NLP

From Words To Insights

Unstructured Text Analysis for Anticipating AI shift in Job Industry

~ a project work by GARIMA SOHI

TABLE OF CONTENTS

Executive Summary

01

Executive Summary

OVERVIEW:

The project conducted an in-depth analysis of the current state of AI technology and its potential impact on various industries, jobs, and workforce at large. Utilizing a vast corpus of around 200,000 news articles, it investigated trends, sentiments, geographical distribution, potential benefits, and risks associated with AI.

INSIGHTS:

Large Language Model, 'chatgpt', received predominantly positive sentiment due to its role in sparking an AI revolution. However, negative sentiments were associated with its outages, potential impact on student learning, job elimination, and potential facilitation of hate speech. AI has been trending consistently since 2020, with a peak in interest from Google in 2023, and the most engagement in the US, followed by India and China.

Several business lines have emerged as primary beneficiaries of AI investment. These include healthcare, defense, media, education, finance, AI chipsets, cyber security, music/entertainment, air quality checks, food/beverages, and automotive. However, the rise of AI is not without potential downfalls. Concerns include unauthorized data collection, political manipulation, deepfakes, cheating in educational settings, health data breaches, and general data privacy breaches.

CONCLUSION:

The insights gleaned from this project present a nuanced picture of AI's role in our future. On the one hand, AI continues to revolutionize industries, offering the potential to significantly improve productivity and solve complex problems. On the other hand, the misuse of AI technologies, particularly around data privacy and security, must be addressed.

Recommendations 02

A. Application Success in AI

- Some lines of business that should invest in AI are as follows:
  - Healthcare
  - Defense
  - Media
  - Education
  - Finance
  - AI Chipsets
  - Cyber Security
  - Music/Entertainment
  - Air Quality Check
  - Food/beverages
  - Automotive

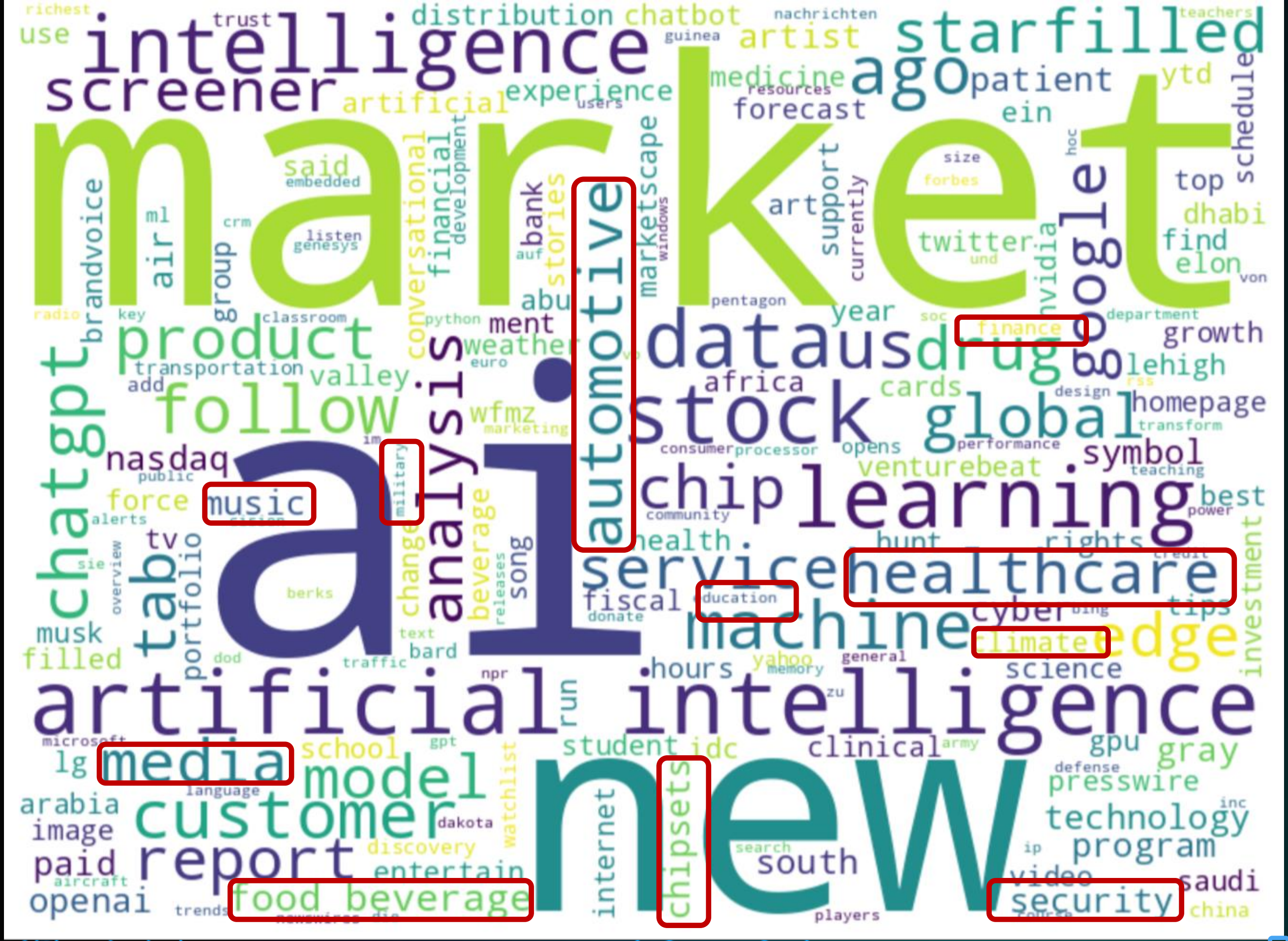
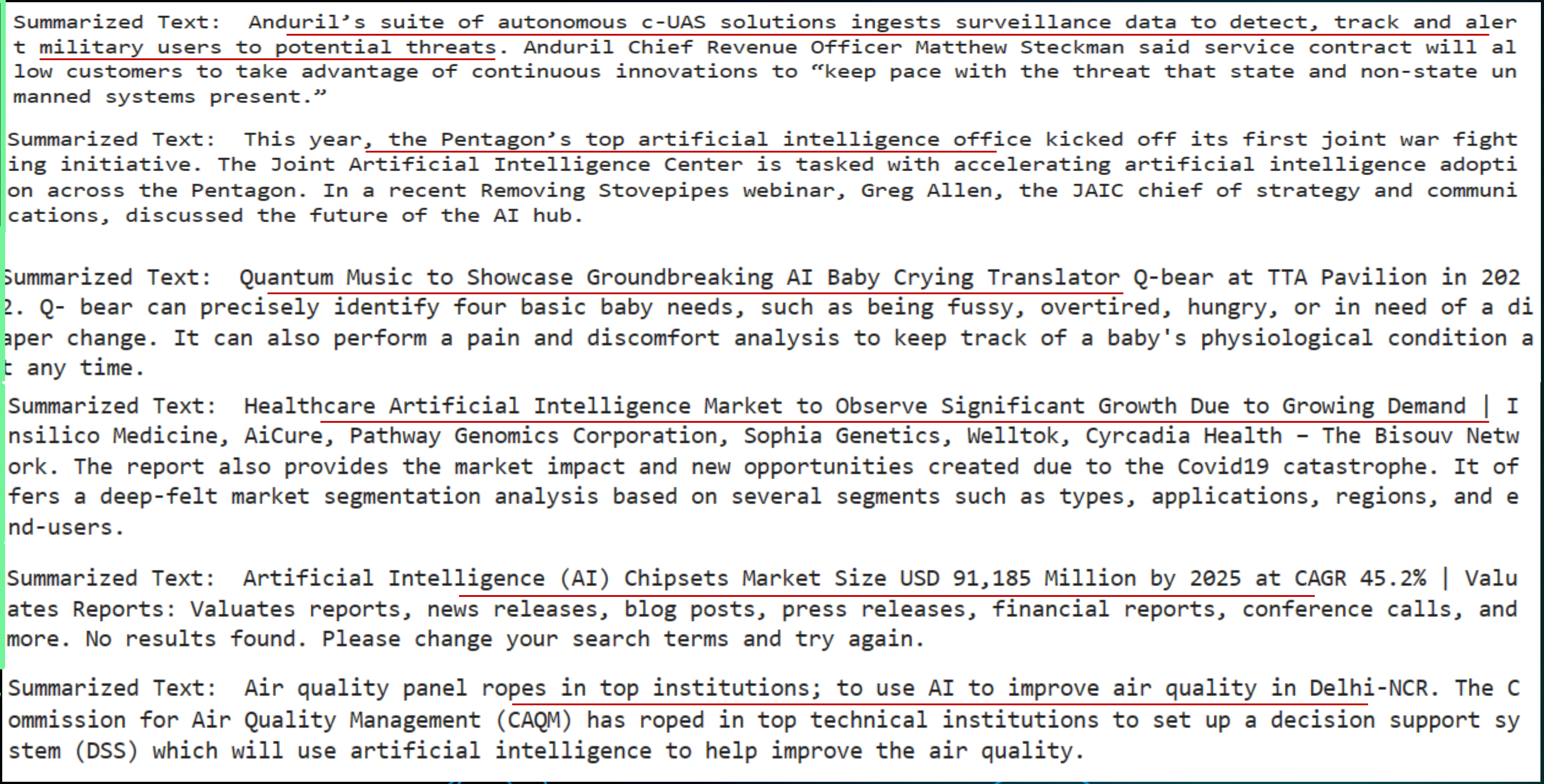


Figure 1

Sample News Articles highlighting success stories of AI (Using K-train Summarization):



