News Article Analysis - AI Impacts on Jobs & Industries

University of Chicago

ADSP 32018 Natural Language Processing and Cognitive Computing

Lexi Lin

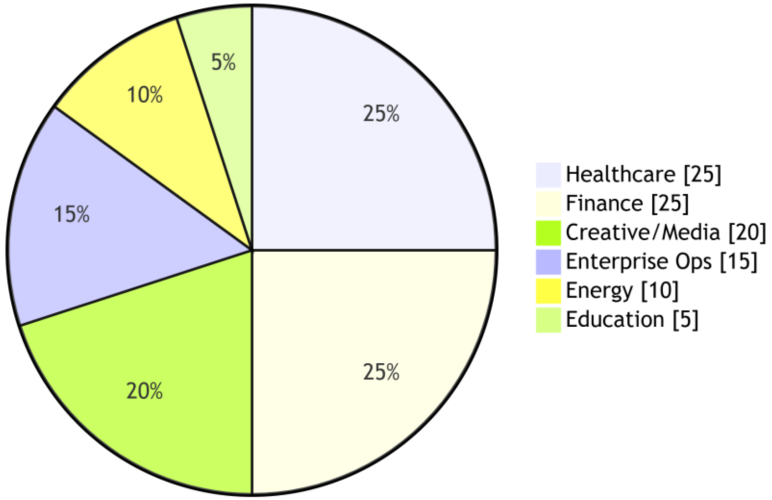
Executive Summary

Based on Analysis of 150K News Articles, 421 Topics & 400K+ Entities…

Key Insight: **82% of entities and 93% of topics show net-positive sentiment**, but media underrepresents privacy/job risks—creating an "optimism bubble.“ Media's AI optimism bias creates a self-reinforcing cycle: positive coverage drives audience engagement, amplifying the 'bubble.’ This makes rare critical voices vital—they pierce the hype to surface real risks like privacy erosion and job disruption that mainstream narratives overlook.

Strategic Outlook: AI’s biggest impacts will be in healthcare, finance, and creative sectors—where **automation augments human capability**. Success requires balancing efficiency gains with proactive ethics and reskilling. Monitor entity-level sentiment (e.g., privacy/tech terms) to anticipate backlash and align adoption with public optimism cycles.

Industries Most Impacted

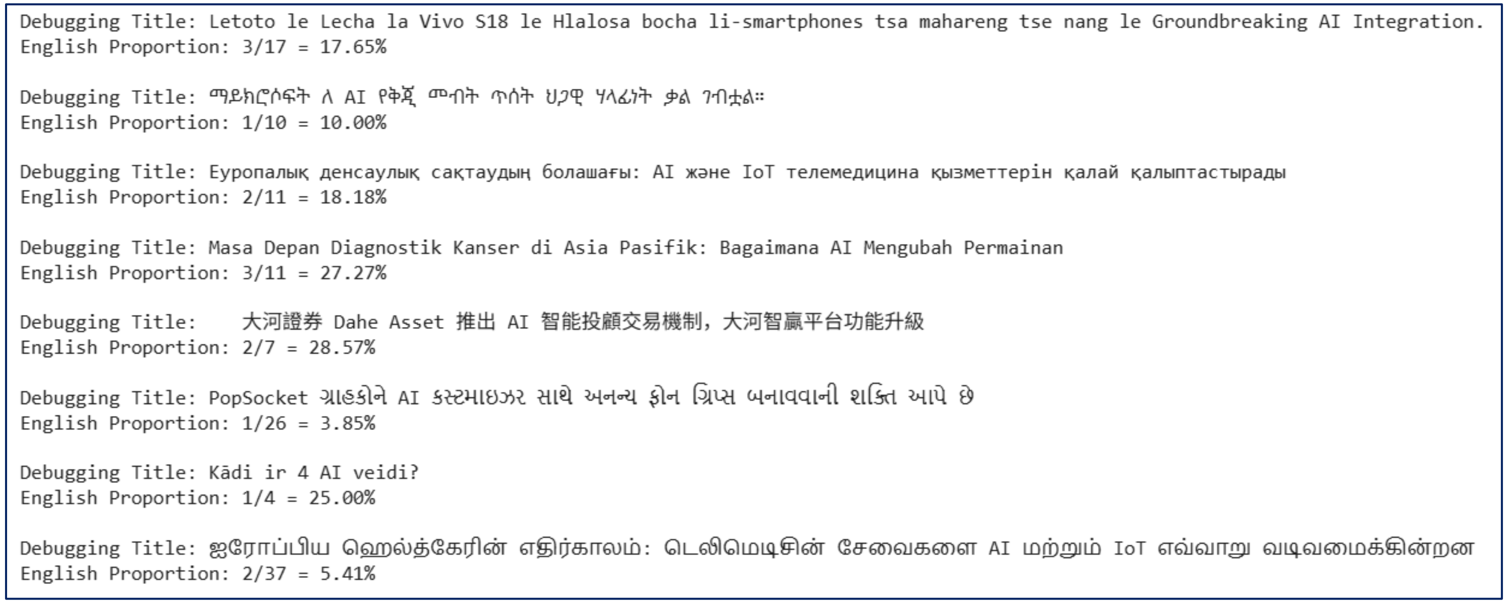


Actionable Recommendations

1. Automate Strategically, Not Broadly
   - ➢ Target: Routine cognitive tasks, such as deploy LLMs for document processing, or computer vision for quality control
   - ➢ Outcome: Free 20%-30% of employee time for high-value work
2. Boost Productivity via Augmentation
   - ➢ Healthcare: AI-diagnostic tools (e.g., radiology assistants)
   - ➢ Creative: GenAI for rapid prototyping
   - ➢ Energy: AI sensors for real-time infrastructure monitoring
3. Upskill Vulnerable Roles
   - ➢ Media: Train journalists in AI fact-checking tools
   - ➢ Finance: Upskill analysts for AI-augmented forecasting
4. Address Hidden Risks
   - ➢ Implement privacy-preserving AI and establish ethics board for AI use

Data Cleaning - High Level Filtering

- Removed rows where **Text** length is less than 100 words
- **Text** with duplicate content are identified, and only the first occurrence is retained
- Removed data where less than 30% of the **Title** words are English

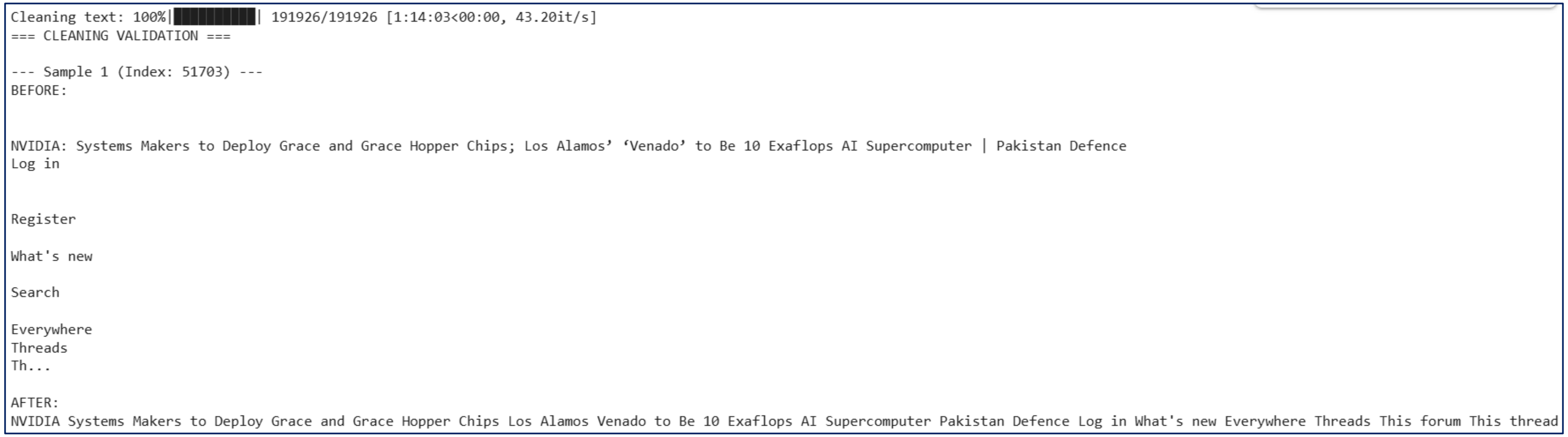


```
Debugging Title: Letoto le Lecha la Vivo S18 le Hlalosa bocha li-smartphones tsa mahareng tse nang le Groundbreaking AI Integration.
English Proportion: 3/17 = 17.65%

Debugging Title: ማይክሮሶፍት ለ AI የቅጂ መብት ጥሰት ህጋዊ ሃላፊነት ቃል ገብቷል።
English Proportion: 1/10 = 10.00%

Debugging Title: Еуропалық денсаулық сақтаудың болашағы: AI және IoT телемедицина қызметтерін қалай қалыптастырады
English Proportion: 2/11 = 18.18%

Debugging Title: Masa Depan Diagnostik Kanser di Asia Pasifik: Bagaimana AI Mengubah Permainan
English Proportion: 3/11 = 27.27%

Debugging Title:    大河證券 Dahe Asset 推出 AI 智能投顧交易機制，大河智赢平台功能升級
English Proportion: 2/7 = 28.57%

Debugging Title: PopSocket ગ્રાહકોને AI કસ્ટમાઇઝર સાથે અનન્ય ફોન ગ્રિપ્સ બનાવવાની શક્તિ આપે છે
English Proportion: 1/26 = 3.85%

Debugging Title: Kādi ir 4 AI veidi?
English Proportion: 1/4 = 25.00%

Debugging Title: ஐரோப்பிய ஹெல்த்கேரின் எதிர்காலம்: டெலிமெடிசின் சேவைகளை AI மற்றும் IoT எவ்வாறு வடிவமைக்கின்றன
English Proportion: 2/37 = 5.41%
```

The above image shows examples of rows that are removed as the titles contain English proportion less than 30%.

Data Cleaning – Detailed Text Cleaning

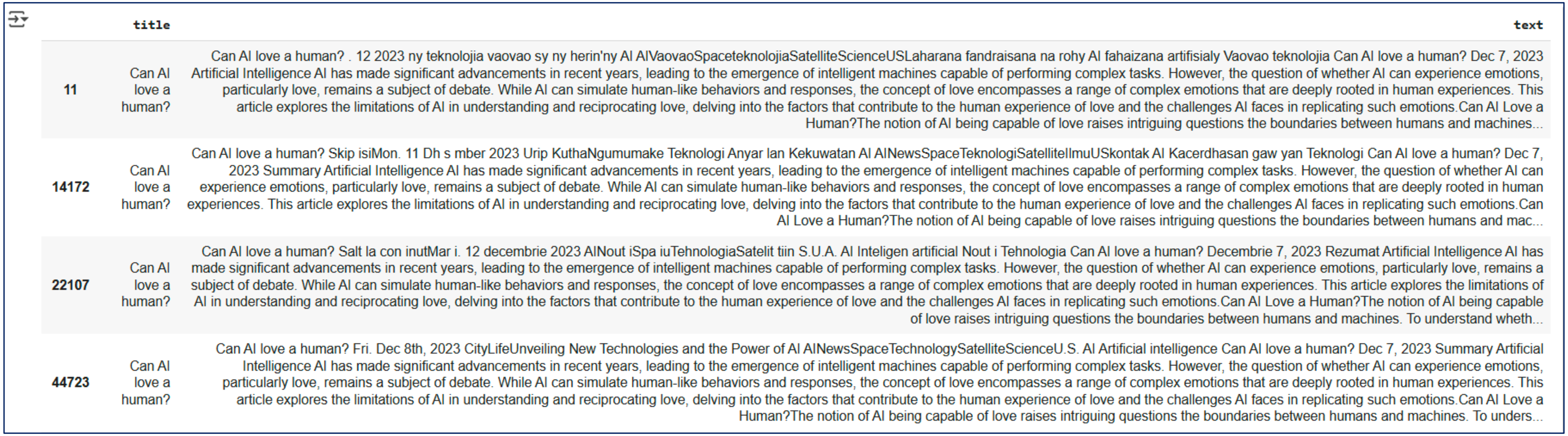
- Removed HTML tags using Beautiful Soup
- Remove common web navigation and UI elements
- Remove specific boilerplate content identified in the sample using regex
- Remove non-English text patterns (while being careful not to remove English content)
- Remove timestamps and dates that are not part of article content
- Remove promotional and marketing content
- Remove repeated category tags and navigation elements
- Clean up weather, location, and metadata
- Remove special characters but preserve basic punctuation
- Clean up spacing and normalize whitespace
- Remove very short fragments (likely remnants)
- Remove standalone category words and navigation remnants
- Removed extra newlines and excessive whitespaces
- Removed special characters (keeping alphanumeric, basic punctuation, and spaces)



```
Cleaning text: 100%|██████████| 191926/191926 [1:14:03<00:00, 43.20it/s]
=== CLEANING VALIDATION ===

--- Sample 1 (Index: 51703) ---
BEFORE:


NVIDIA: Systems Makers to Deploy Grace and Grace Hopper Chips; Los Alamos’ ‘Venado’ to Be 10 Exaflops AI Supercomputer | Pakistan Defence
Log in


Register

What's new

Search

Everywhere
Threads
Th...

AFTER:
NVIDIA Systems Makers to Deploy Grace and Grace Hopper Chips Los Alamos Venado to Be 10 Exaflops AI Supercomputer Pakistan Defence Log in What's new Everywhere Threads This forum This thread
```

Data Cleaning – Removed Duplicate Text

- 47706 rows of Text have duplicate Titles
- Goal: Preserve only the best version of Text and discard the duplicates
  - Group duplicate Text by Title.
  - For Text within a group, compute the **detect language score** (a score between 0-1 where 1 is most likely English) and **content quality score** (higher score indicates better quality/more complete content).
  - Use Smart deduplication that selects the best version of each title. Prioritizes English content and higher quality text.



