NLP Final Project:

Impact of AI

Luke Schwenke
University of Chicago
March 2024

Contents

1. Executive Summary
2. Data Preparation
3. Zero Shot Topic Modeling
4. Sentiment Analysis
   a. Generative AI
5. Named Entity Recognition (NER)
   a. Organizations
   b. People
   c. Locations
6. Business Recommendations

Executive Summary

Objective: Identify the impact of Artificial Intelligence (AI) by performing modern NLP techniques on thousands of unstructured news articles. Use this data to glean insights on how AI will influence jobs.

Zero Shot Topic Modeling analysis was used across 3 AI-relevant topics (Industry and Sector, Technology, and Societal). The frequency of articles discussing sectors like manufacturing, utilities, transportation, and agriculture was low, indicating AI will not disrupt manual sectors as much as the more frequent technology and healthcare focused areas. Generative AI was identified as the primary technology topic by a wide margin indicating a lot of impact. Employment and jobs were also most frequently talked about in the society topic meaning the primary concern of news discussions with AI is people's livelihoods.

Sentiment Analysis & Summarization on Generative AI was implemented to further explore the most commonly identified Technology topic with Zero Shot learning. The results showed primarily positive sentiment though negative sentiments increased in 2023 indicating excitement but also concerns for the new technology. A lot of the concerns are related to competition between the US and China with more extreme cases highlighting human extinction and the abstract relationship of Man vs. AI. Most positive articles reveal financial promise with growth surges, particularly as GenAI relates to the financial markets, technology, and healthcare.

NER on Organizations recognized various organizations including the White House, Bard, Meta, Apple, and more. A deeper dive on Meta, a generally more controversially regarded company, showed primarily neutral sentiments, with the positives largely outweighing the negatives. This indicates Meta's rebranding and reputation has improved, particularly around the companies involvement and implementation of AI.

NER on People revealed people like Elon Musk, Sam Altman, Donald Trump, Xi Jinping, and more are discussed frequently in AI articles. Extended analysis on Sam Altman showed a neutral sentiment; however, Sam's negative-rated articles (5.6%) outweighed his positives (4.3%) meaning there are concerns about his role, or OpenAI's, in the impact of AI.

NER on Locations showed the United States, China, Japan, and Canada are taking the most advantage of new AI technologies. This suggests any investment into AI should most likely touch one of these regions to have the highest reach/impact.

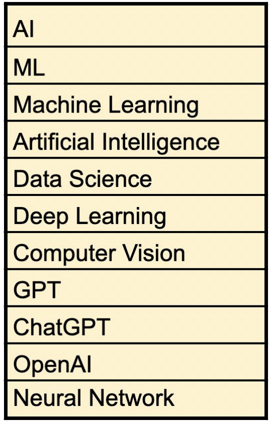
Article Filtering & Text Cleaning

The original dataset contained 200,000+ rows → Filtered dataset contains 26,000+ rows

Filtering Methods

1. Removed **duplicates**
2. Removed articles **less than 30 characters**
3. Removed articles that **do not discuss AI** topics by creating a list of 150+ common AI keywords

