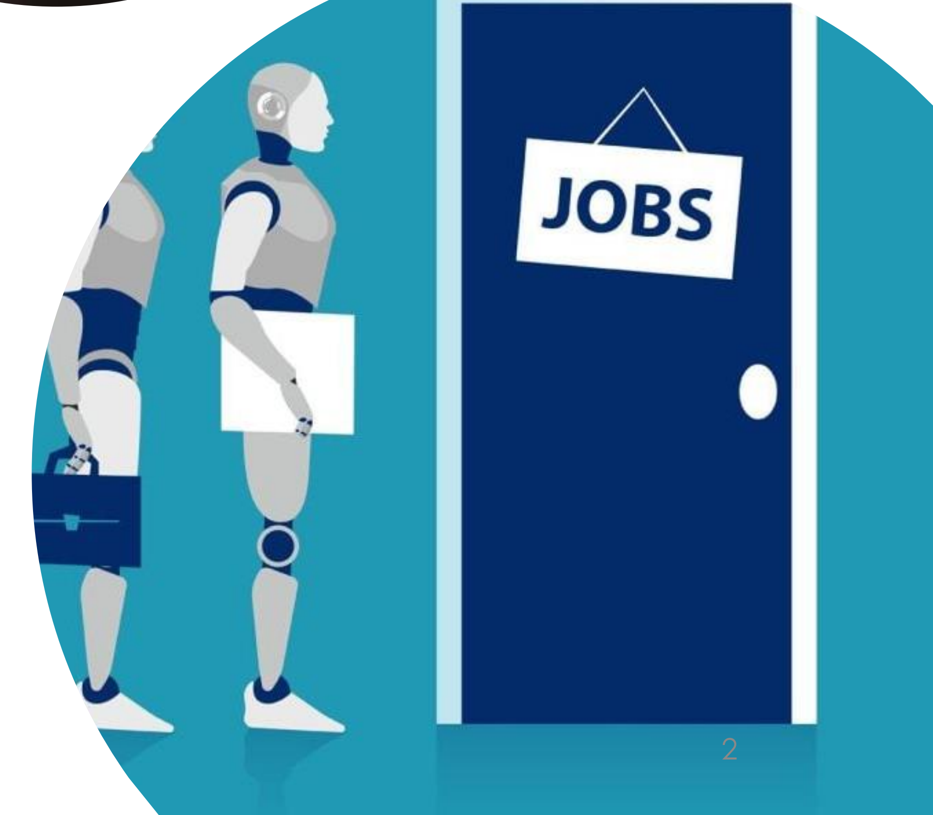
Final Project Data Science – NLP

Neal Xu

Table of Content

- **Executive Summary**
- **Source Data Overview**
- **Article Clean-Up and Filtering**
- **Topic Detection**
  - I. Methodology
  - II. Summary
- **Sentiment Analysis**
  - I. Methodology & Summary
  - II. Sentiment Over Time
- **Entity identification – Organizations & People**
- **Targeted Entity Sentiment**
- **Appendix (High Resolution Figures)**



Executive Summary

Problem Statement:

This project is designed to identify what types of tasks and jobs are most likely to see the biggest impact from AI by extracting meaningful insights from unstructured text and hereby provide recommendations

Background:

According to a report from Goldman Sachs, specific job categories such as office tasks, legal, architecture, and social sciences have the potential for over 30% automation. This means that a significant portion of tasks within these fields could be automated using AI and other technologies.

Assumption:

Not all jobs will be affected by the AI, positions such as construction, installation, and building maintenance are projected to be largely unaffected by automation. These roles typically involve hands-on work that requires physical presence and manual skills, which are harder to automate using current technologies.

Executive Summary – Continued

Findings

Most of **news** article has a slightly **positive** tongue when talking about jobs and data science AI Technology. While **AI** technology has **shown significant** advancements in most **industries** and positive sentiment overall, there are certain types of applications that **currently face limitations** in terms of AI transformation based on the news analyzed in this project.

Two industries that **stand out** are **stock** market analysis and **agriculture**. In stock market analysis, factors such as quarter results, ETFs, and stock prices still heavily rely on human decision-making, making it challenging for AI to accurately predict and transform this domain. Similarly, the agriculture industry requires an understanding of growth analysis, revenue, key players, or the actual labor which makes it difficult for AI to fully transform this sector.

Actionable recommendations

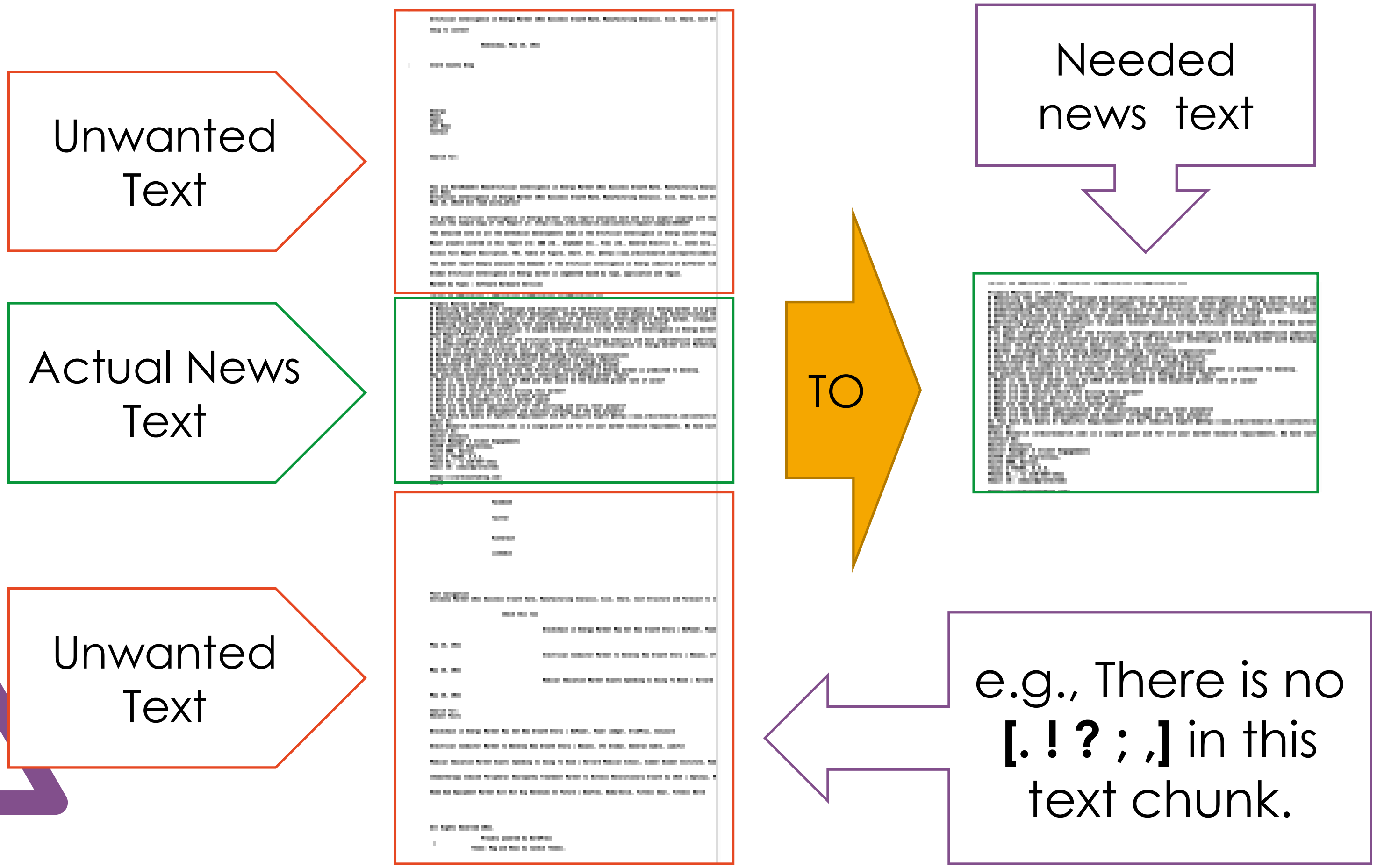
The **recommendation** about what can be done with AI to automate jobs and/or improve employee productivity. Companies looking to invest in data science initiatives today or in the near future focus on **industries** where **AI has demonstrated success and significant potential**.

Technical companies, in particular, can **benefit** from **investing** in data science AI initiatives due to **advantages** such as competitive advantage, enhanced decision-making, improved customer experience, higher operational efficiency, cost reduction, innovation, and new revenue streams.

However, it is crucial to still **consider** the **limitations** of **AI** in certain industries and **prioritize** areas where **human** expertise is **indispensable** or where AI may not yet have reached its full potential. Such as the **stock** market and **agriculture**.

Source data overview

In this project, there are 200,000 **unstructured** news articles along with additional information such as **URL**, **date**, **language**, **title**, and **text**.



Unwanted Text

Actual News Text

Unwanted Text

TO

Needed news text

e.g., There is no **[. ! ? ; ,]** in this text chunk.

