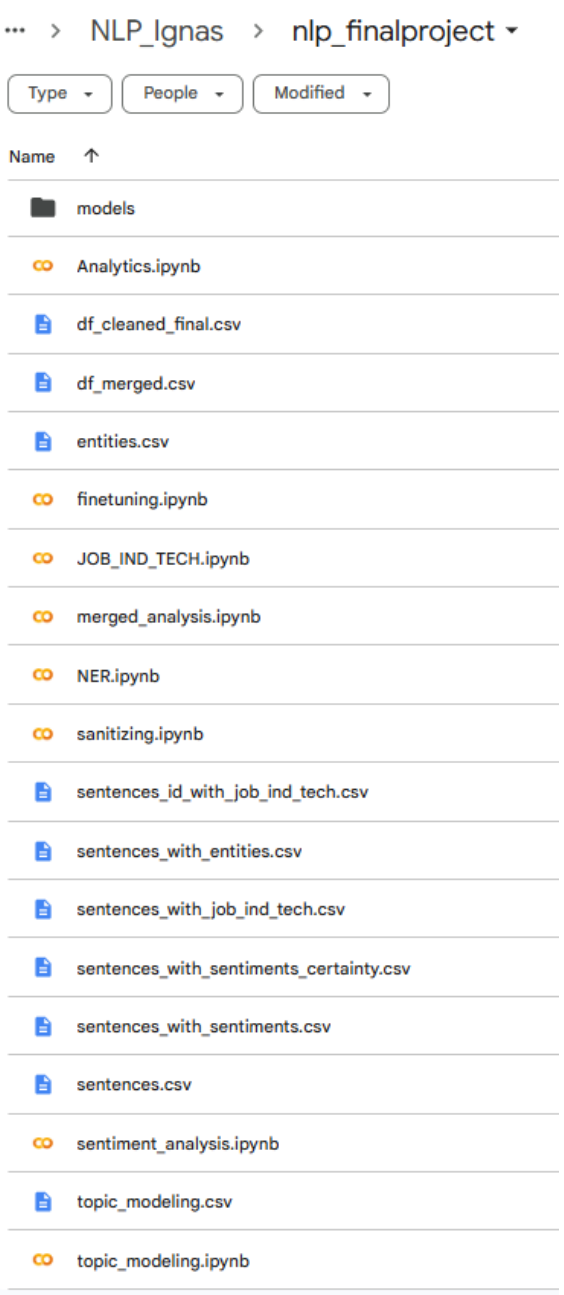
NATURAL LANGUAGE PROCESSING AND COGNITIVE COMPUTING

Liang Gong

Raw Data and Workflow



Raw Data

```
[ ] # Inspect the First Few Rows:
    df_news_final_project.head()
```

	url	date	language	title	text
0	http://spaceref.com/astronomy/observation-simulation-and-ai-join-forces-to-reveal-a-clear-universe.html	2021-07-05	en	Observation, Simulation, And AI Join Forces To Reveal A Clear Universe - SpaceRef	\n\nObservation, Simulation, And AI Join Forces To Reveal A Clear Universe - SpaceRef\n\n \nHome |\nNASA Watch\nSpaceRef Business\nAstrobiology Web\nAdvertising\nAdd an Event\nSign up for our Daily Newsletter \n\n \n\n\nInternational Space Station\nNASA Hack Space\nCalendar\nMissions\nSpace Weather \n \n \nObservation, Simulation, And AI Join Forces To Reveal A Clear Universe\n\n\n Press Release - Source: NATIONAL INSTITUTES OF NATURAL SCIENCES...
1	http://www.agoravox.it/Covid-19-un-messaggio-dai.html	2020-03-13	en	Covid-19: un messaggio dai ricercatori italiani ai colleghi stranieri - AgoraVox Italia	\n\n\nCovid-19: un messaggio dai ricercatori italiani ai colleghi stranieri - AgoraVox Italia\n\nAccedi\n\nPartecipa ad AgoraVox!\nIscriviti e proponi un articolo\nHome page\n\n\nAttualitÃ \n\n\nAmbiente\nCronaca\nCronaca Locale\nCultura\nEconomia\nEuropa\nMedia\nIstruzione\nMondo\nPolitica\nSalute\nReligione\nSocietÃ \nScienza e Tecnologia\nTribuna Libera\n\n\tDi' la tua\n\n\nTempo Libero\n\n\nGossip\nRedazionali\nConcorsi\nCinema\nFame&Tulipani\nIncredibile ma vero!\nLa vignetta del giorno...
2	http://www.dataweek.co.za/21690r	2024-04-05	en	Flash for AI - 28 March 2024 - EBV Electrolink - Dataweek	\nFlash for AI - 28 March 2024 - EBV Electrolink - Dataweek\nHome\nAbout us\nBack issues / E-book / PDF\nSubscribe\nAdvertise\n\n \n\nEMP Handbook\n\nCategories\n\n ▸ Editor's Choice\n ▸ Multimedia, Videos\n ▸ AI & ML\n ▸ Analogue, Mixed Signal, LSI\n ▸ Circuit & System Protection\n ▸ Computer/Embedded Technology\n ▸ Design Automation\n ▸ DSP, Micros & Memory\n ▸ Edge Computing & IIoT\n ▸ Electronics Technology\n ▸ Enclosures, Racks, Cabinets & Panel Products\n ▸ Events\n ▸ Interconnection\n ▸ Manufacturi...

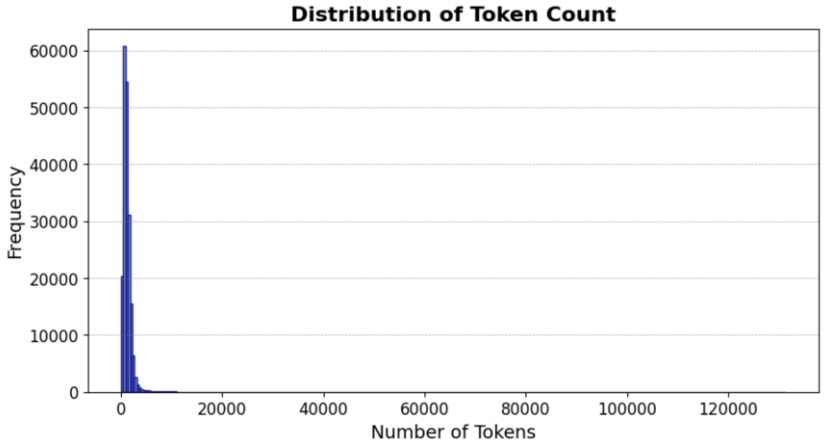
Workflow Structure

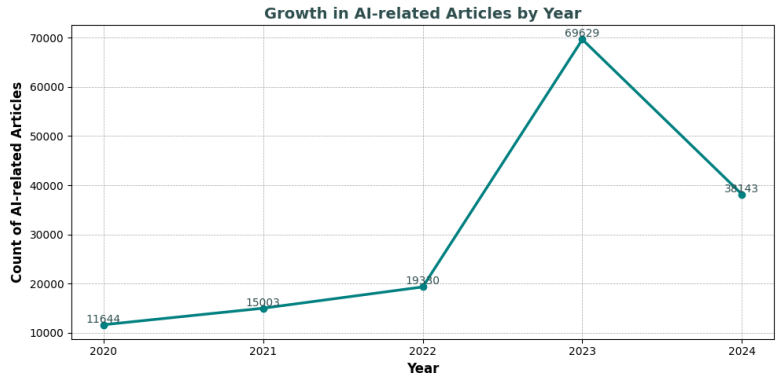
purpose	dataframe	dataframe shape	notebooks
raw table: each row is an article	news_final_project.parquet	(200588, 5)	sanitizing.ipynb
each row is an article	df_cleaned_final.csv	(153749, 7)	sanitizing.ipynb
article-level	topic_modeling.csv	(153749, 11)	topic_modeling.ipynb
break articles into sentences	sentences.csv	(6523835, 3)	NER.ipynb
sentence-level NER using spaCy	sentences_with_entities.csv	(6523835, 3)	NER.ipynb
sentence-level NER using JOB_IND_TECH	sentences_with_job_ind_tech.csv	(6523835, 3)	JOB_IND_TECH.ipynb
fine-tune a sentiment analytis model	prepare for sentiment analysis	model.save_pretrained('./financial-distilbert-lg')	finetuning.ipynb
sentence-level sentiment analysis using fine-tuned model	sentences_with_sentiments_certainty.csv	(6523802, 7)	sentiment_analysis.ipynb
level dataframe including all necessary columns	df_merged.csv	(6523802, 10)	merged_analysis.ipynb
			Analytics.ipynb

SANITIZING

- Removing URLs from the text.
- Keeping only specific characters
- Removing extra whitespace
- Exclude failed web crawl patterns
- Exclude non-English articles

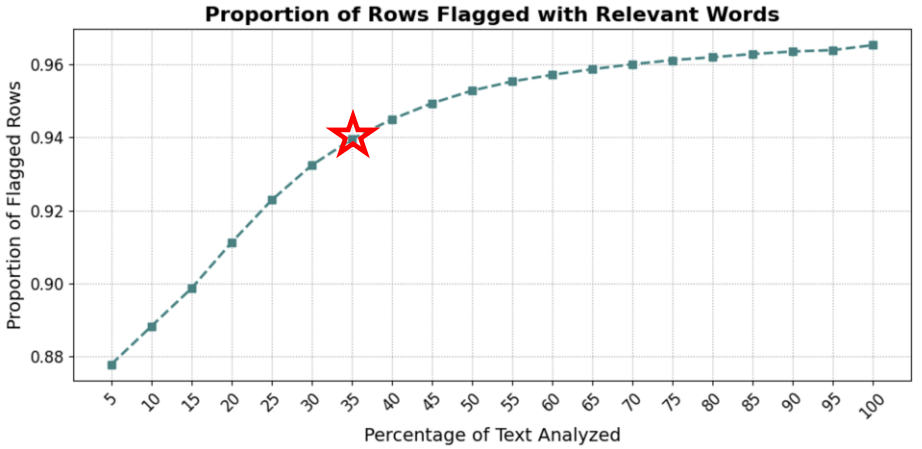
top and bottom 5% of articles by Token Count are removed from the analysis.





Notice that the number of articles provided are declining after 2023

rel_words = ['Artificial Intelligence', 'AI', 'Data Science', 'DS', 'Machine Learning', 'ML']



Among all the rows (articles), 94% of all the rows (articles) include at least one of the rel_words at the first 35% part of the article. For a row (article), if no rel_words appears within first 35% of the text, I exclude the row (article) in my analytics.

TOPIC MODELING

Cleans Text

- **Removing special characters and numbers** to keep only alphabetic characters.
- **Converting text to lowercase** for uniformity.
- **Removing stopwords** (common words like "and", "the") to focus on meaningful words.
- **Lemmatizing** words (reducing them to their base form, e.g., "running" → "run").

Use Latent Dirichlet Allocation (LDA) from ktrain

When a document is analyzed, the topic model determines the proportion of the document that belongs to each topic.

```
tm = ktrain.text.get_topic_model(
    texts=texts,
    n_topics=12,
    n_features=10000,
    min_df=0.001,
    max_df=0.5,
    stop_words='english',
    model_type='lda',
    lda_max_iter=5,
    verbose=1)

tm.build(texts, threshold=0.25)
```

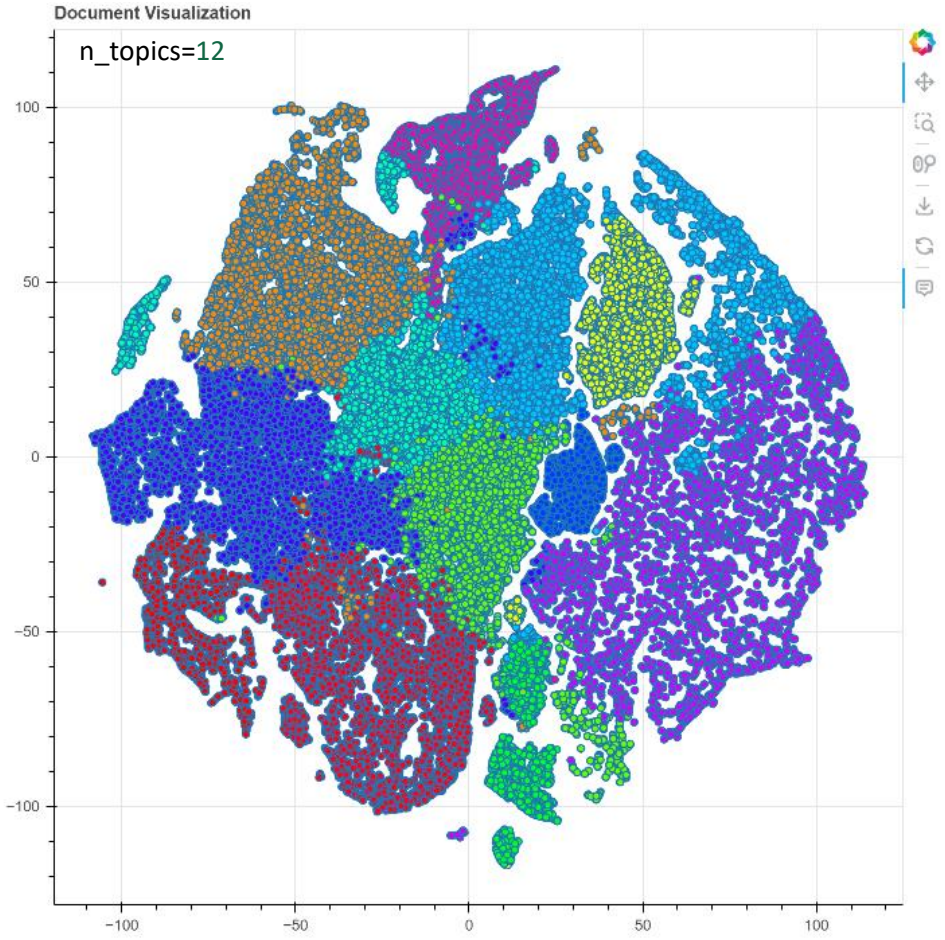
If the most probable topic for a document has a probability equal to or greater than 0.25, that topic is assigned to the document.

```
tm.print_topics(show_counts=True)
```

```
topic:9 | count:26215 | gray group release press solution prnewswire alert statement platform patient
topic:8 | count:19444 | human science like learning machine people used model using health
topic:6 | count:19118 | customer cloud solution platform model enterprise generative application experience digital
topic:1 | count:17554 | google user best feature video apple image tech microsoft search
topic:11 | count:17230 | ago video hour weather sport story day local public search
topic:3 | count:13980 | risk government security job tech law potential generative tool work
topic:5 | count:10742 | share india price day latest bank daily read market subscribe
topic:2 | count:7697 | market global analysis growth research forecast key trend player size
topic:10 | count:6800 | stock market nasdaq investor symbol quote investment fund insurance analyst
topic:0 | count:5472 | music public program radio schedule event art community search npr
topic:4 | count:4393 | release newswires south press north ein distribution presswire island country
topic:7 | count:3717 | product entertainment consumer release resource general health public financial overviewview
```

