UNRAVELING THE DATA SCIENCE REVOLUTION

Sam Ding

MSCA 32018 NLP Final Project

May 26, 2023

EXECUTIVE SUMMARY

Problem Statement:
Identify what types of tasks and jobs are most likely to see the biggest impact from AI by extracting meaningful insights from unstructured text

Data:
We have a data of 200k news articles about Data Science and AI published since 2020. We cleaned the article and removed unrelated contents for our downstream analysis.

Findings:
- Major topics include investments, risks, application in different industries, as well as computing infrastructure and process optimization for companies.
- Sentiment toward AI initiatives are generally positive, though negative sentiments increased since the advent of ChatGPT.
- Data-intensive and technologically-equipped industries and companies are successful in AI-initiatives, whereas creative, strictly-regulated industries, or industries where ethics play a significant role, are facing barriers.

Recommendations:
For companies, academic institution, or governments investing in AI initiatives, we recommend focusing on developing necessary infrastructures, fostering collaboration, and exercising caution on ethics and data privacy.

TABLE OF CONTENTS

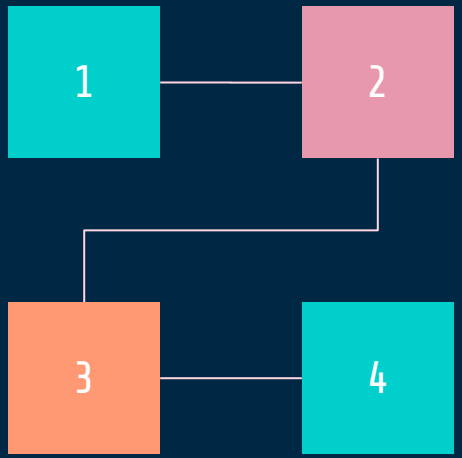
DATA CLEANUP

SOURCE DATA

We have 200k news articles about Data Science and AI, published since Jan 2020

DROPPING IRRELEVANCIES

Remove remaining special characters, links, or other non-ASCII chars.
Filter based on a list of keywords like '[M|m]achine learning' and '[D|d]ata science'.



SIMILARITY TESTING BY CHUNKS

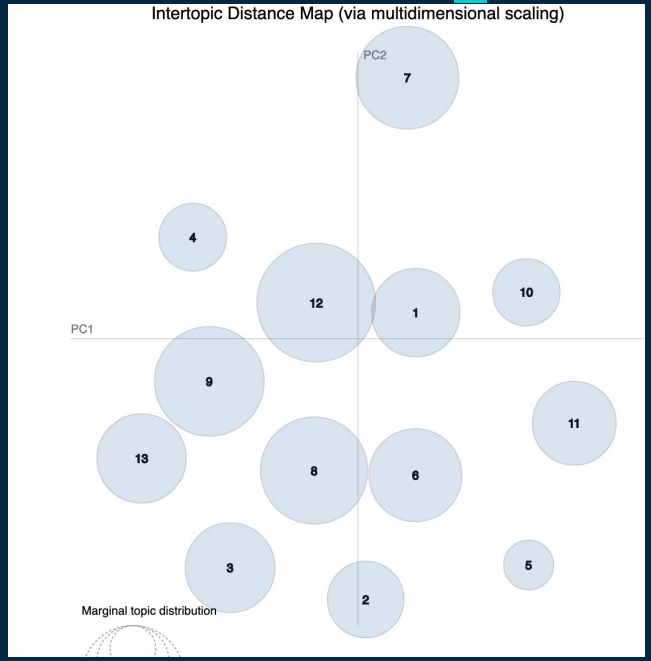
Split each text by occurrences of date or tab/newline chars like \t, \n & \xa0. Tokenize and compare each item with their title using cosine similarity. Dropping items if similarity score is under a threshold, which could be potential web crawl remnants.