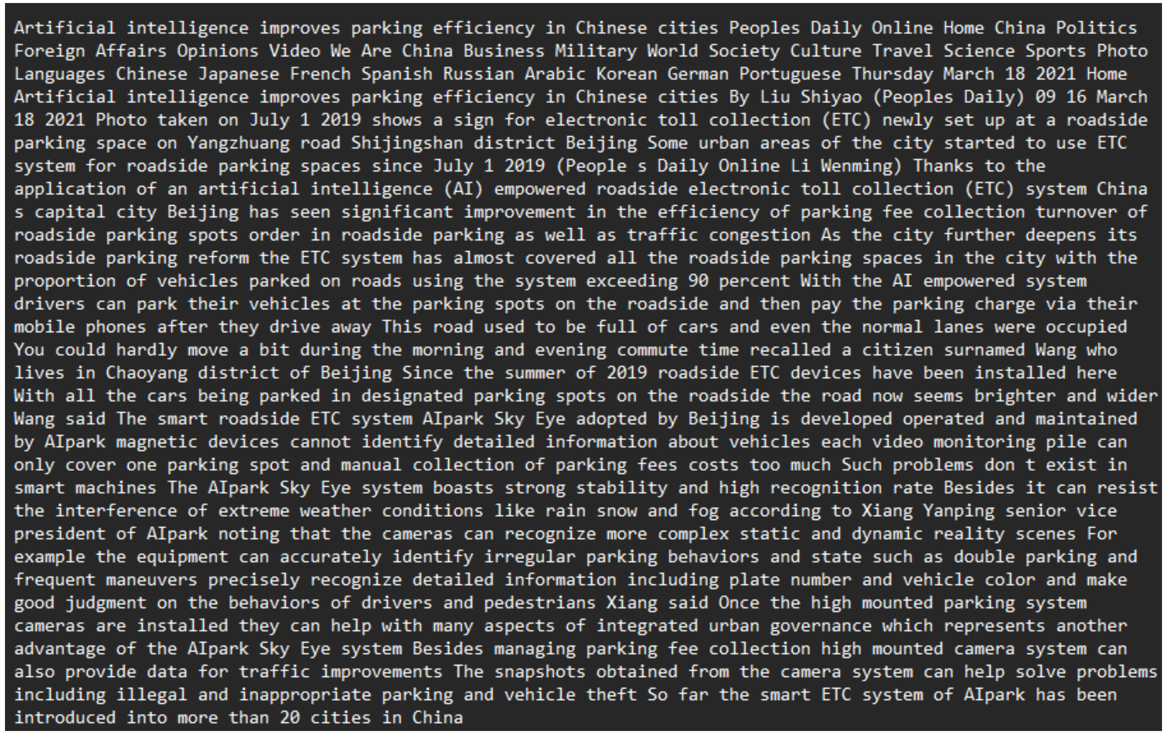
WHY?
It is important to remove irrelevant words before we do any additional analysis such as topic modeling and sentiment analysis.

HOW?
The method used to remove the irrelevant chunk is by utilizing punctuations.

- A news text will have **[. ! ? ; ,]** as a punctuation at most every 60-100 words.
- But for most unwanted text are not separated by **[. ! ? ; ,]**
- So, we split text by **[. ! ? ; ,]** and remove item with more then 400 characters in it to clean chunks.

Article clean-up and filtering

Example of Cleaned New Text by following the Steps on the right-hand side:



```
Artificial intelligence improves parking efficiency in Chinese cities Peoples Daily Online Home China Politics
Foreign Affairs Opinions Video We Are China Business Military World Society Culture Travel Science Sports Photo
Languages Chinese Japanese French Spanish Russian Arabic Korean German Portuguese Thursday March 18 2021 Home
Artificial intelligence improves parking efficiency in Chinese cities By Liu Shiyao (Peoples Daily) 09 16 March
18 2021 Photo taken on July 1 2019 shows a sign for electronic toll collection (ETC) newly set up at a roadside
parking space on Yangzhuang road Shijingshan district Beijing Some urban areas of the city started to use ETC
system for roadside parking spaces since July 1 2019 (People s Daily Online Li Wenming) Thanks to the
application of an artificial intelligence (AI) empowered roadside electronic toll collection (ETC) system China
s capital city Beijing has seen significant improvement in the efficiency of parking fee collection turnover of
roadside parking spots order in roadside parking as well as traffic congestion As the city further deepens its
roadside parking reform the ETC system has almost covered all the roadside parking spaces in the city with the
proportion of vehicles parked on roads using the system exceeding 90 percent With the AI empowered system
drivers can park their vehicles at the parking spots on the roadside and then pay the parking charge via their
mobile phones after they drive away This road used to be full of cars and even the normal lanes were occupied
You could hardly move a bit during the morning and evening commute time recalled a citizen surnamed Wang who
lives in Chaoyang district of Beijing Since the summer of 2019 roadside ETC devices have been installed here
With all the cars being parked in designated parking spots on the roadside the road now seems brighter and wider
Wang said The smart roadside ETC system AIpark Sky Eye adopted by Beijing is developed operated and maintained
by AIpark magnetic devices cannot identify detailed information about vehicles each video monitoring pile can
only cover one parking spot and manual collection of parking fees costs too much Such problems don t exist in
smart machines The AIpark Sky Eye system boasts strong stability and high recognition rate Besides it can resist
the interference of extreme weather conditions like rain snow and fog according to Xiang Yanping senior vice
president of AIpark noting that the cameras can recognize more complex static and dynamic reality scenes For
example the equipment can accurately identify irregular parking behaviors and state such as double parking and
frequent maneuvers precisely recognize detailed information including plate number and vehicle color and make
good judgment on the behaviors of drivers and pedestrians Xiang said Once the high mounted parking system
cameras are installed they can help with many aspects of integrated urban governance which represents another
advantage of the AIpark Sky Eye system Besides managing parking fee collection high mounted camera system can
also provide data for traffic improvements The snapshots obtained from the camera system can help solve problems
including illegal and inappropriate parking and vehicle theft So far the smart ETC system of AIpark has been
introduced into more than 20 cities in China
```

Steps of texting cleaning:

1. **Remove** newlines, tabs, \r, \n.
2. **Remove** URLs, emails, phone numbers, special characters, and other unwanted elements.
3. **Split** the text into sentences using punctuation marks.
4. **Remove** irrelevant chunks or sections that are not necessary for analysis.
5. **Filter out** rows or articles that do not contain keywords from a customized list related to AI tech and jobs.
6. **Tokenize** the text by splitting it into individual words and remove punctuation.
7. **Remove** common stopwords (e.g., articles, prepositions, conjunctions).
8. **Generate** n-grams (word sequences) to capture contextual information if needed.
9. **Lemmatize** words to their base or root form.
10. **Ready for Modeling!**

Topic detection (Methodology)

For the Topic detection, LDA using the **genism** package was first applied with hyperparameter tuning to find out the best choice **of a number of topics is 12** with the best combination of topic overlap and coherence score of 0.434649.

This graph illustrates the topic distribution for K-Train:

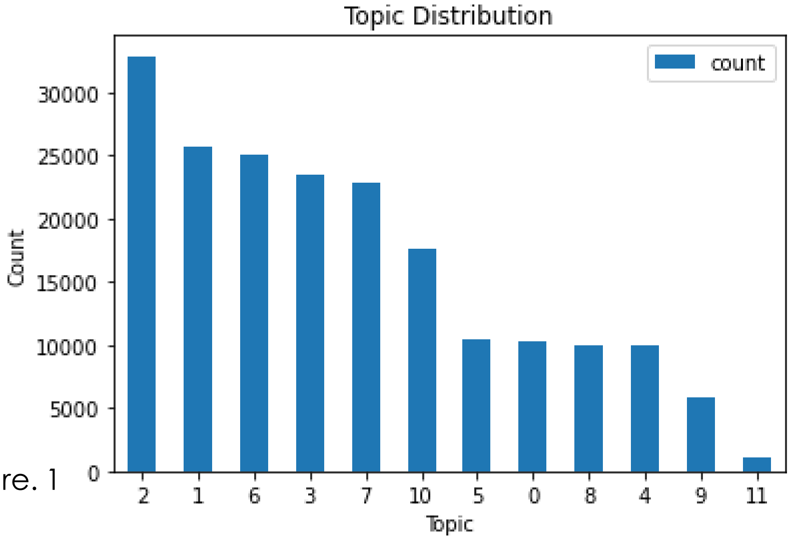


Figure. 1

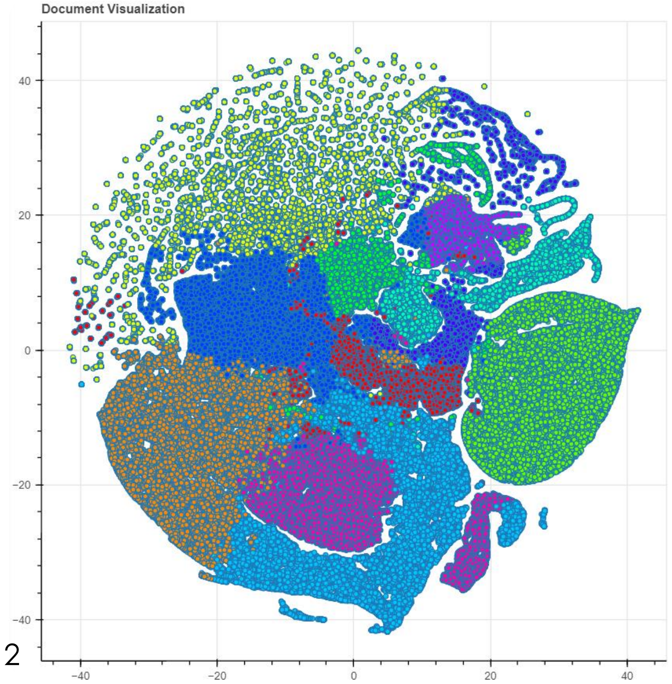
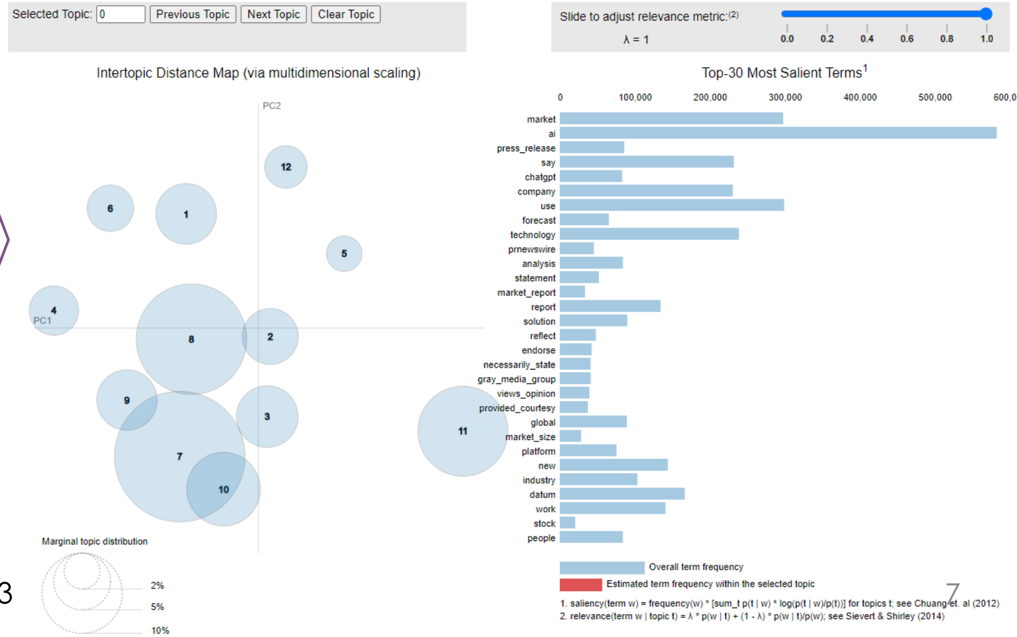


Figure. 2

K-train Topic (Left) and Genism (Right) illustrate the acceptable topic overlaps.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure. 3

Topics Detection (K-train)

Main Topics:

- Press release distribution and **content solutions**
- Human-like methods used to **work with images**.
- **Analysis** of global **markets**, **growth** forecasts, industry **research**, and **key players**.
- Use of **machine learning** by **security companies** for managing digital information.
- **Advancements** in **language** technology by ChatGPT, Google, Microsoft, OpenAI, and others.
- **Insights** into **stock shares**, investment, market outlook, prices, and NASDAQ.
- Research, **education**, and **student** involvement in machine learning and computer science.
- Application of **AI in healthcare**, including medical care, patient management, cancer treatment, drug development, and medical imaging.
- Benefits and **performance optimization** of edge computing, NVIDIA technology, and smart devices.
- **Cloud services**, **energy-related** technologies, **software solutions**, and global companies like IBM and Amazon.