Keywords Sample



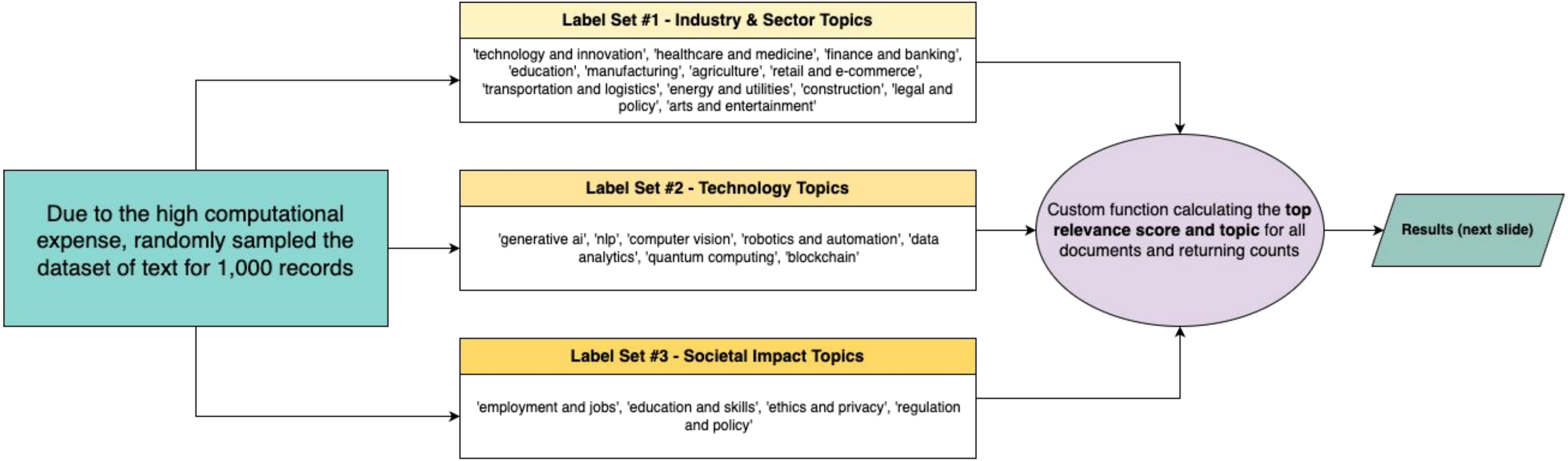
AI
ML
Machine Learning
Artificial Intelligence
Data Science
Deep Learning
Computer Vision
GPT
ChatGPT
OpenAI
Neural Network

Text Cleaning Methods

1. Converted multiple spaces into single spaces
2. Removed:
    a. Non-English Words
    b. Leading & Trailing Whitespaces
    c. Hyperlinks
    d. HTML tags
    e. Newline characters
    f. Carriage returns
    g. Special characters
3. **Stopwords** were removed too but a Raw Text column was also kept to test summarization and other NLP methods on non-altered data
4. **Lemmatization** was applied to simplify words to their base dictionary lemma to reduce noise
5. **Lowercasing** was applied as needed. Capitalization can be used to help identify proper nouns and names

Topic Modeling with Zero Shot Learning

Zero shot learning offers a way to perform various tasks, including topic modeling, without having to train a model. This form of ML allows immediate out-of-the-box results with the expense of being computationally expensive.



Due to the high computational expense, randomly sampled the dataset of text for 1,000 records

Label Set #1 - Industry & Sector Topics

'technology and innovation', 'healthcare and medicine', 'finance and banking', 'education', 'manufacturing', 'agriculture', 'retail and e-commerce', 'transportation and logistics', 'energy and utilities', 'construction', 'legal and policy', 'arts and entertainment'

Label Set #2 - Technology Topics

'generative ai', 'nlp', 'computer vision', 'robotics and automation', 'data analytics', 'quantum computing', 'blockchain'