DEFENSE

```
Summarized Text:  Anduril’s suite of autonomous c-UAS solutions ingests surveillance data to detect, track and aler
t military users to potential threats. Anduril Chief Revenue Officer Matthew Steckman said service contract will al
low customers to take advantage of continuous innovations to “keep pace with the threat that state and non-state un
manned systems present.”
```

```
Summarized Text:  This year, the Pentagon’s top artificial intelligence office kicked off its first joint war fight
ing initiative. The Joint Artificial Intelligence Center is tasked with accelerating artificial intelligence adopti
on across the Pentagon. In a recent Removing Stovepipes webinar, Greg Allen, the JAIC chief of strategy and communi
cations, discussed the future of the AI hub.
```

MUSIC

```
Summarized Text:  Quantum Music to Showcase Groundbreaking AI Baby Crying Translator Q-bear at TTA Pavilion in 202
2. Q- bear can precisely identify four basic baby needs, such as being fussy, overtired, hungry, or in need of a di
aper change. It can also perform a pain and discomfort analysis to keep track of a baby's physiological condition a
t any time.
```

HEALTH

```
Summarized Text:  Healthcare Artificial Intelligence Market to Observe Significant Growth Due to Growing Demand | I
nsilico Medicine, AiCure, Pathway Genomics Corporation, Sophia Genetics, Welltok, Cyrcadia Health – The Bisouv Netw
ork. The report also provides the market impact and new opportunities created due to the Covid19 catastrophe. It of
fers a deep-felt market segmentation analysis based on several segments such as types, applications, regions, and e
nd-users.
```

CHIPS

```
Summarized Text:  Artificial Intelligence (AI) Chipsets Market Size USD 91,185 Million by 2025 at CAGR 45.2% | Valu
ates Reports: Valuates reports, news releases, blog posts, press releases, financial reports, conference calls, and
more. No results found. Please change your search terms and try again.
```

```
Summarized Text:  Air quality panel ropes in top institutions; to use AI to improve air quality in Delhi-NCR. The C
ommission for Air Quality Management (CAQM) has roped in top technical institutions to set up a decision support sy
stem (DSS) which will use artificial intelligence to help improve the air quality.
```

B. Application Failures in AI

- Applications that may result in downfall of AI are:
- Data Collection without permissions (images, videos etc.)
- Politics
- Deepfakes
- Learning for students as students may cheat to pass the exams in schools
- Health Data breaches
- Data Privacy breaches

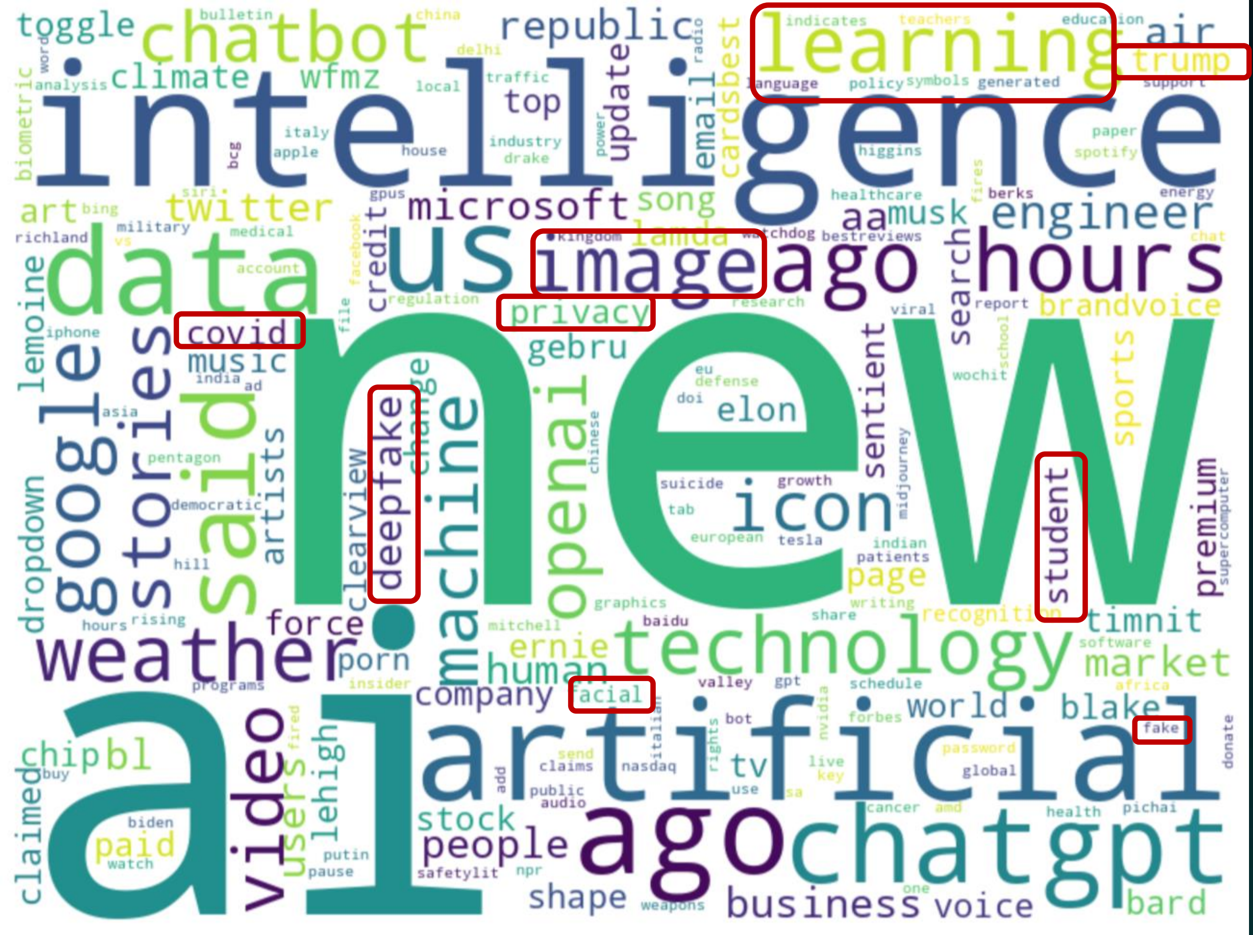
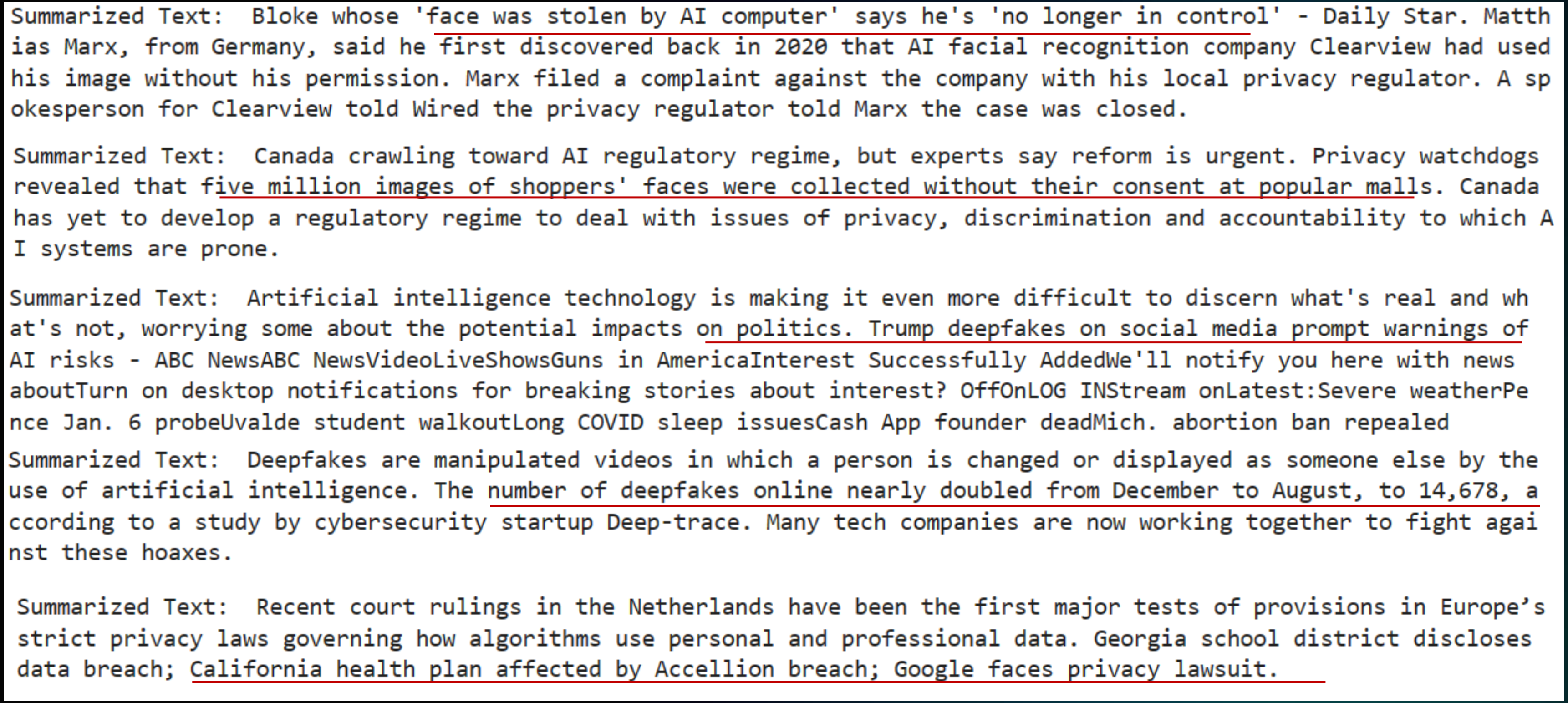