The topics on the left hand cover various main topics in the field of technology and business. It covers press release distribution, human-like image processing, global market analysis, machine learning in security, advancements in language technology, stock market insights, research in machine learning and computer science, AI in healthcare, edge computing and smart devices, and cloud services by companies like IBM and Amazon.

Sentiment Analysis (Methodology)
-- *Using a combination of **Transformer + TextBlob***

The **absence** of a pre-trained **model tailored to sentiment analysis of news** articles pertaining to AI technology and jobs presents a **significant challenge**. Furthermore, the **impracticability of human labeling** due to the substantial volume of labels required exacerbates the problem. In light of these limitations, this study proposes a **novel methodology** that **integrates TextBlob and a transformer** model sourced **from Hugging Face**, incorporating a total of 215 sentiment analysis models. Leveraging the polarity and subjectivity metrics provided by TextBlob, the accuracy of the model's predictions is assessed. Remarkably, the findings demonstrate a **high** degree of **congruence** and **accuracy** between the combined approach and TextBlob's results. To enhance the validity of the outcomes, a random **sample of 50 news articles** was selected, and the sentiment analysis generated by the model was compared against human judgments, yielding **consistent** and accurate **outcomes**.

Figure. 4

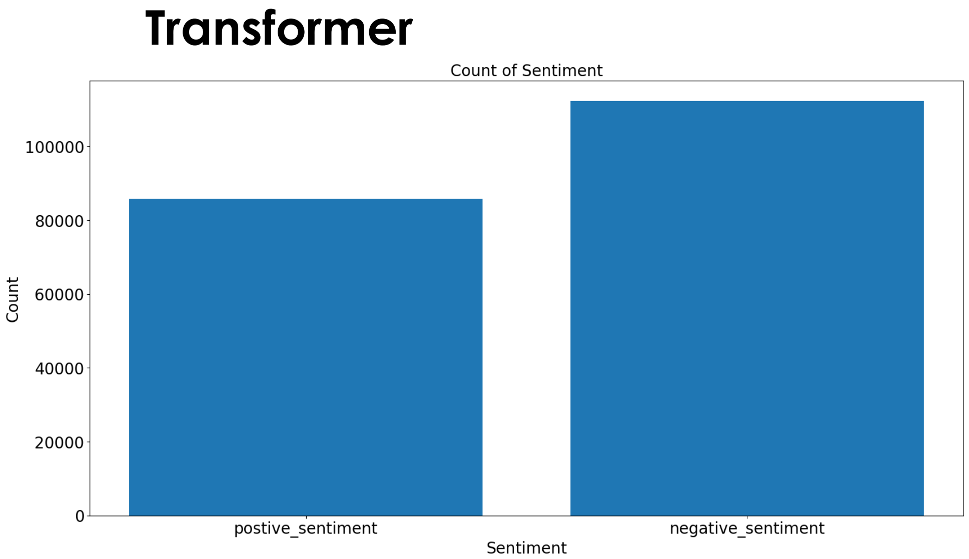

Transformer

Figure. 5

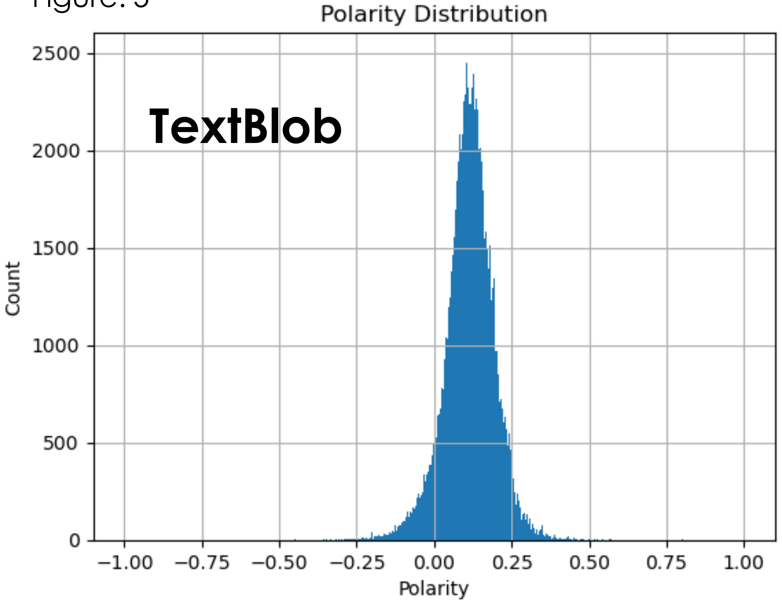
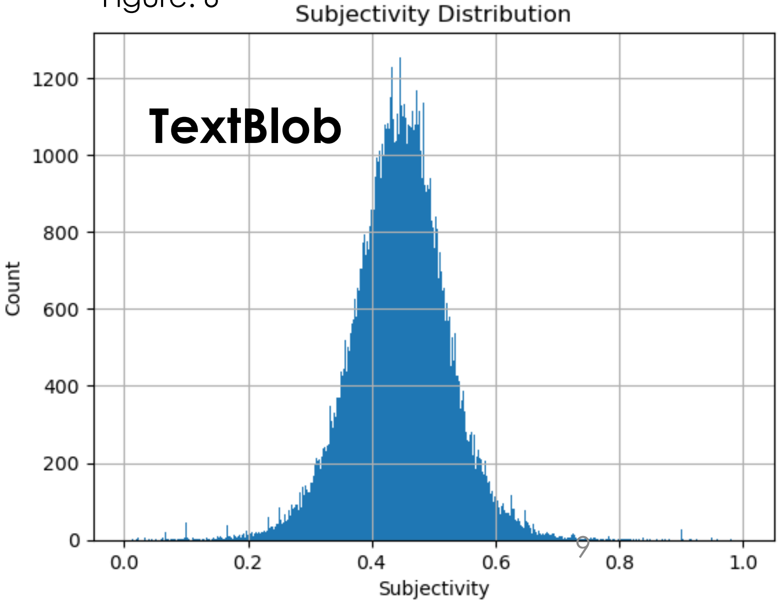
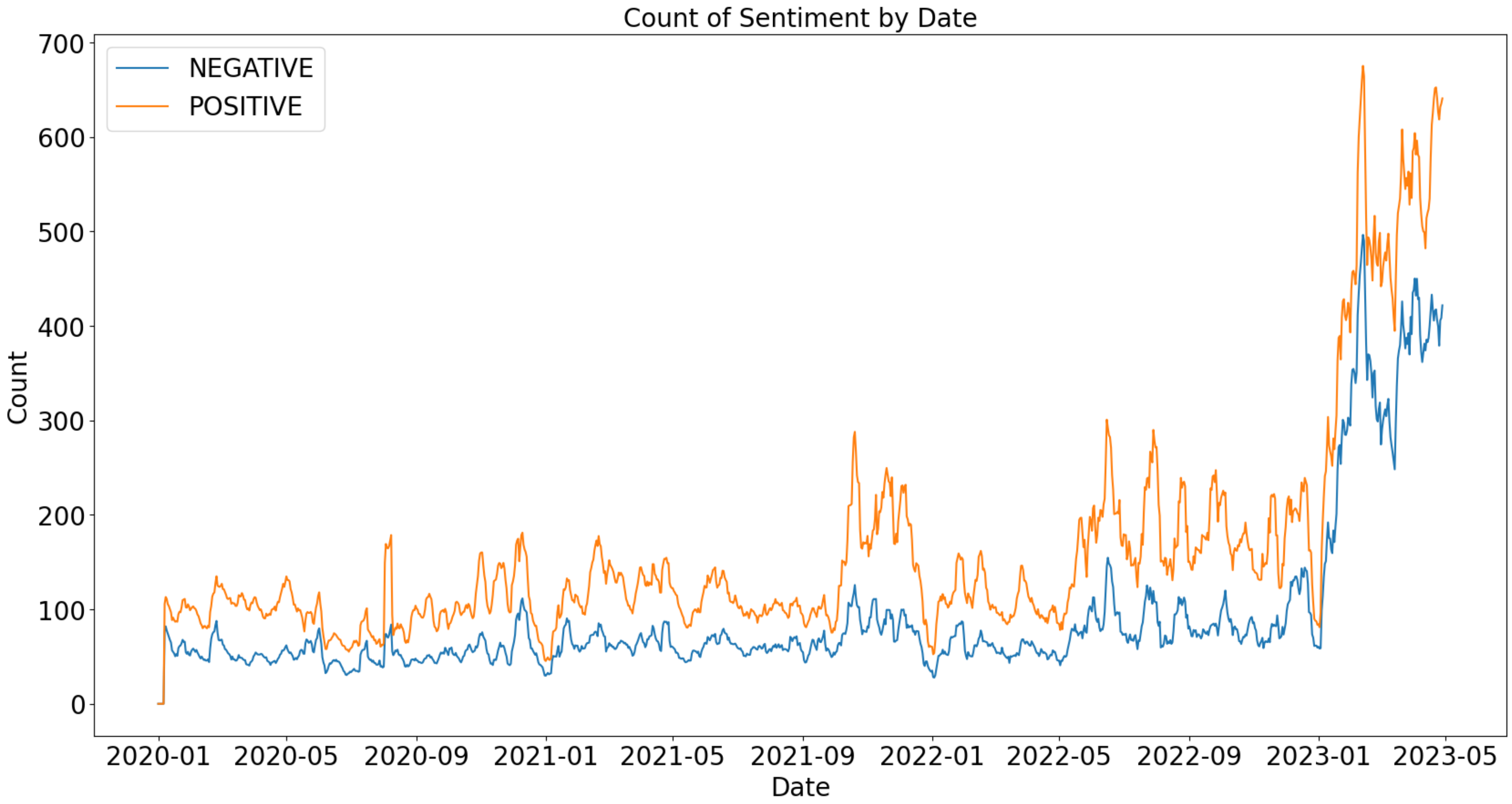

TextBlob

Figure. 6


TextBlob

Sentiment over time analysis and visualization



*(**Note**: This analysis excludes sentiments close to neutral, where the subjectivity score is near 0.)*

On the left, you can observe the overall sentiment towards AI technology over time. It is evident that the positive sentiment has consistently remained higher than the negative sentiment for the majority of the duration.