Topic Modeling ktrain LDA

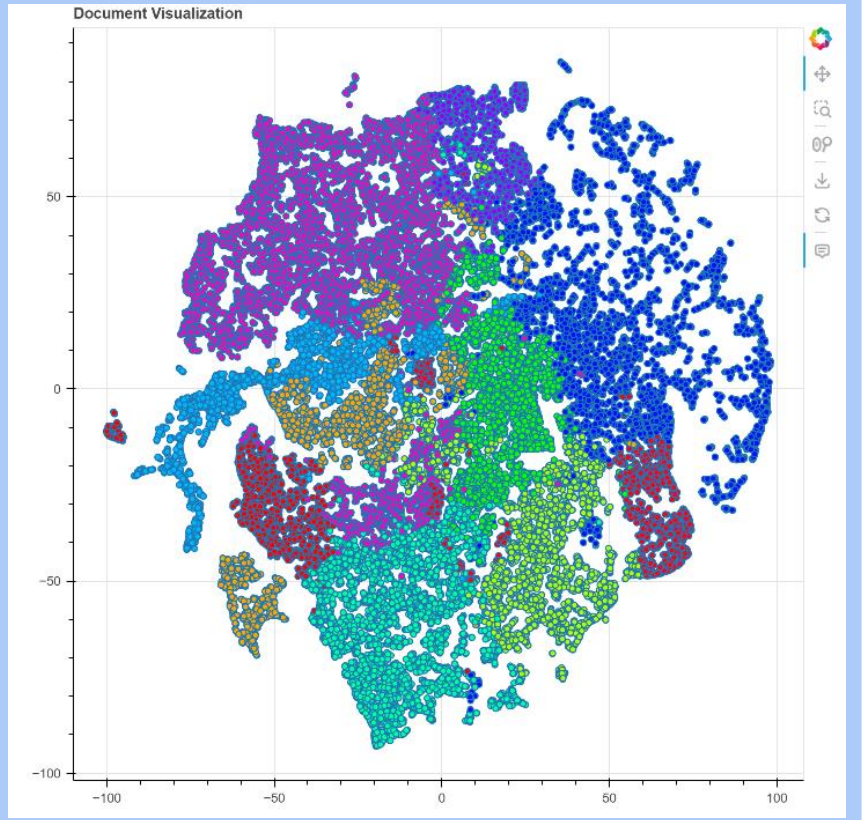
Visualize the topic distributions for documents based on their topic distributions after training a topic model (tm) in ktrain LDA. Tried different n_topics and 12 has the best clustering.



Art: Topic 0
MarketResearch: Topic 2, Topic 5
Media: Topic 4
PolicyandLegal: Topic 3
Finance: Topic 10
Topics that are related to tech:
Topic 6, Topic 8, and Topic 1
Keywords in the topic seem to have no relevance (other): Topic 7, Topic 9, and Topic 11

Perform LDA again on the articles under these tech topics to narrow down tech topics

EntertainmentTech: Topic 0, Topic 8
HealthcareTech: Topic 1
AIResearch: Topic 5
TechProducts: Topic 4
CloudComputing: Topic 6
AIinContentCreation: Topic 2
Keywords in the topic seem to have no relevance (other): Topic 7, Topic 3, Topic 9



Reduces the dimensionality of the doc_topics matrix (e.g., using t-SNE or UMAP) to project documents into 2D.

- Each document is represented as a point in this 2D space, with colors or labels indicating the dominant topic for each document.
- Clusters of Documents: Documents with similar topic distributions will be closer to each other.
- Dominant Topics: Points are color-coded by the most probable topic, helping identify clusters for each topic.
- Documents in the same region are likely to share similar dominant topics. Overlaps in clusters indicate multi-topic documents.

Topic Modeling

Detect major topics and draw connections to different industries

```python
# Define a dictionary to map topic numbers to topic names
topic_mapping = {
    '0.0': 'Art',
    '1.0': 'topic_2',
    '2.0': 'MarketResearch',
    '3.0': 'PolicyandLegal',
    '4.0': 'Media',
    '5.0': 'MarketResearch',
    '6.0': 'topic_2',
    '7.0': 'other',
    '8.0': 'topic_2',
    '9.0': 'other',
    '10.0': 'Finance',
    '11.0': 'other',
    'other': 'other'
}

# Map the topic numbers to topic names in the 'topic' column
df['topic'] = df['topic'].map(topic_mapping)
```

Perform LDA again on the articles under these tech topics to narrow down tech topics

```python
# Define a dictionary to map topic numbers to topic names
topic2_mapping = {
    "0.0": "EntertainmentTech",
    "1.0": "HealthcareTech",
    "2.0": "AIinContentCreation",
    "3.0": "other",
    "4.0": "TechProducts",
    "5.0": "AIResearch",
    "6.0": "CloudComputing",
    "7.0": "other",
    "8.0": "EntertainmentTech",
    "9.0": "other",
    "other": "other"
}

# Map the topic numbers to topic names in the 'topic2' column
filtered_df['topic_2'] = filtered_df['topic_2'].map(topic2_mapping)
```

‘topic’ column is updated with the corresponding values from the ‘topic_2’ column for specific rows, where rows have values in column ‘topic_2’

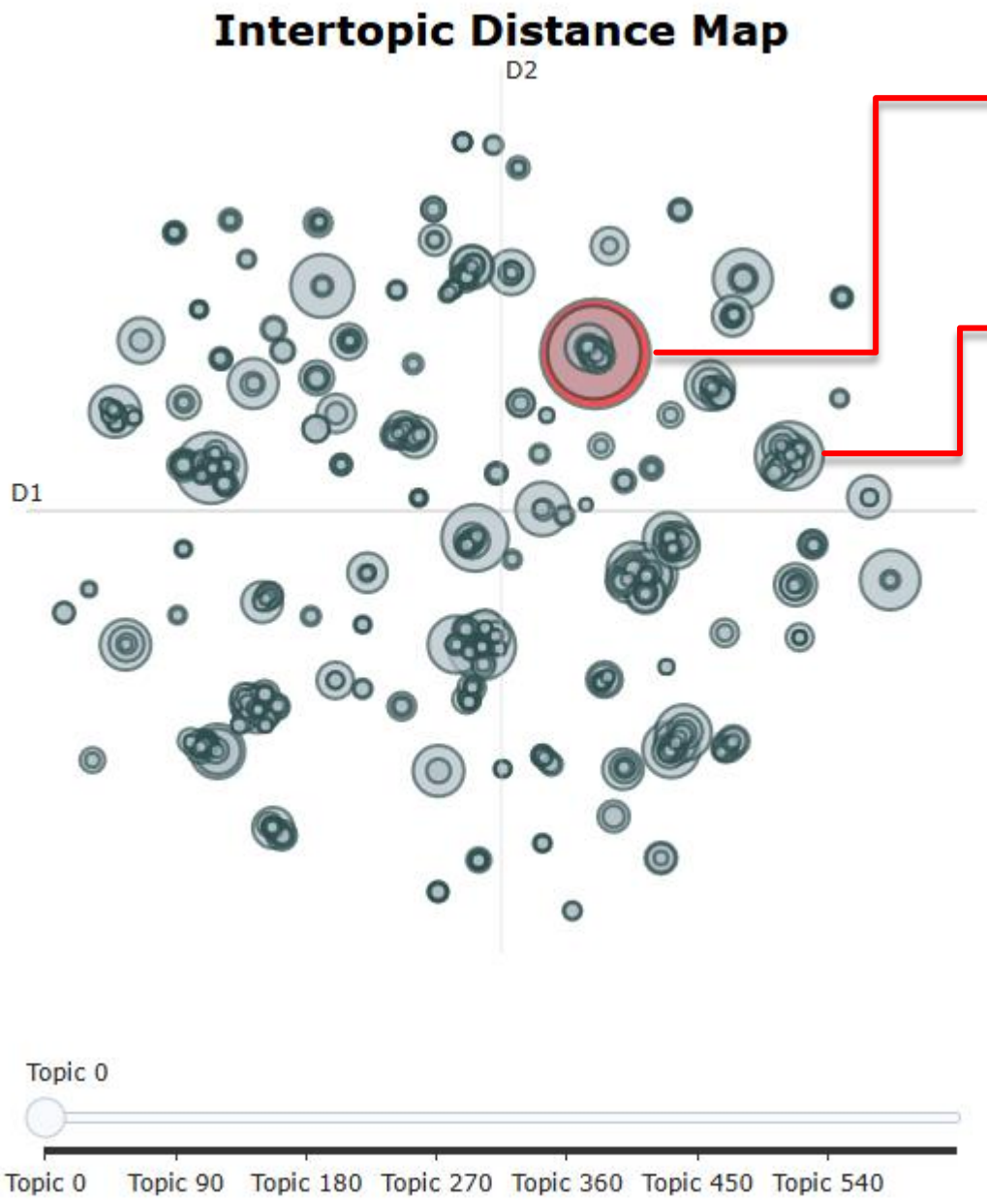
```python
# Topic Count
merged_df['topic'].value_counts()
```

topic	count
other	58991
MarketResearch	18439
EntertainmentTech	14882
PolicyandLegal	13980
CloudComputing	10652
Finance	6800
TechProducts	6585
Art	5472
AIinContentCreation	5000
HealthcareTech	4474
Media	4393
AIResearch	4081

Final topics and counts

topic_modeling.csv

Topic Modeling - Try BERTopic



['overviewview', 'cision', 'entertainment medium', 'general', 'send release', 'entertainment', 'people culture', 'consumer', 'resource', 'resource blog']

['market', 'analysis', 'artificial intelligence', 'report', 'artificial', 'growth', 'intelligence', 'forecast', 'global artificial', 'global']

Here are the key disadvantages of BERTopic compared to LDA:

- **Complexity**: BERTopic relies on transformers, which are computationally expensive and slower than LDA, especially for large datasets.
- **Resource Requirements**: BERTopic requires GPUs or powerful hardware for efficient processing, whereas LDA can run on modest setups.
- **Dependency on Pre-trained Models**: BERTopic's performance depends on pre-trained embeddings, which might not align perfectly with specific domains. LDA works directly on text without pre-trained dependencies.

I choose LDA results in this research.

NER (NAMED ENTITY RECOGNITION)

Clean text for NER

- Removing URLs.
- Removing or replacing certain special characters.

Split document into sentences

```
nlp = spacy.load("en_core_web_sm", disable=["ner", "tagger", "lemmatizer", "attribute_ruler"])
```

Converts the nested structure (articles containing multiple sentences) into a flat structure where each row represents a single sentence along with its article_id and sentence_id.

Define a function to perform NER using spaCy

```
def spacy_ner(text):
    doc = nlp(text)
    # Return a list of entities found in the text along with their labels
    return [(ent.text, ent.label_) for ent in doc.ents]
```

```
sentences_df.head(3)
```

	article_id	sentence_id	sentence	entities
0	1	1	Observation Simulation And AI Join Forces To Reveal A Clear Universe - SpaceRef Home NASA Watch SpaceRef Business Astrobiology Web Advertising Add an Event Sign up for our Daily Newsletter International Space Station NASA Hack Space Calendar Missions Space Weather Observation Simulation And AI Join Forces To Reveal A Clear Universe Press Release - Source NATIONAL INSTITUTES OF NATURAL SCIENCES Posted July 4 2021 1000 PM View Comments Using AI driven data analysis to peel back the noise and f...	[(July 4 2021, DATE), (AI, PRODUCT), (Universe, ORG)]
1	1	2	CREDIT The Institute of Statistical Mathematics Japanese astronomers have developed a new artificial intelligence AI technique to remove noise in astronomical data due to random variations in galaxy shapes.	[(CREDIT The Institute of Statistical Mathematics Japanese, ORG), (AI, PRODUCT)]
2	1	3	After extensive training and testing on large mock data created by supercomputer simulations they then applied this new tool to actual data from Japan's Subaru Telescope and found that the mass distribution derived from using this method is consistent with the currently accepted models of the Universe.	[(Japan, GPE), (Subaru Telescope, PRODUCT), (Universe, ORG)]

NER (named entity recognition)

	LOC	ORG	PRODUCT	GPE	PERSON	LANGUAGE	EVENT	LAW
0	Europe (14785)	AI (288028)	AI (659762)	US (211148)	Biden (13699)	English (15892)	the European Economic Area (1599)	the Terms Conditions and Privacy Policy (1490)
1	Africa (6765)	Google (86002)	YouTube (6099)	India (68629)	Elon Musk (9106)	Spanish (1235)	World Cup (1539)	the AI Act (1206)
2	North America (6470)	Microsoft (68703)	Android (5379)	U.S. (53004)	Musk (8987)	Arabic (725)	Olympics (1456)	the Securities Act (528)
3	Middle East (5216)	ChatGPT (66490)	UsMeet (4234)	China (50751)	Trump (8703)	Hindi (581)	World Sunrise Inside (1089)	our Visitor Agreement Privacy Policy (491)
4	Asia (5081)	Gray Media Group Inc. (48833)	Twitter (3711)	PRNewswire (42050)	Sam Altman (7616)	Mandarin (483)	Black Friday (833)	Privacy Policy Terms Conditions Advertise With Us (457)
5	Silicon Valley (4212)	Gray Media Group (32754)	Windows (3598)	UK (31996)	GPT-4 (6705)	French (196)	Series (816)	Chapter 3 (446)
6	Earth (3768)	Gray Television Inc. (32272)	JavaScript (3425)	Japan (22564)	Altman (5822)			
7	Asia Pacific (2380)	Nvidia (30745)	Galaxy (3141)	France (21444)	CaptioningAudio DescriptionAt (5325)			
8	the Middle East (2306)	Apple (30265)	Facebook (3111)	Russia (20785)	Joe Biden (5061)			
9	Latin America (2191)	OpenAI (30228)	Windows 11 (3031)	California (19423)	CaptioningAudio (4336)			
10	North America Europe (1874)	Facebook (25118)	Google Cloud (2953)	Canada (18721)	Donald Trump (4126)			
11	Mars (1696)	Amazon (24676)	Bing (2857)	Israel (18037)	Bing (3432)			
12	Americas (1491)	IBM (20069)	HelpSupport (2532)	Texas (18007)	AdvertisingAt Gray (3366)			

Count and display the top N most common entities

Full list of **spaCy Named Entity Recognition (NER) labels** with their descriptions:

1. **PERSON**: People, including fictional characters.
2. **NORP**: Nationalities, religious groups, or political groups.
3. **FAC**: Buildings, airports, highways, bridges, and other man-made structures.
4. **ORG**: Organizations, companies, institutions, or government agencies.
5. **GPE**: Countries, cities, states—geopolitical entities.
6. **LOC**: Non-political locations, such as mountains, bodies of water, and other geographical features.
7. **PRODUCT**: Objects, vehicles, foods, and other tangible products.
8. **EVENT**: Named events of historical significance, including wars, sports events, and natural disasters.
9. **WORK_OF_ART**: Titles of creative works like books, songs, and paintings.
10. **LAW**: Named legal documents and legislation.
11. **LANGUAGE**: Named languages.
12. **DATE**: Dates or periods, including absolute and relative dates or periods.
13. **TIME**: Times smaller than a day, such as specific times of day.
14. **PERCENT**: Percentage values, including the percent sign.
15. **MONEY**: Monetary values, including currency units.
16. **QUANTITY**: Measurements, weights, or distances.
17. **ORDINAL**: Terms that denote an order or rank in a sequence.
18. **CARDINAL**: Numerals that do not fall under another type.