Figure 2

Sample News Articles highlighting failures in AI (Using K-train Summarization):



```
Summarized Text:  Bloke whose 'face was stolen by AI computer' says he's 'no longer in control' - Daily Star. Matth
ias Marx, from Germany, said he first discovered back in 2020 that AI facial recognition company Clearview had used
his image without his permission. Marx filed a complaint against the company with his local privacy regulator. A sp
okesperson for Clearview told Wired the privacy regulator told Marx the case was closed.

Summarized Text:  Canada crawling toward AI regulatory regime, but experts say reform is urgent. Privacy watchdogs
revealed that five million images of shoppers' faces were collected without their consent at popular malls. Canada
has yet to develop a regulatory regime to deal with issues of privacy, discrimination and accountability to which A
I systems are prone.

Summarized Text:  Artificial intelligence technology is making it even more difficult to discern what's real and wh
at's not, worrying some about the potential impacts on politics. Trump deepfakes on social media prompt warnings of
AI risks - ABC NewsABC NewsVideoLiveShowsGuns in AmericaInterest Successfully AddedWe'll notify you here with news
aboutTurn on desktop notifications for breaking stories about interest? OffOnLOG INStream onLatest:Severe weatherPe
nce Jan. 6 probeUvalde student walkoutLong COVID sleep issuesCash App founder deadMich. abortion ban repealed
Summarized Text:  Deepfakes are manipulated videos in which a person is changed or displayed as someone else by the
use of artificial intelligence. The number of deepfakes online nearly doubled from December to August, to 14,678, a
ccording to a study by cybersecurity startup Deep-trace. Many tech companies are now working together to fight agai
nst these hoaxes.

Summarized Text:  Recent court rulings in the Netherlands have been the first major tests of provisions in Europe’s
strict privacy laws governing how algorithms use personal and professional data. Georgia school district discloses
data breach; California health plan affected by Accellion breach; Google faces privacy lawsuit.
```

Data Pre-Processing

03

UNDERSTANDING TOKEN DISTRIBUTION IN NEWS ARTICLES

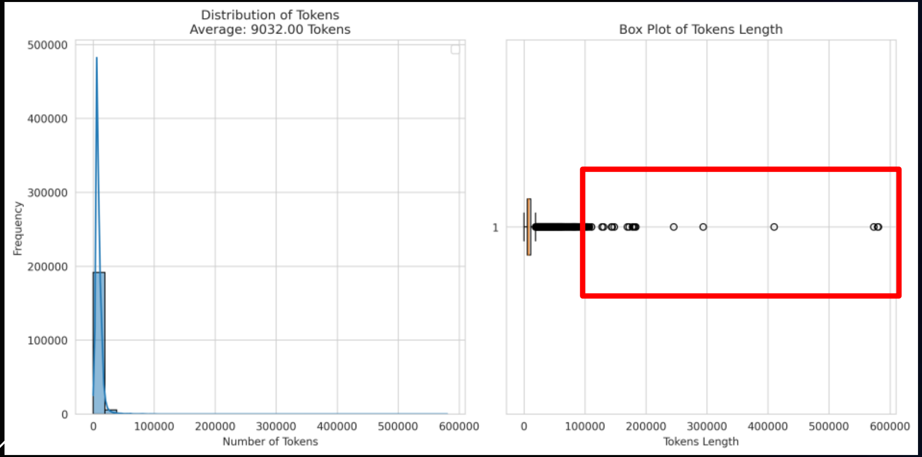


Figure 3

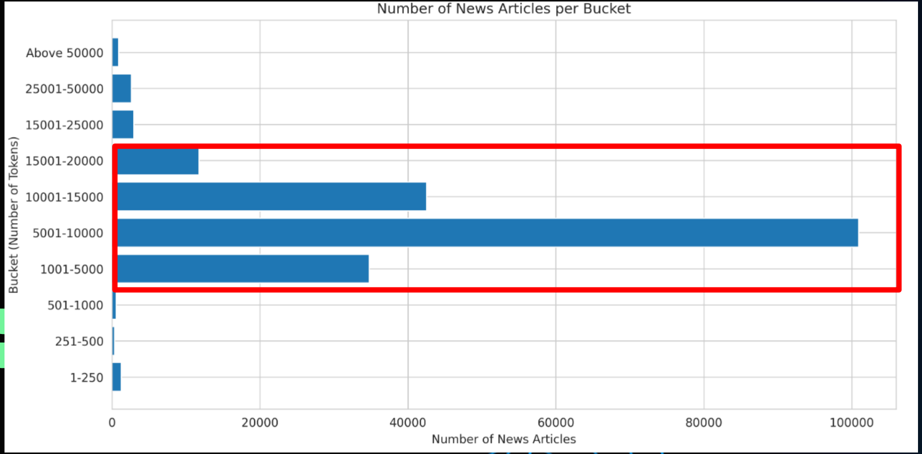


Figure 4

- Initial dataset of ~200K articles with zero missing values.
- Token Distribution in the news articles is highly skewed in nature because of the possible outliers in news articles having number of tokens over 100K. (Figure 1)
- Thus, bucketized the distribution of tokens in news articles to understand underlying the proportion.
- It was found that ~96% of news articles have token distribution between 1000 and 20000. (Figure 2)

DATA PROCESSING FLOW

STAGE 1

200K

data points

Initial dataset of News Articles

STAGE 2

198K

data points

Implemented following steps:

1. Data Cleaning
2. Lemmatization
3. Removing Duplicates

STAGE 3

190K

data points

- Bucketized the news articles based on number of tokens in text
- Limited the News Articles with minimum number of tokens = 1K & maximum number of tokens = 20K in text

STAGE 4

190K

data points

Prepared a list of relevant keywords using K-means, Word2Vec & Bertopic

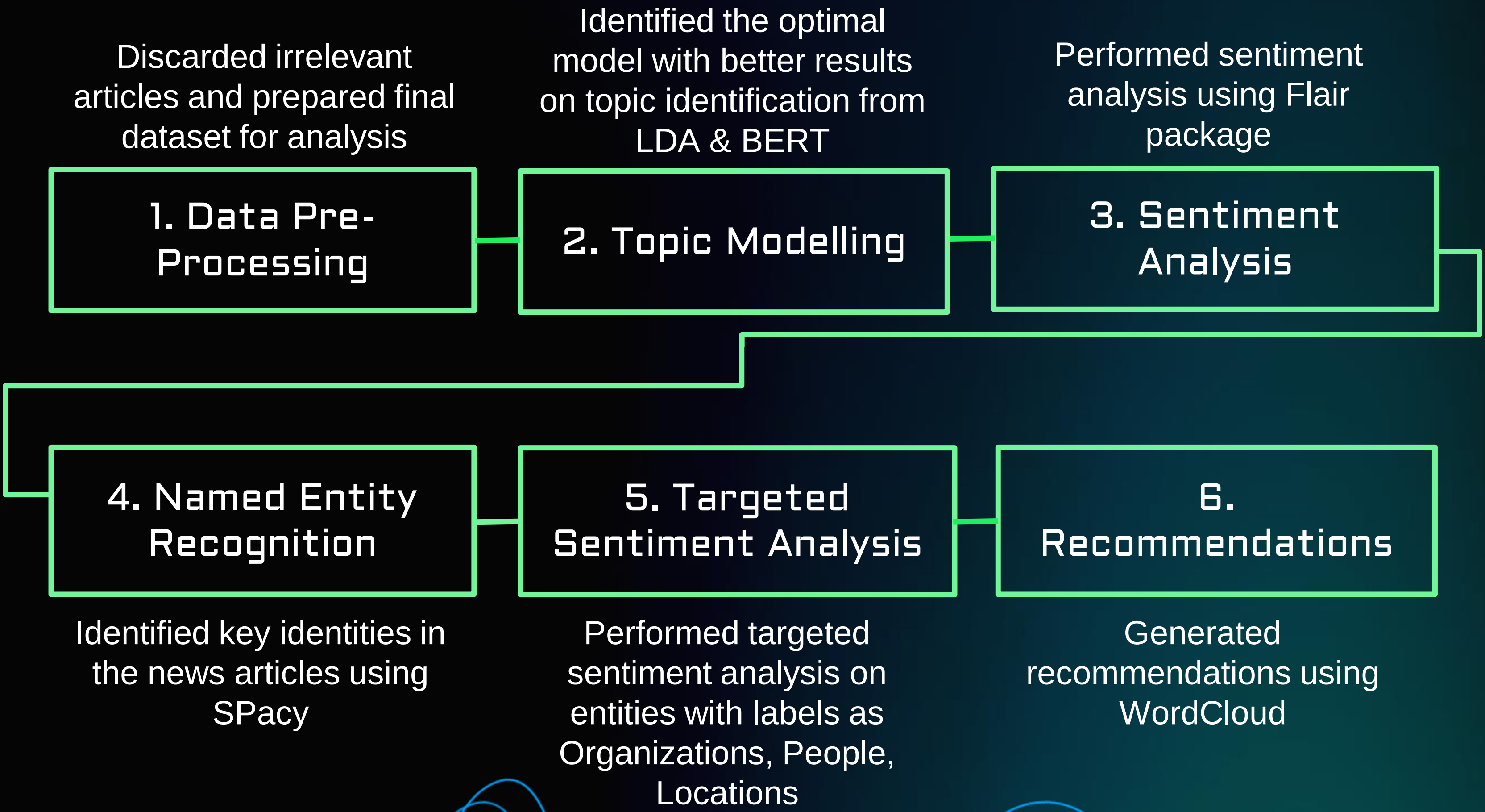
STAGE 5

160K

data points

Filtered the dataset on the prepared keyword list to retrieve the final set of news articles

ANALYSIS PIPELINE



Discarded irrelevant articles and prepared final dataset for analysis

1. Data Pre-Processing

Identified the optimal model with better results on topic identification from LDA & BERT

2. Topic Modelling

Performed sentiment analysis using Flair package

3. Sentiment Analysis

4. Named Entity Recognition

Identified key identities in the news articles using SPacy

5. Targeted Sentiment Analysis

Performed targeted sentiment analysis on entities with labels as Organizations, People, Locations

6. Recommendations

Generated recommendations using WordCloud

Topic Detection

04

IDENTIFYING KEY TOPICS
A. USING LDA

- Discovered the optimal number of topics = 12 using the default parameters of LDA model.