However, starting in **January 2023**, there was a noticeable increase in both positive and negative sentiments. One possible reason for this shift could be attributed to the launch of ChatGPT 3.5 in November 2022, which gained significant attention from the public by January 2023. As people became more aware of the advancements in AI technology, it sparked both positive and negative reactions.

Sentiment Analysis:

Q: Why do certain types of applications succeed, and others failed?

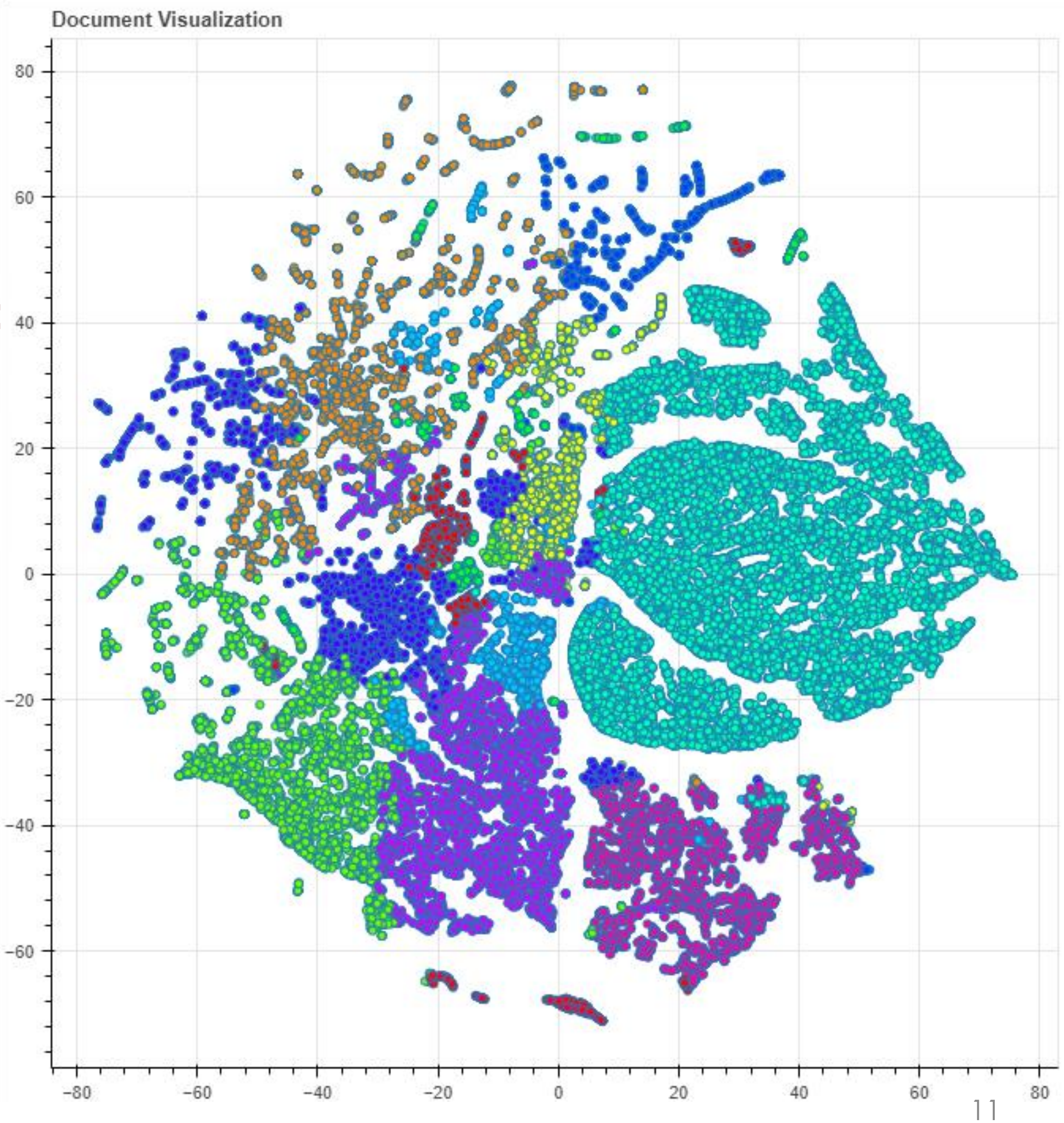
A: To answer, K-train topic modeling is used twice for both succeed and failed news articles for answer:

Successful AI, data science, and tech applications:

- Improved user experience (Topic 8)
- Reliable news and information (Topic 3)
- Market analysis and research (Topic 7)
- Cloud solutions and customer service (Topic 11)
- Advancements in healthcare (Topic 2)

Failed AI, data science, and tech applications:

- Privacy and legal concerns (Topic 0)
- Ineffective work and education solutions (Topic 1)
- Inadequate healthcare solutions (Topic 2)
- Biased or inaccurate media (Topic 3)
- Inadequate hardware and software solutions (Topic 4)



Entity identification
– Technology, Organizations & People

Note: These organizations have overall !!!POSITIVE!!! sentiments toward AI technology

Figure .7 below is a timeline for the introduction of the new technology and AI solutions that might be affecting the landscape of data science applications, the most frequently mentioned are **ChatGPT**, **Biometrics**, **Chatbot**, **Alexa**, **DeepMind**, **Alibaba**, **Deepbrain AI**, **AI Assist**, **Quantum Computing**, **Siri**, **Xiaomi**, **Quantum**, and **Virtual Agent**.

Figure. 7

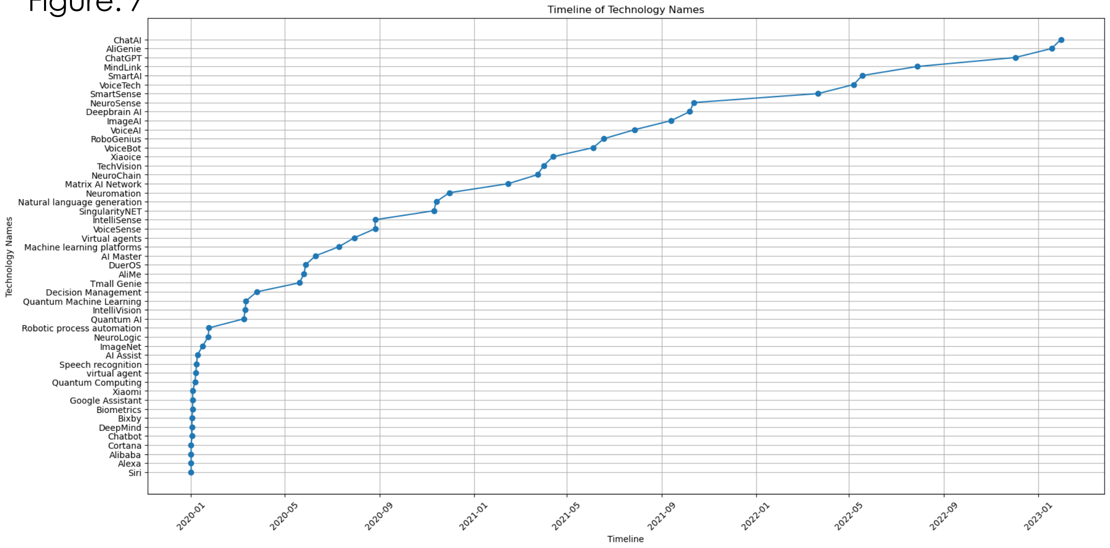


Figure. 8 (Organization)

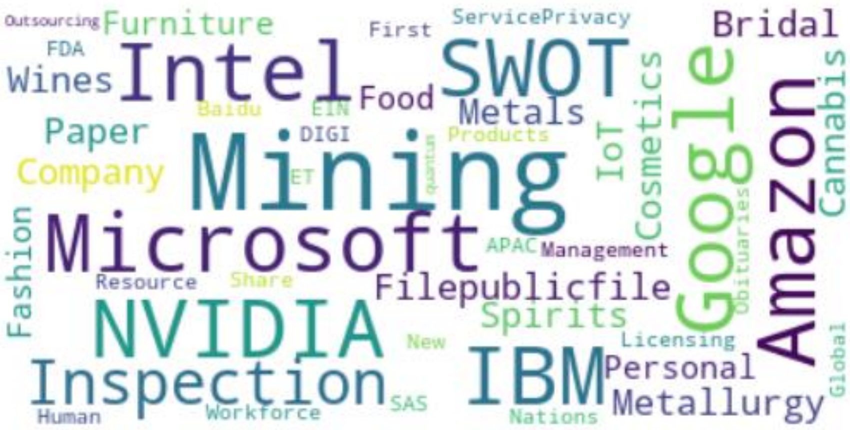


Figure. 9 (Persons)

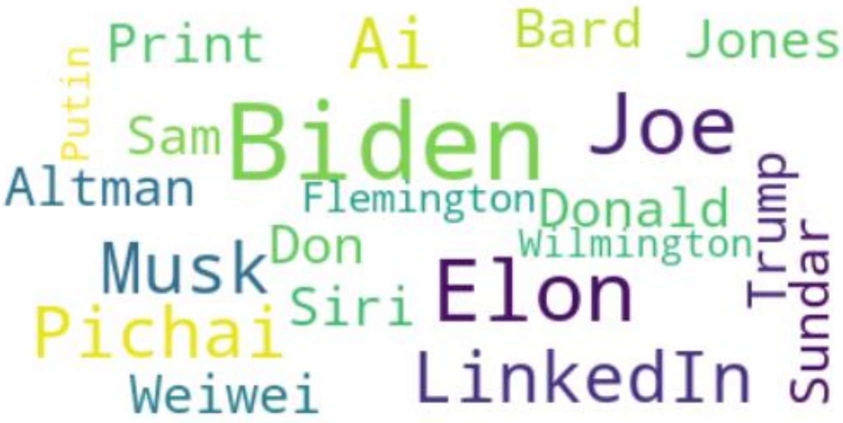


Figure. 8 and Figure. 9 on the lefthand are two Word Clouds depict the most commonly mentioned **organizations** and **persons** by utilizing **spaCy**. In the organization Word Cloud, it is noticeable that it includes both companies and some AI applications. Among the frequently mentioned companies, **a majority of them** are **technology-focused.** Regarding **individuals**, Elon Musk, the CEO of Tesla, Sam Altman, the CEO of OpenAI, and President Biden are prominent figures in this word cloud. This occurrence is likely **due to** their close involvement with the **development** and **regulation** of **AI** applications.