NER (named entity recognition)

	Industry
0	Aerospace
1	Agriculture
2	Airline
3	Apparel
4	Automotive
5	Biotech
6	Biotechnology
7	Chemical
8	Communication Service
9	Construction
10	Consulting
11	Consumer

	Job
0	Engineer
1	Account Executive
2	Copywriter
3	Graphic Designer
4	Market Research Analyst
5	Product Manager
6	Agricultural Specialist
7	Biochemist
8	Biomedical Engineer
9	Research Coordinator
10	Microbiologist
11	Architect

	Technology
0	GenAI
1	Gen AI
2	Generative AI
3	GPT
4	GPT3.5
5	GPT-3.5
6	GPT4
7	GPT-4
8	ChatGPT
9	LLM
10	LLMs
11	Transformer

- Use pre-fined list of Industry, Job and Technology.
- Add manual entities in the same format as NER.
- Define a function to detect if keywords are present in a sentence.
- Merge 3 entities columns to 'entities_combined'

	IND	JOB	TECH
0	Software (128671)	Analyst (19776)	Cloud (122182)
1	Financial (100165)	Editor (17298)	Generative AI (120941)
2	Sports (87475)	Professor (15782)	Machine Learning (101190)
3	Entertainment (78697)	Scientist (10016)	ChatGPT (84544)
4	Healthcare (71748)	Writer (9737)	OpenAI (58864)
5	Education (70685)	Artist (8752)	Chatbot (37512)
6	Energy (70631)	Engineer (7149)	Cybersecurity (26310)
7	Government (63607)	Athlete (6204)	GPT (23726)
8	School (53135)	Teacher (4984)	ML (23338)
9	Consumer (49468)	Designer (3556)	Blockchain (21954)
10	Insurance (47628)	Data Scientist (2222)	IoT (15508)

Count and display the top N most common entities

	article_id	sentence_id	sentence	entities_ind	entities_job	entities_tech
95	3	1	Flash for AI - 28 March 2024 - EBV Electrolink - Dataweek Home About us Back issues E-book PDF Subscribe Advertise EMP Handbook Categories Editor's Choice Multimedia Videos AI ML Analogue Mixed Signal LSI Circuit System Protection ComputerEmbedded Technology Design Automation DSP Micros Memory Edge Computing IIoT Electronics Technology Enclosures Racks Cabinets Panel Products Events Interconnection Manufacturing Production Technology Hardware Services News Opto-Electronics Passive Components...	[[Hardware, IND], [Manufacturing, IND]]	[[Editor, JOB]]	[[ML, TECH], [Edge Computing, TECH], [IoT, TECH]]
1914	46	3	The study led by Professor Brian Lucey and Professor Michael Dowling utilized the AI model known as ChatGPT to draft an academic paper in finance.	[[Finance, IND]]	[[Professor, JOB]]	[[ChatGPT, TECH]]

FINE-TUNING

(fine-tuning distilbert-base-uncased using the Financial PhraseBank dataset from the Hugging Face Datasets Repository)

- Load the dataset from Hugging Face

```
dataset = load_dataset('takala/financial_phrasebank', 'sentences_allagree')
```

- Load the pre-trained distilbert-base-uncased model and modify it for sentiment

```
classificationmodel = DistilBertForSequenceClassification.from_pretrained(
   'distilbert-base-uncased', num_labels=3
)
```

- Use the DistilBERT tokenizer to preprocess the text data for training

```
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')
```

- Use the Hugging Face Trainer for training and evaluation

- After training, save the model and tokenizer for future use

```
model.save_pretrained('./financial-distilbert-lg’)
tokenizer.save_pretrained('./financial-distilbert-lg')
```

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.481200	0.356668	0.896247	0.893115	0.894153	0.896247
2	0.157800	0.182963	0.935982	0.935819	0.936635	0.935982
3	0.065300	0.165968	0.938190	0.937874	0.937645	0.938190
4	0.048200	0.170022	0.938190	0.938298	0.938419	0.938190

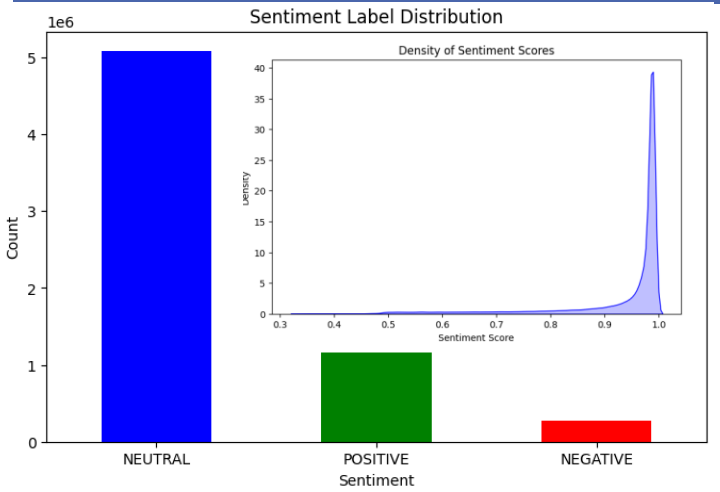
```
Article: The company announced record-breaking profits this quarter.
Sentiment: POSITIVE, Confidence: 0.9697
Article: The market crash led to widespread panic among investors.
Sentiment: NEGATIVE, Confidence: 0.9198
Article: The new product launch received mixed reviews.
Sentiment: NEGATIVE, Confidence: 0.6330
```

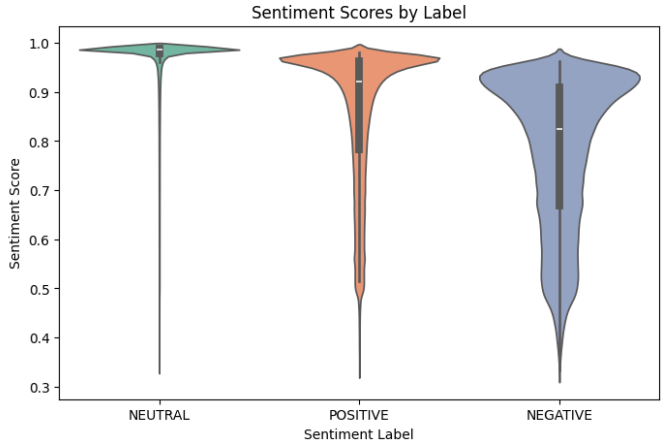
A high confidence score doesn't necessarily mean the sentiment is strongly positive or negative. It reflects the model's certainty in its prediction.

However, it can loosely be interpreted as the strength of the sentiment. If the confidence is close to 1.0, we can interpret it as the sentiment being strongly aligned with the predicted label.

Dataset - https://huggingface.co/datasets/takala/financial_phrasebank

SENTIMENT ANALYSIS ('./FINANCIAL-DISTILBERT-LG')





	article_id	sentence_id	__index_level_0__	sentiment_score_DistilBERT	sentiment_label_certainty
count	6.523802e+06	6.523802e+06	6.523802e+06	6.523802e+06	6.523802e+06
mean	7.689930e+04	2.641915e+01	3.261915e+06	9.307047e-01	1.202929e-01
std	4.437603e+04	1.979123e+01	1.883271e+06	1.138193e-01	3.822494e-01
min	1.000000e+00	1.000000e+00	0.000000e+00	3.338542e-01	-9.631420e-01
25%	3.843800e+04	1.100000e+01	1.630951e+06	9.416891e-01	0.000000e+00
50%	7.695000e+04	2.200000e+01	3.261916e+06	9.836959e-01	0.000000e+00
75%	1.153300e+05	3.700000e+01	4.892876e+06	9.884683e-01	0.000000e+00
max	1.537490e+05	2.450000e+02	6.523834e+06	9.916066e-01	9.797616e-01

Lexicon vs distilbert

Lexicon-based tools like VADER primarily analyze words individually, relying on predefined sentiment scores for each word or phrase. They do not deeply understand the connections between words or the context of the sentence.

Limitations:

- No nuanced understanding
- Lack of contextual relationships
  For deeper contextual understanding, models like BERT are better suited because they use contextual embeddings to analyze the relationships between words.

So, I chose my fine tuned distilbert model rather than Lexicon-based tools.

	article_id	sentence_id	sentence	__index_level_0__	sentiment_score_DistilBERT	sentiment_label_DistilBERT	sentiment_label_certainty
0	1	1	Observation Simulation And AI Join Forces To R...	0	0.989424	NEUTRAL	0.000000
1	1	2	CREDIT The Institute of Statistical Mathematic...	1	0.541715	POSITIVE	0.541715
2	1	3	After extensive training and testing on large ...	2	0.933930	NEUTRAL	0.000000

sentences_with_sentiments_certainty.csv

- Apply my fine-tuned model (financial-distilbert-lg) to sentences.csv
- Assign the sentiment_score_DistilBERT as-is for POSITIVE labels.
- Assign 0 for NEUTRAL labels, since they don't contribute strongly to sentiment.
- Assign a negative version of sentiment_score_DistilBERT to NEGATIVE labels.
- Use "sentiment_label_certainty" to tell how certain the sentiment_label_DistilBERT is for a sentence, treating sentiment_label_certainty as a proxy for strength.

MERGING DATAFRAMES

Merge the dataframes on sentence level
on=['article_id', 'sentence_id']

df_sentences

df_sentences_with_entities

df_sentences_with_job_ind_tech

df_sentences_with_sentiments_certainty

Merge article level info into sentence level dataframe
on=['article_id']