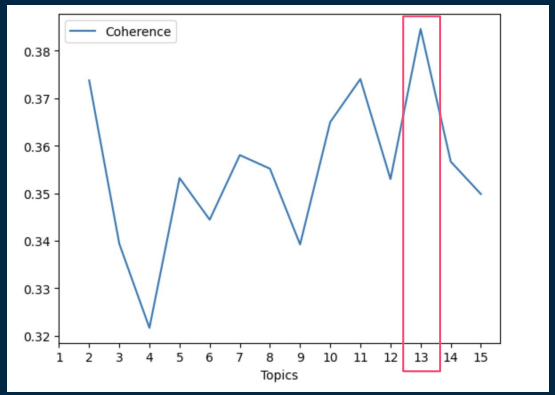
FILTERED DATA

Contains ~100k articles, their date published and other relevant info

MAJOR TOPICS DISCUSSED

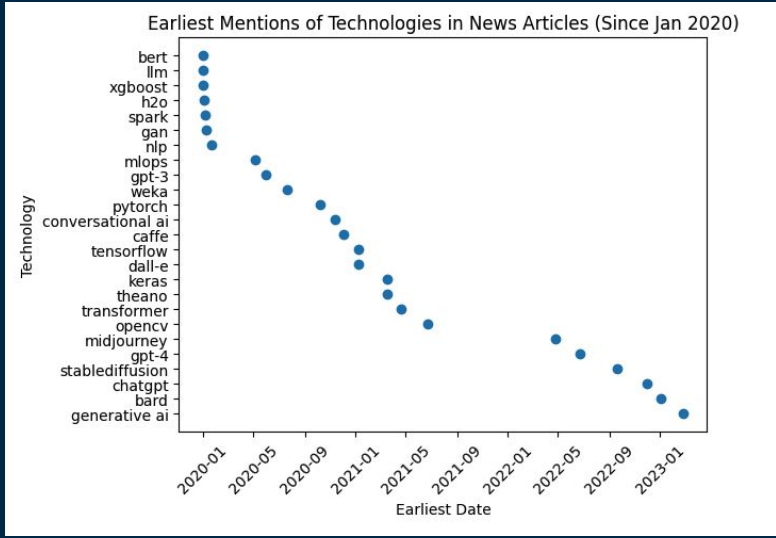
1. Stock Fund Investment and Trading Related to Data Science
2. Governance and Security Risks
3. AI-Enabled Medical Research, Care and Imaging Innovations
4. AI-Powered Energy Management and Supply Chain Optimization
5. Cloud Computing and Computer Vision
6. Data-Driven Business Insights and Revenue Forecasting
7. AI-Driven Analytics and Process Optimization



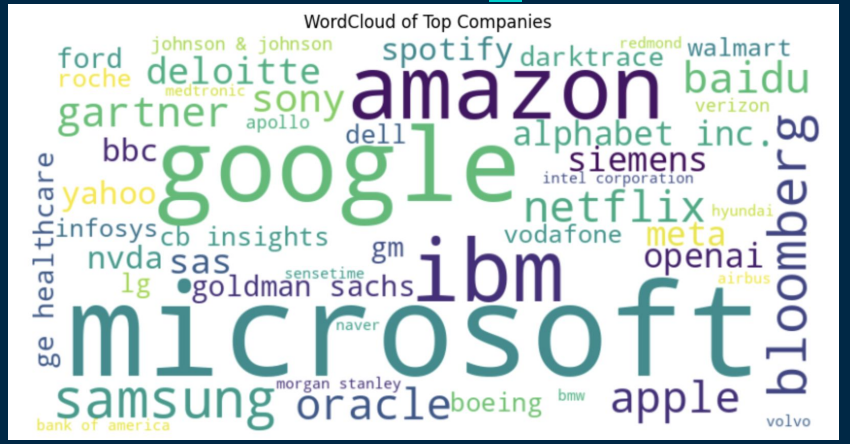


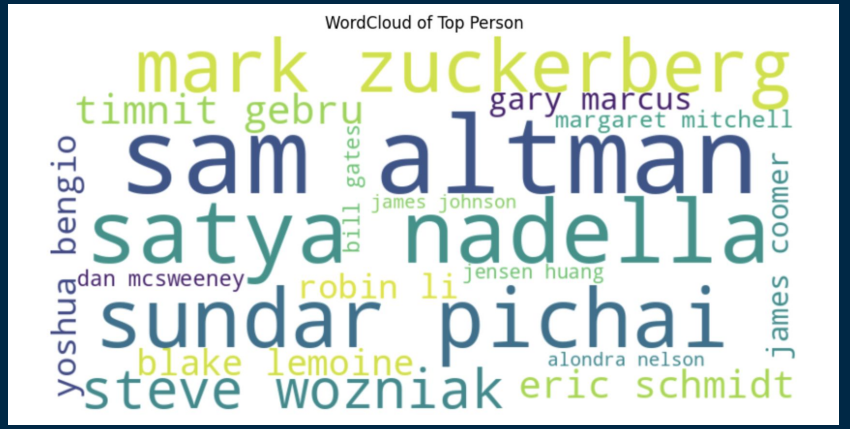
I used LDA topic modeling on a sample for visualization, then I used ktrain topic modeling on all articles for faster performance.
These are the top topics selected due to limited space.
Hyperparameter tuning shows best performance at 13 topics. Topics are evenly distributed, not much overlapping.