Continue:
Targeted (entity) sentiment identification

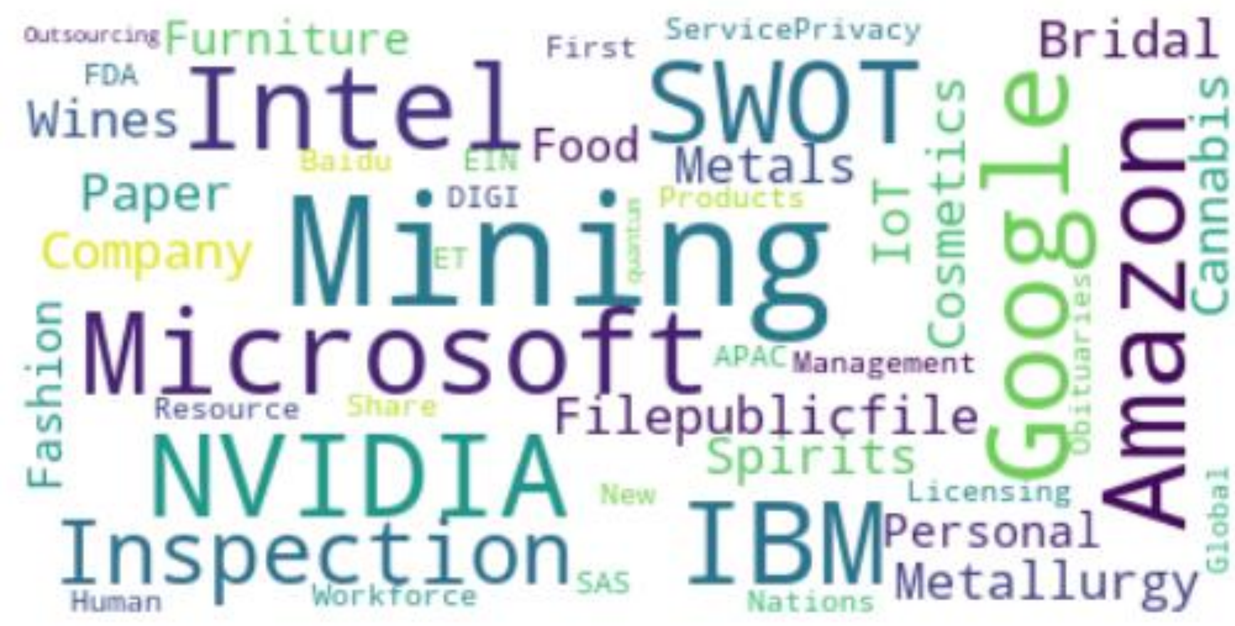
Q: What types of companies should invest in data science initiatives today or near future?

A: As shown in **Figure. 8** which has companies that generally have positive sentiments in the news article, technical companies should invest in data science Ai initiatives today or near future.

It is intuitive since there are such advantages for them as shown:

1. Competitive Advantage
2. Enhanced Decision Making
3. Improved Customer Experience:
4. Higher Operational Efficiency
5. Cost Reduction
6. Innovation and New Revenue Streams
7. Future-Proofing

Figure. 8 (Organization)



Continue: Targeted (entity) sentiment identification

Q: What types of applications cannot currently be transformed by AI, based on today's state of technology (failures)?

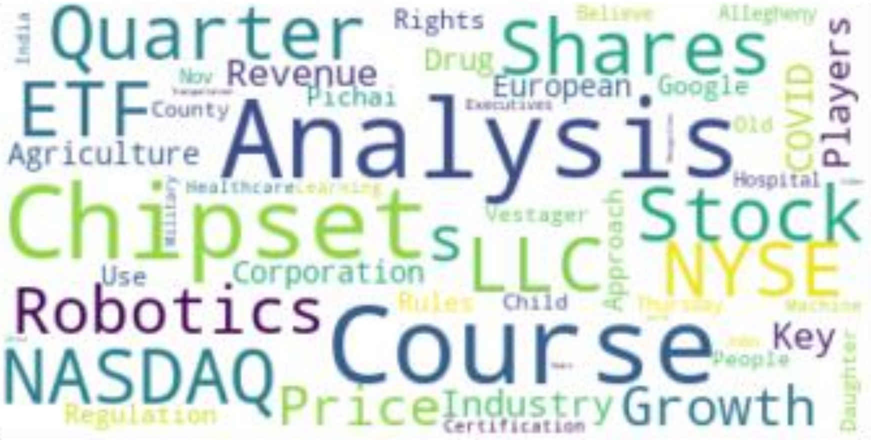
A:

- **Stock Market Analysis (MOST Mentioned as shown in Figure.9):** While AI has been used in analyzing stock market trends, factors such as quarter results, ETFs, and stock prices are still predominantly influenced by human decision-making. The complexities of the stock market make it challenging for AI to accurately predict and transform this domain.
- **Agriculture Industry:** Although there have been advancements in using AI for agriculture, there are still limitations. Factors such as growth analysis, revenue, and key players in the industry require a deep understanding of the agriculture sector, making it difficult for AI to fully transform this industry.

The industry below are less mentioned (For more information, refer to appendix page 25):

- **News and Politics:** AI can assist in providing news and information, but understanding nuanced topics like politics, public opinion, and regulations can be challenging. Issues related to regulation, European rules, and the approach of companies like Google may require human judgment and interpretation.
- **Healthcare Education:** While AI has been used in healthcare for various applications, certain aspects, such as certification courses, executive-level training, and online learning, may require human expertise, especially when it comes to providing personalized education and understanding complex medical scenarios.
- **Military and Transportation:** The analysis of military strategies, transportation systems, video recognition, and chipsets for military use requires highly specialized knowledge and expertise. AI may have limitations in fully transforming these domains due to the complexity and sensitivity of the subject matter.

Figure. 9 Negative Application Sentiment



*This targeted word cloud was created by focusing on **negative sentiment**. The **topic modeling** technique was used **for** identifying **keywords** associated with failures, and the sentiment was **adjusted** accordingly to present the final results.*

THANK YOU!

Appendix

Page 16 – Page 25

Figure. 1

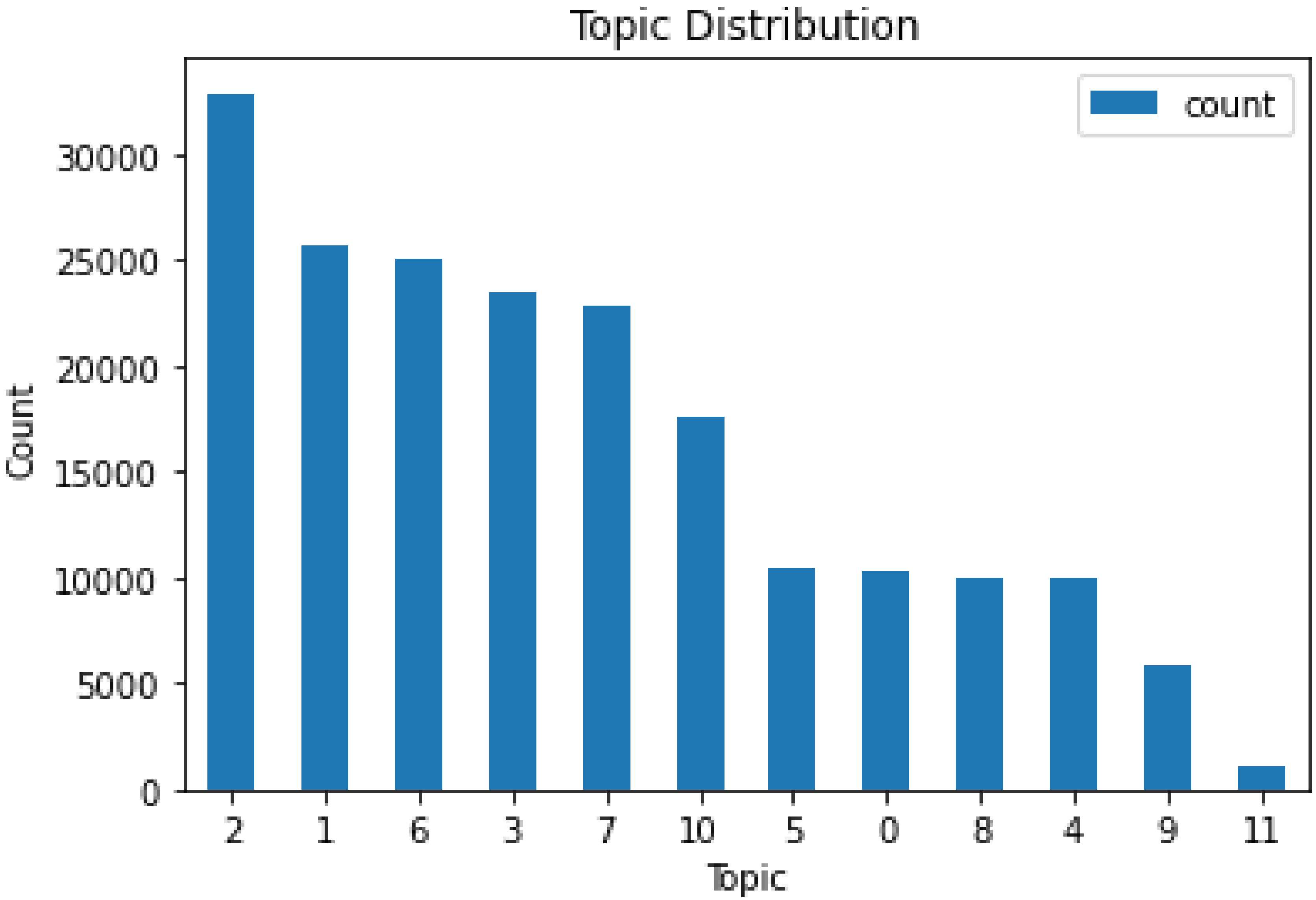

Topic Distribution
count
Count
30000
25000
20000
15000
10000
5000
0
2
1
6
3
7
10
5
0
8
4
9
11
Topic