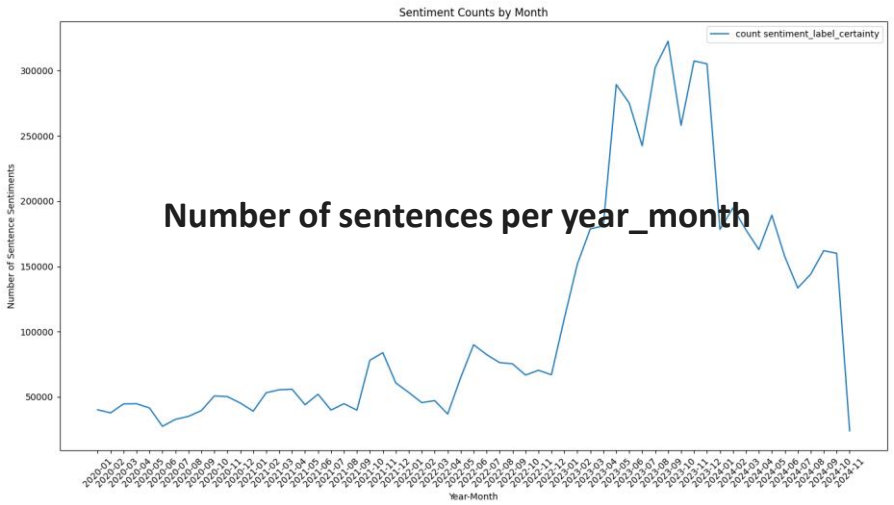
df_topic_modeling

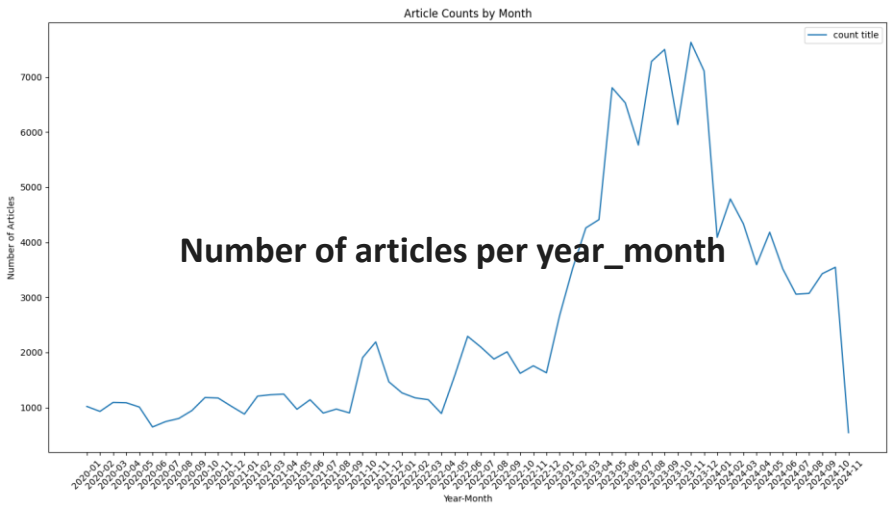
Final merged dataframe

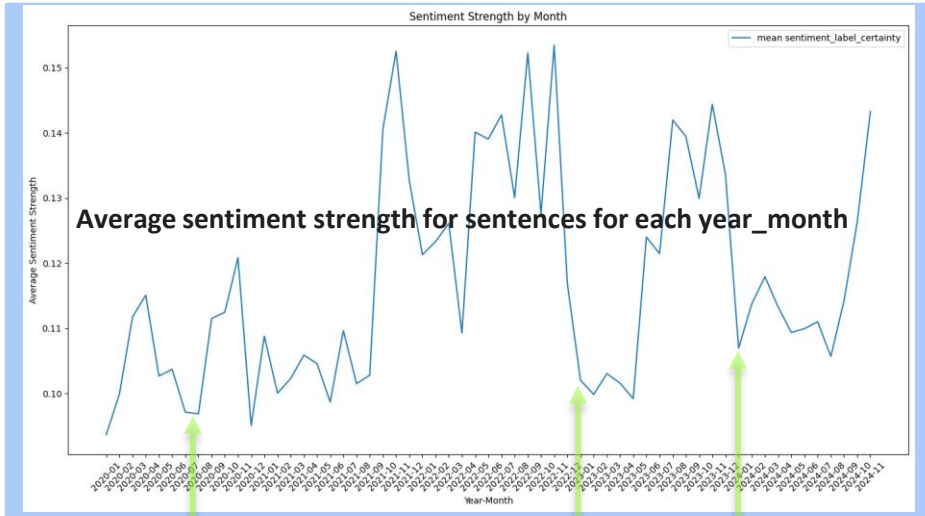
	article_id	sentence_id	sentence	entities	entities_combined	sentiment_score_DistilBERT	sentiment_label_DistilBERT	sentiment_label_certainty	date	topic
0	1	1	Observation Simulation And AI Join Forces To Reveal A Clear Universe - SpaceRef Home NASA Watch SpaceRef Business Astrobiology Web Advertising Add an Event Sign up for our Daily Newsletter International Space Station NASA Hack Space Calendar Missions Space Weather Observation Simulation And AI Join Forces To Reveal A Clear Universe Press Release - Source NATIONAL INSTITUTES OF NATURAL SCIENCES Posted July 4 2021 1000 PM View Comments Using AI driven data analysis to peel back the noise and f...	[('July 4 2021', 'DATE'), ('AI', 'PRODUCT'), ('Universe', 'ORG')]	[]	0.989424	NEUTRAL	0.000000	2021-07-05	AIResearch
1	1	2	CREDIT The Institute of Statistical Mathematics Japanese astronomers have developed a new artificial intelligence AI technique to remove noise in astronomical data due to random variations in galaxy shapes.	[('CREDIT The Institute of Statistical Mathematics Japanese', 'ORG'), ('AI', 'PRODUCT')]	[]	0.541715	POSITIVE	0.541715	2021-07-05	AIResearch
2	1	3	After extensive training and testing on large mock data created by supercomputer simulations they then applied this new tool to actual data from Japan's Subaru Telescope and found that the mass distribution derived from using this method is consistent with the currently accepted models of the Universe.	[('Japan', 'GPE'), ('Subaru Telescope', 'PRODUCT'), ('Universe', 'ORG')]	[]	0.933930	NEUTRAL	0.000000	2021-07-05	AIResearch
3	1	4	This is a powerful new tool for analyzing big data from current and planned astronomy surveys.	[]	[['Big Data', 'TECH']]	0.789955	NEUTRAL	0.000000	2021-07-05	AIResearch
4	1	5	Wide area survey data can be used to study the large-scale structure of the Universe through measurements of gravitational lensing patterns.	[('Universe', 'ORG')]	[]	0.990115	NEUTRAL	0.000000	2021-07-05	AIResearch

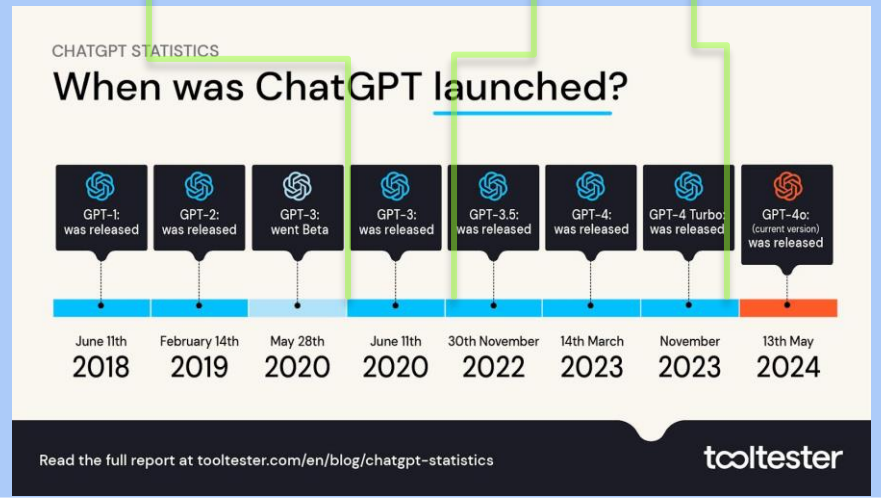
OVERALL ANALYTICS

The number of provided articles AND parsed sentences after 2023 are declining.
It does not necessarily mean the interest in AI is declining.







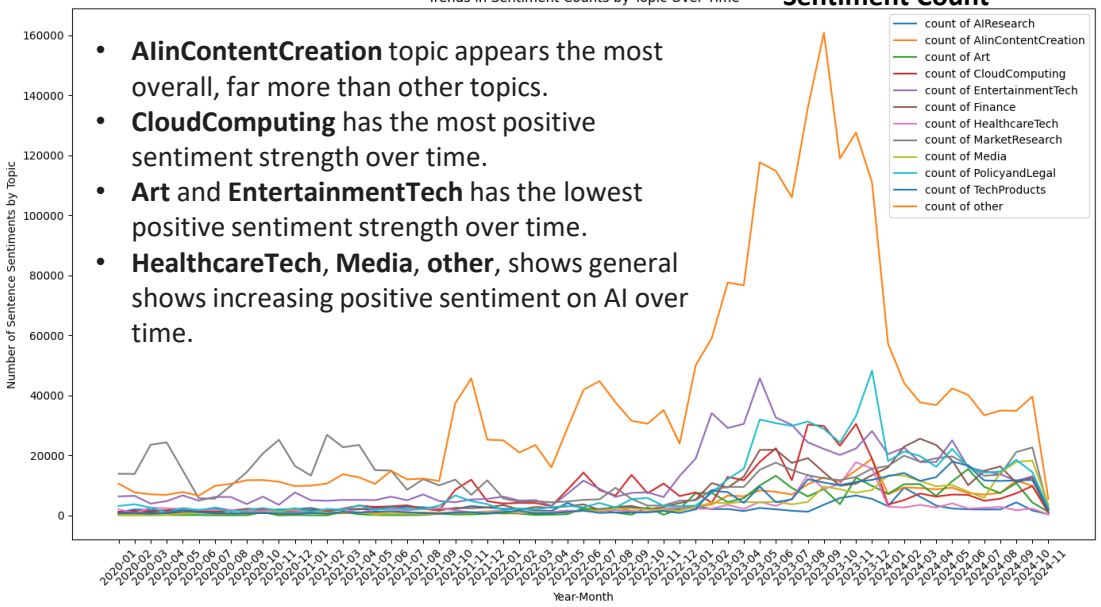


Interestingly, the major releases of ChatGPT **coincide** with periods of relatively lower average sentiment strength on the 'Sentiment Strength by Month' chart. This **inverse relationship** may suggest a shift in focus or public sentiment during these key milestones.

Overall Sentiment

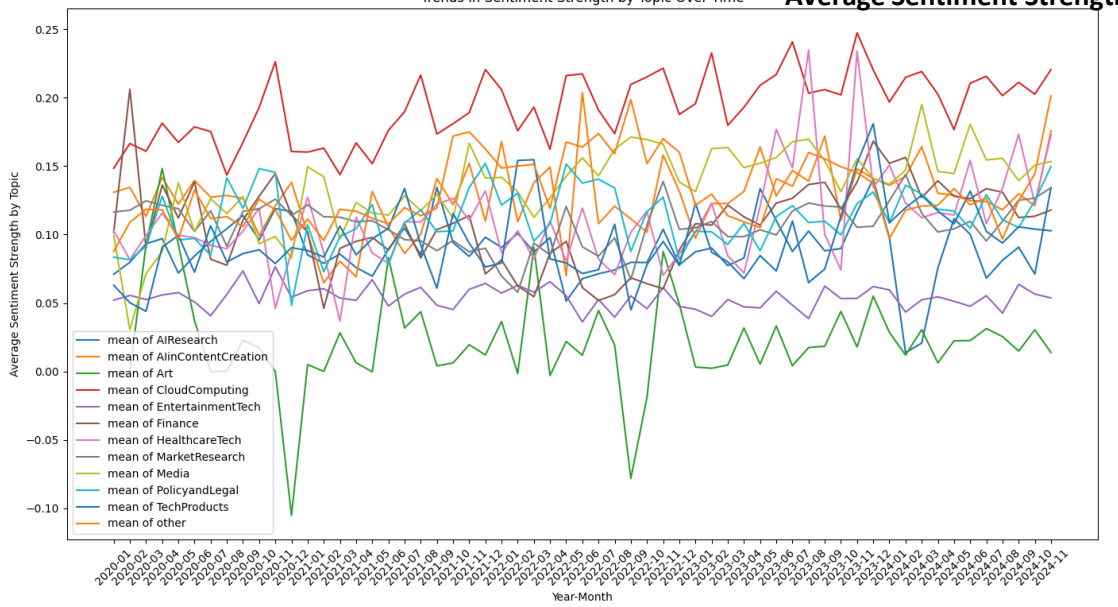
Topic-level sentiment analysis (customized)

"Trends in Sentiment Counts by Topic Over Time" **- Sentiment Count**



- **AIinContentCreation** topic appears the most overall, far more than other topics.
- **CloudComputing** has the most positive sentiment strength over time.
- **Art** and **EntertainmentTech** has the lowest positive sentiment strength over time.
- **HealthcareTech**, **Media**, **other**, shows general shows increasing positive sentiment on AI over time.

"Trends in Sentiment Strength by Topic Over Time" **- Average Sentiment Strength**



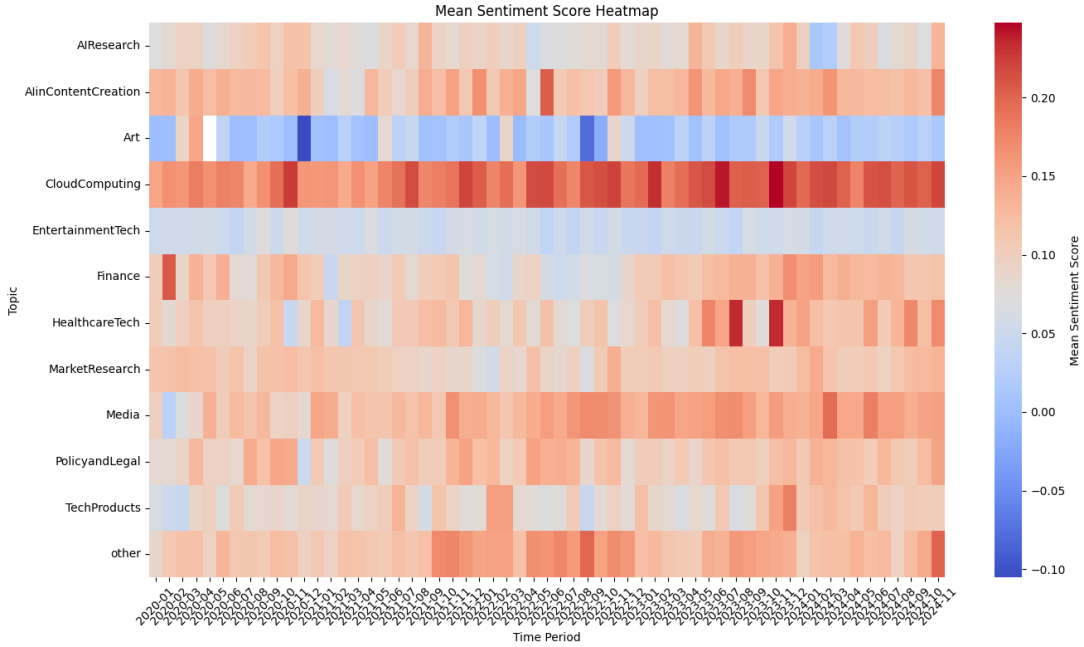
Text extracts shows strong negative sentiment on AI in Art and Entertainment related to job losses.

"The conversation surrounding the use of AI in art has been exacerbated by the Writers Guild of America and Screen Actors Guild-American Federation of Television and Radio Artists strikes with their concerns including the threat of artificial intelligence."

"Ziv Epstein a researcher at the MIT Media Lab's Human Dynamics Group says the advancement of AI image generators complicates notions of ownership in the art industry."

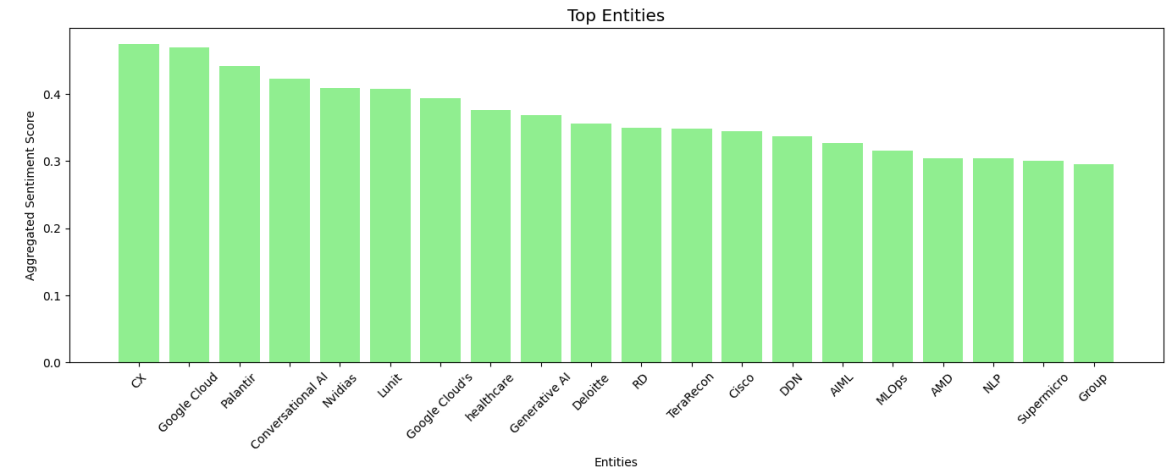
"And in the entertainment industry Merchant notes writers and actors have been striking to protest attempts by studios to use AI to degrade their pay and job stability."

"Generative AI raises job loss concerns in entertainment industry Generative AI particularly in film and animation is anticipated to result in more job losses according to CVL Economics."