ENTITY IDENTIFICATION



Even though ChatGPT is one of the latest AI products unveiled in this chart, it actually is the **most mentioned entity** across the board at 14k+ times.

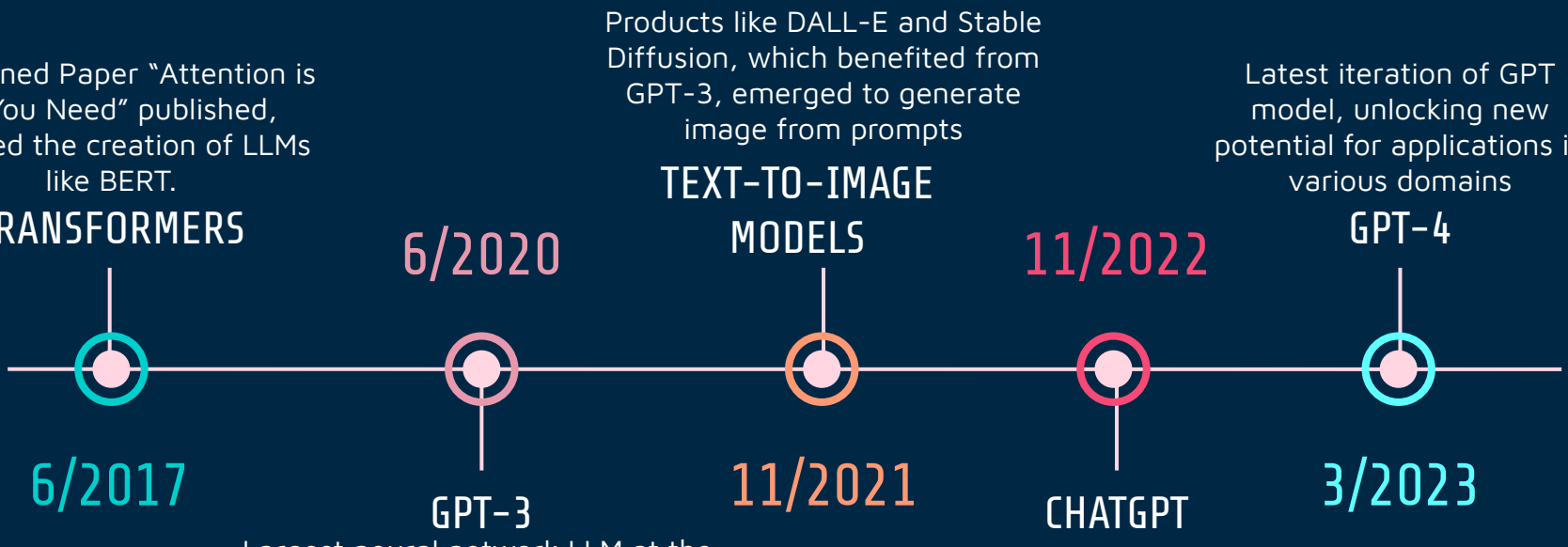




*For the Wordclouds, I selected top orgs and persons and cleaned the results so that they only reflect related orgs and person.

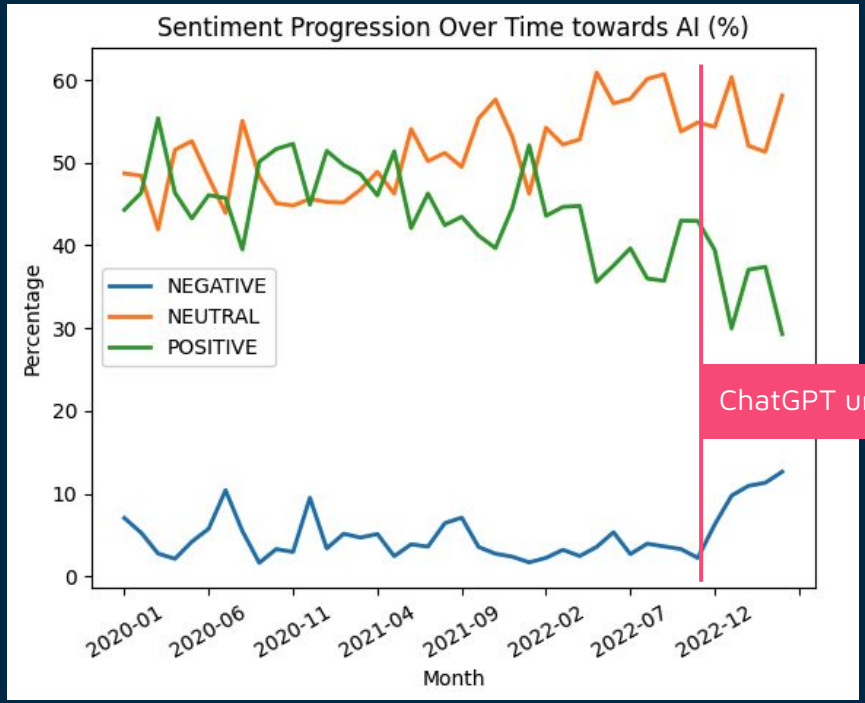
(SIMPLIFIED) TIMELINE OF AI DEVELOPMENTS