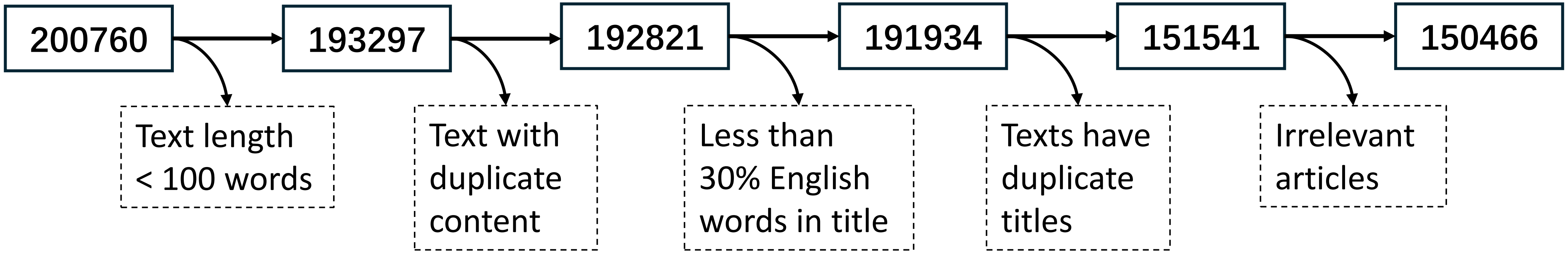
	title	text
11	Can AI love a human?	Can AI love a human? . 12 2023 ny teknolojia vaovao sy ny herin'ny AI AIVaovaoSpaceteknolojiaSatelliteScienceUSLaharana fandraisana na rohy AI fahaizana artifisialy Vaovao teknolojia Can AI love a human? Dec 7, 2023 Artificial Intelligence AI has made significant advancements in recent years, leading to the emergence of intelligent machines capable of performing complex tasks. However, the question of whether AI can experience emotions, particularly love, remains a subject of debate. While AI can simulate human-like behaviors and responses, the concept of love encompasses a range of complex emotions that are deeply rooted in human experiences. This article explores the limitations of AI in understanding and reciprocating love, delving into the factors that contribute to the human experience of love and the challenges AI faces in replicating such emotions.Can AI Love a Human?The notion of AI being capable of love raises intriguing questions the boundaries between humans and machines...
14172	Can AI love a human?	Can AI love a human? Skip isiMon. 11 Dh s mber 2023 Urip KuthaNgumumake Teknologi Anyar lan Kekuwatan AI AINewsSpaceTeknologiSatelliteIlmuUSkontak AI Kacerdhasan gaw yan Teknologi Can AI love a human? Dec 7, 2023 Summary Artificial Intelligence AI has made significant advancements in recent years, leading to the emergence of intelligent machines capable of performing complex tasks. However, the question of whether AI can experience emotions, particularly love, remains a subject of debate. While AI can simulate human-like behaviors and responses, the concept of love encompasses a range of complex emotions that are deeply rooted in human experiences. This article explores the limitations of AI in understanding and reciprocating love, delving into the factors that contribute to the human experience of love and the challenges AI faces in replicating such emotions.Can AI Love a Human?The notion of AI being capable of love raises intriguing questions the boundaries between humans and mac...
22107	Can AI love a human?	Can AI love a human? Salt la con inutMar i. 12 decembrie 2023 AINout iSpa iuTehnologiaSatelit tiin S.U.A. AI Inteligen artificial Nout i Tehnologia Can AI love a human? Decembrie 7, 2023 Rezumat Artificial Intelligence AI has made significant advancements in recent years, leading to the emergence of intelligent machines capable of performing complex tasks. However, the question of whether AI can experience emotions, particularly love, remains a subject of debate. While AI can simulate human-like behaviors and responses, the concept of love encompasses a range of complex emotions that are deeply rooted in human experiences. This article explores the limitations of AI in understanding and reciprocating love, delving into the factors that contribute to the human experience of love and the challenges AI faces in replicating such emotions.Can AI Love a Human?The notion of AI being capable of love raises intriguing questions the boundaries between humans and machines. To understand wheth...
44723	Can AI love a human?	Can AI love a human? Fri. Dec 8th, 2023 CityLifeUnveiling New Technologies and the Power of AI AINewsSpaceTechnologySatelliteScienceU.S. AI Artificial intelligence Can AI love a human? Dec 7, 2023 Summary Artificial Intelligence AI has made significant advancements in recent years, leading to the emergence of intelligent machines capable of performing complex tasks. However, the question of whether AI can experience emotions, particularly love, remains a subject of debate. While AI can simulate human-like behaviors and responses, the concept of love encompasses a range of complex emotions that are deeply rooted in human experiences. This article explores the limitations of AI in understanding and reciprocating love, delving into the factors that contribute to the human experience of love and the challenges AI faces in replicating such emotions.Can AI Love a Human?The notion of AI being capable of love raises intriguing questions the boundaries between humans and machines. To unders...

This image shows a few examples of Text with the same Title “Can AI love a human?”

Data Cleaning – Article Relevancy Validation

- Checks if an article is relevant based on keywords in title and text.
- keywords = ['AI', 'artificial intelligence', 'technology', 'chatbot', 'data science', 'data', 'machine learning', 'neural network', 'generative’]
- Additional checks for common irrelevant topics that might contain keywords incidentally
- If an irrelevant pattern is found, apply stricter filtering - check if AI/tech keywords appear frequently enough in the context of irrelevant topics
- Removed 1075 irrelevant rows

Removing process for 200760 rows to 150466 rows:



200760 → 193297 → 192821 → 191934 → 151541 → 150466

Text length < 100 words

Text with duplicate content

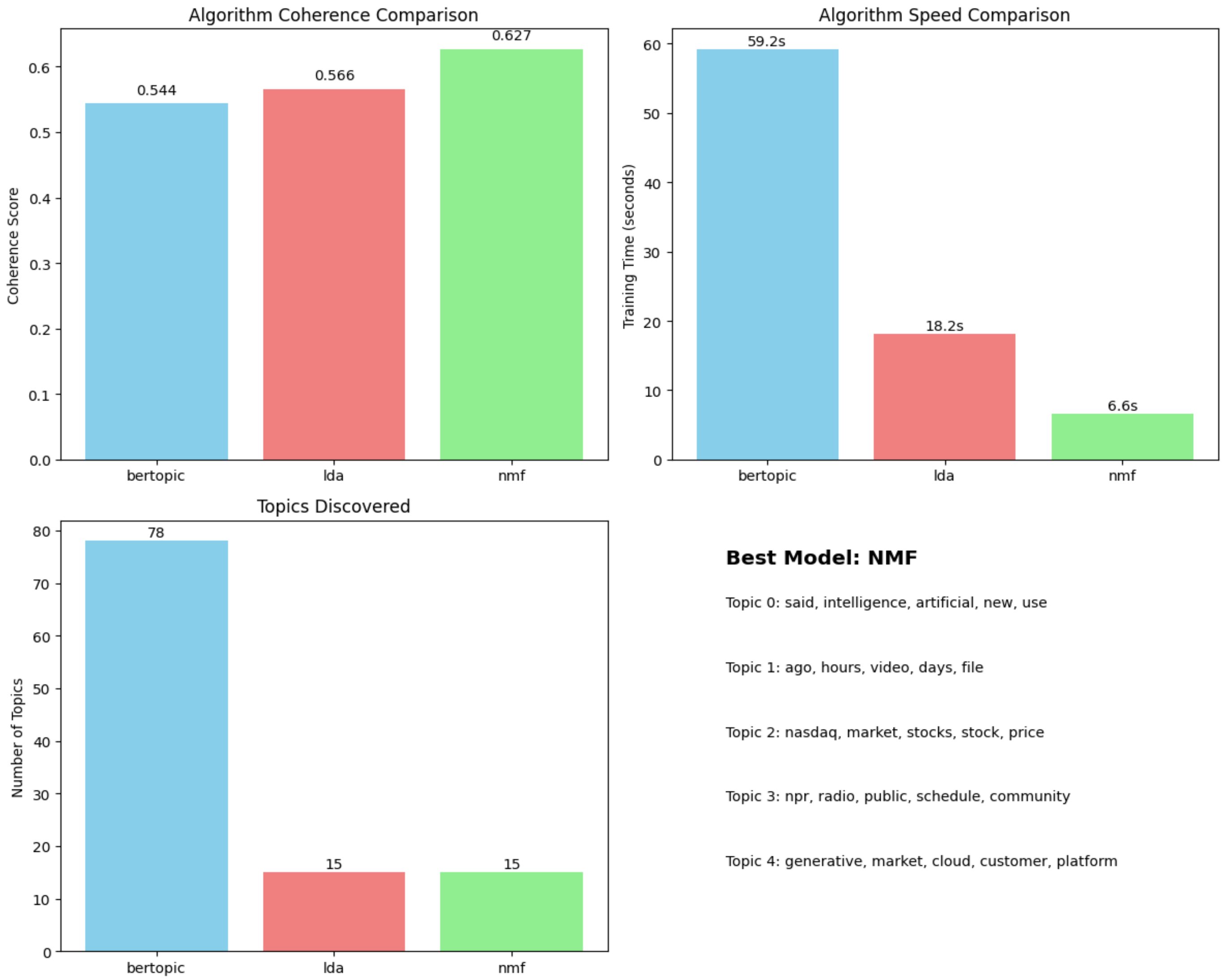
Less than 30% English words in title

Texts have duplicate titles

Irrelevant articles

Topic Modeling

- Three topic modeling algorithms (**LDA, NMF, and BERTopic**) are compared on 5000 randomly selected samples
- Number of topics for LDA and NMF are set to 15 for simplicity
- Combinations of different hyperparameters are evaluated and only the best hyperparameter is retained and compared with other models
- Coherence scores are calculated for each model
- The best model identified is Non-Negative Matrix Factorization (NMF)



Best Model: NMF

Topic 0: said, intelligence, artificial, new, use

Topic 1: ago, hours, video, days, file

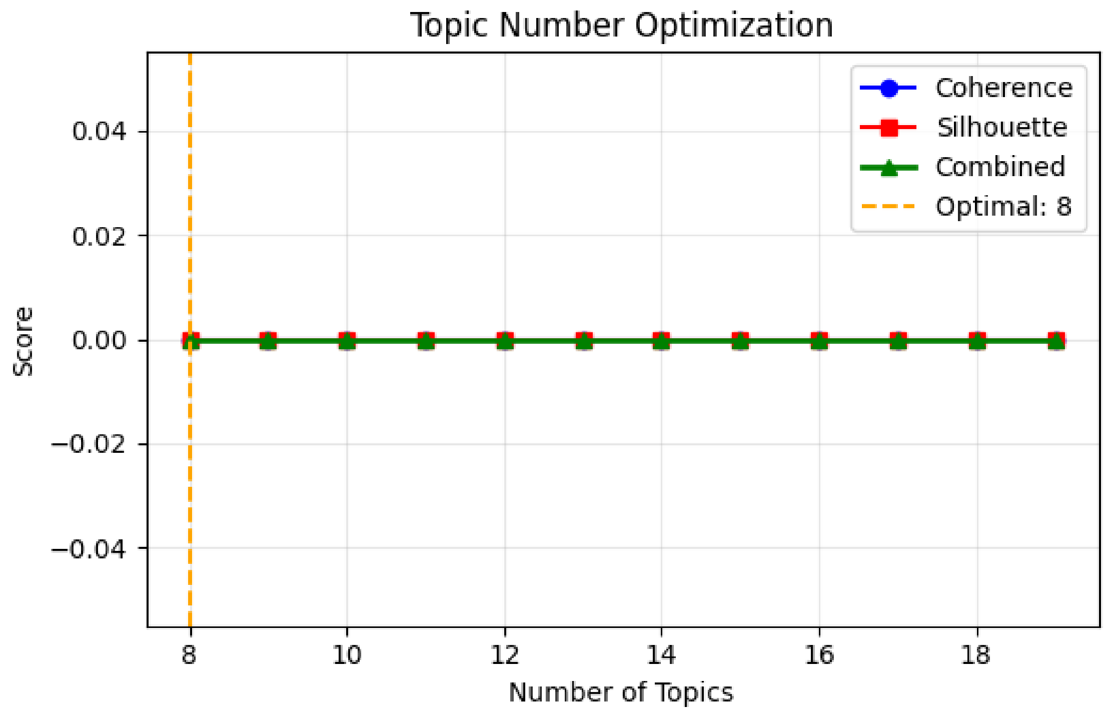
Topic 2: nasdaq, market, stocks, stock, price

