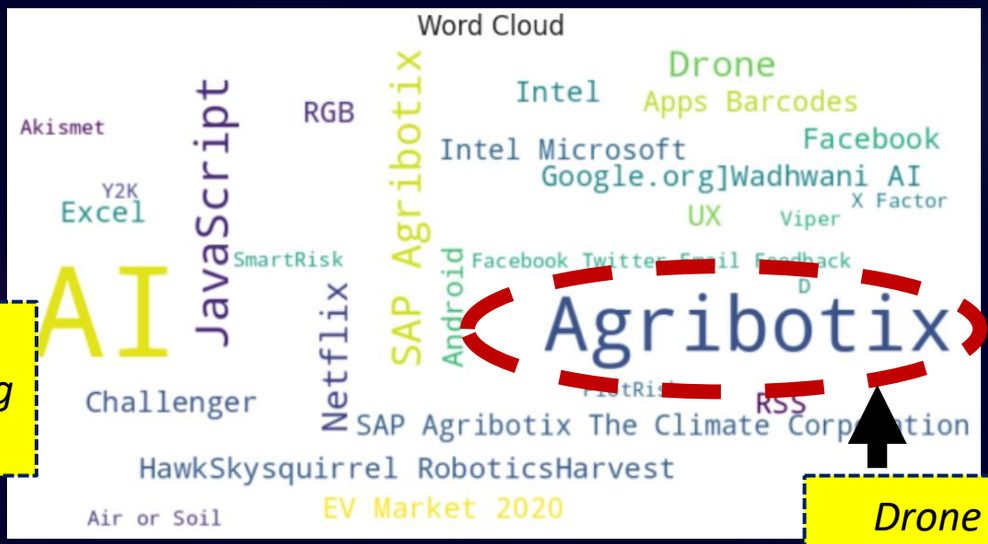
*Word Cloud based on SpacyLarge NER

Negative Sentiment Case Study - Agriculture

Word Cloud of Top Persons in Negative Sentiment Articles in Agriculture



Word Cloud of Top Products in Negative Sentiment Articles in Agriculture



Farm Monitoring Solution

Drone enabled software company

1

Successful Agriculture Products

Focuses on drone technology and farm monitoring solutions that help geotag farms / digitize farm records / monitor crop productivity and boost field officer productivity

2

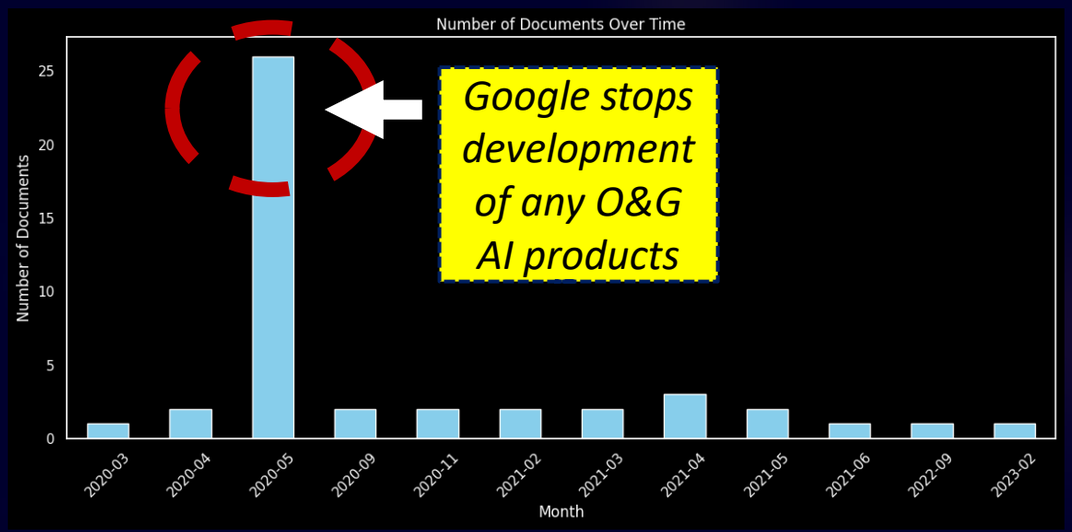
Challenges Facing Adoption

Low count of agriculture related articles suggest that adoption of AI and data science is low in the industry. Challenges include: data collection, high cost of change, and talent acquisition