Sentiment by ORG

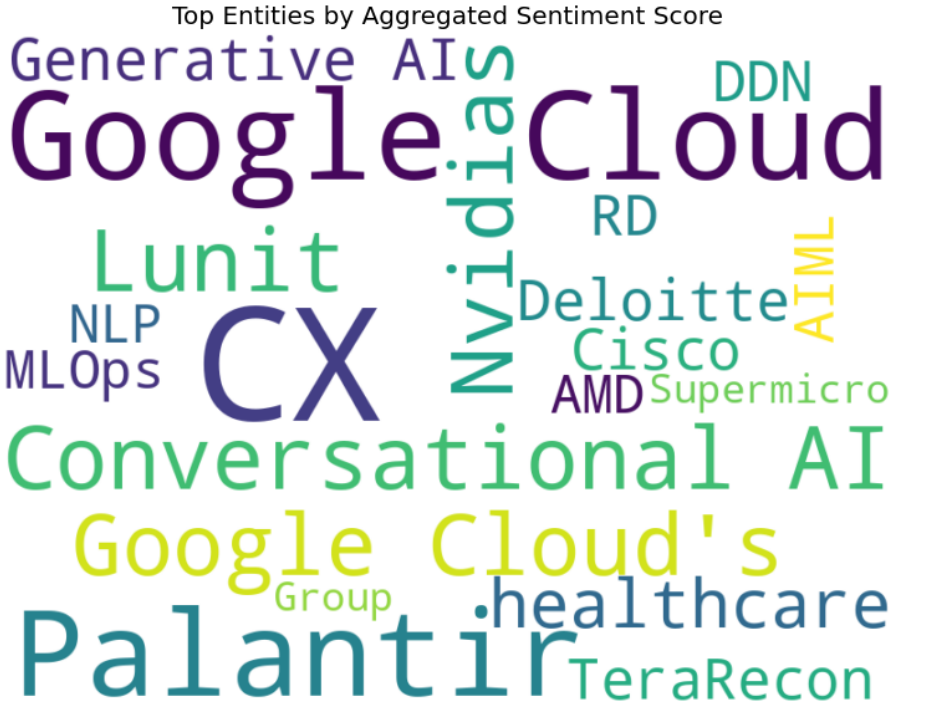
What types of companies are planning to invest in these technologies today or near future (success stories)



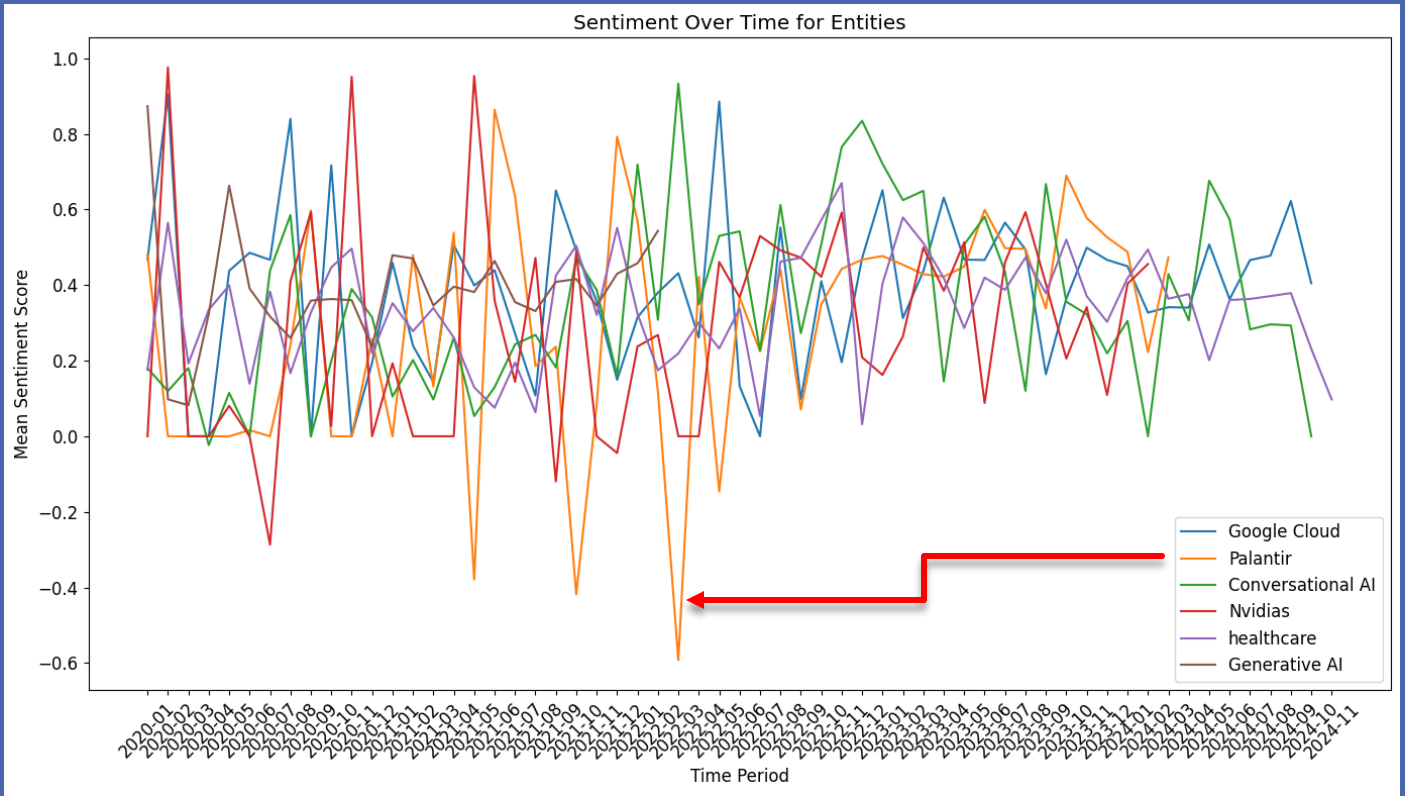
To accelerate transformative capabilities, companies, academic institutions, and governments should invest in AI-driven healthcare, develop advanced conversational AI tools, enhance cloud and computing infrastructure, and promote collaborations with leading organizations like Google Cloud, Palantir, and Nvidia.

Google Cloud, Palantir, Conversational AI, Nvidia, healthcare, and Generative AI are leading organizations gaining positive sentiment for their significant investments in AI, particularly in advancing transformative capabilities.

Demonstrate what companies, academic institutions, and government entities can do to accelerate the development of these transformative capabilities



Sentiment by ORG



Palantir's sentiment instability reflects its position as a **polarizing company** in the tech and analytics space.

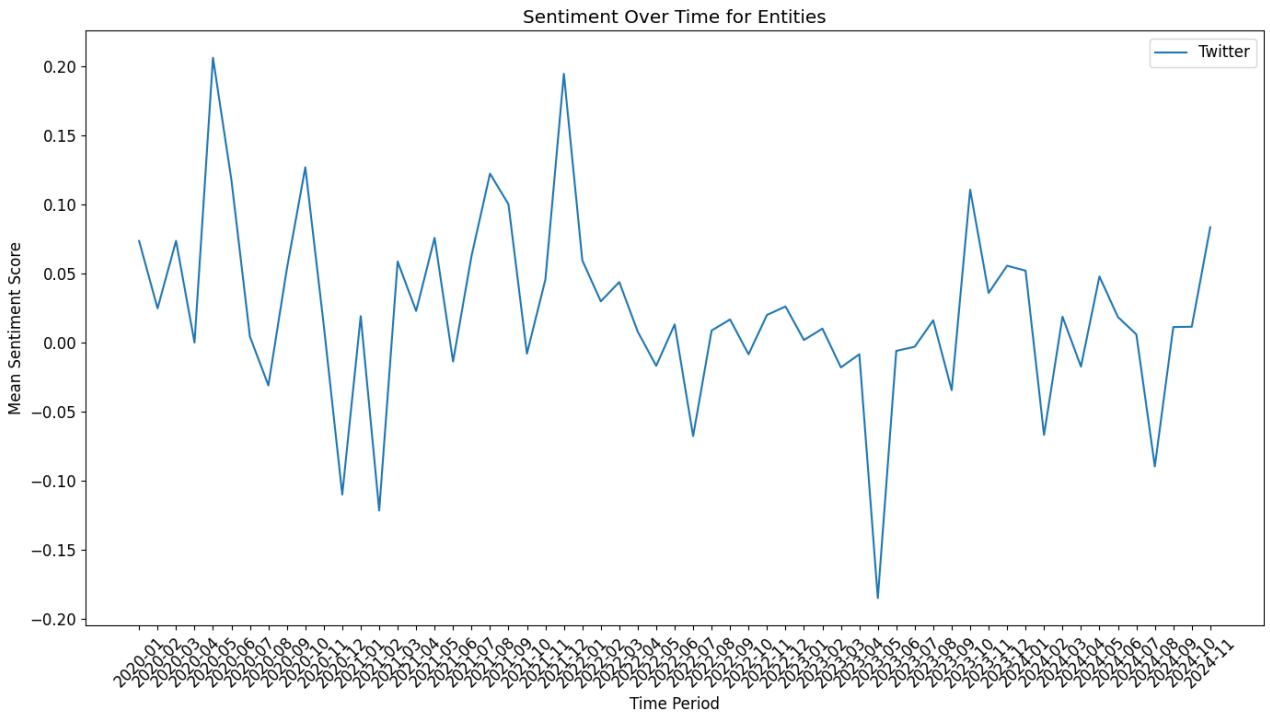
Events like **financial disclosures**, **contract announcements**, and **public debates** can lead to sharp fluctuations in sentiment over time.

This variability is common for companies operating in controversial and high-impact industries.

What happened to Palantir in 2022-02?

	year_month	sentence	sentiment_label_certainty
1085625	2022-02	After extensive experience building data infrastructure and managing data science teams at Two Sigma Palantir and Google co-founders Jakub Jurovych Jan Matas and Filip Stollar set out to build an innovative data science notebook.	0.000000
2214323	2022-02	After adjusting for stock-based compensation and other expenses Palantir earned 2 cents a share down from 3 cents a share a year earlier while analysts trac57m agoBarrons.comPalantirs	-0.959173
2214326	2022-02	The data analytics software company also issues first-quarter revenue guidance better than Wall Street estimates.37m agoReutersSoftware firm Palantir boosts revenue view on commercial strengthPalantir Technologies forecast current-quarter sales above estimates on Thursday after a steady flow of government contracts and a growing commercial portfolio boosted the data analytics software firm's fourth-quarter revenue.	0.971964
2214327	2022-02	Known for its work with the U.S. Army the Central Intelligence Agency and other government bodies Palantir's next leg of growth is widely expected to come from commercial contracts with large businesses.1h agoTipRanksJ.P. Morgan Says Buy These 2 Stocks as They Are Oversold2022 has started out with a marked increase in market volatility accompanied by a sharp reversal of last years bullish trend.	0.935774

Sentiment by ORG



Twitter

The **huge drop** in sentiment in **April 2023** can be attributed primarily to:

- **Public backlash against the paid verification system.**
- **Concerns about AI risks raised by Elon Musk and Steve Wozniak.**

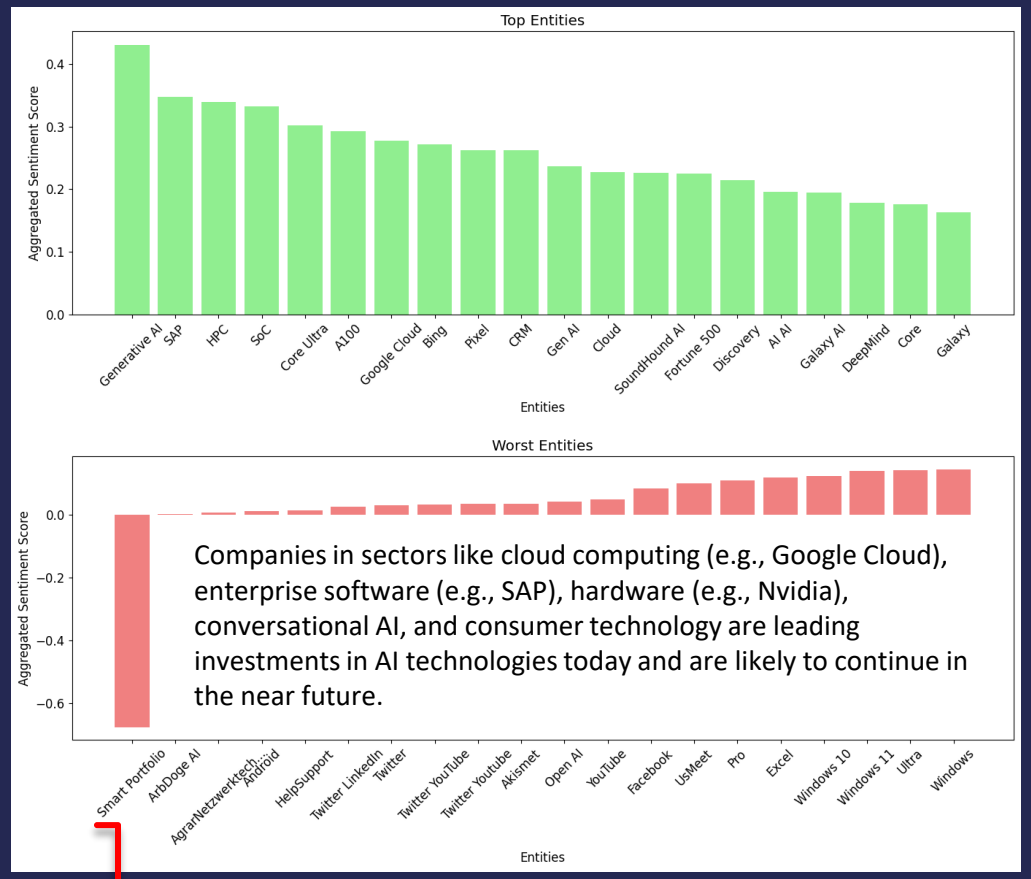
These events, combined with ongoing controversies about Musk's leadership and changes to Twitter, amplified the negative sentiment.

What happened to Twitter in 2023-04?