- As can be seen in WordCloud of LDA (Figure 3), main topics were around:
  - ➢ Artificial Intelligence
  - ➢ ChatGPT
  - ➢ Market Growth etc.

- Although topics were related to AI, however, I found it difficult to locate specific industries in the topics.

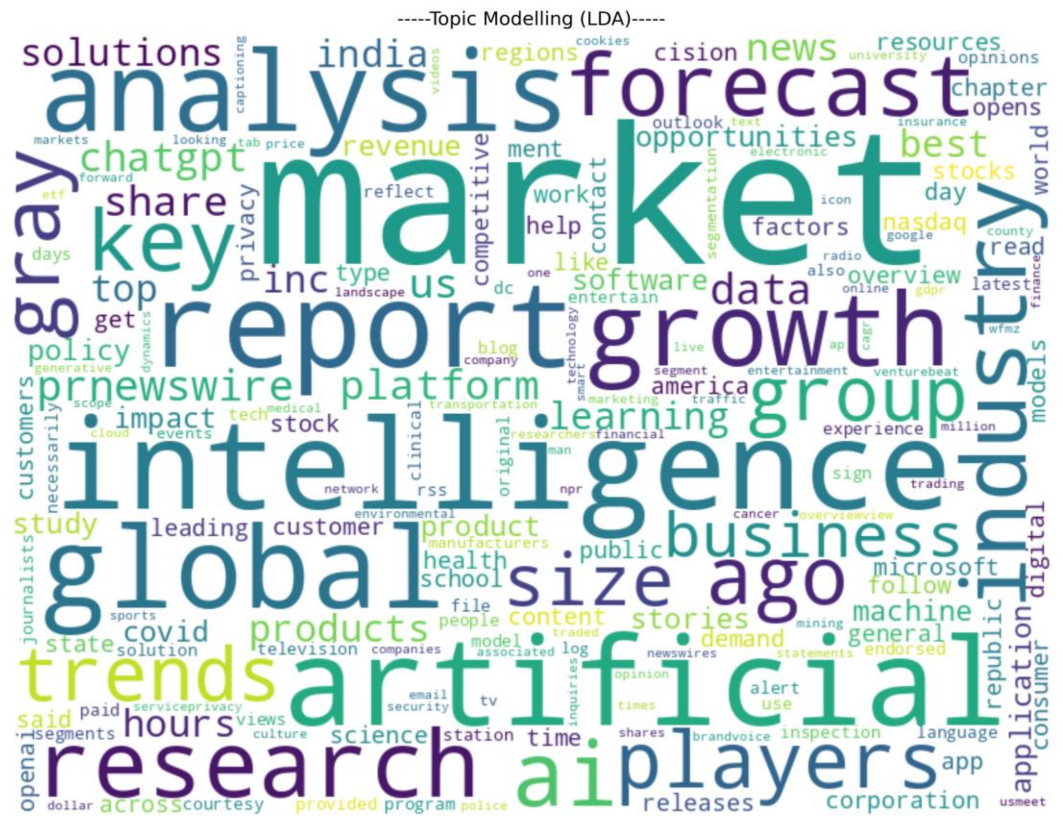


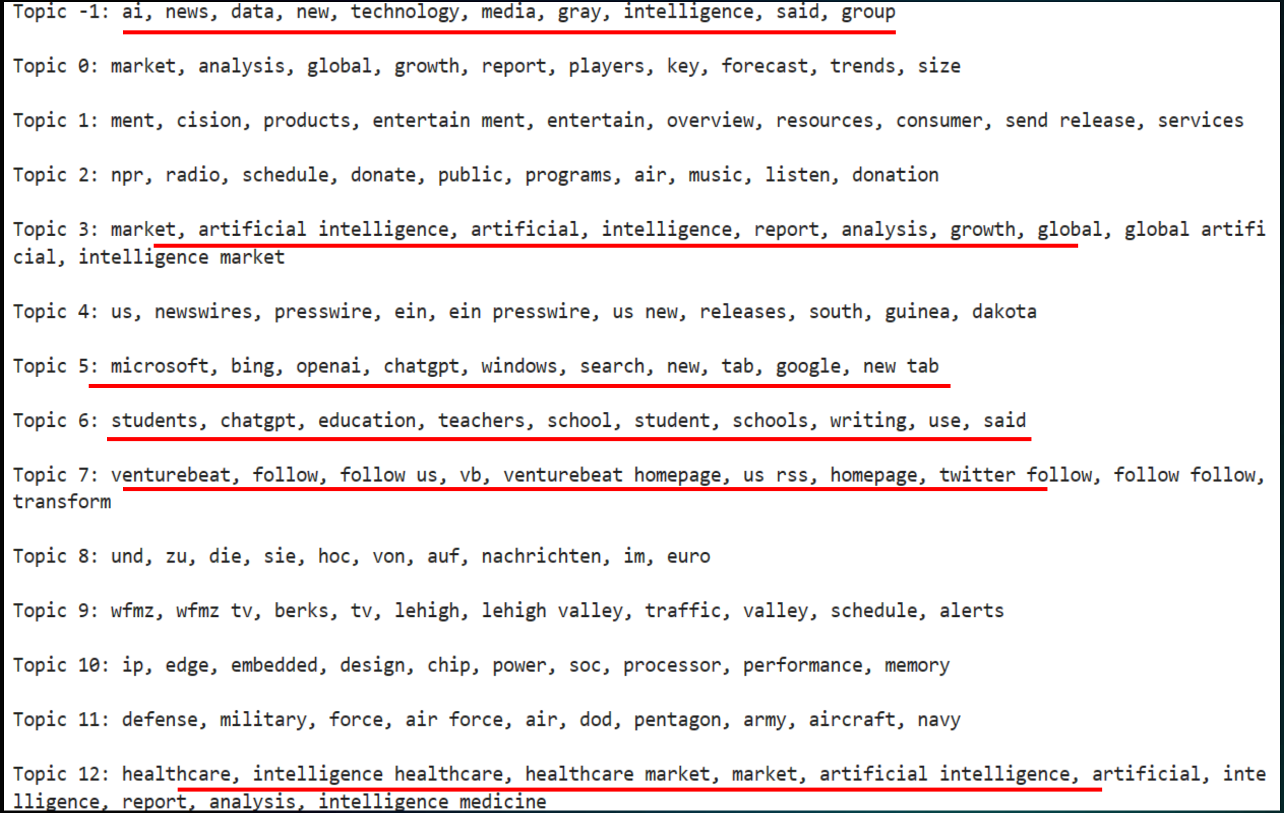
Figure 5

B. USING BERTOPIC

- Another method I explored to identify key topics was using BERT.

- The number of total topics found in the entire corpus of news articles came out to be 267.

- Figure 4 shows some key topics from the result, and it can be noticed that BERT performed better in identifying relevant industries in addition to AI related keywords.



```
Topic -1: ai, news, data, new, technology, media, gray, intelligence, said, group

Topic 0: market, analysis, global, growth, report, players, key, forecast, trends, size

Topic 1: ment, cision, products, entertain ment, entertain, overview, resources, consumer, send release, services

Topic 2: npr, radio, schedule, donate, public, programs, air, music, listen, donation

Topic 3: market, artificial intelligence, artificial, intelligence, report, analysis, growth, global, global artifi
cial, intelligence market

Topic 4: us, newswires, presswire, ein, ein presswire, us new, releases, south, guinea, dakota

Topic 5: microsoft, bing, openai, chatgpt, windows, search, new, tab, google, new tab

Topic 6: students, chatgpt, education, teachers, school, student, schools, writing, use, said

Topic 7: venturebeat, follow, follow us, vb, venturebeat homepage, us rss, homepage, twitter follow, follow follow,
transform

Topic 8: und, zu, die, sie, hoc, von, auf, nachrichten, im, euro

Topic 9: wfmz, wfmz tv, berks, tv, lehigh, lehigh valley, traffic, valley, schedule, alerts

Topic 10: ip, edge, embedded, design, chip, power, soc, processor, performance, memory

Topic 11: defense, military, force, air force, air, dod, pentagon, army, aircraft, navy

Topic 12: healthcare, intelligence healthcare, healthcare market, market, artificial intelligence, artificial, inte
lligence, report, analysis, intelligence medicine
```

Figure 6

TOPIC DETECTION USING BERT

- Identified topics are as follows:
- ➢ artificial intelligence
- ➢ chatgpt
- ➢ bard
- ➢ healthcare
- ➢ openai
- ➢ microsoft
- ➢ google
- ➢ machine learning
- ➢ covid
- ➢ media
- ➢ market
- ➢ education
- ➢ elon musk etc.