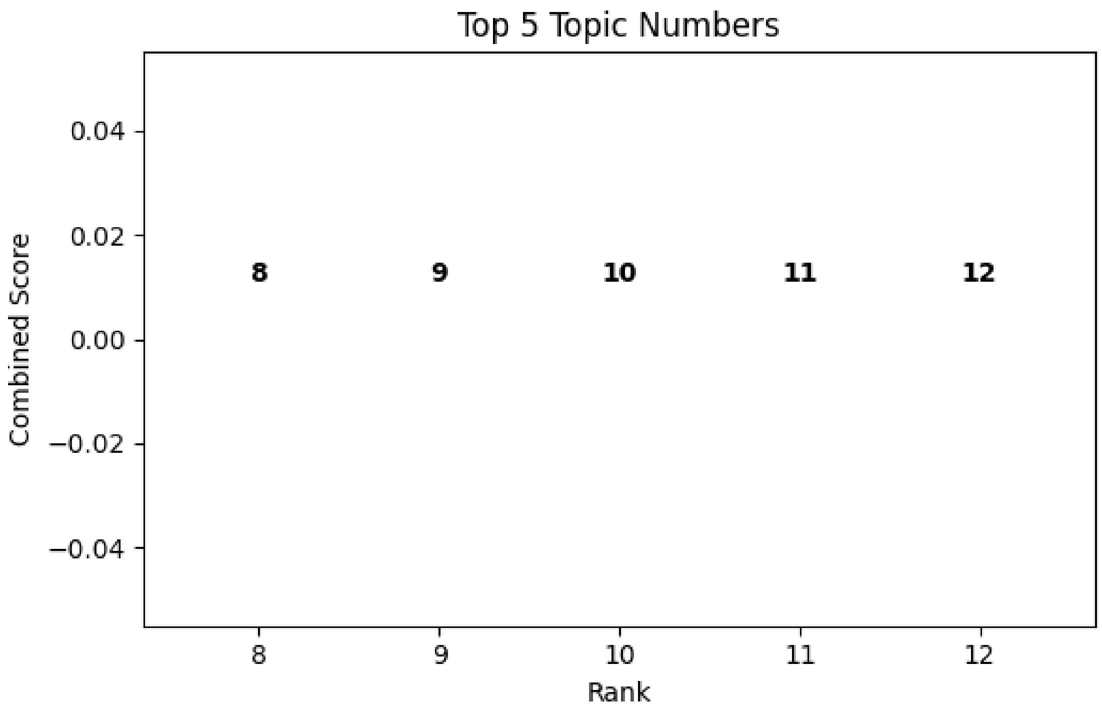
Topic 3: npr, radio, public, schedule, community

Topic 4: generative, market, cloud, customer, platform

Topic Modeling – Initial Experimentation

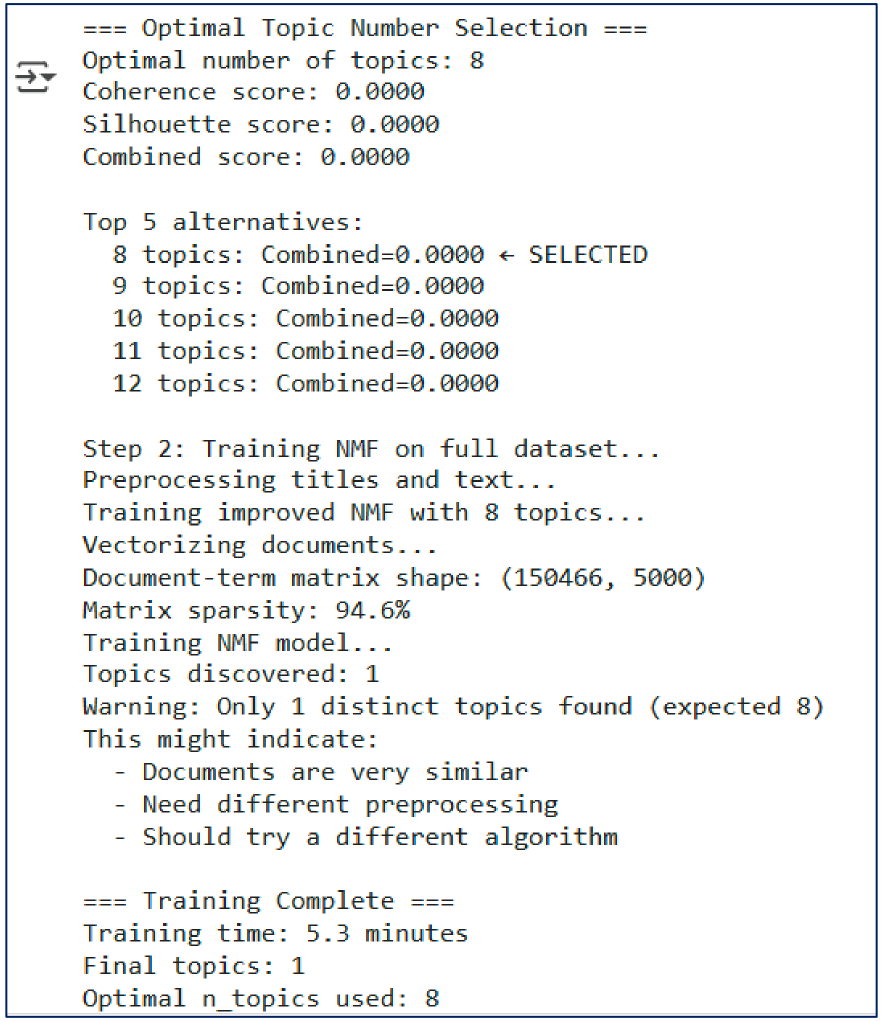
- To find the optimal number of topics for NMF, I tested various topic numbers on 1000 documents
- Tested topic numbers: 8 9 10 11 12 13 14 15 16 17 18 19
- Surprisingly, the various number of topics resulted in scores close to 0 for all metrics (image on left), but the system still determined that the best topic number is 8, as shown in the right image





Topic Modeling – NMF Modeling Failure

- A close inspection into the optimal number of topic – 8 - revealed that the coherence, silhouette, and combined scores of the NMF model on 1000 sample data are all zero
- Note that the other alternative topic numbers (e.g., 9, 10, 11, 12) also produced zero metric scores
- Implementing NMF on the full dataset resulted in utter failure – only 1 topic was found, indicating a complete inability to differentiate underlying themes
- Possible Reason: High document similarity. The documents in the dataset appear highly similar, causing the model to collapse all content into a single topic rather than identifying diverse patterns
- **Due to the inadequate performance by NMF and uncertainty of manual selection of topic numbers, I decided to perform topic modeling on BERTopic as it can capture semantic relationships, effective at handling noisy data, and the topic numbers are selected automatically**



```
=== Optimal Topic Number Selection ===
Optimal number of topics: 8
Coherence score: 0.0000
Silhouette score: 0.0000
Combined score: 0.0000

Top 5 alternatives:
  8 topics: Combined=0.0000 ← SELECTED
  9 topics: Combined=0.0000
  10 topics: Combined=0.0000
  11 topics: Combined=0.0000
  12 topics: Combined=0.0000

Step 2: Training NMF on full dataset...
Preprocessing titles and text...
Training improved NMF with 8 topics...
Vectorizing documents...
Document-term matrix shape: (150466, 5000)
Matrix sparsity: 94.6%
Training NMF model...
Topics discovered: 1
Warning: Only 1 distinct topics found (expected 8)
This might indicate:
  - Documents are very similar
  - Need different preprocessing
  - Should try a different algorithm

=== Training Complete ===
Training time: 5.3 minutes
Final topics: 1
Optimal n_topics used: 8
```

Topic Modeling – BERTopic Training

BERTopic model is creating using

- Sentence Transformer *all-MiniLM-L6-v2* for embedding model
- UMAP for dimensionality reduction
- HDBSCAN for clustering
- Count Vectorizer for topic representation

Dataset is split into

Training set: 108334 samples

Validation set: 12038 samples

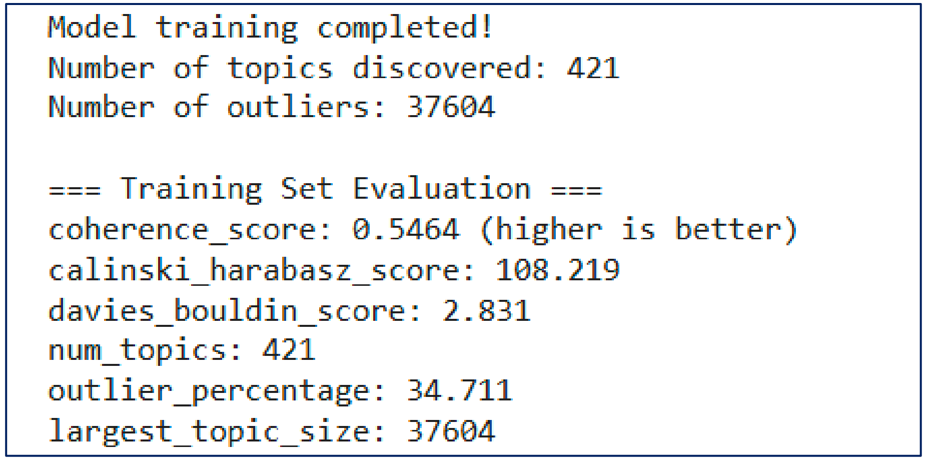
Test set: 20094 samples

The evaluation results (right) shows strong performance in

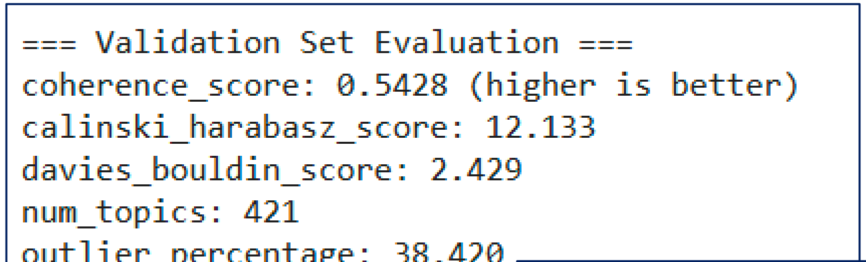
✓ Coherence Score (Test performs comparable with Train)

✓ Davies Bouldin Score (Test performs better than Train)

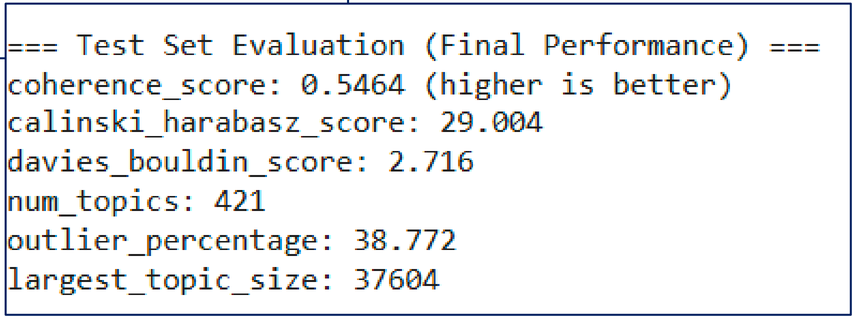
✓ Calinski Harabasz Score (Test performs better than Validation)



```
Model training completed!
Number of topics discovered: 421
Number of outliers: 37604

=== Training Set Evaluation ===
coherence_score: 0.5464 (higher is better)
calinski_harabasz_score: 108.219
davies_bouldin_score: 2.831
num_topics: 421
outlier_percentage: 34.711
largest_topic_size: 37604
```