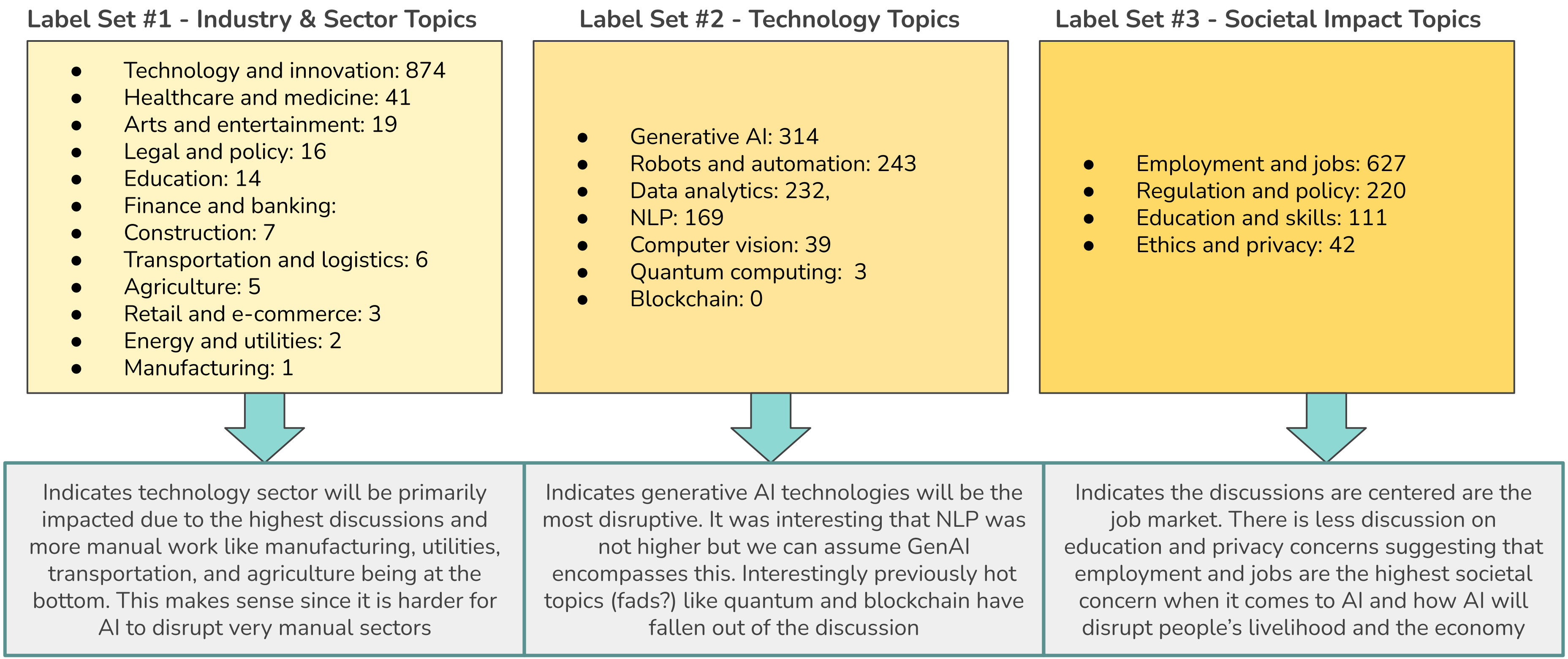
Label Set #3 - Societal Impact Topics

'employment and jobs', 'education and skills', 'ethics and privacy', 'regulation and policy'

Custom function calculating the **top relevance score and topic** for all documents and returning counts

Results (next slide)

Topic Modeling with Zero Shot Learning Results



Label Set #1 - Industry & Sector Topics

- Technology and innovation: 874
- Healthcare and medicine: 41
- Arts and entertainment: 19
- Legal and policy: 16
- Education: 14
- Finance and banking:
- Construction: 7
- Transportation and logistics: 6
- Agriculture: 5
- Retail and e-commerce: 3
- Energy and utilities: 2
- Manufacturing: 1

Indicates technology sector will be primarily impacted due to the highest discussions and more manual work like manufacturing, utilities, transportation, and agriculture being at the bottom. This makes sense since it is harder for AI to disrupt very manual sectors

Label Set #2 - Technology Topics

- Generative AI: 314
- Robots and automation: 243
- Data analytics: 232,
- NLP: 169
- Computer vision: 39
- Quantum computing: 3
- Blockchain: 0

Indicates generative AI technologies will be the most disruptive. It was interesting that NLP was not higher but we can assume GenAI encompasses this. Interestingly previously hot topics (fads?) like quantum and blockchain have fallen out of the discussion

Label Set #3 - Societal Impact Topics

- Employment and jobs: 627
- Regulation and policy: 220
- Education and skills: 111
- Ethics and privacy: 42

Indicates the discussions are centered are the job market. There is less discussion on education and privacy concerns suggesting that employment and jobs are the highest societal concern when it comes to AI and how AI will disrupt people’s livelihood and the economy

Sentiment Analysis, After Zero Shot

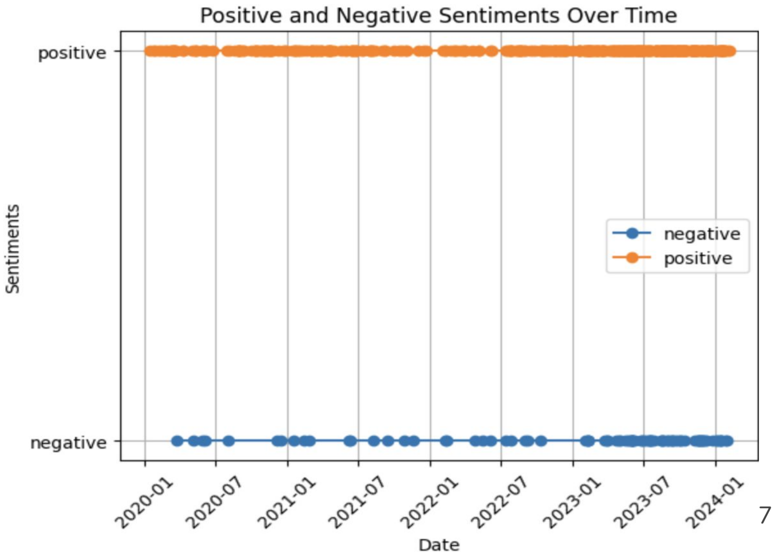
Now that we have performed Zero Shot learning and seeing that Generative AI is the most relevant Technology topic mentioned in a 3,000 record random sample, we want to see the sentiment around the topic