Figure .2

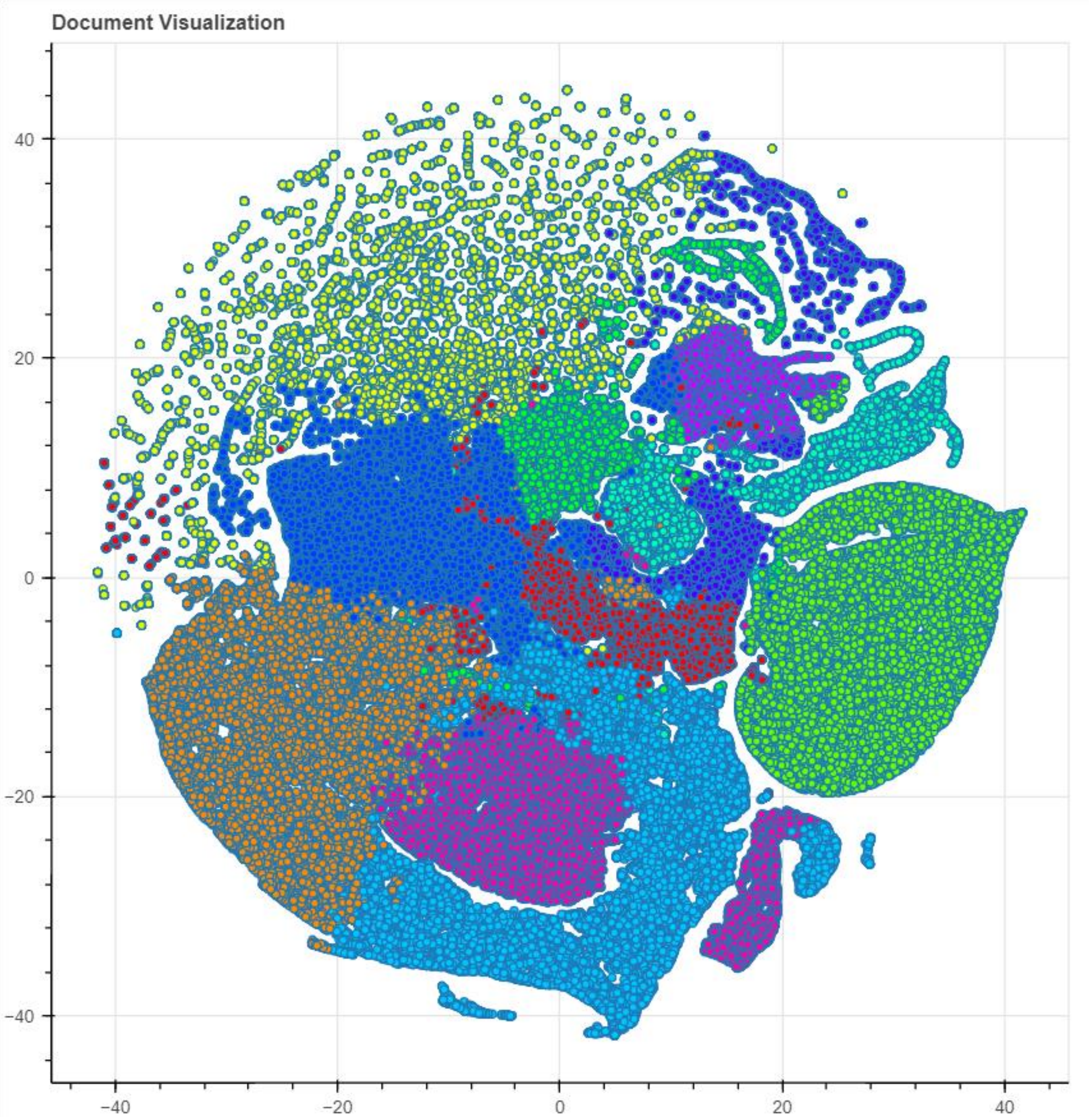
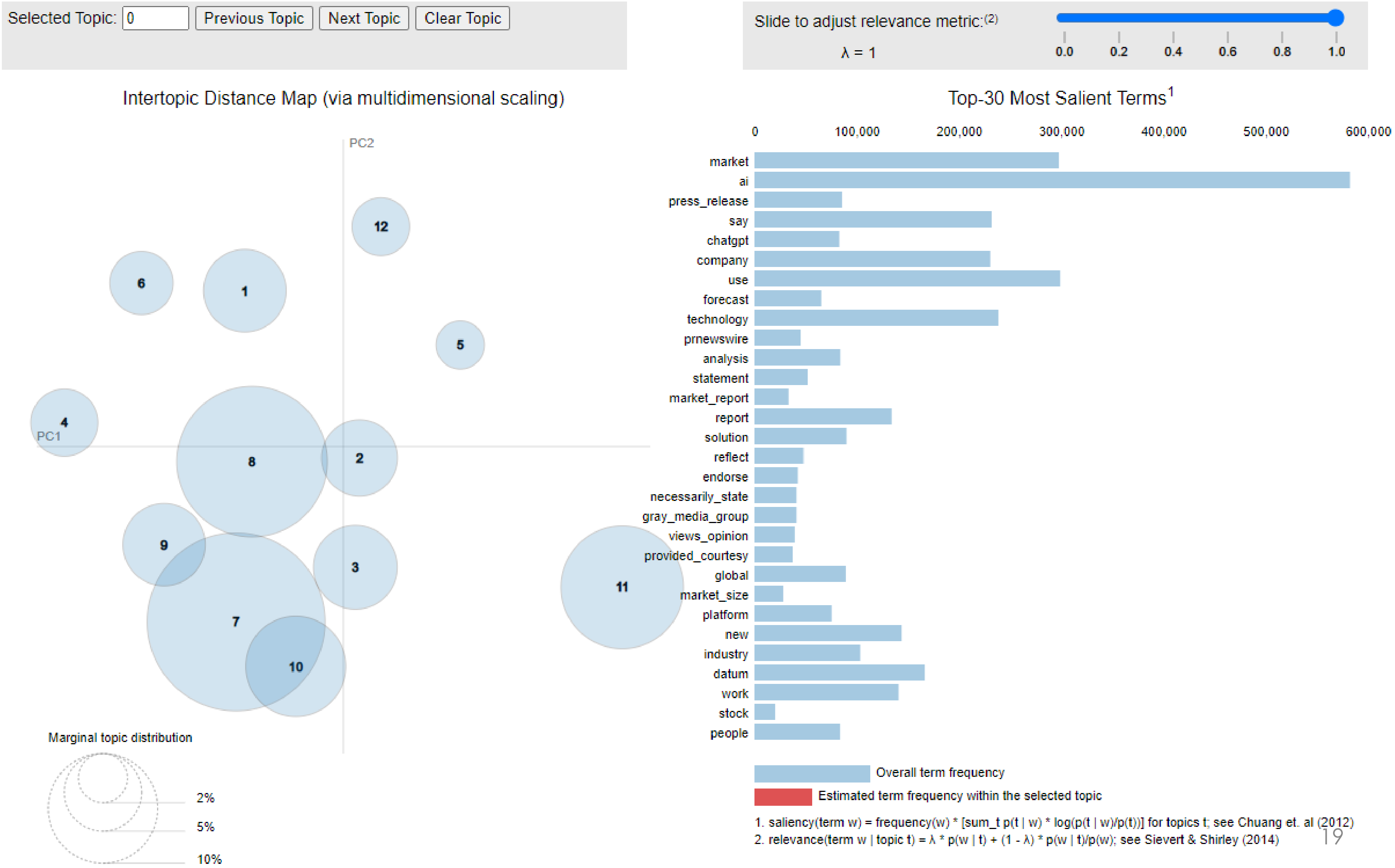