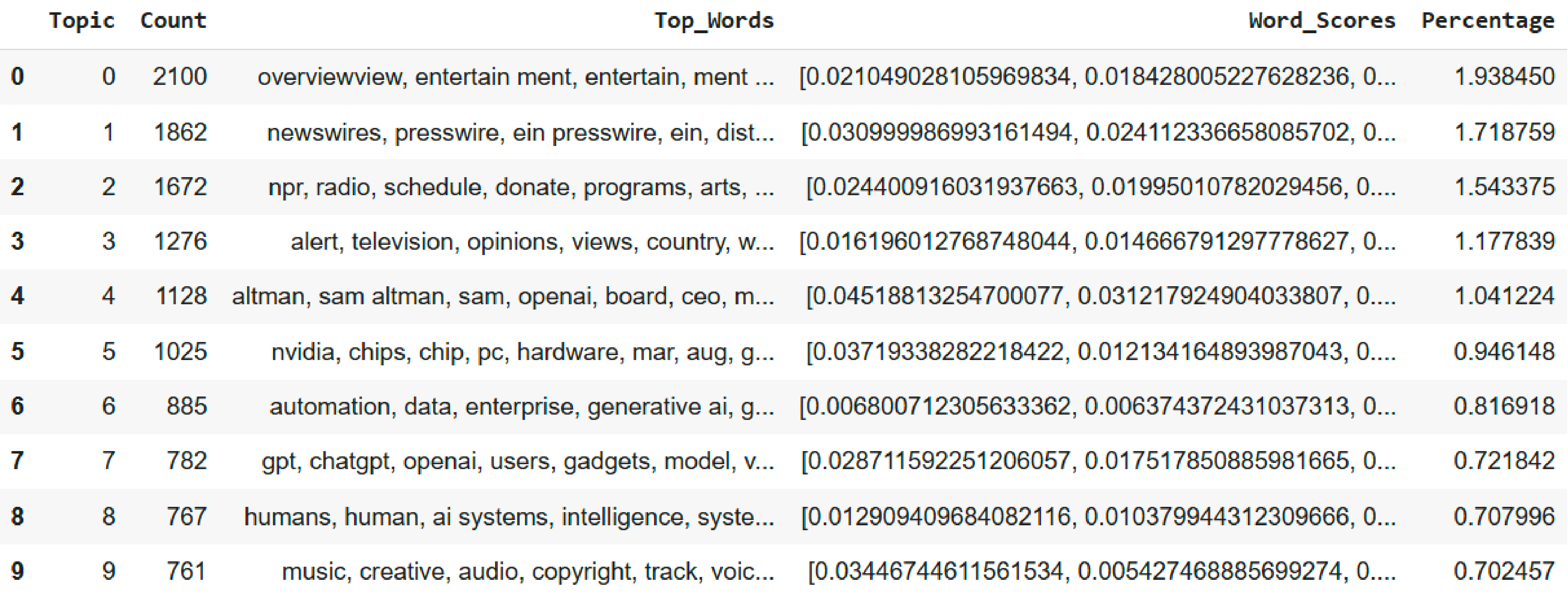
```
=== Validation Set Evaluation ===
coherence_score: 0.5428 (higher is better)
calinski_harabasz_score: 12.133
davies_bouldin_score: 2.429
num_topics: 421
outlier_percentage: 38.420
largest_topic_size: 37604
```



```
=== Test Set Evaluation (Final Performance) ===
coherence_score: 0.5464 (higher is better)
calinski_harabasz_score: 29.004
davies_bouldin_score: 2.716
num_topics: 421
outlier_percentage: 38.772
largest_topic_size: 37604
```

Topic Modeling – BERTopic Results

Topic information of the first 10 topics



	Topic	Count	Top_Words	Word_Scores	Percentage
0	0	2100	overviewview, entertain ment, entertain, ment ...	[0.021049028105969834, 0.018428005227628236, 0...	1.938450
1	1	1862	newswires, presswire, ein presswire, ein, dist...	[0.030999986993161494, 0.024112336658085702, 0...	1.718759
2	2	1672	npr, radio, schedule, donate, programs, arts, ...	[0.024400916031937663, 0.01995010782029456, 0....	1.543375
3	3	1276	alert, television, opinions, views, country, w...	[0.016196012768748044, 0.014666791297778627, 0...	1.177839
4	4	1128	altman, sam altman, sam, openai, board, ceo, m...	[0.04518813254700077, 0.031217924904033807, 0....	1.041224
5	5	1025	nvidia, chips, chip, pc, hardware, mar, aug, g...	[0.03719338282218422, 0.012134164893987043, 0....	0.946148
6	6	885	automation, data, enterprise, generative ai, g...	[0.006800712305633362, 0.006374372431037313, 0...	0.816918
7	7	782	gpt, chatgpt, openai, users, gadgets, model, v...	[0.028711592251206057, 0.017517850885981665, 0...	0.721842
8	8	767	humans, human, ai systems, intelligence, syste...	[0.012909409684082116, 0.010379944312309666, 0...	0.707996
9	9	761	music, creative, audio, copyright, track, voic...	[0.03446744611561534, 0.005427468885699274, 0....	0.702457

Topic Modeling – BERTopic Results

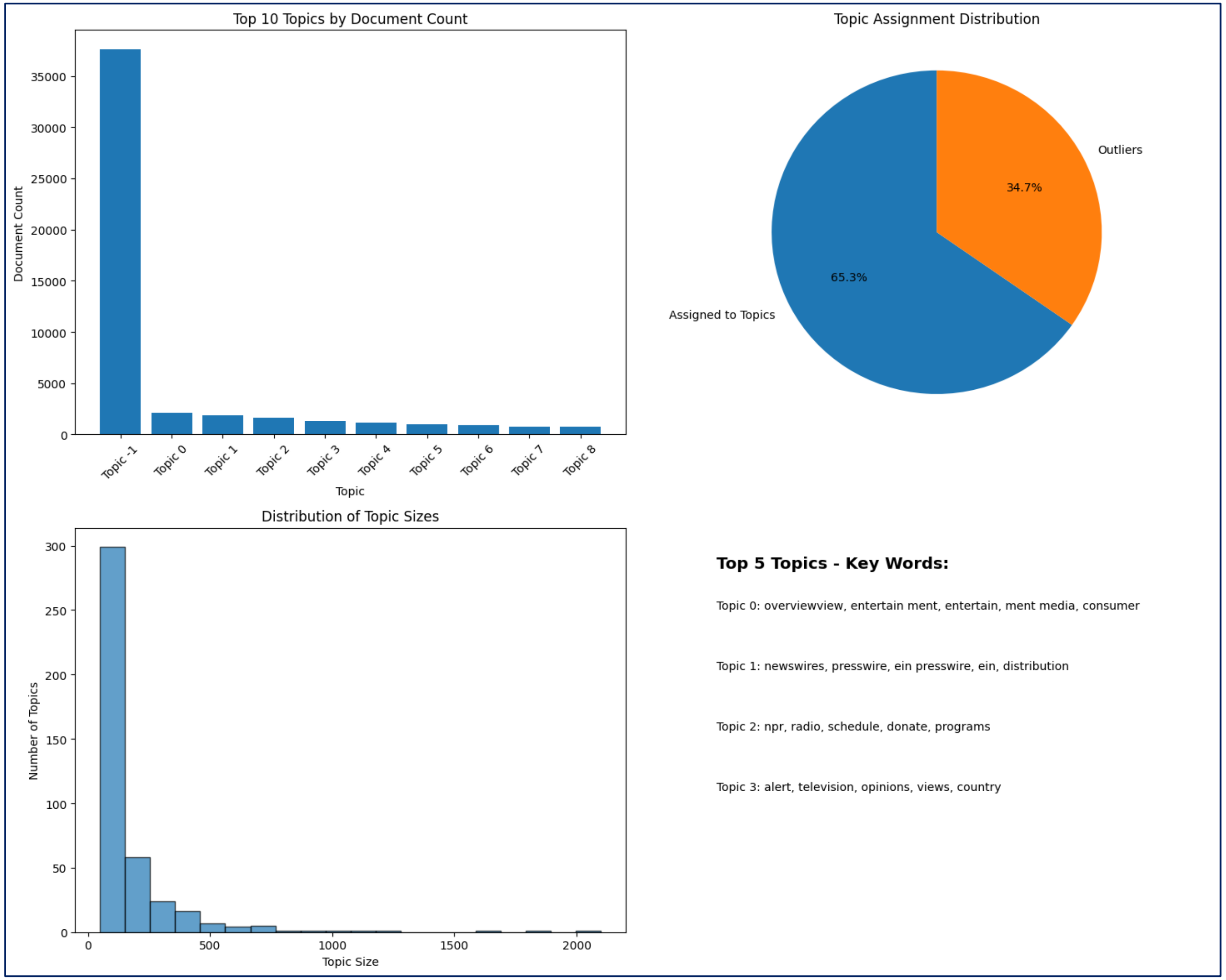
A total of 421 topics are extracted

Topic -1 (not assigned to any topics) dominates with over 35,000 documents, far exceeding other topics, indicating a significant portion of outliers

Most topics contain less than 100 documents, thus leading to the extraction of large number of topics

As the topic number increases, the document count decreases, suggesting that the **topics appearing first are more informative and important than the topics towards the end**

For this reason, I decided to categorize and summarize the major themes in the first 100 topics



Top 5 Topics - Key Words:

Topic 0: overviewview, entertain ment, entertain, ment media, consumer

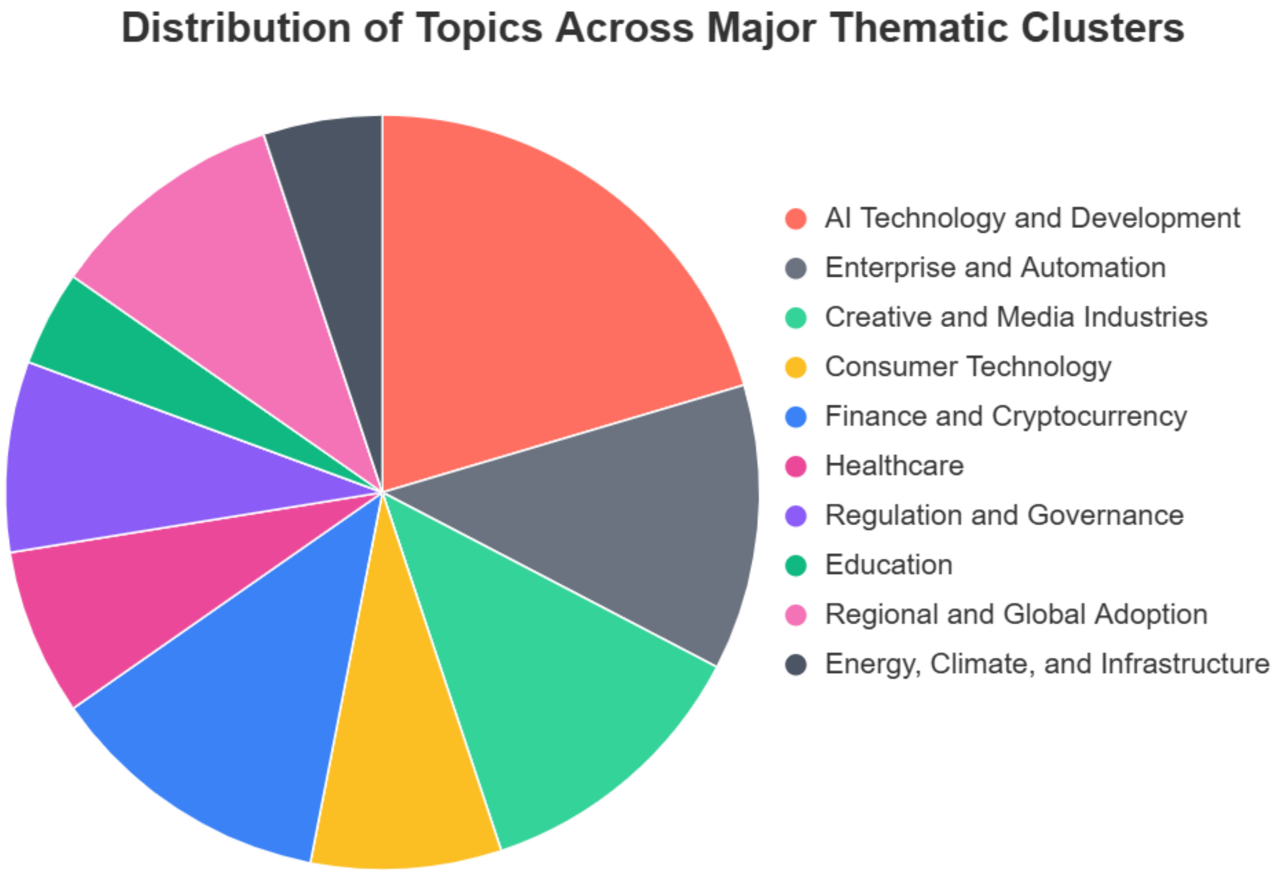
Topic 1: newswires, presswire, ein presswire, ein, distribution

Topic 2: npr, radio, schedule, donate, programs

Topic 3: alert, television, opinions, views, country

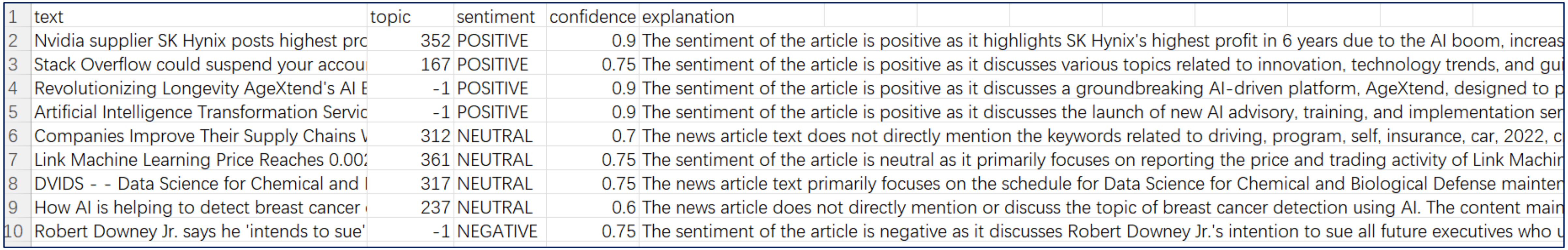
BERTopic – Major Themes in the First 100 Topics

1. AI Technology and Development
2. AI in Enterprise and Automation
3. AI in Creative and Media Industries
4. AI in Consumer Technology
5. AI in Finance and Cryptocurrency
6. AI in Healthcare
7. AI Regulation and Governance
8. AI in Education
9. Regional and Global AI Adoption
10. AI in Energy, Climate, and Infrastructure
11. Miscellaneous and Emerging Topics