A highly popular **LLM tuned on articles/news** was downloaded from Hugging Face and run on the articles pinpointed by Zero Shot learning to be relevant to “Generative AI”

mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

12.5% - Positive
84.6% - Neutral
2.8% - Negative

Negative sentiments have increased in 2023 but the majority are still positive



The sentiment of the most popular technology topic identified, Generative AI, has primarily neutral and some positive sentiment. This can indicate that people are excited for the technology, even if it is disruptive to jobs

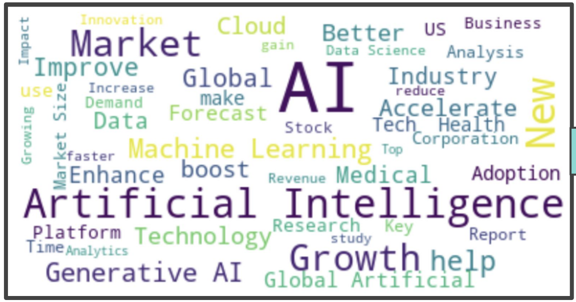
What are people saying about Generative AI?

Word Clouds (Top 50 Words)

Summary of Articles

Interpretation

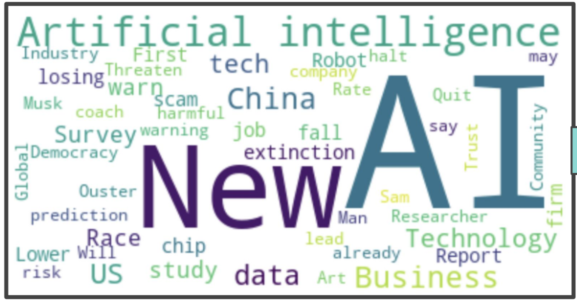
Positive Sentiment



AI sector saw big job gains recently wont unscathed downturn report. AI sector has potential to Surge Big This Summer. These AI Stocks Have Potential Surge Big this Summer. AI Solution Enhance Wind Solar Storage Asset Performance Increase Energy Production.

Interpreting the word cloud and summary on **positive views** of GenAI indicates it will have promise in the fields of medicine, business, the financial markets, and global growth. It will boost and accelerate industries and grow in demand. GenAI has huge potential to surge and enhance business, especially in the energy and technology sectors.

Negative Sentiment



Artificial intelligence could lead human extinction. US AI chip China would cause permanent loss. US export ban advanced AI chips hit almost China tech. Man AI falling best relationship ever.

Interpreting the word cloud and summary on **negative views** indicates extremes with mentions of extinction as well as intense competition with China and the US. There are also risks of threats, scams, warnings, race issues, losses, and more mentioned.

Named Entity Recognition (NER) with Organizations

Top 20 Organization Entities Detected

Organization	Count
AI	8,929
Times	1,246
Daily Mail	1,077
Artificial Intelligence	632
EU	622
The Economic Times	507
Pledge Times	503
Meta	500
Bard	480
White House	458
Financial Analysis	423
Apple	416
Tech News	327
Technology AI Business	310
AI News	290
Fox News	281
Congress	272
UN	272
Journal	244
Digital	237

Meta generally has a negative reputation due to privacy distrust issues in the past – they even rebranded away from Facebook because of this! Let's analyze Meta further to see if the data confirms this:

First off, what is the sentiment across the 500 Meta documents? **12.5% - Positive, 84.6% - Neutral, 2.8% - Negative**