Figure. 3



Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure. 4

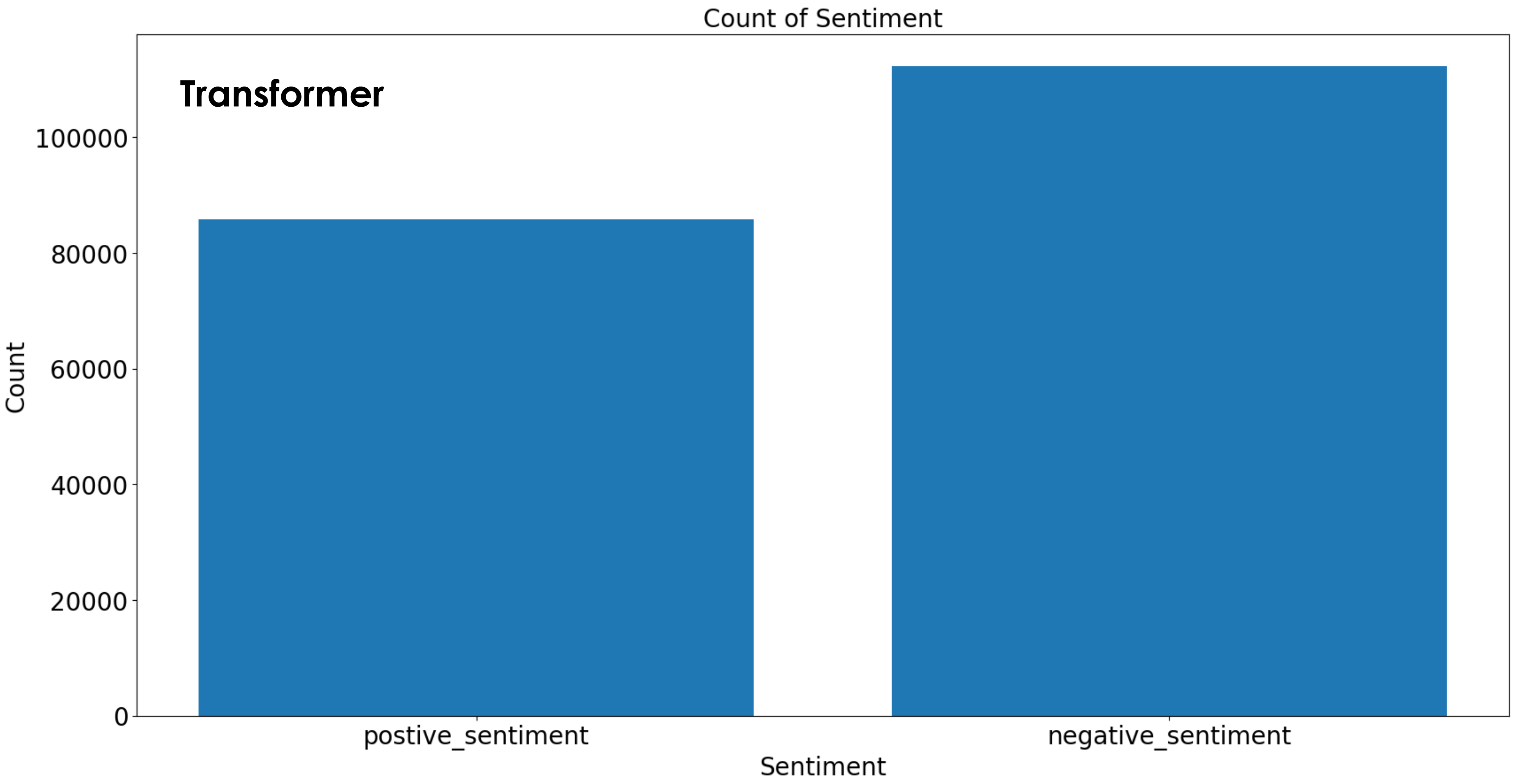

Count of Sentiment
Transformer
100000
80000
60000
40000
20000
0
Count
postive_sentiment
negative_sentiment
Sentiment

Figure.5 and Figure. 6

Figure. 5 Polarity Distribution

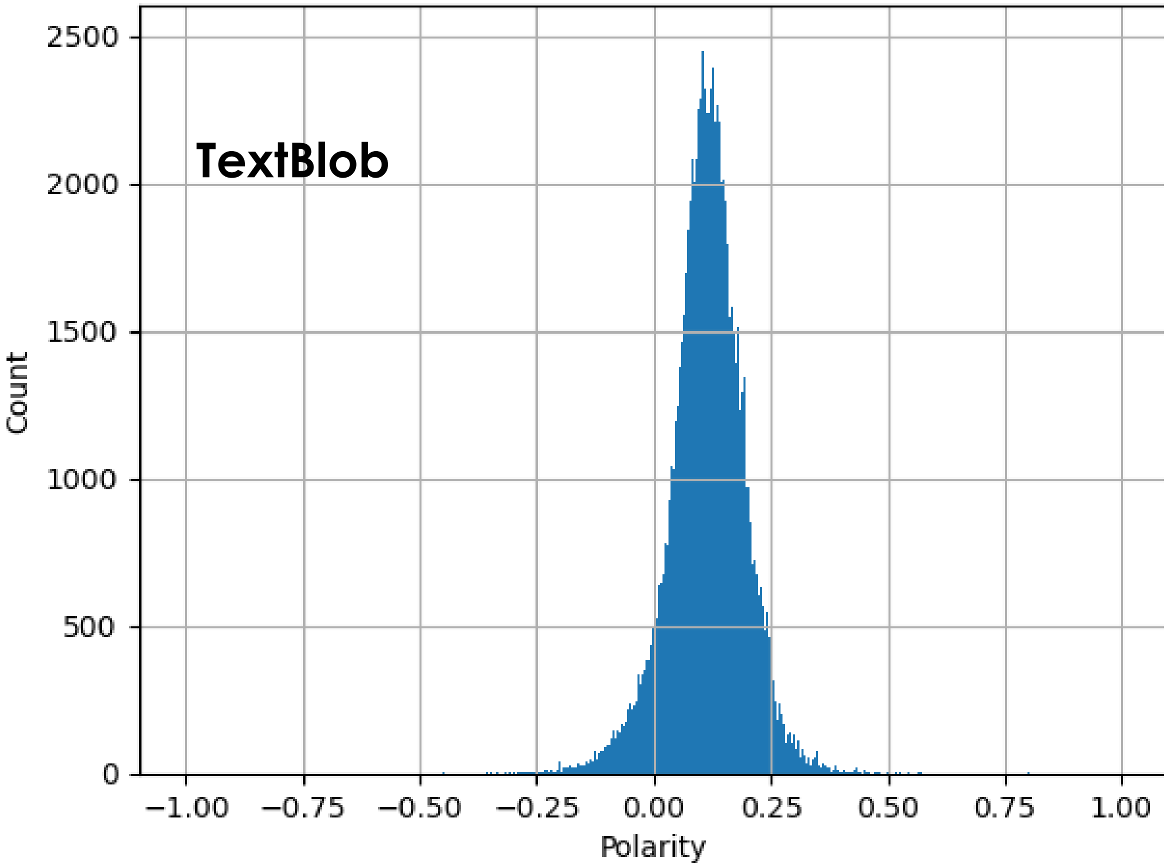


Figure. 6 Subjectivity Distribution

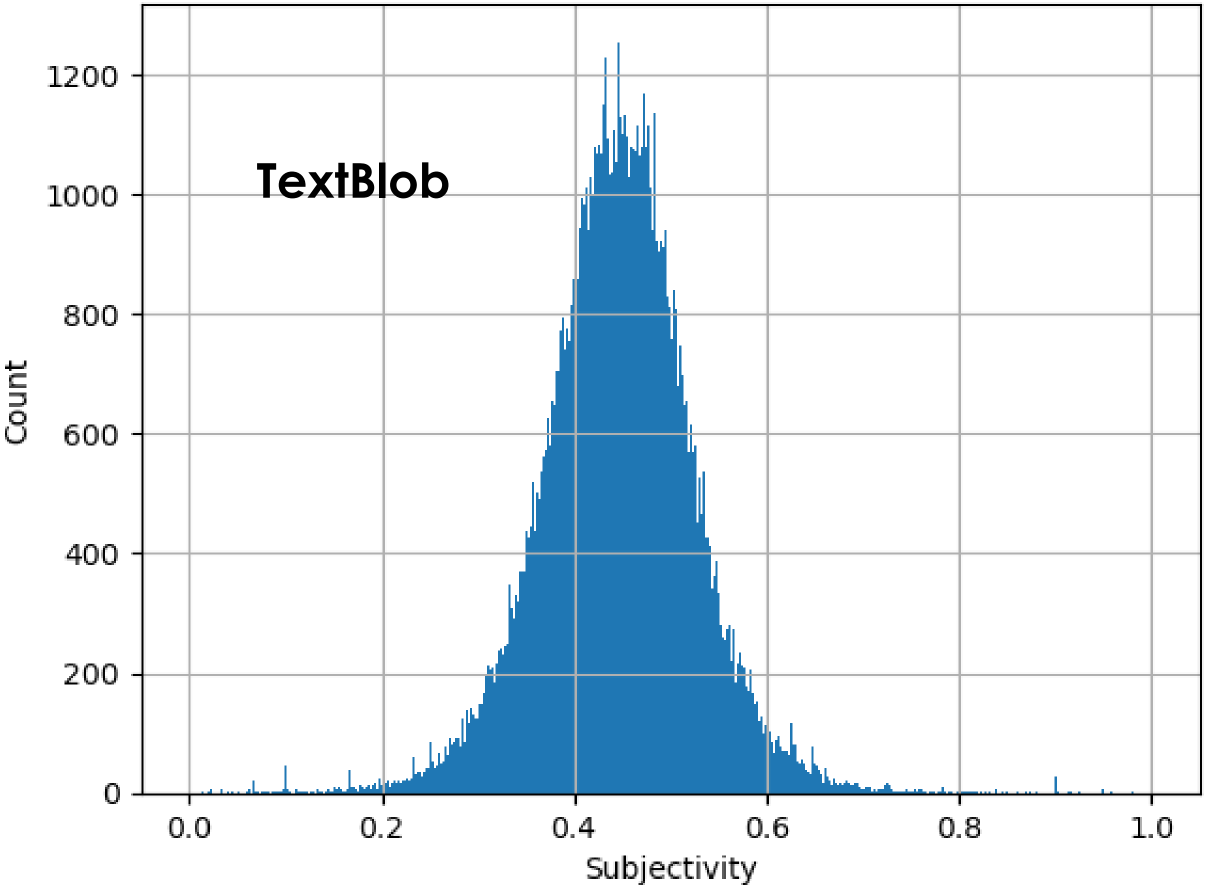


Table. 1 (Detailed topic summarization)

Based on the information you provided, it seems that the topics trained by ktrain on your dataset of news articles related to Data Science, Machine Learning, and Artificial Intelligence are as follows:

1. Topic: Gray Media Group PRNewswire Platform Press Release Content Solutions Customers | Count: 32,868
    - This topic seems to be related to press releases and content solutions provided by Gray Media Group's PRNewswire platform for its customers.
2. Topic: People Making Way Images Work Using Human-like Methods | Count: 25,749
    - This topic focuses on how people employ human-like methods to work with images.
3. Topic: Global Market Analysis, Growth Forecast, Industry Research, Key Players | Count: 23,469
    - This topic involves market analysis, growth forecasts, industry research, and discussions about key players in the field.
4. Topic: Security Companies Utilizing Machine Learning Systems to Help with Digital Information | Count: 22,793
    - This topic explores how security companies are utilizing machine learning systems to assist with managing digital information.
5. Topic: ChatGPT, Google, Microsoft, OpenAI, Search, Chatbot, GPT, Language Tech | Count: 17,603
    - This topic covers discussions about ChatGPT, Google, Microsoft, OpenAI, search engines, chatbots, GPT models, and language technology.
6. Topic: Stock Shares, Investment, Market Outlook, Price, NASDAQ | Count: 10,430
    - This topic revolves around discussions related to stock shares, investments, market outlook, prices, and NASDAQ.
7. Topic: Machine Learning Research, Education, Students, University, Computer Science | Count: 10,292
    - This topic focuses on machine learning research, education, students, universities, and computer science.
8. Topic: Healthcare, Medical Care, Patients, Clinical, Cancer, Drug, Imaging | Count: 9,967
    - This topic involves discussions related to healthcare, medical care, patients, clinical practices, cancer treatments, drugs, and medical imaging.
9. Topic: Edge Computing, NVIDIA, Smart Devices, Performance, Software, Hardware | Count: 9,938
    - This topic explores edge computing, NVIDIA technology, smart devices, performance optimization, software, and hardware.
10. Topic: Republic, Email Policy, Travel, Password, Sports, American Services | Count: 5,837
    - This topic covers discussions related to republics, email policies, travel, passwords, sports, and American services.
11. Topic: Cloud Services, Energy, Software, Global, IBM, Amazon, Supply | Count: 1,055
    - This topic involves discussions about cloud services, energy-related topics, software, global perspectives, IBM, Amazon, and supply chains.