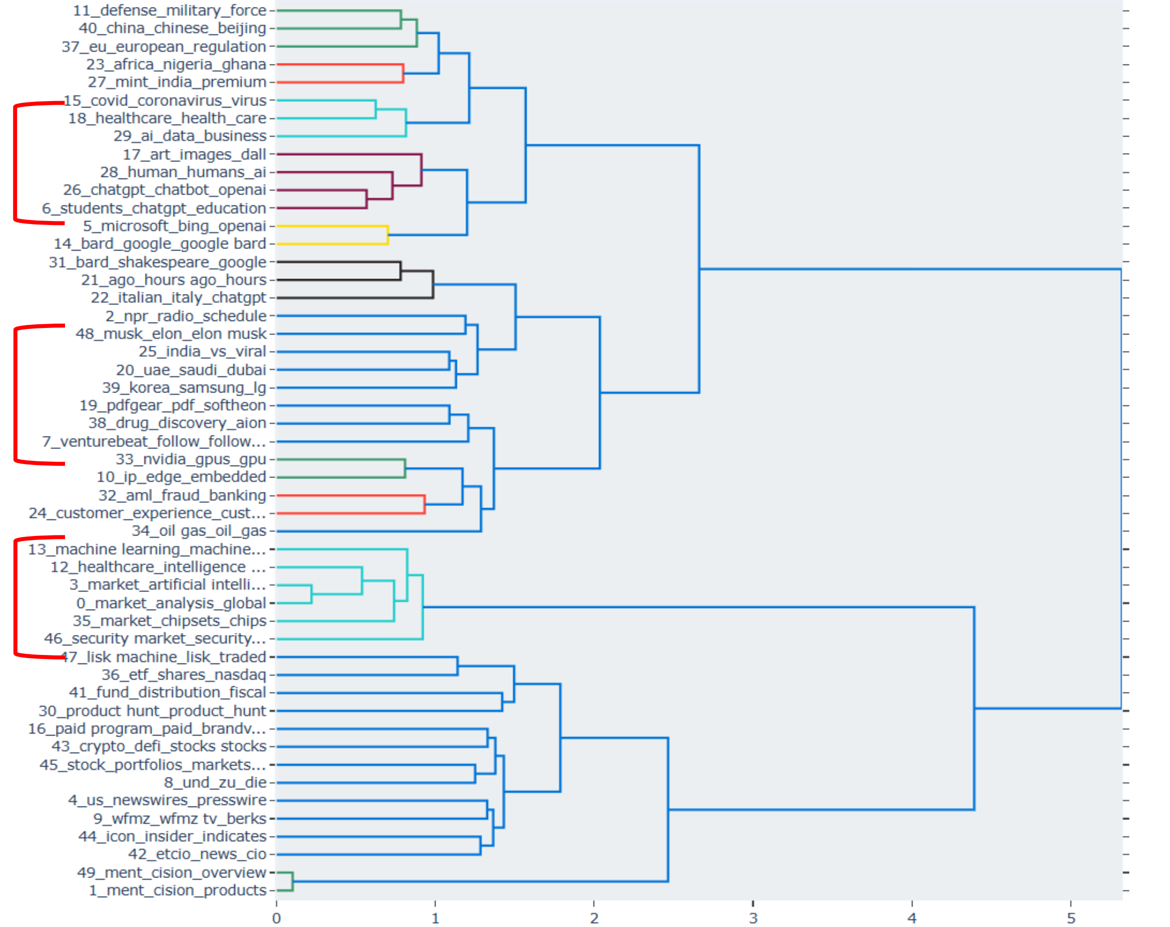
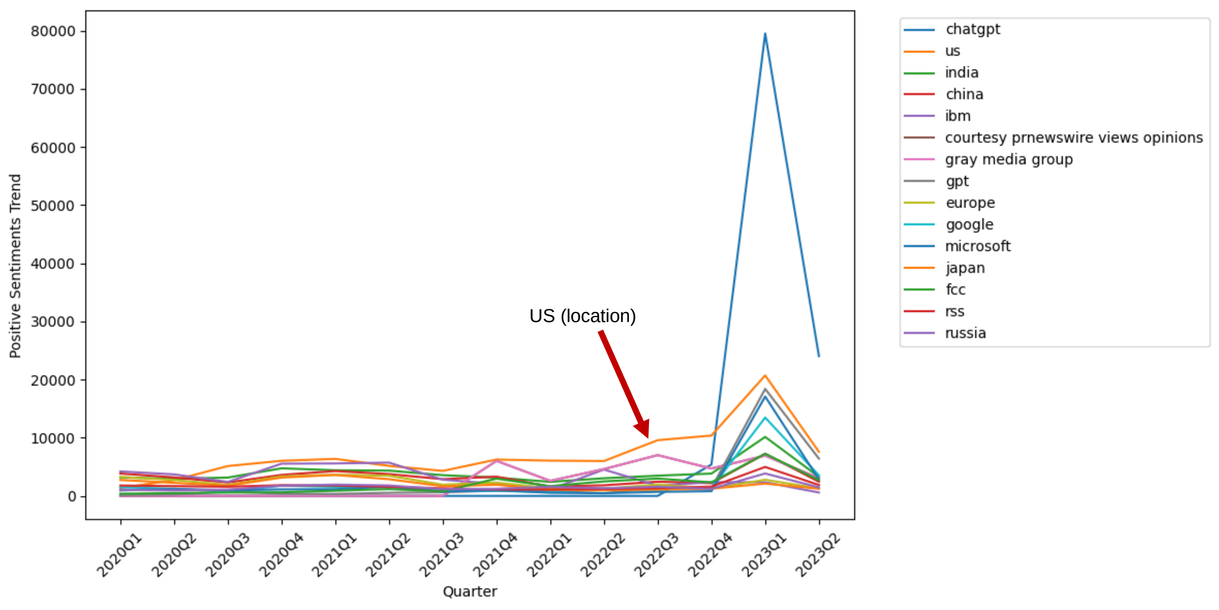


Figure 7

Sentiment Analysis

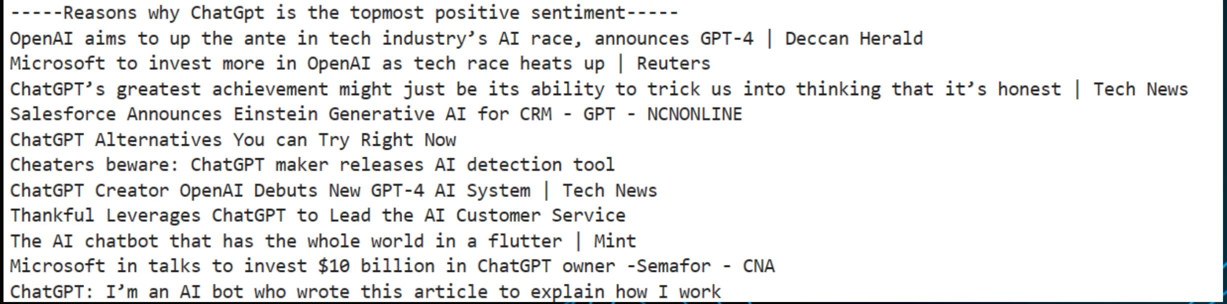
05

A. POSITIVE SENTIMENTS



US (location)

Figure 8



```
-----Reasons why ChatGpt is the topmost positive sentiment-----
OpenAI aims to up the ante in tech industry’s AI race, announces GPT-4 | Deccan Herald
Microsoft to invest more in OpenAI as tech race heats up | Reuters
ChatGPT’s greatest achievement might just be its ability to trick us into thinking that it’s honest | Tech News
Salesforce Announces Einstein Generative AI for CRM - GPT - NCNONLINE
ChatGPT Alternatives You can Try Right Now
Cheaters beware: ChatGPT maker releases AI detection tool
ChatGPT Creator OpenAI Debuts New GPT-4 AI System | Tech News
Thankful Leverages ChatGPT to Lead the AI Customer Service
The AI chatbot that has the whole world in a flutter | Mint
Microsoft in talks to invest $10 billion in ChatGPT owner -Semafor - CNA
ChatGPT: I’m an AI bot who wrote this article to explain how I work
```

Figure 9

- Using Flair Package, ~83% articles appeared to have positive sentiments and remaining 17% had negative sentiments.

- In the positive sentiments' dataset, it was found that ‘chatgpt’ is the topmost positive sentiment talked about specially between the 4th quarter of the year 2022 and 2nd quarter of year 2023.

- Further, US location started to trend from similar timeframe. (Figure 6)

- There are multiple reasons as can been in Figure 7 for chatgpt to appear as most positive sentiment, which mainly talk about the AI revolution it brought in the whole world.