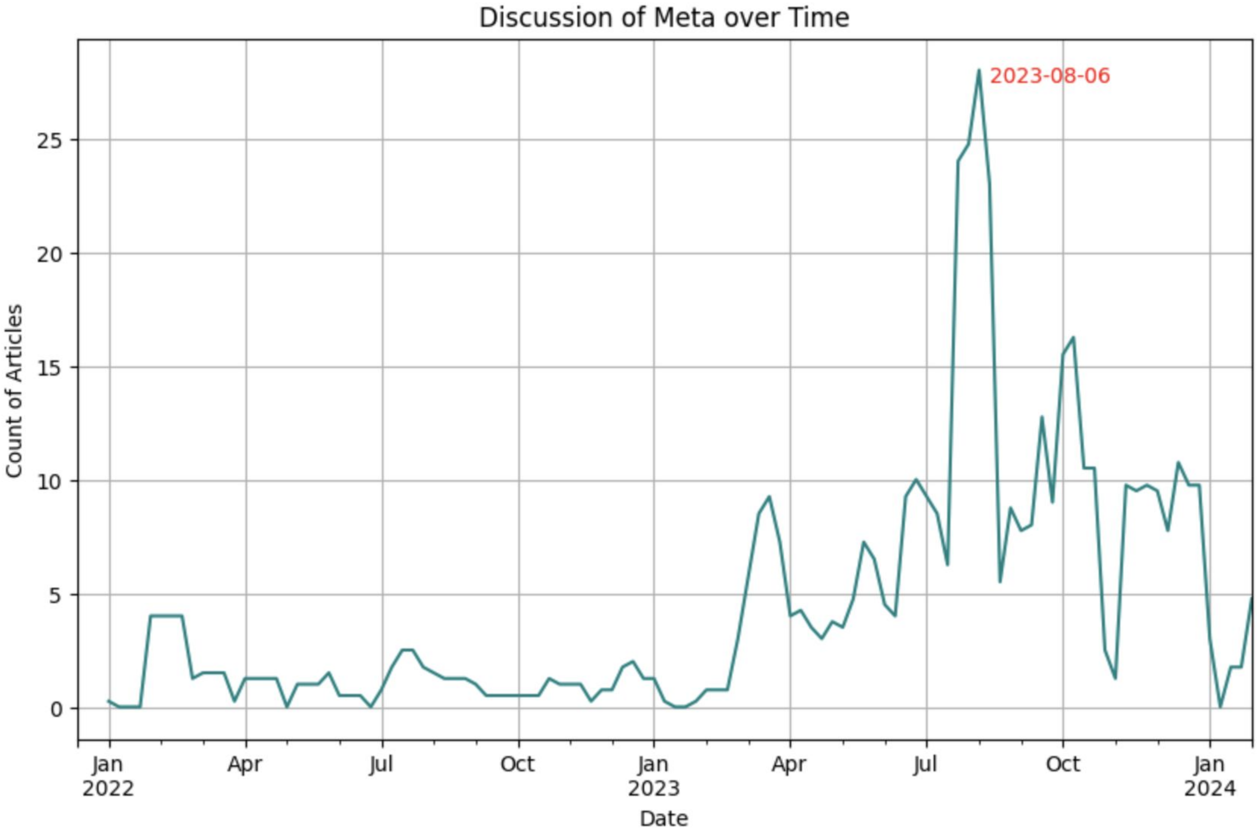
Now, let's summarize the articles using the **bart-large-cnn** LLM from Hugging Face: **The Meta launches AI Enhanced Workplace Efficiency. Meta promises White House they'll develop AI responsibly. The Hill Sri government data following attack. Twitter throttle New York Times Sri government Data.**

What is the sentiment of the overall summary generated? **Neutral - indicates Meta's reputation in these articles is not as negative as other media may suggest**

NER was accomplished with **spaCy's Large English Language Model** and the "**ORG**" entity label

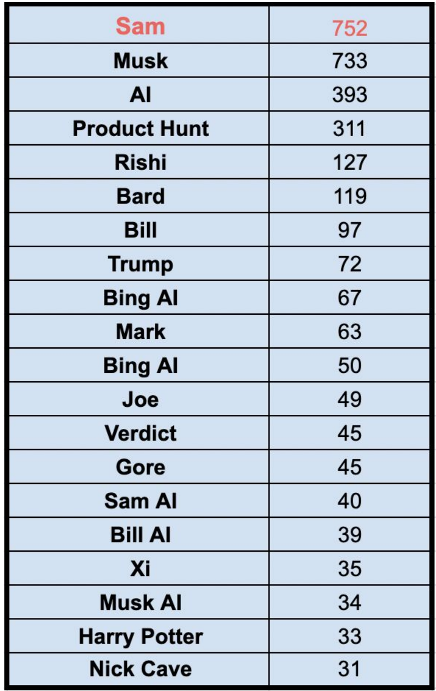
Named Entity Recognition (NER) with Orgs. Cont.