Among all the rapid developments in AI and data science, I picked several significant events and/or products to illustrate the timeline.



SENTIMENT CHANGE OVER TIME

- I used ktrain topic modeling that is based off of BERT.
- Hand-labeled sample of articles show **85%+ accuracy** of labeling
- We see decreasing positive sentiment since 2022
- We also see significant rising negative sentiment since Nov 2022, roughly the time **ChatGPT rolled out**. This is potentially concern about the product and its use, so we dig in deeper.



TARGETED SENTIMENT

Doing targeted sentiment analysis for industries and person helps us better understand where the AI revolution is now.

I found that:

Industries associated with higher positive sentiments often are data-intensive, see AI as an enhancement to their decision-making process, and have manual, repetitive tasks that are replaceable by AI systems.

Industries associated with higher negative sentiments are generally creative or strictly-regulated sectors.

Detailed examples of industries on the next page.

Persons associated with higher positive sentiments are generally reporting on **executives leading AI initiatives**, including Jensen Huang (NVIDIA), Satya Nadella (Microsoft), and Sundar Pichai (Google).

Persons associated with higher negative sentiments are cited **raising concerns about AI ethics or lack of transparency**, including Gary Marcus, Blake Lemoine, and Timnit Gebru.

*Marcus is the founder of Robust.ai, and the latter two are ex-employees of Google.

In conclusion, persons with higher positive and higher negative sentiments are at odds with each other. These differing perspectives highlight the ongoing debate surrounding AI and the need to ensure the technology is harnessed for the benefit of society.

SOME INDUSTRIES TO INVEST OR TO BE CAUTIOUS ABOUT

Some Successful Industries in AI Transformation:

Healthcare: AI can aid researchers in identifying pathological patterns, help doctors improve medical diagnosis, enable personalized treatment plans for patients, and support telemedicine initiatives.
Consulting Services: AI-powered analytics and automation can enhance consulting services by providing data-driven insights, predictive modeling, and process optimization.
Retail: AI technologies enable personalized customer experiences, demand forecasting, inventory management, and targeted marketing campaigns.
Insurance: AI system are capable of risk assessment, fraud detection, expedite claim processing, and provide personalized customer experiences.

Some Industries to be watchful about:

Academic Institutions: while AI innovation helps research purposes, **academic plagiarism** is also on the rise due to the development of Generative AI platforms.
Journalism & Media: there are concerns related to the **reliability and credibility** of AI-generated content, the potential for **misinformation** and **deepfakes**, and the impact on j**ob security for journalists**.
Legal Services: the adoption of AI in legal services raises concerns about **job displacement** for paralegals and junior lawyers, **potential biases** in AI algorithms that may lead to unjust verdicts, and ethical implications related to **data privacy and confidentiality**.
Governmental Affairs: AI applications may raise concerns about **citizen privacy**, state **surveillance**, **fairness and equity** of the system, and **cybersecurity**.

RECOMMENDATIONS FOR STAKEHOLDERS

For Companies:

Invest in AI R&D: Allocate resources to AI research and development, fostering innovation within the organization.
Focus on Ethical AI Practices: Develop and implement ethical guidelines for AI development, addressing issues like bias, fairness, and transparency.
Foster Collaboration: Collaborate with education institutions, other companies, and AI startups to share collective resources and knowledge.
Data Management and Privacy: Establish robust data management practices, ensuring the responsible collection, storage, and use of data.
Upskill Employees: Provide training opportunities to employees to foster a workforce equipped with AI knowledge and skills.