Topic Sentiment Analysis – Labeling

- Assigned topics to all articles
- Using stratified sampling to randomly select 1000 samples from the full dataset
    - Ensured each topic gets minimum representation
    - Filled the remaining slots proportionally to the overall topic distribution
- Labeled sentiment on 1000 text articles using GPT-3.5 with the following prompt:
    - You are a sentiment analysis expert. Analyze the sentiment of the following news article text, focusing on how the content relates to this topic
    - Determine the sentiment of this article as it relates to the topic keywords above
        - Consider the overall tone, not just individual words
        - Classify as: POSITIVE, NEGATIVE, or NEUTRAL
        - Provide confidence score (0.0-1.0)



1	text	topic	sentiment	confidence	explanation
2	Nvidia supplier SK Hynix posts highest prc	352	POSITIVE	0.9	The sentiment of the article is positive as it highlights SK Hynix's highest profit in 6 years due to the AI boom, increas
3	Stack Overflow could suspend your accou	167	POSITIVE	0.75	The sentiment of the article is positive as it discusses various topics related to innovation, technology trends, and gui
4	Revolutionizing Longevity AgeXtend's AI E	-1	POSITIVE	0.9	The sentiment of the article is positive as it discusses a groundbreaking AI-driven platform, AgeXtend, designed to p
5	Artificial Intelligence Transformation Servic	-1	POSITIVE	0.9	The sentiment of the article is positive as it discusses the launch of new AI advisory, training, and implementation ser
6	Companies Improve Their Supply Chains V	312	NEUTRAL	0.7	The news article text does not directly mention the keywords related to driving, program, self, insurance, car, 2022, c
7	Link Machine Learning Price Reaches 0.00	361	NEUTRAL	0.75	The sentiment of the article is neutral as it primarily focuses on reporting the price and trading activity of Link Machin
8	DVIDS - - Data Science for Chemical and I	317	NEUTRAL	0.75	The news article text primarily focuses on the schedule for Data Science for Chemical and Biological Defense mainten
9	How AI is helping to detect breast cancer	237	NEUTRAL	0.6	The news article does not directly mention or discuss the topic of breast cancer detection using AI. The content main
10	Robert Downey Jr. says he 'intends to sue'	-1	NEGATIVE	0.75	The sentiment of the article is negative as it discusses Robert Downey Jr.'s intention to sue all future executives who u

This image shows the first rows of labeled data, including sentiment category, confidence score, and explanation generated by GPT-3.5

Topic Sentiment – Customize Model

Data Preparation for Model Training

- Labeled data (with sentiments as 'NEGATIVE', 'NEUTRAL', or 'POSITIVE') was mapped to numerical labels (0, 1, 2 respectively) and split into training and validation sets using an 80/20 ratio, stratified by labels to maintain class balance, via scikit-learn's train_test_split
- The resulting sets were converted to Hugging Face Datasets.

Model Selection and Fine-Tuning

- A pre-trained DistilBERT-base-uncased model was loaded using Hugging Face's AutoModelForSequenceClassification with 3 output labels
- The model was fine-tuned on the tokenized training dataset (using truncation, padding, and max length of 512) for 3 epochs with a batch size of 16, warmup steps of 500, weight decay of 0.01, and evaluation/saving per epoch via the Trainer API, selecting the best model based on validation loss

Lexicon-Based Sentiment Components

- Incorporated rule-based lexicons including VADER (compound score), AFINN (normalized score to [-1, 1]), and SentiWordNet (average positive-negative score per tagged word using NLTK for tokenization and POS tagging).

Ensemble Prediction Method

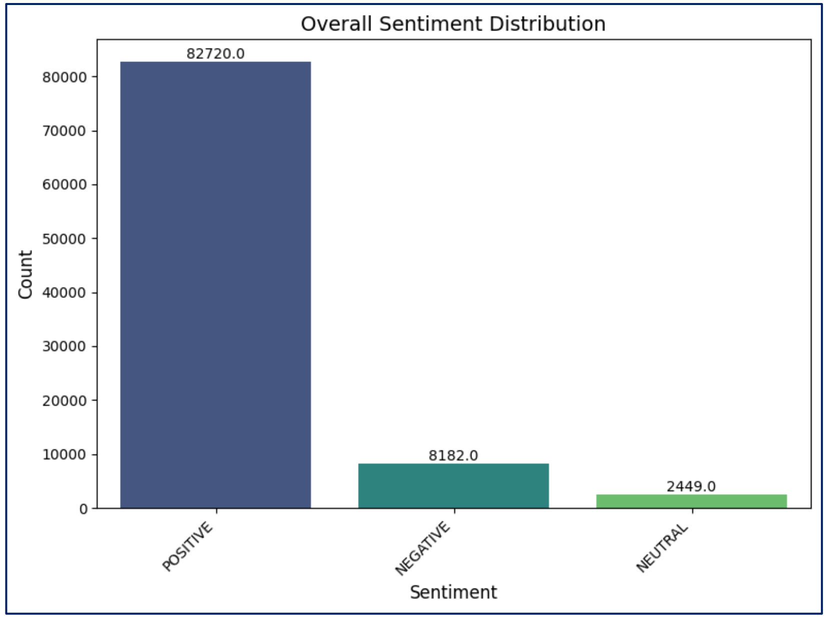
- For each text input (e.g., article), sentiment was predicted via an ensemble combining lexicon scores (VADER, AFINN, SentiWordNet) with the fine-tuned DistilBERT model's softmax probabilities (mapped to scores of -1, 0, or 1)
- A weighted average score was computed (weights: VADER 0.3, AFINN 0.2, SentiWordNet 0.2, DistilBERT 0.3), clamped to [-1, 1], and divided into classes (>0.1 for 'POSITIVE', <-0.1 for 'NEGATIVE', else 'NEUTRAL')
- Confidence was derived from the absolute score value, optionally blended with DistilBERT's max probability.

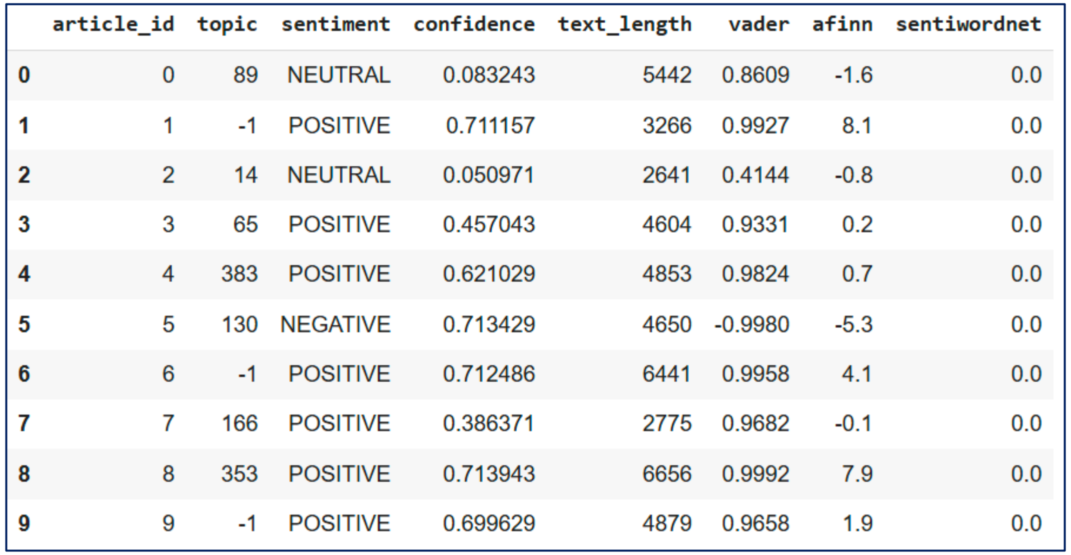
Sentiments on Articles Connected to Topics

Topic-Level Application

Sentiment analysis was applied to all articles in the cleaned dataset with pre-assigned topics (from BERTopic), processing each text individually via the ensemble method and recording per-article results including topic ID, sentiment label, confidence, and individual component scores for potential aggregation by topic. A sample of the results is shown on the lower right.

By aggregating the sentiments (image on lower left), we can see the positive rate provided by the model is astonishingly high across more than 150,000 articles. Most articles show positive attitude towards their own topics which cover from AI, data science, tech companies, hardware companies, gaming industry, etc., implying that during the past four years, the medias had consistent, strong confidence in the high-speed development of AI related areas. The new generation of AI tools including GPT, Deepseek, Gemini, and many other chatbots are playing pivotal roles in these topics.





	article_id	topic	sentiment	confidence	text_length	vader	afinn	sentiwordnet
0	0	89	NEUTRAL	0.083243	5442	0.8609	-1.6	0.0
1	1	-1	POSITIVE	0.711157	3266	0.9927	8.1	0.0
2	2	14	NEUTRAL	0.050971	2641	0.4144	-0.8	0.0
3	3	65	POSITIVE	0.457043	4604	0.9331	0.2	0.0
4	4	383	POSITIVE	0.621029	4853	0.9824	0.7	0.0
5	5	130	NEGATIVE	0.713429	4650	-0.9980	-5.3	0.0
6	6	-1	POSITIVE	0.712486	6441	0.9958	4.1	0.0
7	7	166	POSITIVE	0.386371	2775	0.9682	-0.1	0.0
8	8	353	POSITIVE	0.713943	6656	0.9992	7.9	0.0
9	9	-1	POSITIVE	0.699629	4879	0.9658	1.9	0.0

Topic Sentiment – Sentiments on Topics

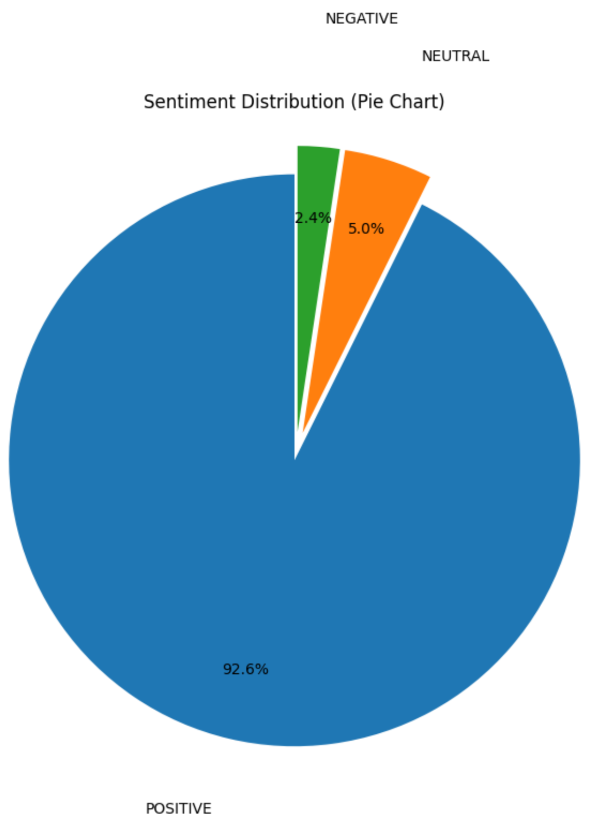
Topic-Level Sentiment Aggregation

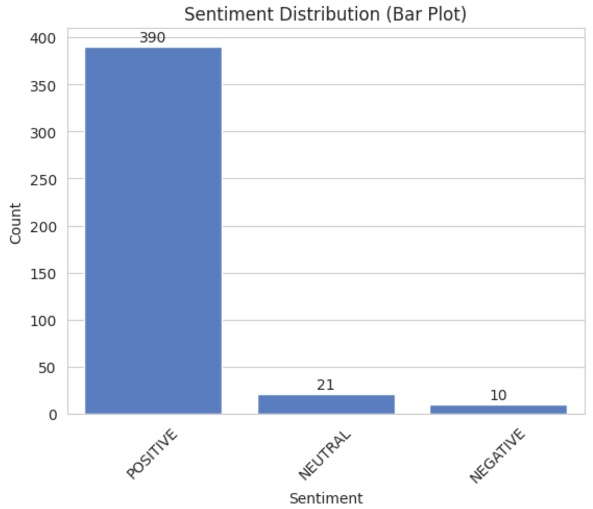
- To determine the overall sentiment for each topic extracted via BERTopic, article-level sentiments are aggregated based on the topic assignments derived from the model's probabilistic clustering.
- This aggregation is weighted by confidence scores, ensuring that more reliable predictions (higher confidence) exert greater influence on the final topic sentiment.

I selected the confidence-weighted aggregation approach because it

- Enhances accuracy and robustness by prioritizing high-confidence predictions, minimizing the impact of noisy or uncertain sentiment labels that could skew results.
- Provides a more nuanced representation of topic sentiment, as it accounts for varying prediction quality across articles, leading to insights that better reflect underlying data reliability.

The results of sentiments for topics are displayed on the right. As in the article-level, the sentiments expressed towards the topics are overwhelmingly positive, reflecting media’s optimistic and tech-leaning attitude towards AI. Recall that the first 100 topics are summarized into major themes in slide 13, including industries such as finance, healthcare, education, and creative media. The major themes highlight AI's permeation across diverse domains, from core technology to specialized applications, indicating AI is no longer siloed in tech but is becoming ubiquitous. **With the positive-skew of sentiment, AI is portrayed as a net positive force, fostering innovation and economic growth in every sector**.