The large spike in discussions of Meta in early August 2023 is likely due to **Meta's decision to block news on its platforms in Canada**. Though this does not directly relate to AI, topics like these have brought into question the company's ethical decisions in AI and other areas, contributing to their high discussion counts according to spaCy



Named Entity Recognition (NER) with People

Top 20 Person Entities Detected



Person	Count
Sam	752
Musk	733
AI	393
Product Hunt	311
Rishi	127
Bard	119
Bill	97
Trump	72
Bing AI	67
Mark	63
Bing AI	50
Joe	49
Verdict	45
Gore	45
Sam AI	40
Bill AI	39
Xi	35
Musk AI	34
Harry Potter	33
Nick Cave	31

Sam Altman is one of the most discussed people within the AI space since he is CEO and cofounder of OpenAI (ChatGPT). Sam was dismissed from OpenAI for a period of time due to internal disputes before coming back, but we can suspect there are some strong opinions, whether positive or negative, around Sam.

What is the sentiment around Sam's 752 articles?

4.3% - Positive, 90.2% - Neutral, 5.6% - Negative

Now, let's summarize the articles. Here is the summary:

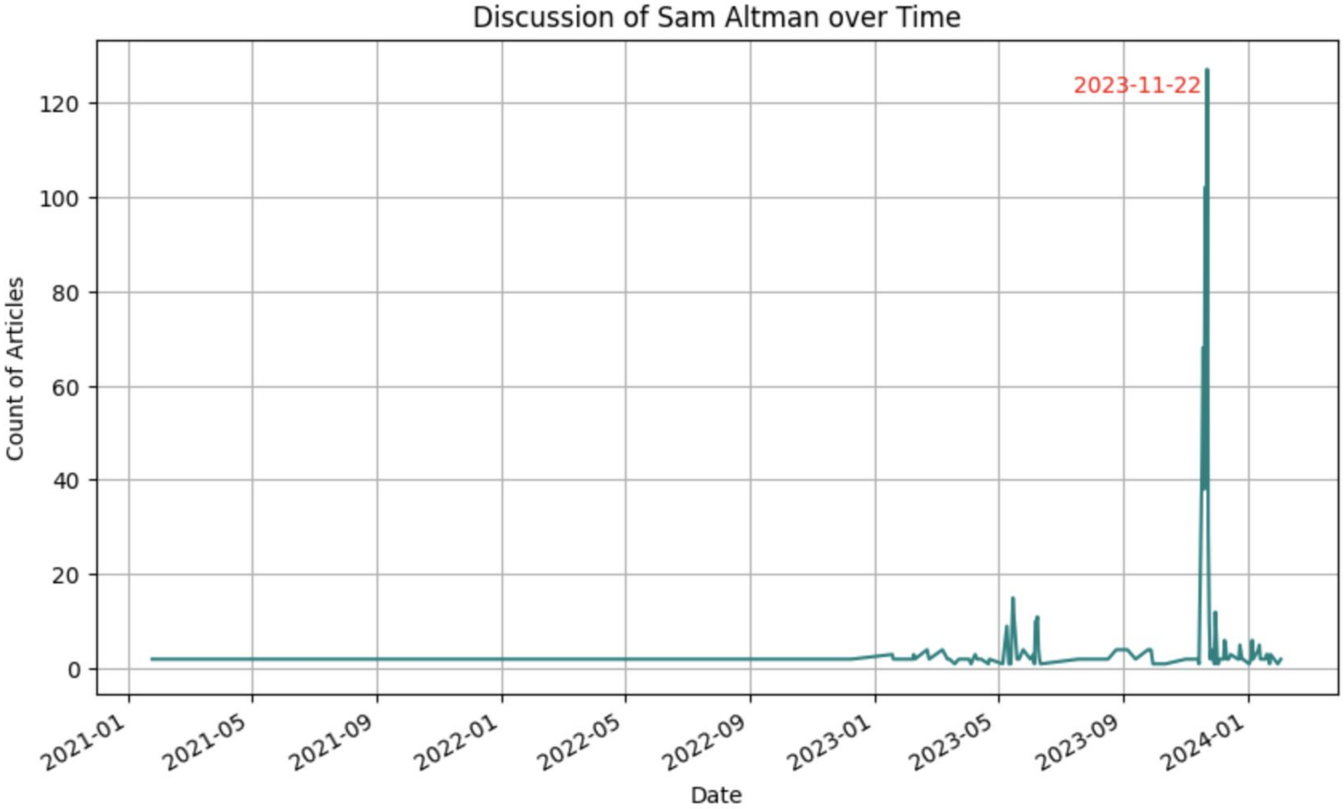
How Sam fired snapped return along new board. Sam could lead global around AI regulation. You could parachute onto island full come back five king. Artificial intelligence profit power is Sam apocalypse.