B. NEGATIVE SENTIMENTS

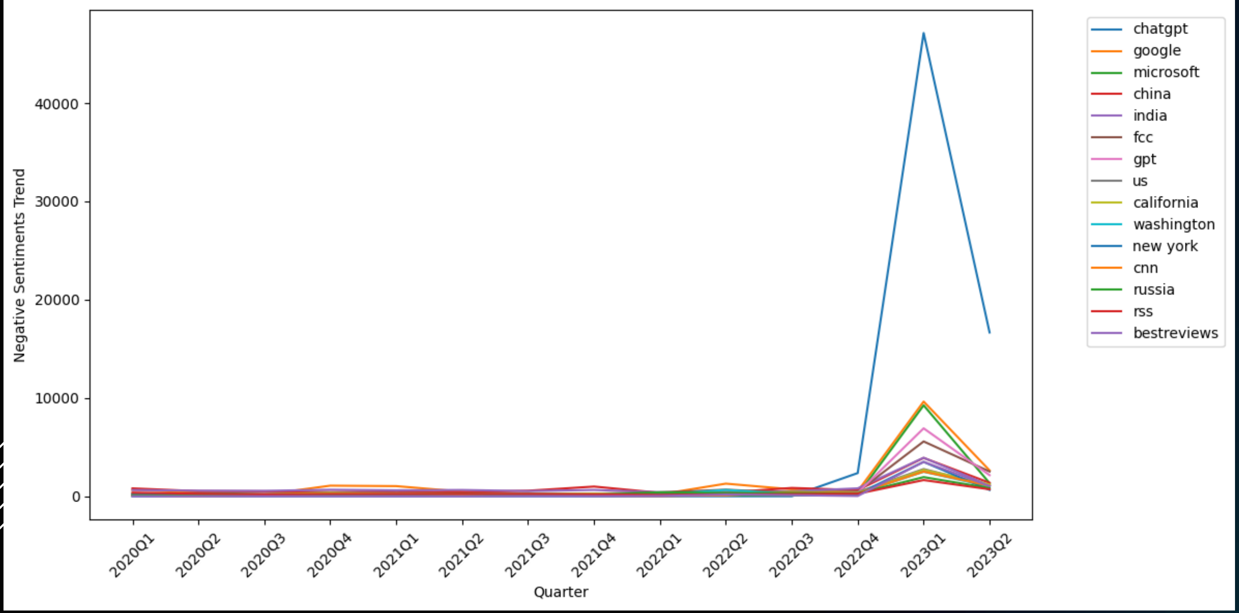
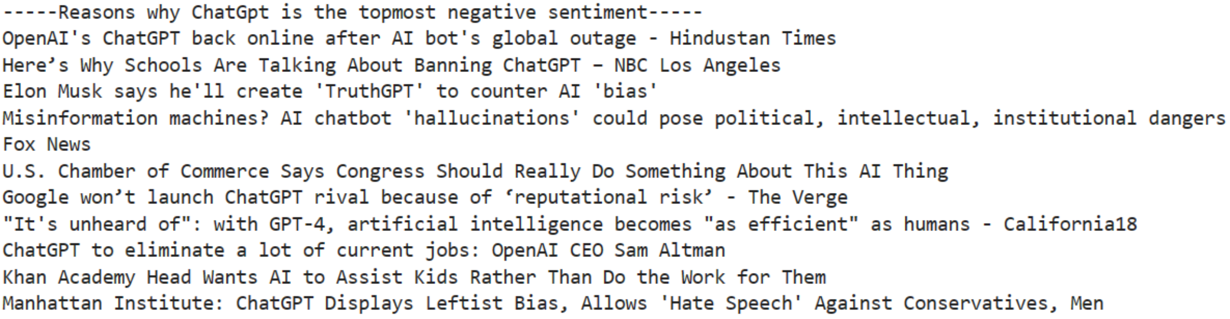


Figure 10



```
-----Reasons why ChatGpt is the topmost negative sentiment-----
OpenAI's ChatGPT back online after AI bot's global outage - Hindustan Times
Here’s Why Schools Are Talking About Banning ChatGPT – NBC Los Angeles
Elon Musk says he'll create 'TruthGPT' to counter AI 'bias'
Misinformation machines? AI chatbot 'hallucinations' could pose political, intellectual, institutional dangers
Fox News
U.S. Chamber of Commerce Says Congress Should Really Do Something About This AI Thing
Google won’t launch ChatGPT rival because of ‘reputational risk’ - The Verge
"It's unheard of": with GPT-4, artificial intelligence becomes "as efficient" as humans - California18
ChatGPT to eliminate a lot of current jobs: OpenAI CEO Sam Altman
Khan Academy Head Wants AI to Assist Kids Rather Than Do the Work for Them
Manhattan Institute: ChatGPT Displays Leftist Bias, Allows 'Hate Speech' Against Conservatives, Men
```

Figure 11

- In the negative sentiments' dataset, again ‘chatgpt’ is the topmost sentiment talked about specially between the 4th quarter of the year 2022 and 2nd quarter of year 2023. (Figure 8)

- Negative sentiment for chatgpt were associated with its outage issues, how it might impact student’s learning and schools thinking of banning its use, elimination of current jobs with chatgpt, allowing hate speeches etc. (Figure 9)

C. SENTIMENTS OVER TIME

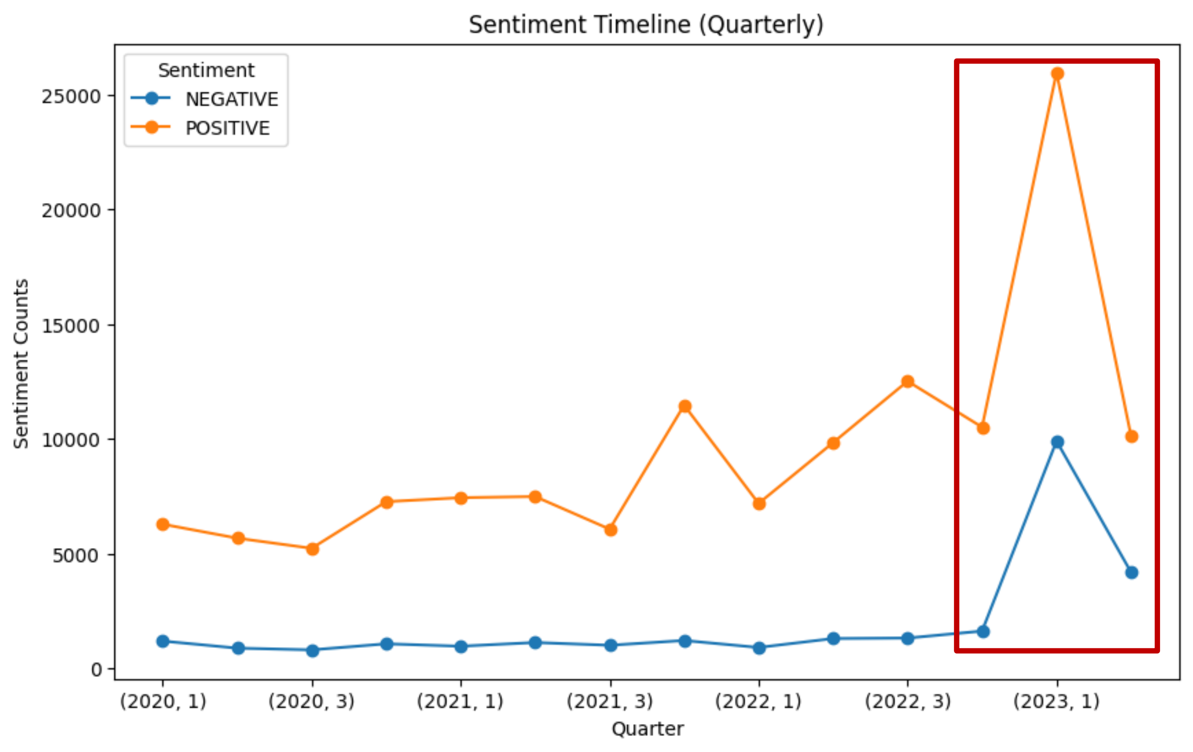


Figure 12

- Overall sentiments were trending low until the boom of chatgpt in the quarter 4 of year 2022, for which the sentiments were expressed more both in positive and negative aspect.

AI Solutions Trends over Time

06

AI Solutions Trends over Time

- AI solutions that are trending from 2020 can be seen in Figure 11. All products are showing stable trend, except Bing which peaked in the first quarter of 2023.

- This maybe due to several reasons (like in Figure 12), however, the major reason could be the chatgpt revolution, in which Microsoft invested and added AI features to Bing.

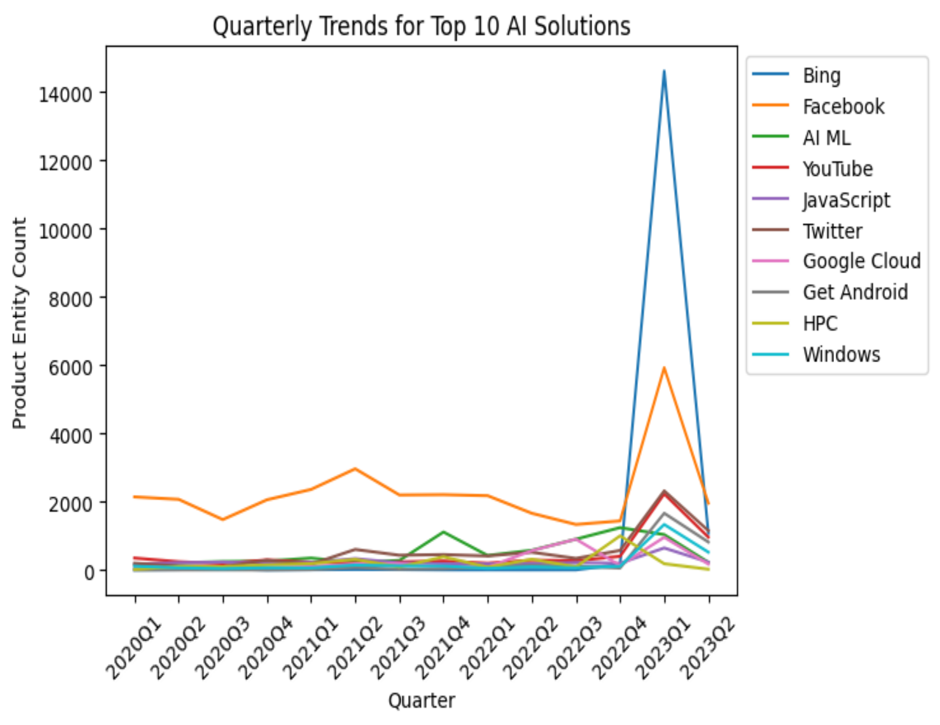
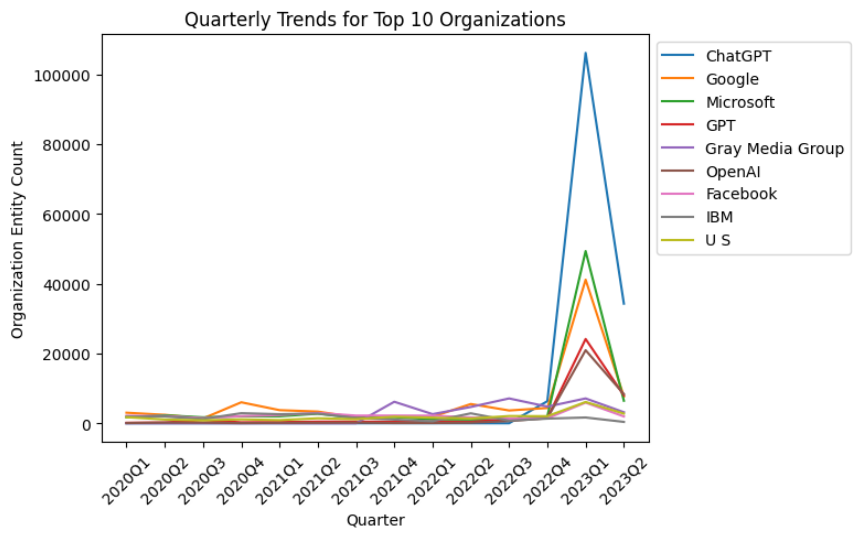