Sentiment Distribution (Pie Chart)


Sentiment Distribution (Bar Plot)

Topic Sentiment – Insights & Findings

Overall, the topic-level sentiment results paint AI as a maturing technology on the cusp of mass scalability, but the sentiment imbalance warns of potential over-optimism, echoing historical tech bubbles.

Based on the major themes, we can infer that **AI's impacts are concentrated in sectors where news coverage clusters**, as these reflect real-world applications and discussions. The positive sentiment implies that impacts are often viewed as enhancements (e.g., efficiency gains, new capabilities) rather than disruptions, though, this could mask job losses or skill shifts.

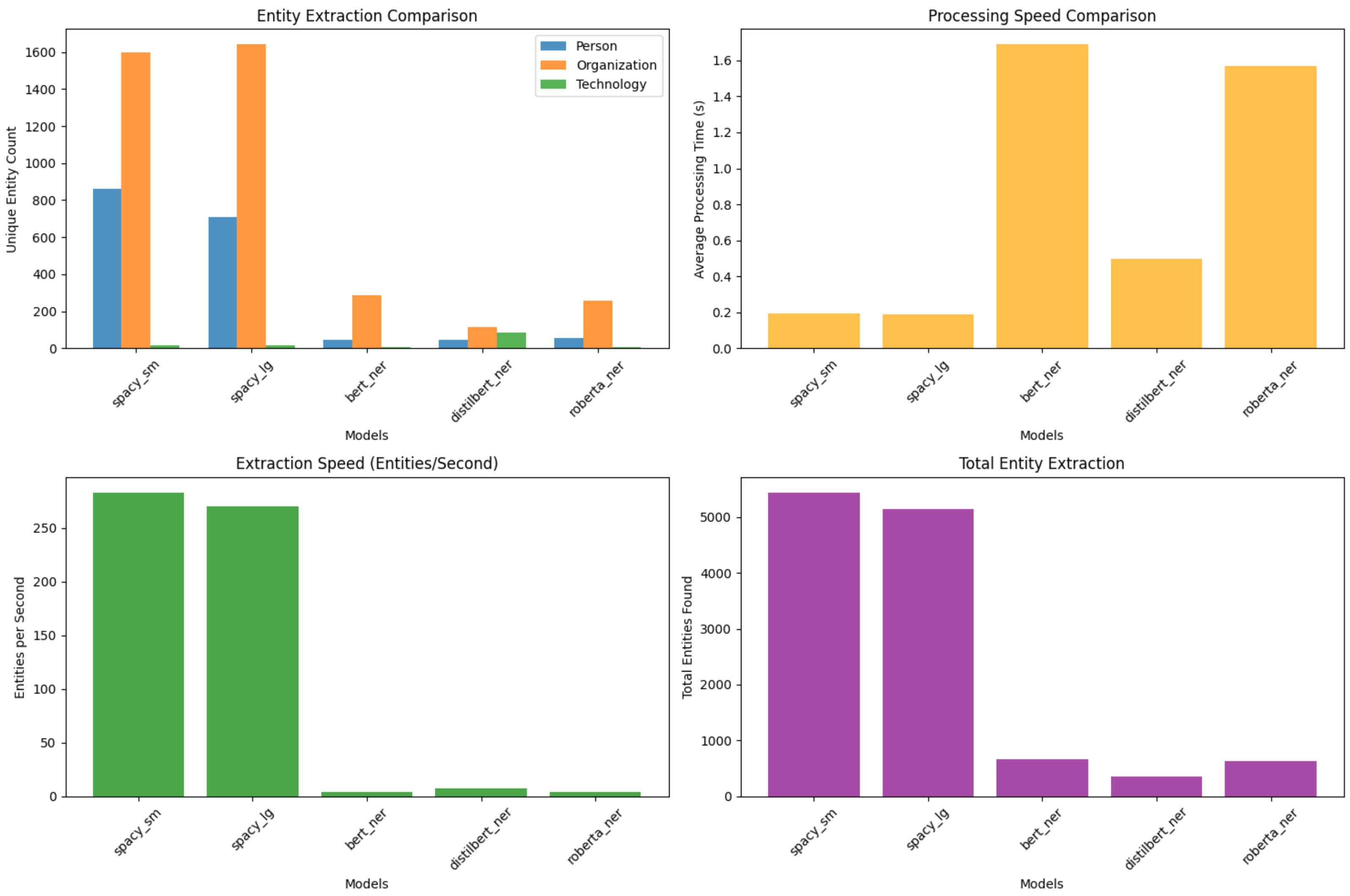
Industries with dedicated themes are likely most affected due to high topic density and sentiment confidence. To structure this inference, I will use a table (on the right) summarizing key industries derived from the major themes, AI's impacts on these industries, potentially affected jobs, and emerging opportunities.

INDUSTRY	IMPACT DRIVERS	JOBS AT GREATEST TRANSFORMATION RISK	EMERGING OPPORTUNITIES
Healthcare	Diagnostics, drug discovery, patient monitoring	Medical imaging analysts, routine diagnostics	AI-medicine specialists, data curators
Finance	Fraud detection, robo-advisors, algorithmic trading	Traditional analysts, risk assessors	Quantitative modelers, compliance strategists
Creative Industries	Content generation (art, music, writing), design automation	Graphic designers, copywriters, entry-level media production	AI-content editors, hybrid creative technologists
Enterprise/Operations	Process automation, predictive maintenance, supply chain optimization	Data entry clerks, inventory managers, routine analysts	AI workflow designers, automation overseers
Education	Personalized learning platforms, administrative automation	Graders, basic tutors, curriculum developers	EdTech designers, adaptive learning specialists
Energy/Infrastructure	Smart grid management, predictive maintenance, climate modeling	Manual inspectors, traditional grid operators	Sustainability AI optimizers

Entity Extraction – Model Selection

To select the best named-entity extraction model on the news article datset, comparison among spacy_sm, spacy_lg, bert_ner, distilber_ner, and roberta_ner was conducted on 200 rows of randomly selected samples.

I decided to use **spacy_sm** for the named-entity extraction because it offers the highest performance in the extraction quantity and quality, processing speed, as well as diversity of results.



Entity Extraction – Modeling Process

Utilized spacy's pre-trained NER models (en_core_web_sm) enhanced with custom pattern-matching rules for technology entity extractions and domain-specific accuracy.

Why spacy?

- ✓ Pre-trained accuracy on PERSON/ORG entities
- ✓ Streamlined batch processing
- ✓ Native support for custom entity augmentation

Domain Adaptation

Extended spacy's capabilities with AI/tech-specific terminology capture via:

- ✓ Multi-word phrase matching
- ✓ Acronym coverage

Result

Extracted 400K+ people, organization, and technology entities that are ready to be refined by entity resolution

Key Processing Methods & Functions

Method	Purpose
Entity Ruler	Added custom TECHNOLOGY label (e.g., *"generative AI"*, *"blockchain"*) via 100+ domain-specific patterns
Batch Processing	Split 150K articles into 1K text batches to ensure memory and performance optimization
Text Truncation	Limited texts to 1M characters to prevent model overflow
Test Mode Sampling	Enabled validation on variable-sized subsets before full-scale run

Entity Extraction – Entity Resolution

Goal: Normalize 400K+ entities from 150K articles into canonical forms for consistent analysis
Core Challenge: Balancing accuracy with scalability for large datasets

Critical Methods & Process

Predefined Canonical Mappings

Resolves known entity variations (e.g., "AI" → "Artificial Intelligence")

Optimized String Cleaning

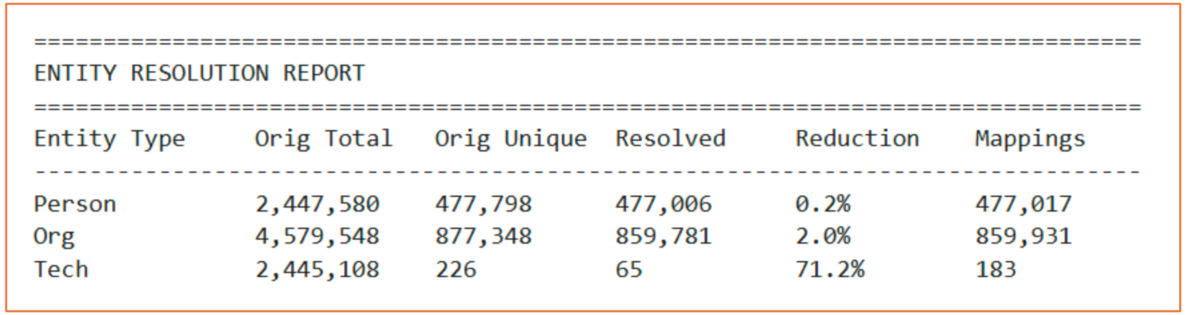
Pre-compiled regex patterns for:
✓ Whitespace normalization
✓ Prefix/suffix removal (e.g., "The Google" → "Google")

Intelligent Similarity Checking

- Early Termination: Skips fuzzy matching if string length ratio < 0.3
- Abbreviation Detection: Instant mapping for acronyms
- Token Set Ration: Uses fuzz.token_set_ratio for order-insensitive comparison

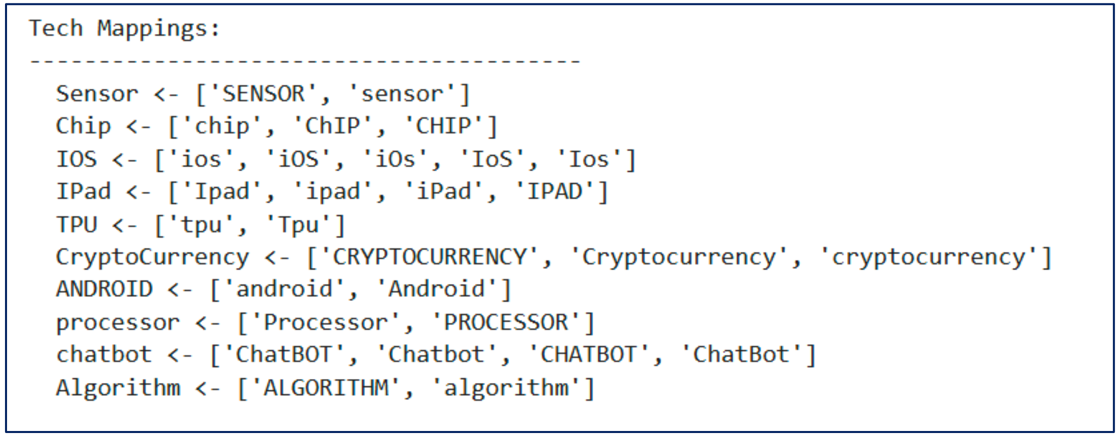
Scalable Clustering

- Prioritizes high-frequency entities to maximize cluster coverage
- Size-Based Strategy
- <10K entities: pairwise similarity checks
- >10K entities: frequency-based canonical mapping



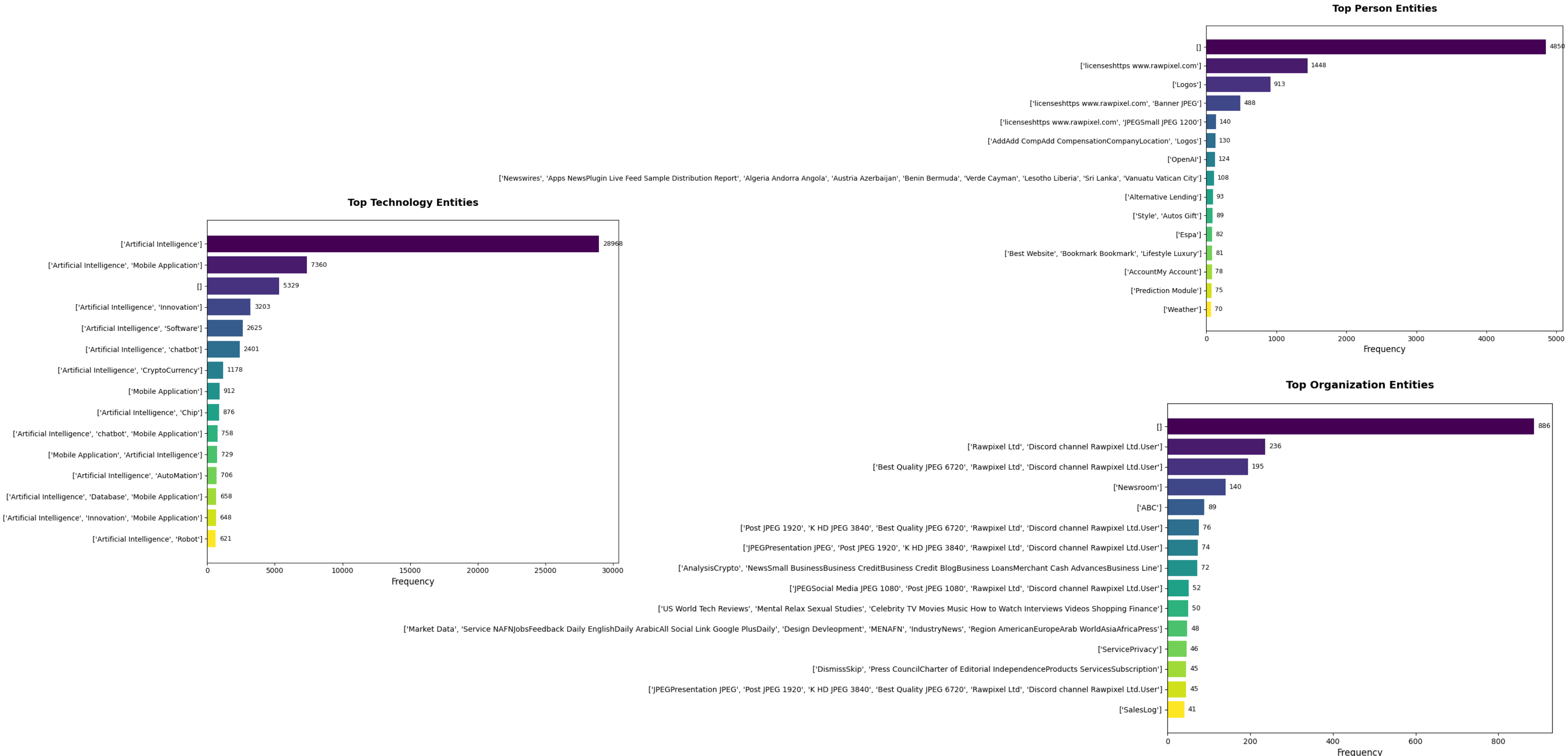
```
================================================================================
ENTITY RESOLUTION REPORT
================================================================================
Entity Type     Orig Total   Orig Unique  Resolved     Reduction    Mappings
--------------------------------------------------------------------------------
Person          2,447,580    477,798      477,006      0.2%         477,017
Org             4,579,548    877,348      859,781      2.0%         859,931
Tech            2,445,108    226          65           71.2%        183
```