If we use the bart-large-cnn LLM from Hugging Face, what is the sentiment of the overall summary generated?

Neutral - indicates Sam has a good reputation that is not overly negative or positive

NER was accomplished with spaCy's Large English Language Model and the "**PERSON**" entity label

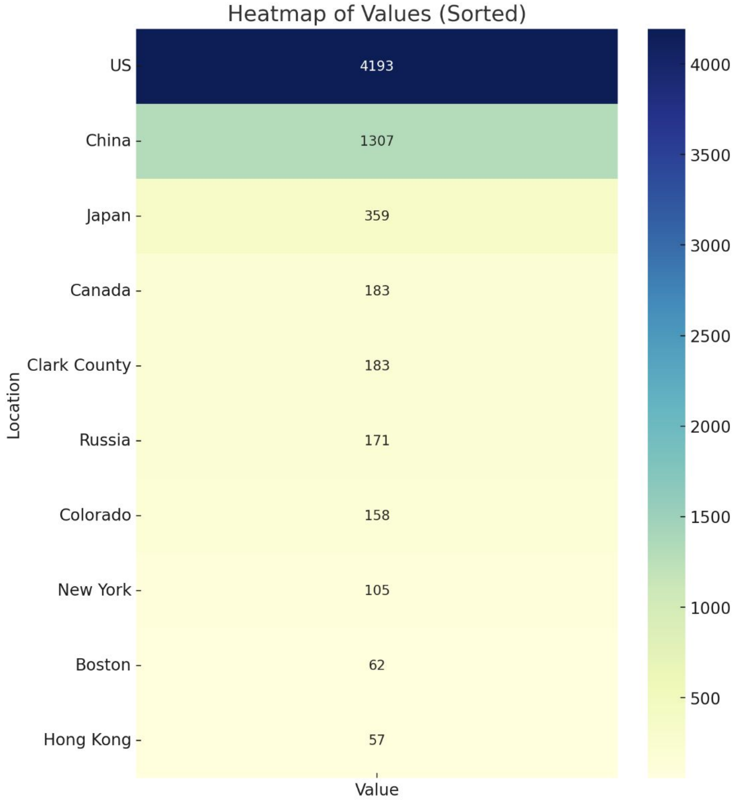
Named Entity Recognition (NER) with People Cont.



We see a large spike in discussions around **Sam Altman around the time he was dismissed (11/17/23)** from OpenAI. We cannot attribute this to being negative sentiment on AI in general, but it does explain why Sam was recognized more through spaCy's NER algorithm

Named Entity Recognition (NER) with Locations/GPE

NER on locations showed AI is most discussed in regards to the **US, China, and Japan**. Interestingly, **Clark County**, also appears on the list which is where Las Vegas is located. Companies looking to leverage AI may want to focus investments in these locations given the majority of news articles are discussing them as being the most impactful regions to AI.



Actionable Recommendations

The extensive analysis on news articles using NLP techniques has proven AI is going to have a large impact. We recommend companies review the data results and take the following actions to stay relevant:

1. **Use generative AI** due to its potential for explosive growth and high-impact on jobs in **technology and healthcare.** Jobs in these industries will be most affected compared to areas like **manufacturing and agriculture** which will have immunity from the effects of AI.
   a. Focus investments and automation strategically in the **US, China, and Japan**
2. Pay close attention to the **financial markets** as many concerns and discussion with AI revolve around these areas. Considering investing in **companies like Meta** who are pouring resources into AI technology
3. Approach business decisions on AI carefully as there is **mixture of both optimism and fear** with its growth