Figure 13

Entity Identification

07

Top Entity Discovery – Organizations & People



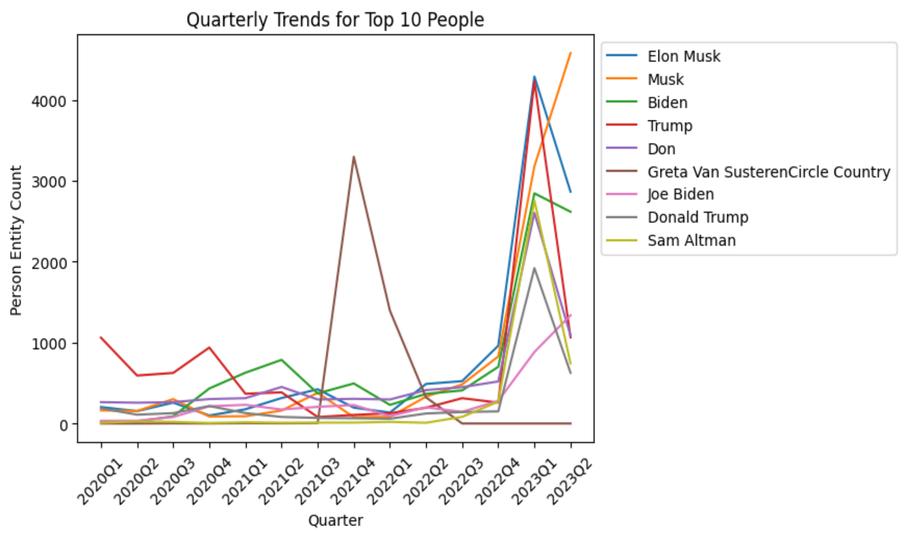


Figure 14

Figure 15

Top organizations (ChatGPT, Google, Microsoft) confirms the presence of AI boom that happened in the last quarter of 2022.

Musk remain on the top since the ChatGPT boom started, along with political personalities and CEO of OpenAI.

Top Entity Discovery - Locations

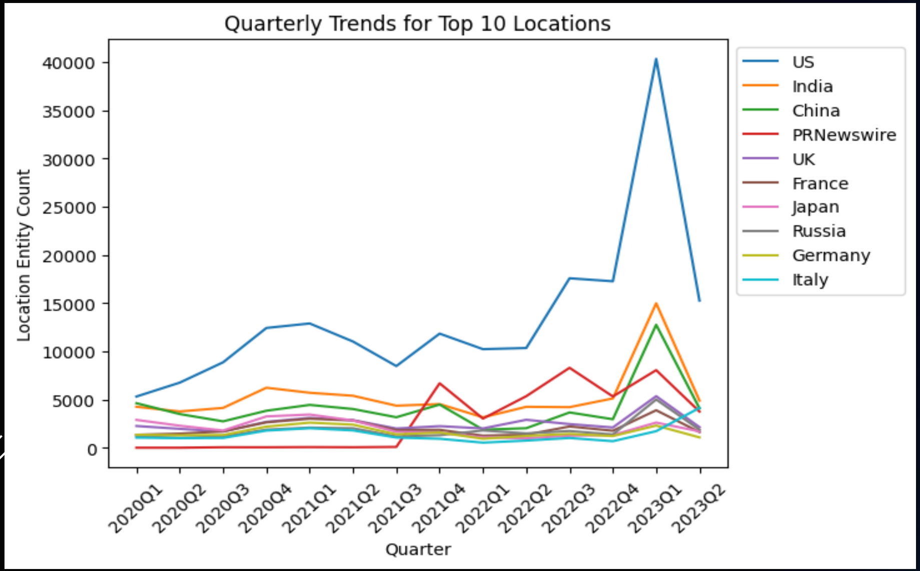
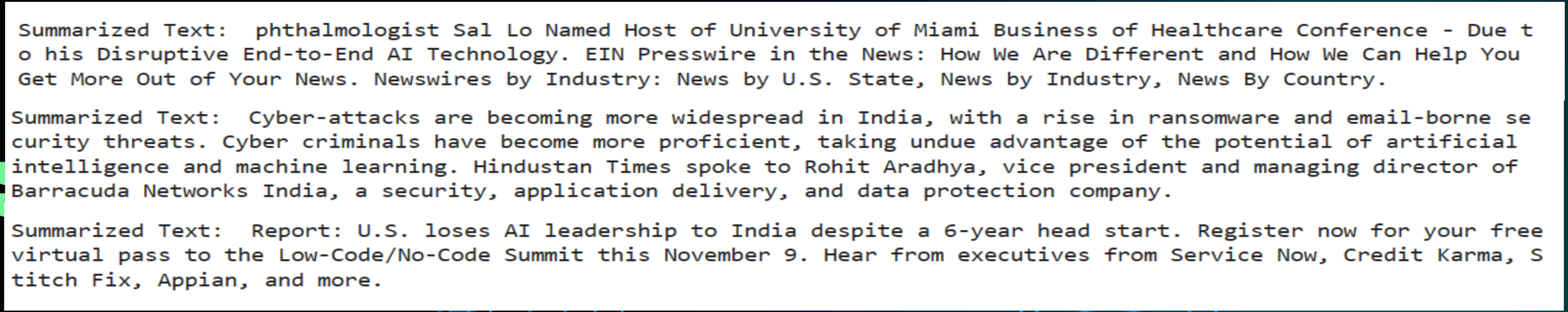


Figure 16

- Top locations include US, India & China with US being on the top. (Figure 17)
- Further, we can see in summarization (Figure 18), how these locations are related to AI technology, cyber attacks etc.

Sample News Articles for Locations (Using K-train Summarization):



```
Summarized Text:  phthalmologist Sal Lo Named Host of University of Miami Business of Healthcare Conference - Due to his Disruptive End-to-End AI Technology. EIN Presswire in the News: How We Are Different and How We Can Help You Get More Out of Your News. Newswires by Industry: News by U.S. State, News by Industry, News By Country.

Summarized Text:  Cyber-attacks are becoming more widespread in India, with a rise in ransomware and email-borne security threats. Cyber criminals have become more proficient, taking undue advantage of the potential of artificial intelligence and machine learning. Hindustan Times spoke to Rohit Aradhya, vice president and managing director of Barracuda Networks India, a security, application delivery, and data protection company.

Summarized Text:  Report: U.S. loses AI leadership to India despite a 6-year head start. Register now for your free virtual pass to the Low-Code/No-Code Summit this November 9. Hear from executives from Service Now, Credit Karma, Stitch Fix, Appian, and more.
```