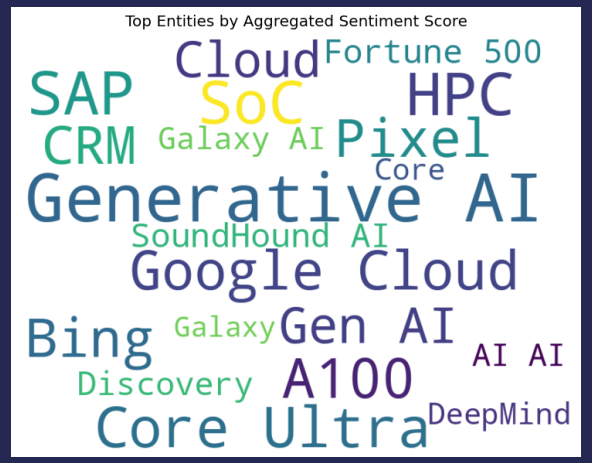
	year_month	sentence	sentiment_label_certainty
6840	2023-04	agoAFPMusk defends paid Twitter as blue tick ultimatum loomsElon Musk on Friday defended his controversial pay model for Twitter claiming that any social media platform that didn't follow suit would fail because they would be swarmed by bots.	-0.490234
16578	2023-04	Did He Keep Any of Them Is There Any Financial Benefit to Paying 8 for Twitter's Blue Check Mark	0.000000
21287	2023-04	Bloomberg Musk joked that his dog is in charge of Twitter.	0.000000
21288	2023-04	WP Were witnessing the brain death of Twitter.	0.000000
67054	2023-04	Tesla and Twitter chief Elon Musk and Apple co-founder Steve Wozniak have warned that human competitive intelligence could pose serious risks to society and humanity.	-0.916336

Sentiment by PRODUCT

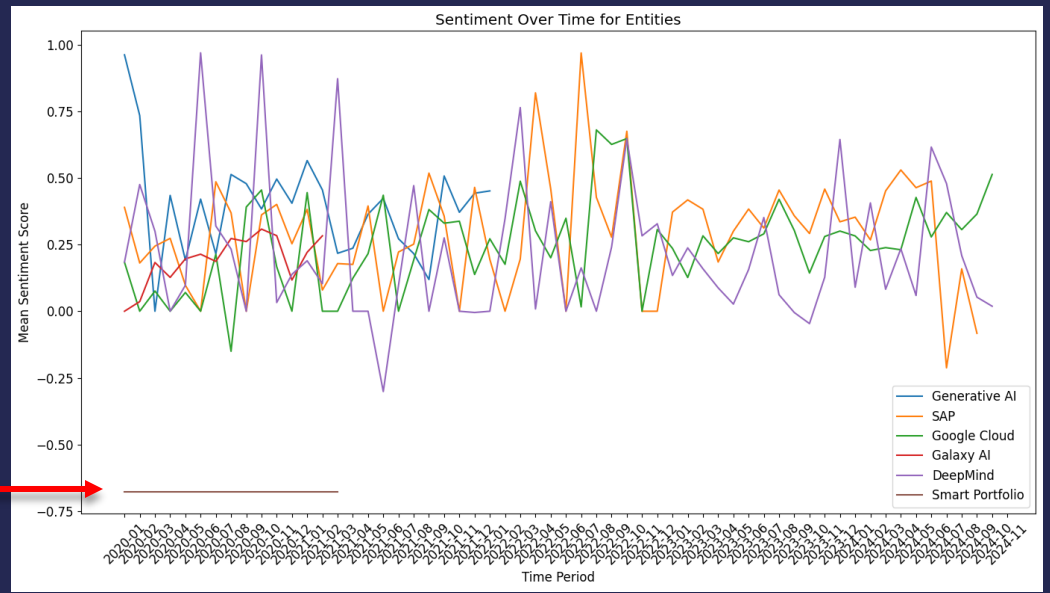
What types of companies are planning to invest in these technologies today or near future (success stories)



Companies in sectors like cloud computing (e.g., Google Cloud), enterprise software (e.g., SAP), hardware (e.g., Nvidia), conversational AI, and consumer technology are leading investments in AI technologies today and are likely to continue in the near future.

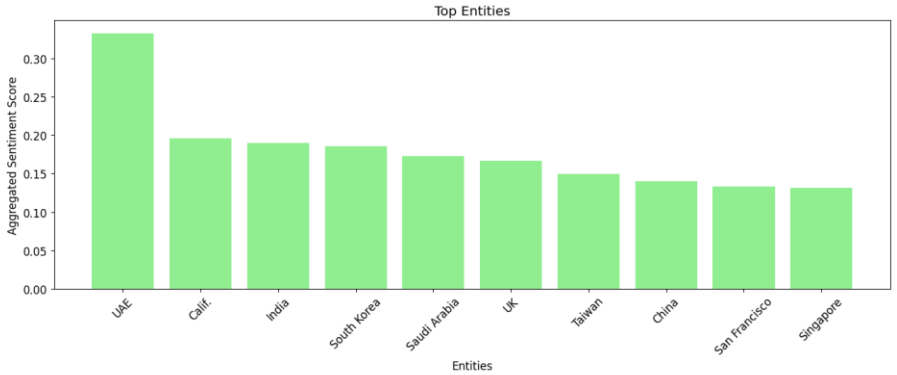


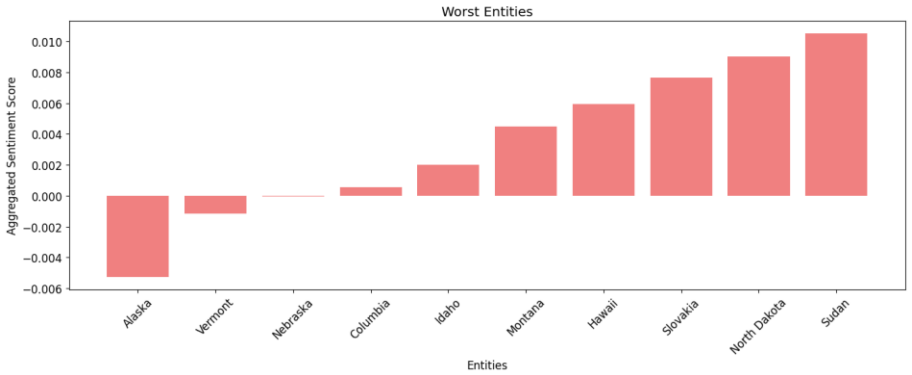
Smart Portfolio has significantly negative sentiment.

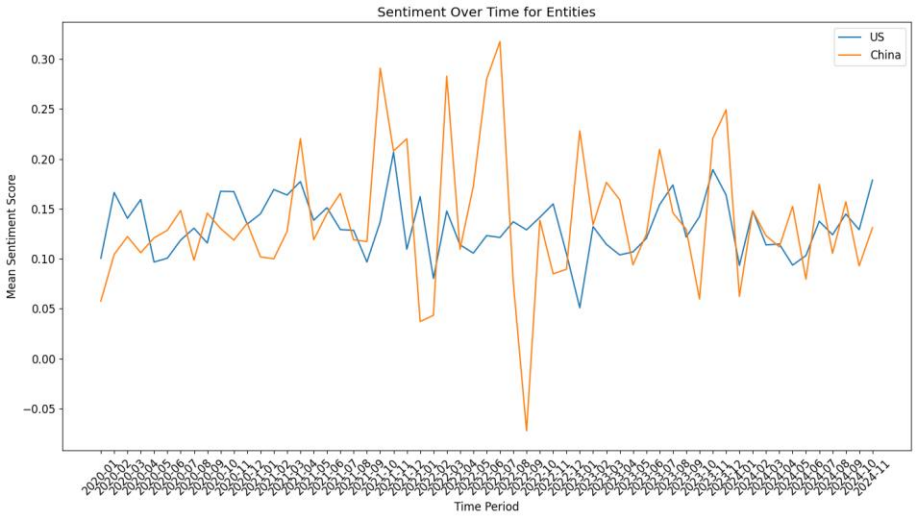


	article_id	sentence_id	sentence	entities	entities_combined	sentiment_score_DistilBERT	sentiment_label_DistilBERT	sentiment_label_certainty	date	topic	year_month
1780000	41957	75	Save data Data is currently not available Opt in to Smart Portfolio We are currently experiencing technical difficulties please try again.	[(Save data Data, ORG), (Smart Portfolio, PRODUCT)]	[]	0.675827	NEGATIVE	-0.675827	2023-09-29	Finance	2023-09
1161688	27442	76	Save data Data is currently not available Opt in to Smart Portfolio We are currently experiencing technical difficulties please try again.	[(Save data Data, ORG), (Smart Portfolio, PRODUCT)]	[]	0.675827	NEGATIVE	-0.675827	2023-09-25	Finance	2023-09

Sentiment by GPE (location)





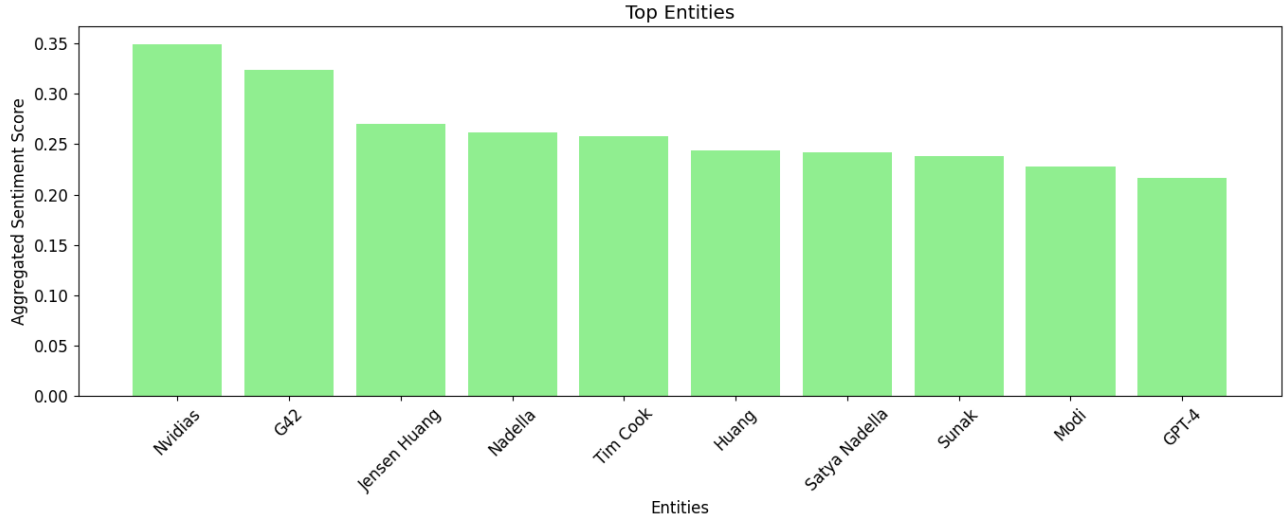


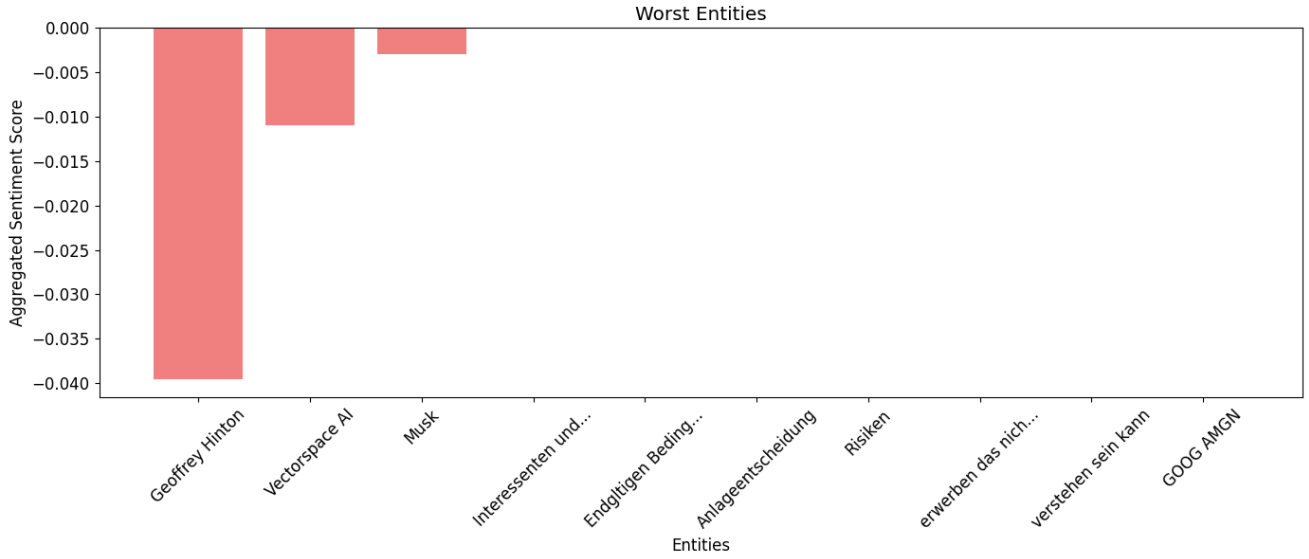
UAE, India and California score high in positive sentiment while Alaska scores low.

The greater fluctuation in sentiment toward **China** compared to the **US** can be summarized as follows:

- **Diverse and Polarizing Topics**: Sentiment toward China is influenced by varied and polarizing issues like IT trade and geopolitics, which evoke strong reactions.
- **Event-Driven Volatility**: Specific geopolitical events (e.g., trade disputes, technology sanctions) often trigger sharp sentiment shifts for China, while the US narrative tends to be steadier.
- **Media Bias**: Western media's polarized reporting on China tech industries amplifies sentiment swings, whereas the US receives more predictable coverage.
- **Mention Volume**: A potentially smaller volume of mentions for China may amplify sentiment variability compared to the US, which is mentioned more frequently.
- **Polarizing Issues**: Topics like plagiarism in technology and regional conflicts drive significant sentiment changes for China, unlike the more consistent themes associated with the US.