Sample Resolution Mapping Results



```
Tech Mappings:
----------------------------------------
  Sensor <- ['SENSOR', 'sensor']
  Chip <- ['chip', 'ChIP', 'CHIP']
  IOS <- ['ios', 'iOS', 'iOs', 'IoS', 'Ios']
  IPad <- ['Ipad', 'ipad', 'iPad', 'IPAD']
  TPU <- ['tpu', 'Tpu']
  CryptoCurrency <- ['CRYPTOCURRENCY', 'Cryptocurrency', 'cryptocurrency']
  ANDROID <- ['android', 'Android']
  processor <- ['Processor', 'PROCESSOR']
  chatbot <- ['ChatBOT', 'Chatbot', 'CHATBOT', 'ChatBot']
  Algorithm <- ['ALGORITHM', 'algorithm']
```

Entity Extraction – Occurrences of Entities



Top Person Entities

Top Technology Entities

Top Organization Entities

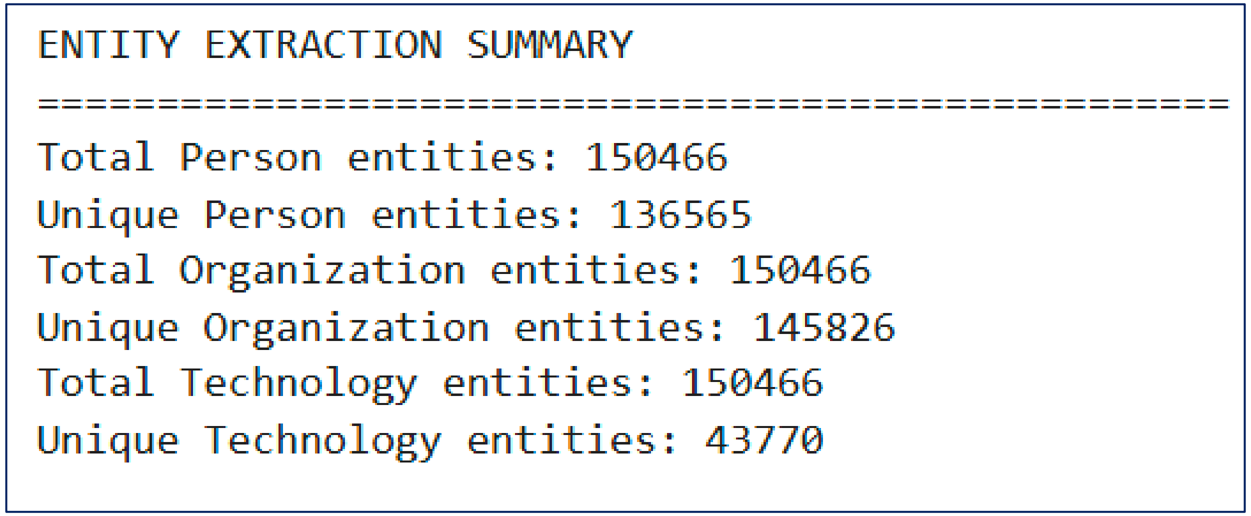
Entity Extraction – Results & Findings

Ubiquity of AI Across Sectors

- High unique organization count (145,826) reflects AI's penetration into diverse industries
- AI is no longer niche – it's actively discussed in non-tech sectors

Human-Centric AI Narrative

- Dominant person entities (136,565 unique counts) outpace technologies (43,770), implying that media frames AI as *human-driven progress*, which reduces fear of new change.



```
ENTITY EXTRACTION SUMMARY
==================================================
Total Person entities: 150466
Unique Person entities: 136565
Total Organization entities: 150466
Unique Organization entities: 145826
Total Technology entities: 150466
Unique Technology entities: 43770
```

Consolidated Technology Terminology

- 3.4 times more mentions per tech term vs. organizations or people (150K mentions / 43.7K unique) explains the uniform positivity (e.g., AI, chatbot, generative AI) that recur in consistently positive contexts.
- There's risk that low tech diversity may mask niche criticisms.

Entity Sentiment Analysis - Methodology

For sentiment analysis, I use a pre-trained aspect-based sentiment analysis (ABSA) model, as it directly handles entity-level (aspect) sentiments without custom training. A strong, ready-to-use option is "yangheng/deberta-v3-base-absa-v1.1" from Hugging Face, which classifies polarity (positive, negative, neutral) for a given entity in the text. It's based on DeBERTa and performs well on news-like data, with high accuracy on benchmarks like SemEval.

Contextual Sentiment Extraction

- ABSA-Specific Input Format: [text] [sep] [entity]
- Entity-aware tokenization (512-token limit)
- Confidence-calibrated predictions via softmax

Targeted Entity Analysis

Top Entity Identification:

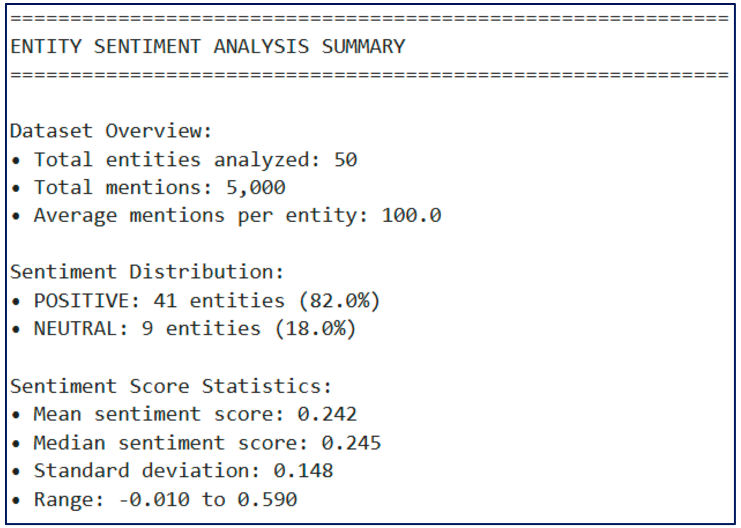
- Extract top 50 most frequent entities from resolved lists (person/org/tech)
- Prioritizes high-impact entities (e.g., "Artificial Intelligence" appeared 850 times)

Mention Retrieval:

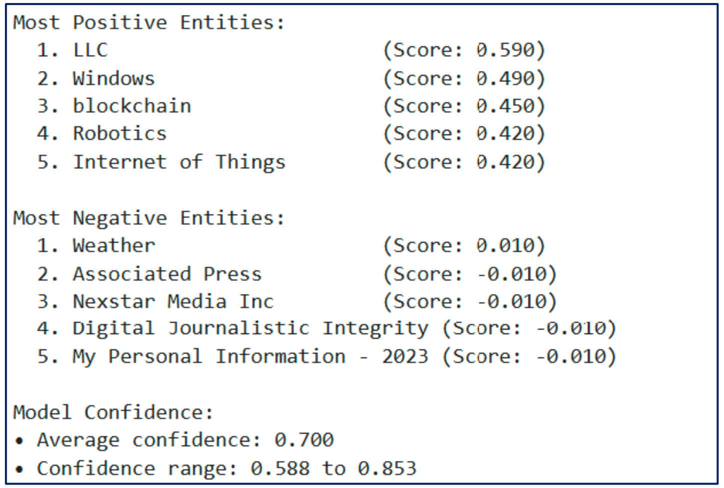
- Finds all articles mentioning each entity via resolved entity columns
- Samples max 100 mentions/entity for efficiency

Metric	Calculation	Decision Threshold
Sentiment Score	(Positive Count – Negative Count) / Total	Positive: > 0.1 Negative: < -0.1
Confidence	Mean prediction probability	Highlight >85% confidence
Overall Label	Score-based assignment	Neutral if between -0.1 and 0.1

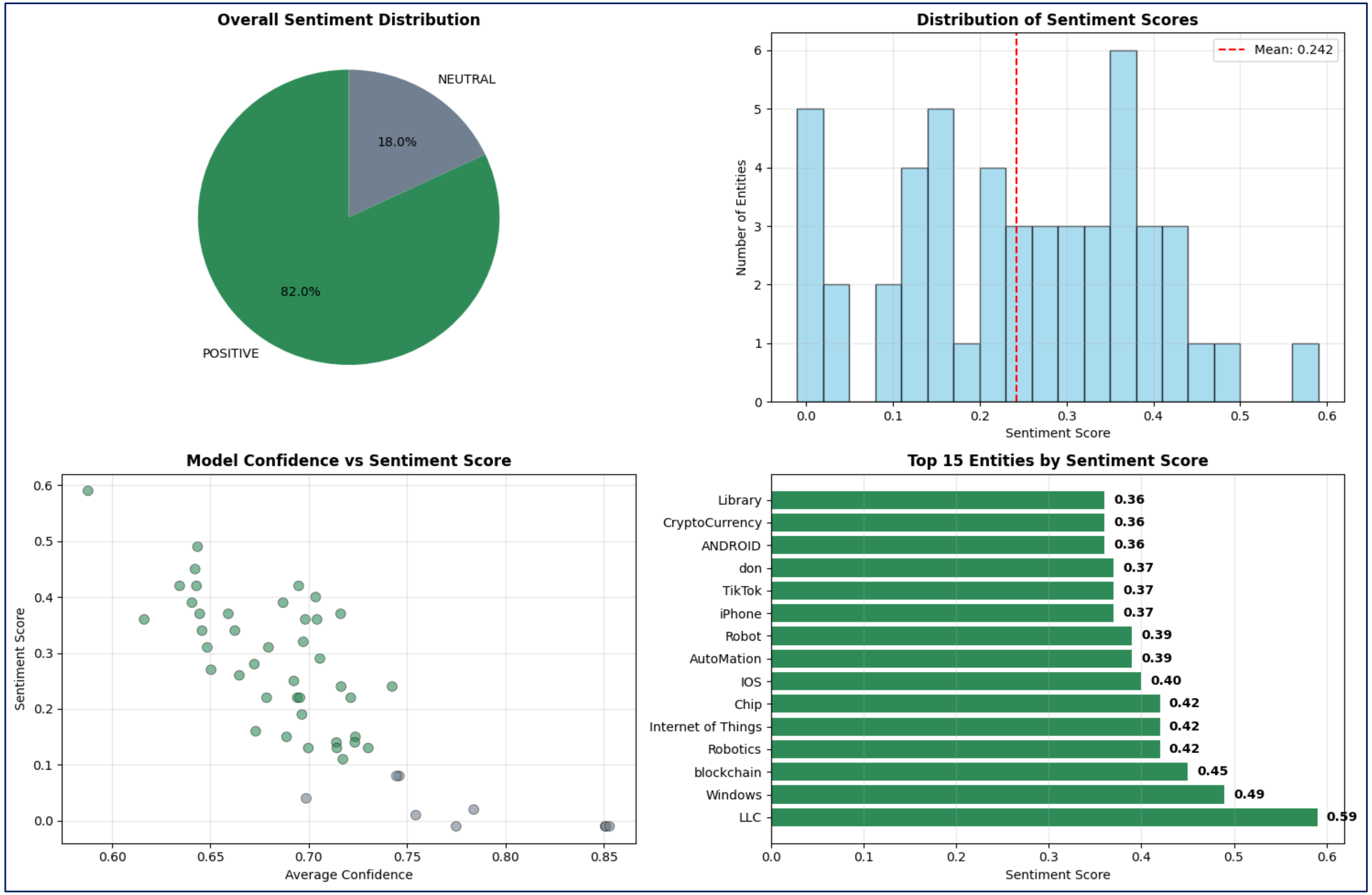
Entity Level Sentiment Analysis - Results