Figure 17

Targeted Sentiment Identification

08

Targeted Sentiment Analysis
Elon Musk | Sam Altman

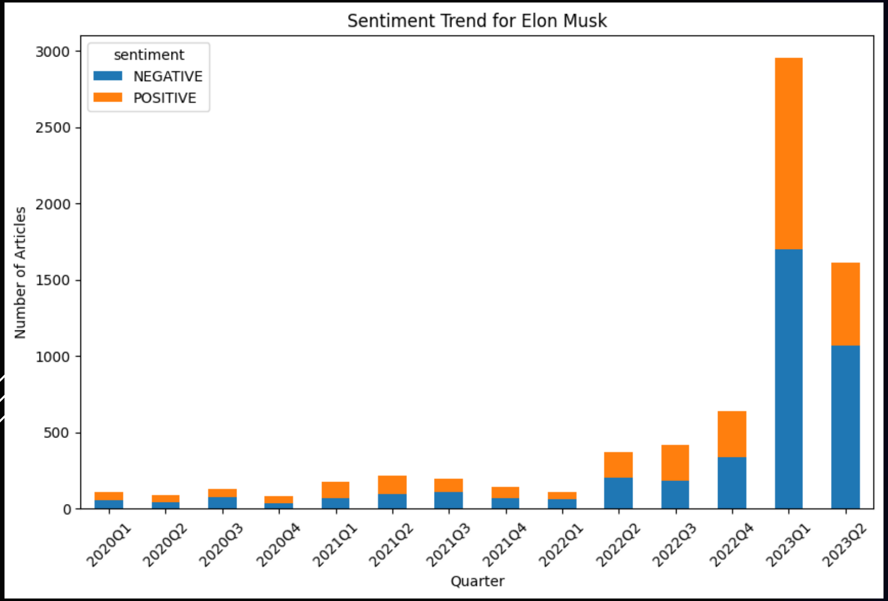


Figure 18

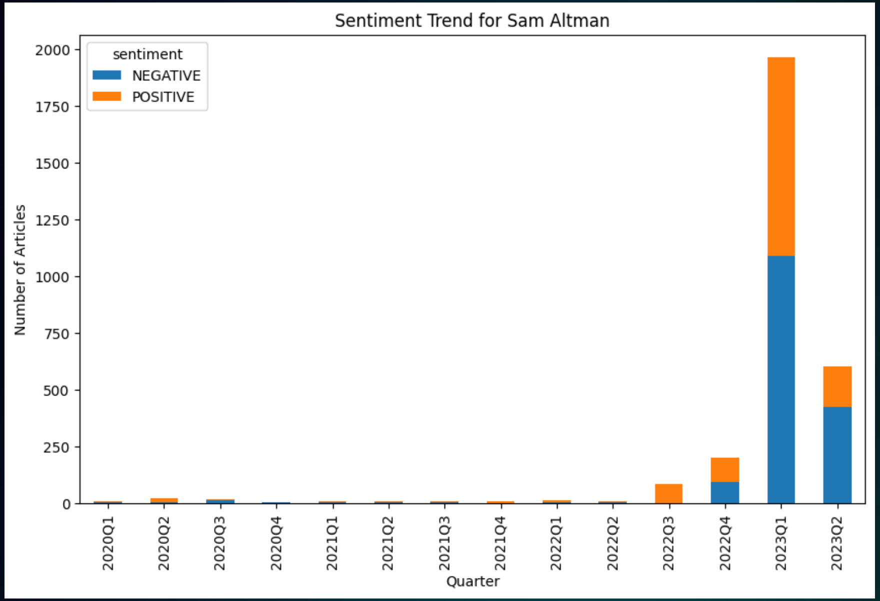


Figure 19

- Elon Musk always trending in news, and sentiments peaked for him in 2023.
- On the other hand, Sam Altman (CEO of OpenAI) was hardly in news before ChatGPT AI revolution. However, with the advent of OpenAI, news articles mentioning both positive & negative sentiments related to him increased exponentially. Overall Sentiments reduced for him in second quarter of 2023, but negative sentiments remained more than positive.

Targeted Sentiment Analysis
Google | Microsoft

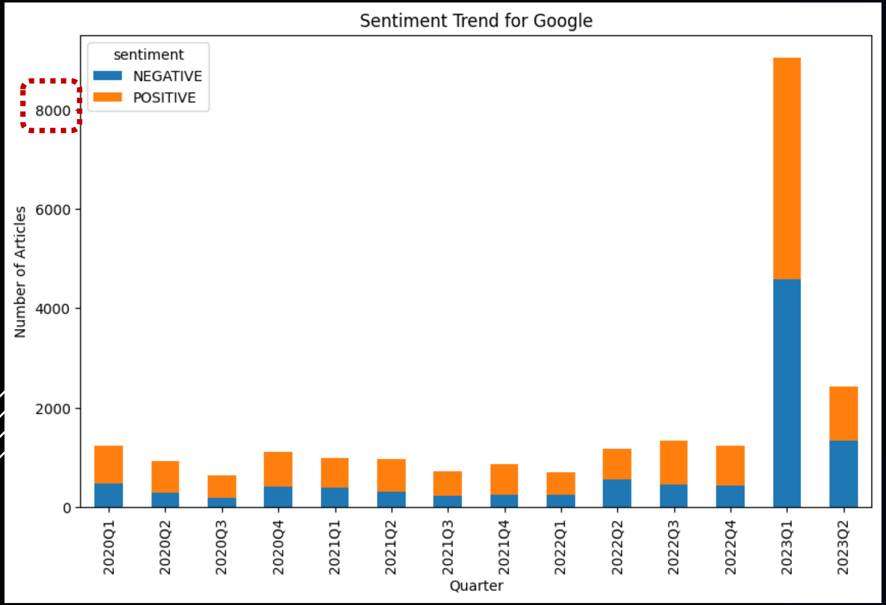


Figure 20

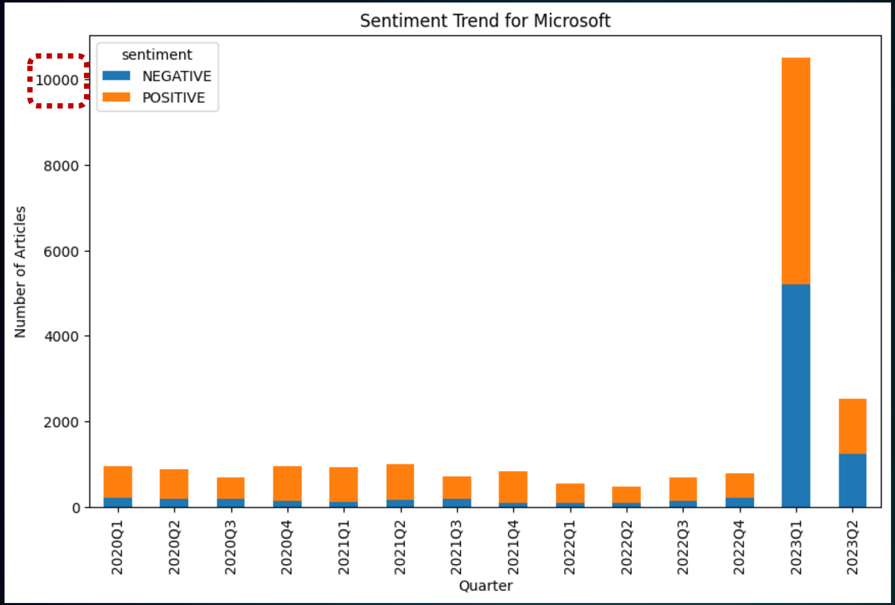


Figure 21

- Microsoft was mentioned a little less than Google in news, but in the first quarter of 2023, its popularity increased because of huge investment in OpenAI and it was mentioned more in ~2k news both for positive and negative sentiments than Google.

Targeted Sentiment Analysis
ChatGPT | Bard

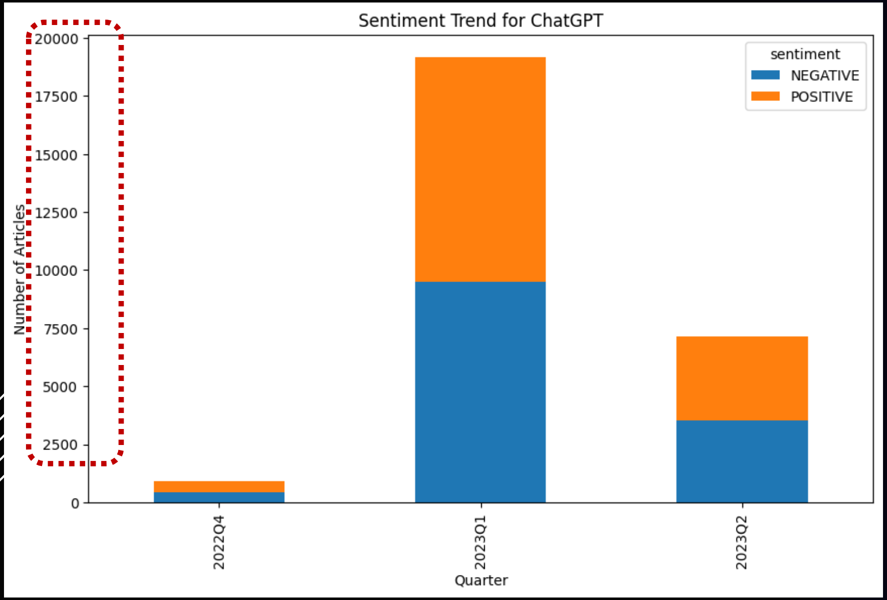


Figure 22

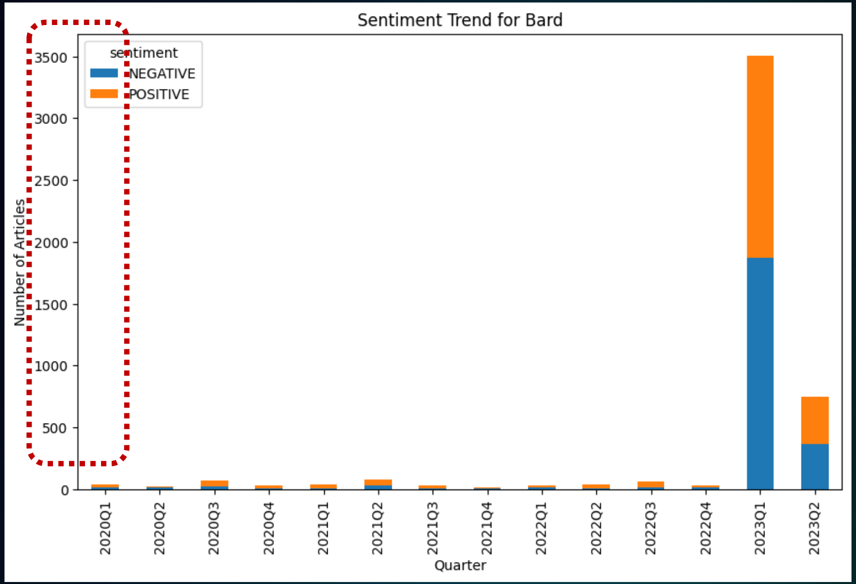


Figure 23

- This comparison is very interesting since ChatGPT was not mentioned at all before 4th quarter of 2022 and as soon as it became a popular demand in 2023, Google released Bard as a response (evident from the sentiments plot).
- Even after the release of Bard, ChatGPT still remained majorly in news articles with upper limit of ~20K news articles.

THANK YOU