Sentiment by PERSON



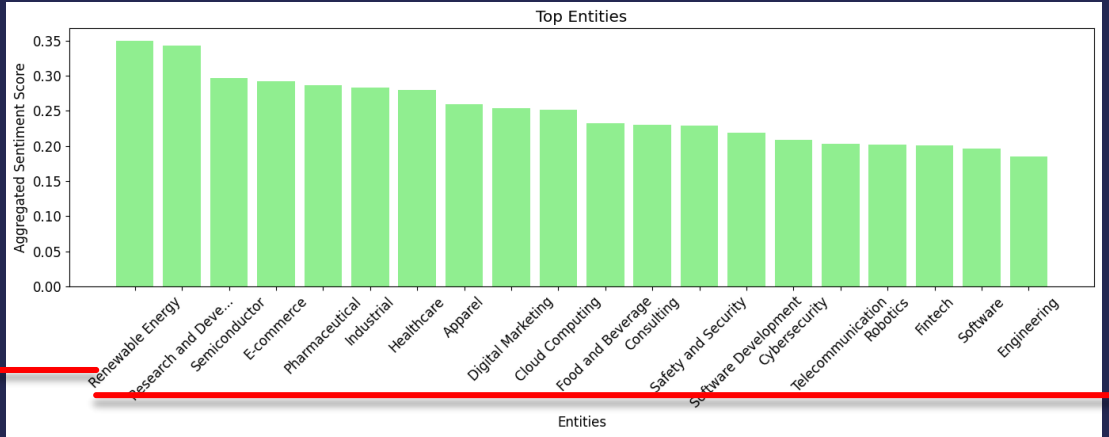


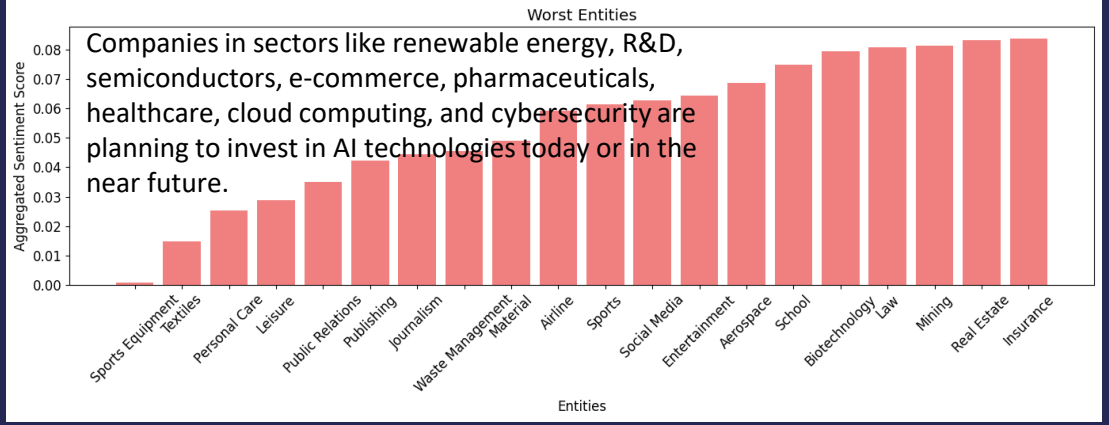
The low sentiment scores for **Geoffrey Hinton** and **Elon Musk** stem from their association with controversial or polarizing topics: Hinton's warnings about AI risks and Musk's actions at Twitter/X often invite criticism.

Conversely, **Jensen Huang** and **Tim Cook** score high due to their association with innovation and stability. Huang is praised for NVIDIA's leadership in AI, while Cook's steady and inclusive management style reinforces Apple's positive image. The difference reflects public perception shaped by their actions, industries, and media portrayal.

Sentiment by IND (Industry)

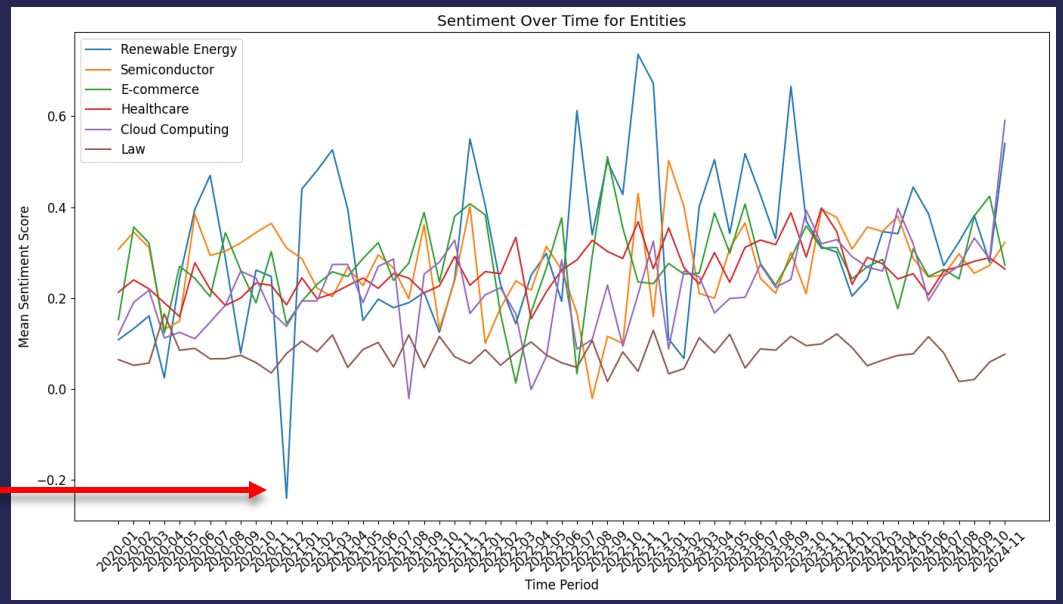
Identify top industries that experienced successful or unsuccessful AI integration What industries are going to be most impacted by AI?
Plot a timeline to illustrate how sentiment is changing over time What types of companies are planning to invest?





Companies in sectors like renewable energy, R&D, semiconductors, e-commerce, pharmaceuticals, healthcare, cloud computing, and cybersecurity are planning to invest in AI technologies today or in the near future.

Industries like renewable energy, healthcare, pharmaceuticals, semiconductors, e-commerce, cloud computing, and cybersecurity are likely to be most impacted by AI over the next several years.

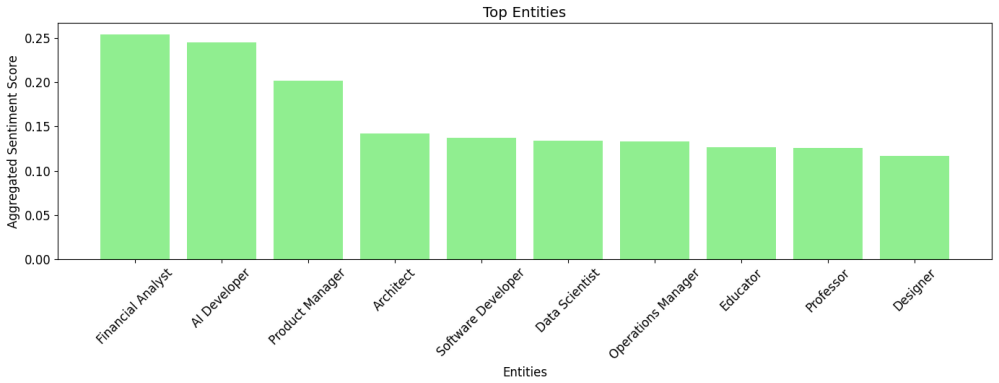


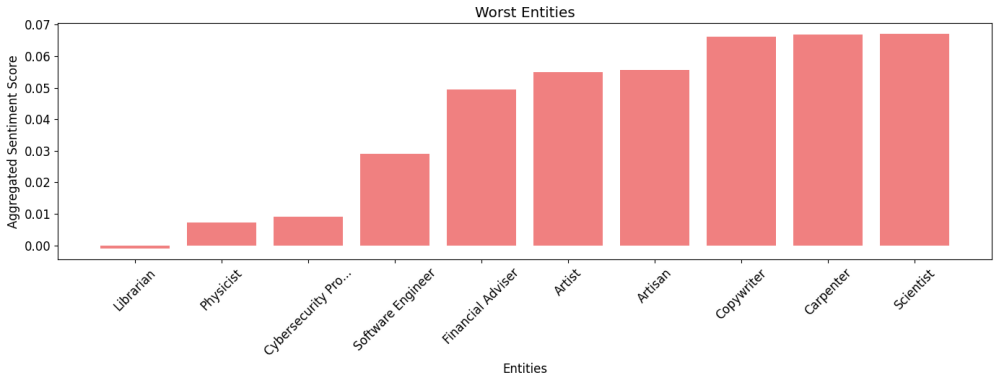
Renewable Energy ranked highest sentiment overall, but it plummet in 2020-12.
Below shows potential AI's negative impact on cutting greenhouse emissions and slowing down climate change.

	year_month	sentence	entities	sentiment_label_certainty
216741	2020-12	Unless we switch to 100 renewable energy sources AI progress may stand at odds with the goals of cutting greenhouse emissions and slowing down climate change.	[[Energy, IND], [Renewable Energy, IND]]	-0.932199

Sentiment by JOB

Identify top industries / job that experienced successful or unsuccessful AI integration





Successful AI Integration:

•Top Professions with Positive Sentiment: Financial Analyst, AI Developer, Product Manager, Architect.

•Reasons for Positive Sentiment:

- **Efficiency Gains:** AI helps automate repetitive and time-consuming tasks, allowing professionals to focus on strategic and creative aspects.
- **Enhanced Decision-Making:** Data-heavy roles like Financial Analyst and AI Developer leverage AI for predictive analytics, insights generation, and innovative problem-solving.
- **Industry Evolution:** These roles are at the forefront of industries that embrace AI as a growth enabler, fostering job security and new opportunities.

Unsuccessful AI Integration:

•Top Professions with Negative Sentiment: Librarian, Artist, Copywriter, Scientist.

•Reasons for Negative Sentiment:

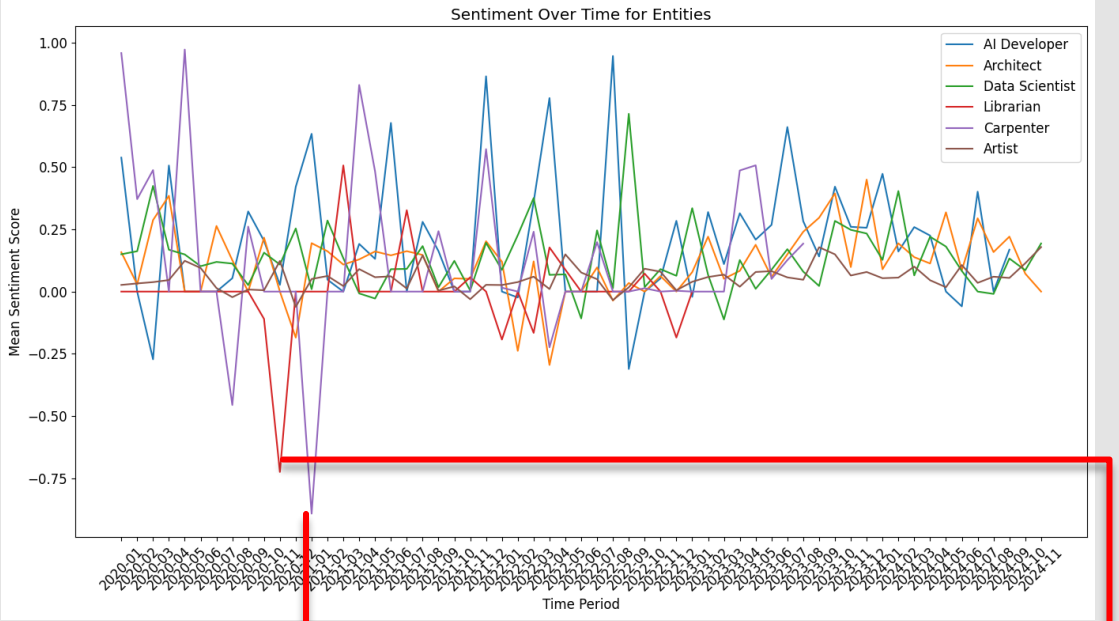
- **Automation Threat**
- **Creative Disruption:** Creative jobs, such as Artist and Copywriter, are impacted by AI-generated content, raising concerns over originality, job relevance, and authenticity.
- **Job Uncertainty:** Perception of AI as a replacement rather than a tool creates resistance and fear among professionals in these fields.

Key Insights:

- **AI as an Enabler vs. Threat:**
  - Positive sentiment reflects roles where AI is viewed as a productivity enabler or collaborator.
  - Negative sentiment indicates fear or skepticism toward AI's impact on traditional skillsets and roles.
- **Creative Industries Face Unique Challenges:**
  - Creative roles require strategies to integrate AI as a tool for amplifying originality rather than replacing it.
- **Call for Reskilling:**
  - Affected professions must adapt to AI-driven changes through upskilling initiatives and exploring hybrid roles that combine human creativity with AI capabilities.
- **Future Opportunities:**
  - Roles embracing AI are likely to shape future industry landscapes, emphasizing adaptability, innovation, and collaboration between humans and AI.

Sentiment by JOB

Plot a timeline to illustrate how sentiment is changing over time
Suggest why certain jobs are more likely to be impacted by AI



The text extracts highlights **AI's impact on librarians** in the context of **reducing CAD model build times** through an **AI-driven part creation methodology**. It suggests that advancements in AI are transforming traditionally manual, time-consuming tasks (like cataloging or model creation) into automated processes.

The **negative sentiment** associated with librarians in the context of AI advancements stems from the following factors:

- **Job Displacement Fears:** AI-driven automation in tasks like cataloging, data management, and digital model creation could lead to concerns about the redundancy of traditional librarian roles, particularly in technical or specialized areas like CAD libraries.
- **Perception of Devaluation:** The emphasis on AI's efficiency might overshadow the human expertise librarians bring, creating a sentiment that their skills are undervalued or replaced.
- **Skill Adaptation Pressure:** Librarians may feel pressured to rapidly adapt to new technologies, which can be seen as a challenge rather than an opportunity, especially for those not well-versed in AI or technical skills.
- **Industry Shift:** AI's ability to handle traditionally manual librarian tasks could lead to the reallocation of resources away from human roles, fostering a perception of reduced job security.

This negative sentiment reflects concerns about the rapid pace of AI adoption and its impact on roles requiring nuanced human judgment and expertise.