These topics provide an overview of the major themes present in your dataset, covering a range of subjects related to data science, machine learning, artificial intelligence, and their potential impacts on various industries and domains.

Figure. 7

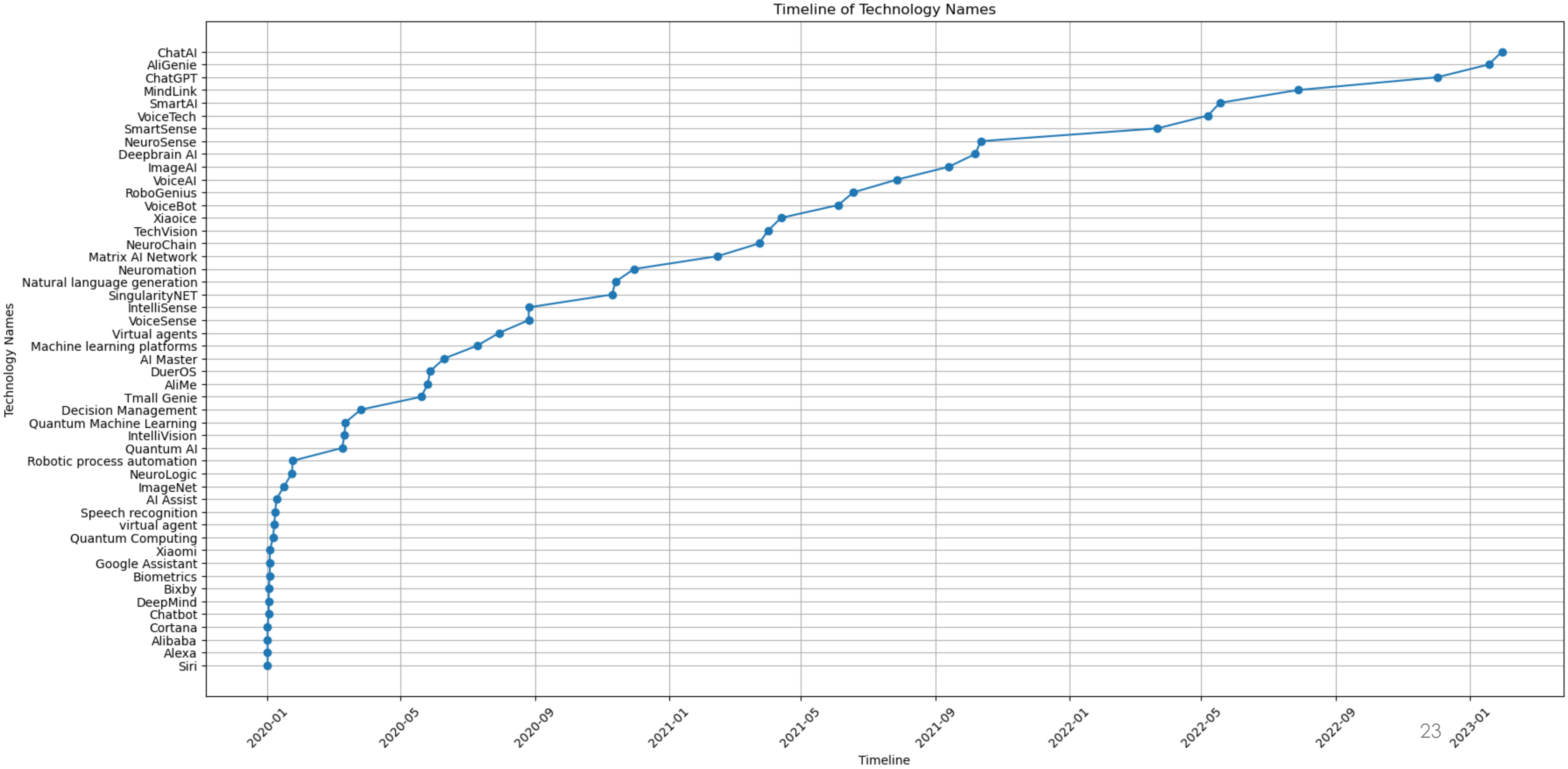
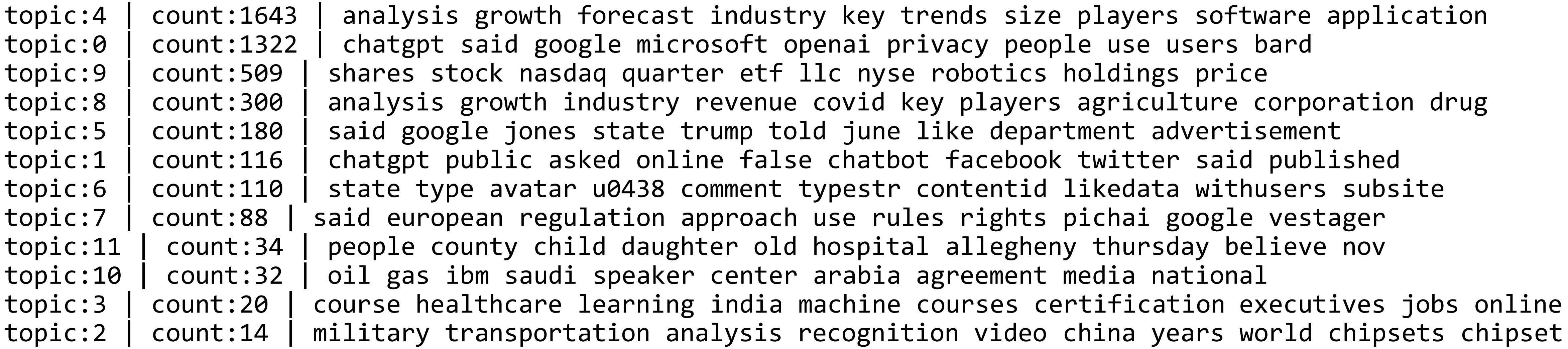

Timeline of Technology Names
Technology Names
ChatAI
AliGenie
ChatGPT
MindLink
SmartAI
VoiceTech
SmartSense
NeuroSense
Deepbrain AI
ImageAI
VoiceAI
RoboGenius
VoiceBot
Xiaoice
TechVision
NeuroChain
Matrix AI Network
Neuromation
Natural language generation
SingularityNET
IntelliSense
VoiceSense
Virtual agents
Machine learning platforms
AI Master
DuerOS
AliMe
Tmall Genie
Decision Management
Quantum Machine Learning
IntelliVision
Quantum AI
Robotic process automation
NeuroLogic
ImageNet
AI Assist
Speech recognition
virtual agent
Quantum Computing
Xiaomi
Google Assistant
Biometrics
Bixby
DeepMind
Chatbot
Cortana
Alibaba
Alexa
Siri
2020-01
2020-05
2020-09
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
Timeline

Table. 2 (Topics for sentiment analysis)

```
topic:5 | count:13131 | report analysis forecast key players trends size opportunities revenue software
topic:9 | count:4714 | said chatgpt year tech microsoft india policy privacy students use
topic:1 | count:4545 | gray media prnewswire nvidia press release solutions platform statements said
topic:3 | count:4101 | like video users time language google search best use people
topic:8 | count:3628 | customer cloud platform digital customers solutions service experience conversational
security
topic:10 | count:3122 | policy travel public american transportation aviation sports real mining beverages
topic:2 | count:1822 | health healthcare clinical care medicine ein patient medical patients presswire
topic:7 | count:1365 | medical imaging healthcare radiology cancer clinical patients gray patient lunit
topic:6 | count:1351 | learning machine science platform analytics education models cloud use model
topic:0 | count:923 | edge smart iot digi communications devices hardware power chip software
topic:4 | count:880 | vision computer health recognition robots robotics conference university people awards
topic:11 | count:302 | let state work union people great americans make address society
```

Table. 3 (In substitute to Slide No. 14)



```
topic:4 | count:1643 | analysis growth forecast industry key trends size players software application
topic:0 | count:1322 | chatgpt said google microsoft openai privacy people use users bard
topic:9 | count:509 | shares stock nasdaq quarter etf llc nyse robotics holdings price
topic:8 | count:300 | analysis growth industry revenue covid key players agriculture corporation drug
topic:5 | count:180 | said google jones state trump told june like department advertisement
topic:1 | count:116 | chatgpt public asked online false chatbot facebook twitter said published
topic:6 | count:110 | state type avatar u0438 comment typestr contentid likedata withusers subsite
topic:7 | count:88 | said european regulation approach use rules rights pichai google vestager
topic:11 | count:34 | people county child daughter old hospital allegheny thursday believe nov
topic:10 | count:32 | oil gas ibm saudi speaker center arabia agreement media national
topic:3 | count:20 | course healthcare learning india machine courses certification executives jobs online
topic:2 | count:14 | military transportation analysis recognition video china years world chipsets chipset
```