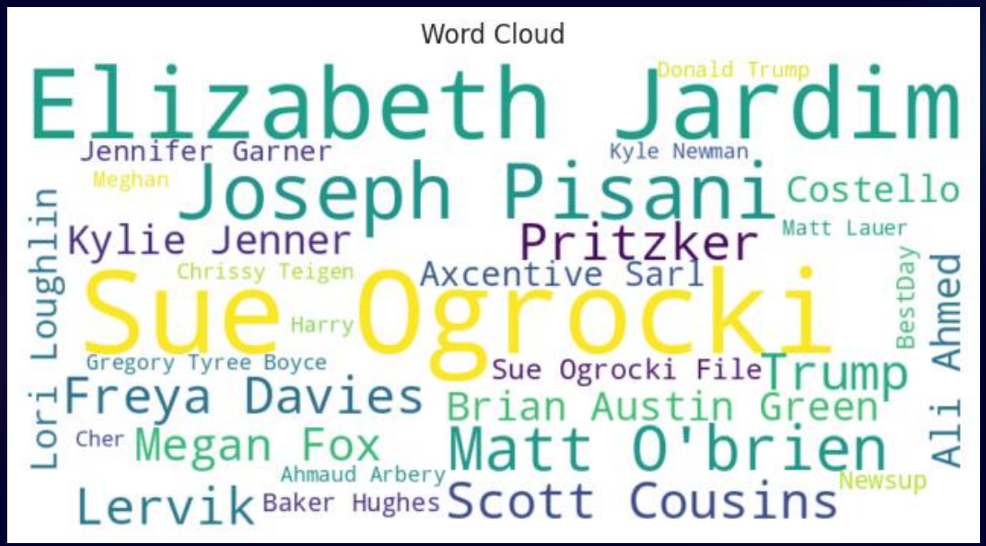
*Word Cloud based on SpacyLarge NER

Negative Sentiment Case Study - Oil & Gas

Negative Sentiment Articles in Oil Over Time



Word Cloud of Top Persons in Negative Sentiment Articles in Oil



1

Successful O&G Products

Microsoft and Amazon focus on pipelines, shipping and fuel storage to speed up shale extraction – however, adoption is low. Future of AI adoption in O&G will be clean energy focused - MBZUAI in collaboration with IBM will develop carbon neutral solutions to existing energy supplies.

2

Challenges Facing Adoption

Key challenges include backlash from climate change organizations like Greenpeace that have led companies like Google to back out of developing O&G AI products.

*Word Cloud based on SpacyLarge NER

#EU proposed ban to use of AI
eu_df_filtered_apr2021 = eu_df_filtered[eu_df_filtered['month_year'] == '2021-04']
#eu_df_filtered_apr2021.head()

Other Negative Sentiment Case Studies

Regulation

- EU continuous to discuss advanced bans on some forms of AI including indiscriminate surveillance or AI intended to manipulate behavior or AI used for "social scoring"

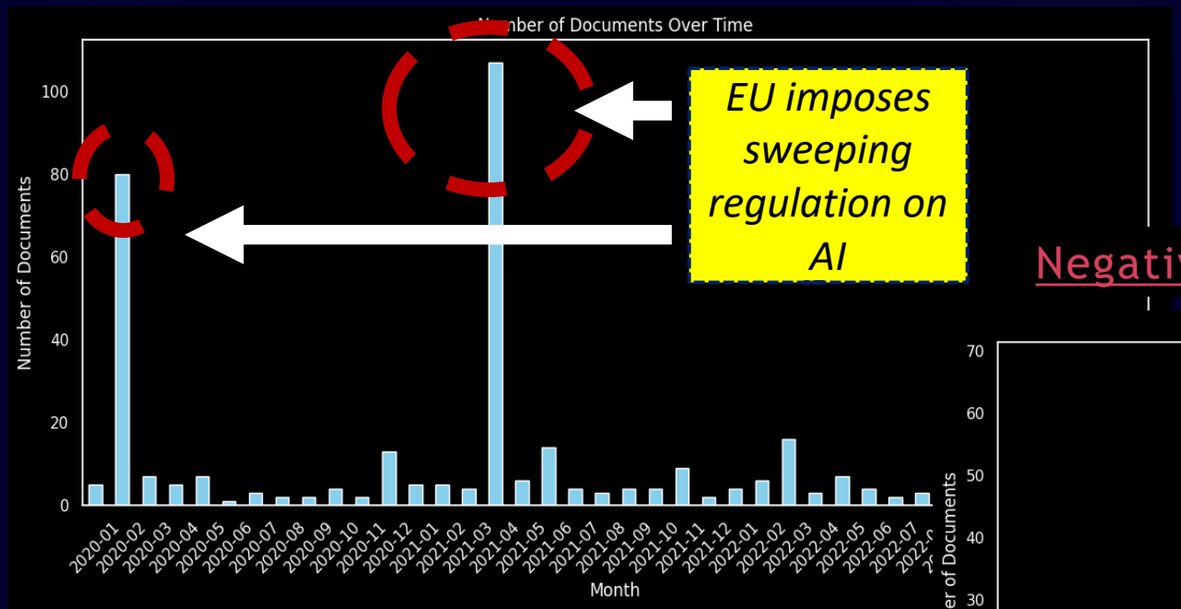
Automotive

- Autonomous cars are struggling to commercialize
- automated driving technology
- Key technologies include ADAS