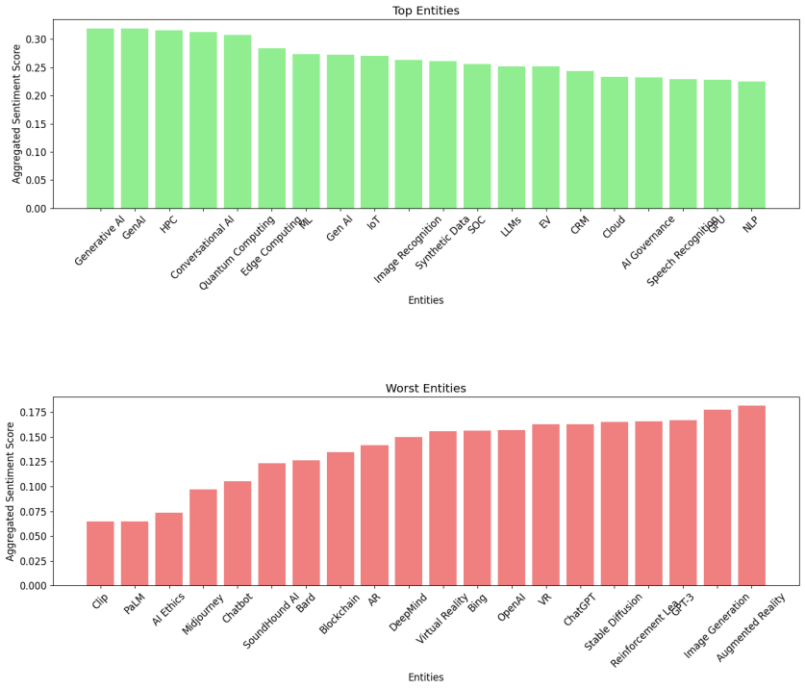
1601653	2023-09	Workflow 1 Workforce 3 Workplace 5 Todays Business NewsSearch IBN Search for Partner SitesIT Business Net Digital Producer Magazine Digital Media Net Media Hub Consumer Electronics Net Health Technology Net Contact About IT Business Net Recent Articles Ultra Librarian to Dramatically Reduce CAD Model Build Times with New AI-Driven Part Creation Methodology Discern Security Lands Funding Round As It Launches	[[Consumer, IND], [Librarian, JOB]]	-0.779318
1601662	2023-09	Social Media Software Software Development Space Industry Space Optics Spatial Intelligence SSD Storage Streaming Supply Chain Technology Threat Detection Touchscreen Training Transformation Video Conferencing Video Intelligence Video Wall VOD Voice Generation VPN VR Wearable Tech Web Web 3.0 Webcam Wi-Fi Windows Wireless Workflow Workforce Workplace You may have missed AI CAD News Ultra Librarian to Dramatically Reduce CAD Model Build Times with New AI-Driven Part Creation Methodology AI Cy...	[[Software, IND], [Social Media, IND], [Cybersecurity, IND], [Software Development, IND], [Librarian, JOB], [Cybersecurity, TECH], [VR, TECH]]	-0.922050

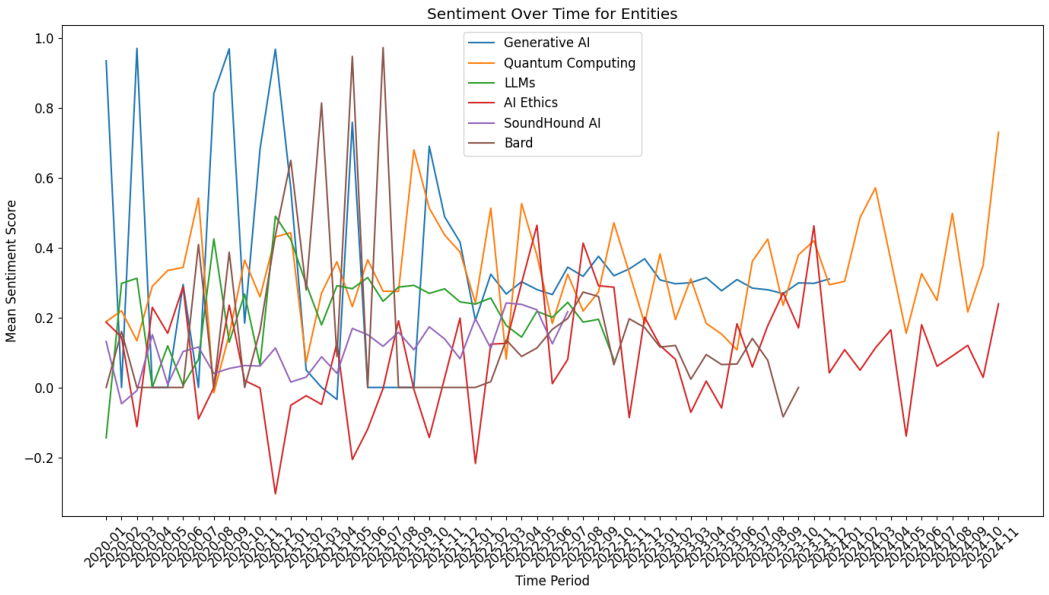
Turns out **Carpenter** is a person name

	year_month	sentence	entities	sentiment_label_certainty
2933157	2021-02	More Coverage L.A. Unified gets 100 doses but needs thousands to reopen schools Signs of hope as virus cases drop in California Entertainment Arts Male Buffy stars back Charisma Carpenter others alleging misconduct by Joss Whedon Entertainment Arts Male Buffy stars back Charisma Carpenter others alleging misconduct by Joss Whedon David Boreanaz Adam Busch Tom Lenk and Danny Strong address Whedons alleged misconduct on social media.	[[Entertainment, IND], [Social Media, IND], [Carpenter, JOB]]	-0.910784
4754195	2021-02	Impact Analysis InForGrowth Search for Recent Posts Global Substation Automation and Integration Market 2020 Share Size Import Export Growth and Outlook by 2025 Surface Mount Technology Market Analysis Type Size Trends Key Players and Forecast 2017 to 2025 Controlled Expansion Alloys Market Strategic Assessment And Forecast Till 2026 Sandvik National Electronic Alloys Hitachi Metals Mitsubishi Materials Nippon Yakin Carpenter Technology Corporation Precious Metal Refining Services Market Siz...	[[Healthcare, IND], [Insurance, IND], [Law, IND], [Metal, IND], [Carpenter, JOB]]	0.000000

Sentiment by TECH

What types of applications cannot by transformed by AI, based on the state of technology (failures)





High Sentiment for Quantum Computing:

- Quantum Computing shows generally high sentiment, with significant spikes, indicating positive coverage likely driven by breakthroughs or advancements.

Fluctuating Sentiments:

- Entities like **Generative AI**, **AI Ethics**, and **Bard** exhibit frequent fluctuations, reflecting both positive and negative coverage. This may stem from controversies, ethical concerns, or shifting public opinion.

AI Ethics Sentiment Challenges:

- AI Ethics consistently struggles with low sentiment scores compared to others. This suggests ongoing debates or criticism regarding fairness, bias, or ethical implementation in AI systems. Cannot be by transformed by AI.

Stable Positive Sentiment for Conversational AI and LLMs :

- **Conversational AI** and **LLMs** maintain a stable sentiment around a positive baseline, reflecting consistent progress and acceptance in the field.

SoundHound AI Sentiment Variability:

- SoundHound AI has lower scores with frequent dips, possibly linked to criticism or struggles in its application or competitive landscape.

Insights:

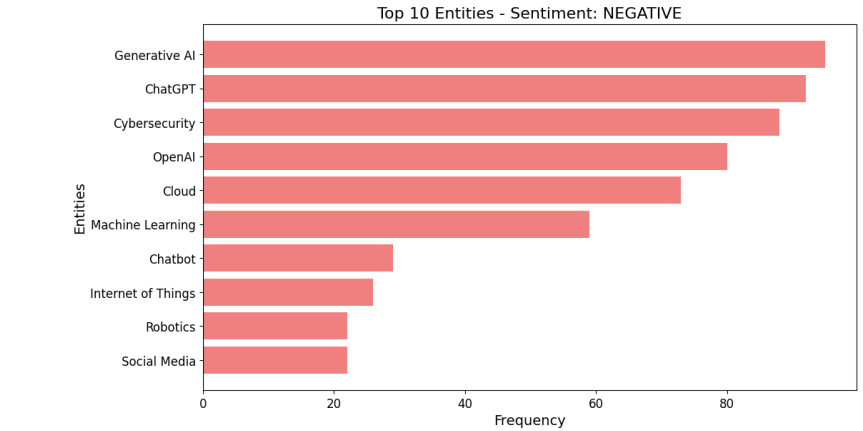
- **Quantum Computing** and **Generative AI** dominate positive sentiment peaks, showing their prominence in breakthroughs.

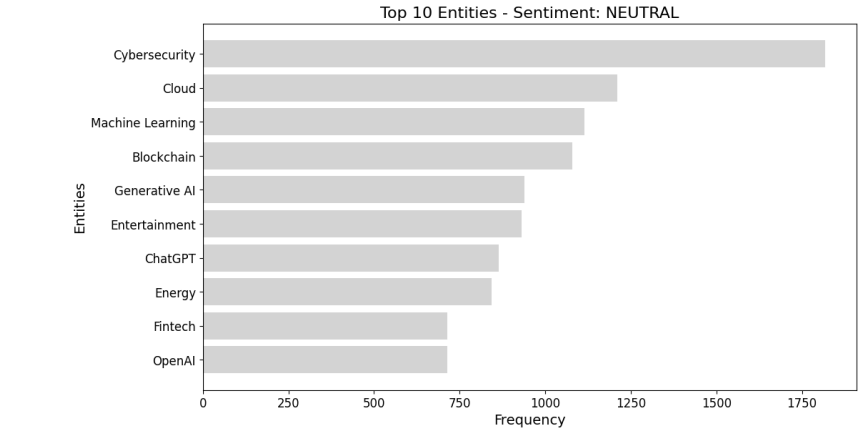
- **AI Ethics** and **SoundHound AI** face more criticism, likely tied to concerns about the societal impact and challenges in adoption.

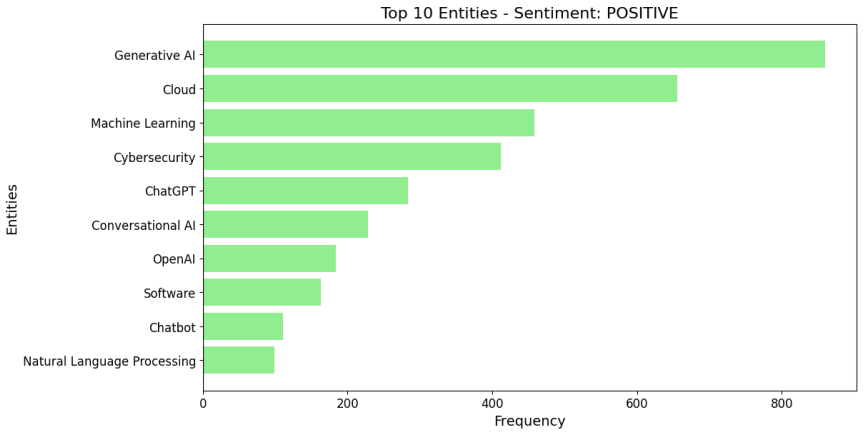
- Sentiment trends align with **technological hype cycles**, reflecting how innovation or controversy drives public and media perception over time.

Sentiment by TECH

Identify technologies and AI solutions that might be affecting the employment landscape







Create a dataframe that every row has TECH under entities AND words_to_search under sentence.

words_to_search = **['employ', 'employed', 'employs', 'employing', 'employment', 'employer', 'employee', 'hire', 'hiring', 'hired', 'hires']**

sentiment_label_DistilBERT	count
NEUTRAL	5811
POSITIVE	3079
NEGATIVE	472

Overall, far more positive than negative
Under employment for TECH entities

The results of entity counts under different sentiment categories (negative, neutral, positive) can provide insights into how technologies are perceived in relation to the employment landscape:

Negative Sentiment:

- **Generative AI**, **ChatGPT**, **Cybersecurity**, and **OpenAI** are frequently mentioned with negative sentiment.
- This suggests these technologies are perceived as threats to employment (e.g., automation replacing jobs, AI displacing human workers, or cybersecurity concerns reducing job stability).

Neutral Sentiment:

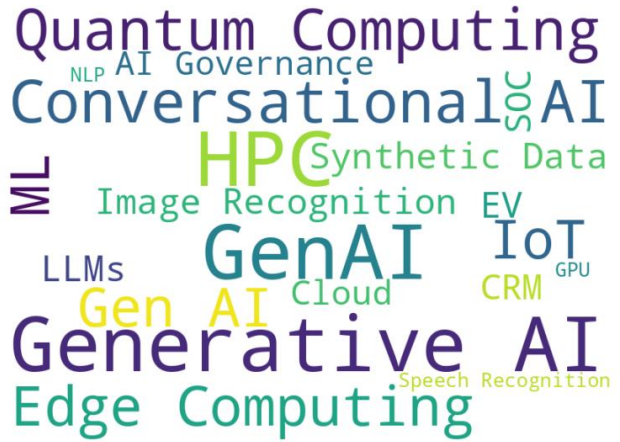
- **Cybersecurity**, **Cloud**, and **Machine Learning** dominate neutral sentiment.
- This indicates these technologies are considered as essential or stable aspects of the employment landscape, with neither strongly positive nor negative impacts.

Positive Sentiment:

- **Generative AI**, **Cloud**, and **Machine Learning** lead in positive sentiment.
- These technologies are seen as enablers of innovation, creating opportunities for specialized jobs (e.g., AI developers, cloud engineers, and data scientists).

EXECUTIVE SUMMARY

- All the filtered articles are related to AI and discuss it extensively.
- Although the articles focus on AI, not every sentence within them is AI-related. Some unrelated sentences carry negative sentiment.
- Sentence-level sentiment analysis is conducted, with results aggregated into entities and topics.
- Duplicate sentences or paragraphs across articles may affect the accuracy of the analytics.



Top technologies receiving positive sentiments

- Strong negative sentiment on AI in Art and Entertainment topics are related to job losses.
- Companies in sectors like cloud computing (e.g., Google Cloud), enterprise software (e.g., SAP), hardware (e.g., Nvidia), generative/conversational AI, healthcare, and consumer technology are leading investments in AI technologies, particularly in advancing transformative capabilities.
- Industries like renewable energy, healthcare, pharmaceuticals, semiconductors, e-commerce, cloud computing, and cybersecurity are likely to be most impacted by AI over the next several years.
- AI Ethics consistently struggles with low sentiment scores compared to others. This suggests ongoing debates or criticism regarding fairness, bias, or ethical implementation in AI systems. Cannot be by transformed by AI.

- Jobs such as Financial Analyst, AI Developer, and Product Manager have experienced successful AI integration, while roles like Librarian and Copywriter have faced challenges with AI integration.
- Generative AI, ChatGPT, Cybersecurity, Machine Learning, and Cloud are identified as technologies significantly affecting the employment landscape. Their impacts vary:
  Positive: Creating new roles in AI, cloud computing, and cybersecurity.
  Negative: Raising concerns about job displacement due to automation and ethical challenges.
  Neutral: reshaping jobs while maintaining overall stability in the employment landscape.
- To accelerate transformative capabilities like Generative AI, HPC, and IoT, companies can increase R&D investments, academic institutions can focus on AI education and innovation, and governments can fund initiatives, establish policies, and foster industry-academic collaborations.