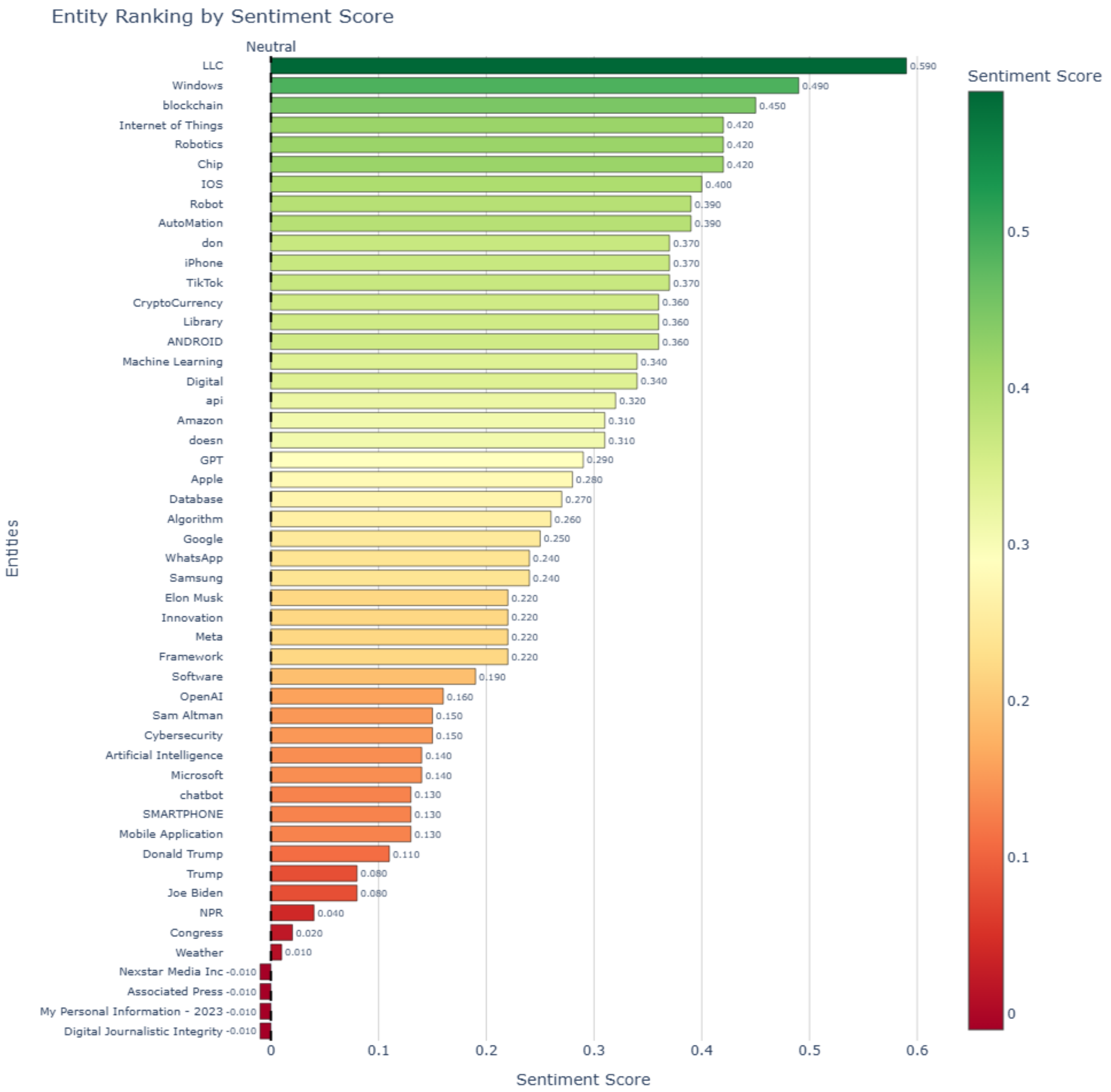
```
============================================================
ENTITY SENTIMENT ANALYSIS SUMMARY
============================================================

Dataset Overview:
• Total entities analyzed: 50
• Total mentions: 5,000
• Average mentions per entity: 100.0

Sentiment Distribution:
• POSITIVE: 41 entities (82.0%)
• NEUTRAL: 9 entities (18.0%)

Sentiment Score Statistics:
• Mean sentiment score: 0.242
• Median sentiment score: 0.245
• Standard deviation: 0.148
• Range: -0.010 to 0.590
```



```
Most Positive Entities:
  1. LLC                        (Score: 0.590)
  2. Windows                    (Score: 0.490)
  3. blockchain                 (Score: 0.450)
  4. Robotics                   (Score: 0.420)
  5. Internet of Things         (Score: 0.420)

Most Negative Entities:
  1. Weather                    (Score: 0.010)
  2. Associated Press           (Score: -0.010)
  3. Nexstar Media Inc          (Score: -0.010)
  4. Digital Journalistic Integrity (Score: -0.010)
  5. My Personal Information - 2023 (Score: -0.010)

Model Confidence:
• Average confidence: 0.700
• Confidence range: 0.588 to 0.853
```



Entity Level Sentiment Analysis – Results



Entity Sentiment Analysis – Insights & Findings

Overwhelming Positivity Confirmed

- 82% of entities how positive sentiment, which aligns with 92.6% topic-level positivity
- Both analyses show <5% negative sentiment, confirming media's tech-optimism bias

New Nuances Revealed by Entity Sentiment

- LLC as most positive entities signals media's focus on commercial operations, validating our topic findings "AI portrayed as economic growth driver"
- The entity "My Personal Information" is among the most negative contrasts with topic-level analysis where data privacy wasn't a major theme. This suggests that critical concerns about privacy exists but are drowned out by dominant media positivity.
- Many media entities (e.g., Associative Press, Nexstar Media, Digital Journalistic Integrity) are marked with negative sentiments explains why "AI in media had lower sentiment confidence in topic-sentiments

Sector	Entities	Topic Alignment
Tech	Windows, IoT, Robotics	"Core Technology" theme with high positivity
Finance	Blockchain	"AI in Finance" topic cluster
Media	Digital Journalistic Integrity (negative)	"Creative/Media" theme's mixed sentiment

This table shows examples of sector impact patterns that are reinforced by the entity-level sentiments

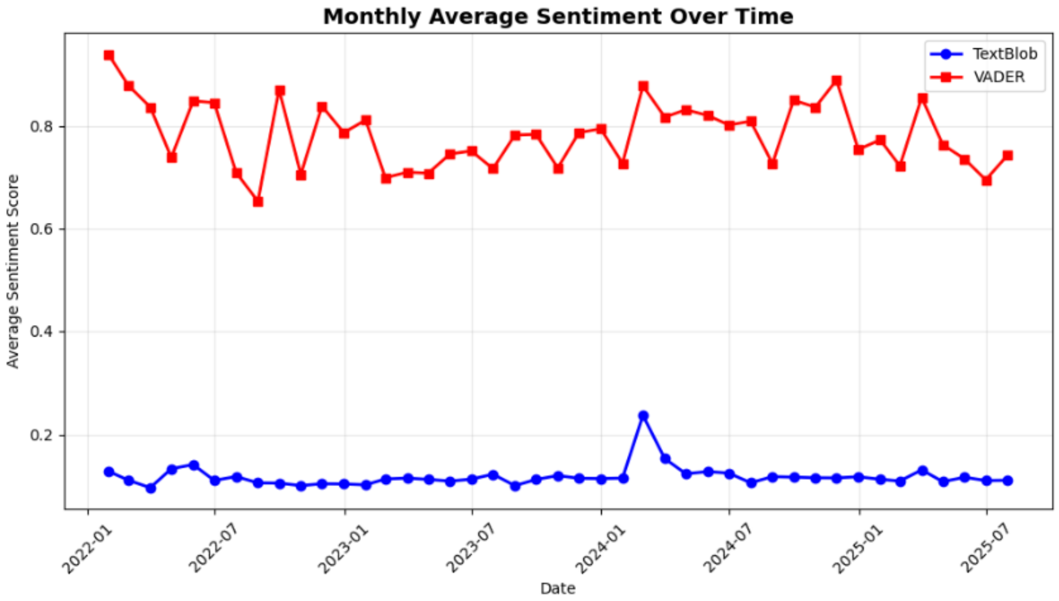
Sentiment Over Time – Methodology & Results

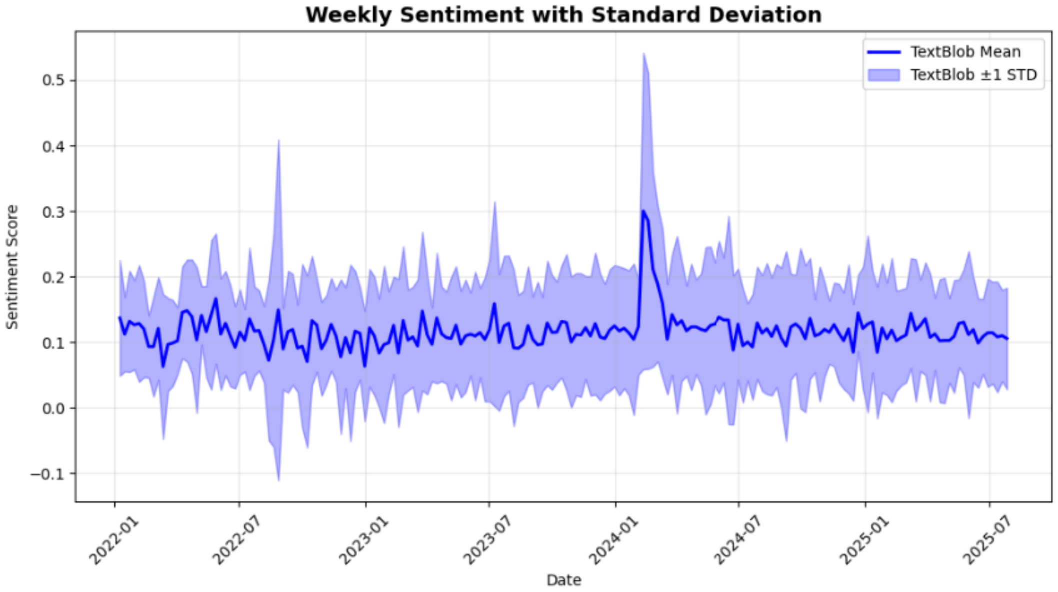
Dual Sentiment Analysis: Uses both Text Blob and VADER sentiment analyzers for more robust analysis

Multiple Time Aggregation: Daily, weekly, and monthly sentiment trends are computed

Why aggregate to month and week?

Given my dataset spans from 20222-2025, monthly aggregation provides about 24-36 data points for trend analysis, which is enough to identify seasonal patterns without being sparse and provides more statistical stability. Weekly produced 100-150 points for more detailed analysis. Additionally, tech news operations on weekly cycles (e.g., major announcements on Mondays, earnings typically on specific weekdays), weekly resolutions syncs with the news cycle sensitivity.





Sentiment Over Time – Discussion & Findings

Summary of Trends in the Sentiment Plots

Monthly VADER Trend: Starts out strong at approximately 0.9 in January 2022, indicating highly positive sentiment. A constant drop in sentiment from until a small upward turning point starting in May/June 2022. Some fluctuations extending to Q1 2023, and the most significant dip in sentiment stretches between Q2 2023 to 2024. In Q1 2024, a strong sharp spike appeared. The score stabilizes afterwards, fluctuates between 0.7 to 0.85 and ends around 0.7 in July 2025.

TextBlob Trend (Monthly and Weekly): Begins around 0.1 in January 2022, hovering low (0.05-0.15) with a significant spike in early 2024, then drops to the original level around 0.1 between Q2 2024 to July 2025.

Impact of Technology Introductions on Sentiment

Based on the plot, no single technology introduction appears to cause a sharp, immediate spike or dip in sentiment scores. Instead, changes seem gradual or tied to broader events.

ChatGPT (Released November 30, 2022): Around December 2022–January 2023, VADER shows a slight dip (~0.7 to 0.65), while Text Blob remains flat at ~0.15. This suggests no sharp positive impact; if anything, initial coverage may have introduced cautious tones amid rapid adoption. Research indicates mixed effects: ChatGPT drove positive social media sentiment and shifted media narratives toward AI leadership and risks, but also led to a decline in general public sentiment toward AI compared to pre-launch levels. No abrupt change is evident in the plot, possibly due to the monthly aggregation smoothing out short-term hype.

The January 2024 spike likely correlates with a cluster of events rather than one technology, including CES 2024 (January 9–12), which featured extensive AI announcements and demos, potentially amplifying positive news sentiment. Other January activities, like the Data Science Salon (January 24) and Enterprise LLM Summit (January 25), highlight a surge in AI-focused gatherings that could have influenced upbeat reporting. Later releases like OpenAI's Sora (February 15, 2024) or Claude 3 (March 2024) post-date the spike and align with the subsequent decline.