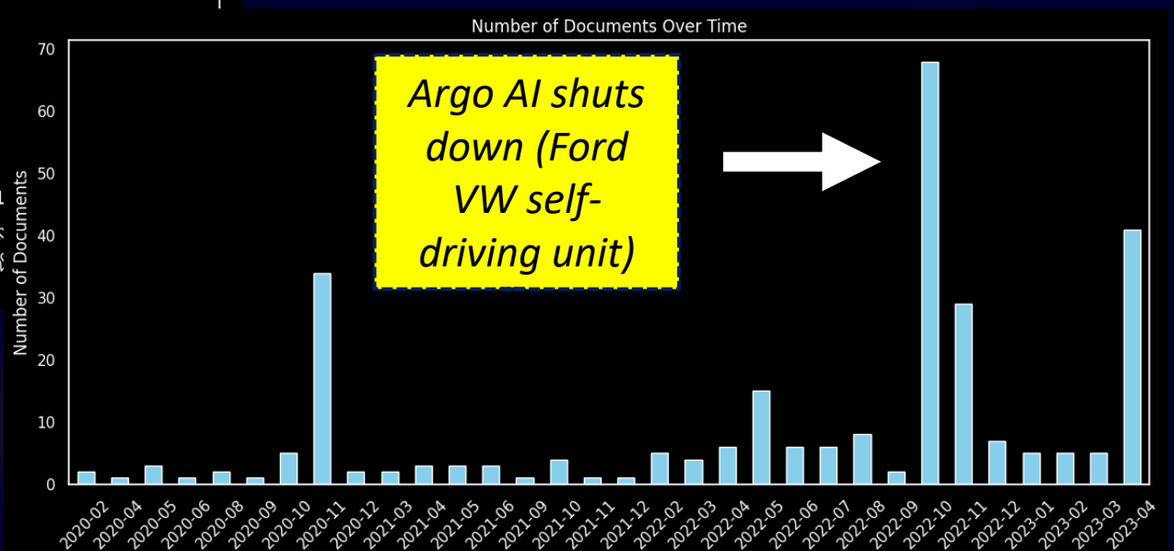
ChatGPT

- Rivals like Google and Musk under pressure to compete with ChatGPT with Bard and TruthGPT
- Regulatory entities like Trump, Biden critical

Negative Sentiment Articles on Regulation Over Time



Negative Sentiment Articles on Auto Over Time



*Word Cloud based on SpacyLarge NER

#EU proposed ban to use of AI
eu_df_filtered_apr2021 = eu_df_filtered[eu_df_filtered['month_year'] == '2021-04']
#eu_df_filtered_apr2021.head(2)

Conclusion

Highly Positive Sentiment around AI

AI has highly influenced societal sentiment in a positive manner and has contributed significantly to advancements in various sectors. This analysis shows there has been a remarkable growth over the past three years in industries such as healthcare, law, and education. We are currently in a 'hype cycle' given the news surge around ChatGPT.

Industries Failing in AI Initiatives

The analysis has also illustrated industries that have yet to continue to advance data science such as agriculture, oil & gas, and insurance.

Multiple Stakeholders in the AI World

Our detailed entity analysis both on the over positive and negative sentiment corpus as well as on targeted topics suggests that governments and various other stakeholders are involved.

#EU proposed ban to use of AI
eu_df_filtered_apr2021 = eu_df_filtered[eu_df_filtered['month_year'] == '2021-04']
#eu_df_filtered_apr2021.head(2)

Future Work and Technical Conclusions

- With more computation power, better results can be achieved through use of sophisticated large language models.
- In terms of topic modeling, BERTopic outperforms traditional methods such as Latent Dirichlet Allocation (LDA). It excels in assembling similar topics into coherent clusters and does so in a computation-efficient manner, consuming less memory.
- Sentiment fine-tuning and extensive manual annotation are crucial for accurately discerning the sentiment inherent in a text. Other methods such as building a custom model on open-source data could be attempted.
- For entity extraction, transformer spacy could be used and compared. Additionally, large language models, or alternatively, developing a NER system on labeled data could potentially offer improvements in entity detection.