For Academic Institutions:

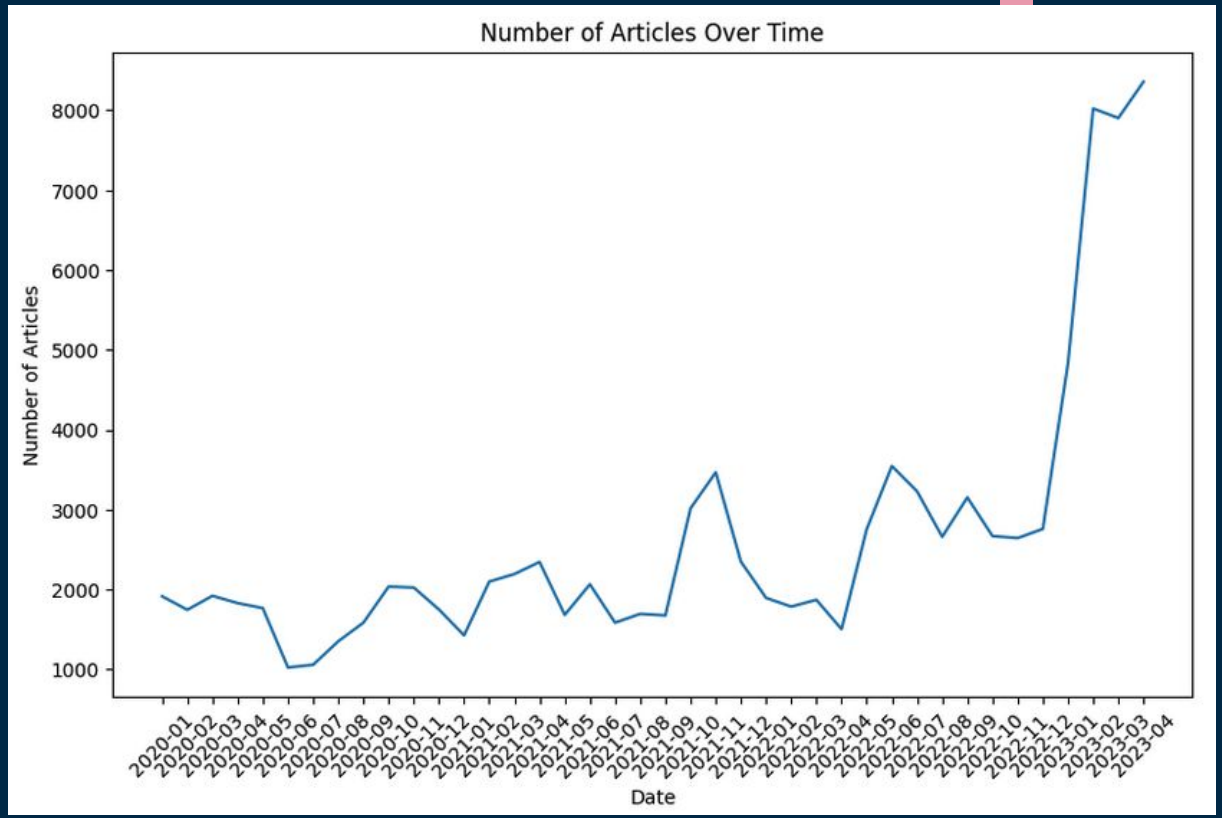
Build Necessary Infrastructure: Allocate funds to build efficient, high-performance computing resources and provide access to software needed.
Advocate AI Education and Training: Offer training programs to students faculty to equip them with the necessary knowledge and skills.
Encourage Research and Collaboration with Industry: Provide funding for research opportunities related to AI. Foster collaboration with industry leaders to gain more exposure in the field for students and faculties.
Prioritize Ethical Considerations: Incorporate ethics education and responsible AI practices into the curriculum. Encourage research that addresses the ethical implications of AI, such as fairness, transparency, data privacy, and bias.

For Governments:

Develop Relevant Regulations: Establish guidelines for responsible AI development, deployment, and usage, ensuring compliance with privacy, security, and fairness standards.
Develop Data Governance and Management: Establish data governance practices to ensure security and privacy of data used in AI.
Collaborate with Industries and Academic Institutions: Foster collaborations to understand industry needs and current trends in research and practice, in order to enhance decision-making process within agencies.
Engage with the Public: Foster open and inclusive public engagement to gather feedback, address concerns, and ensure public acceptance of AI initiatives.

THANKS!

APPENDIX 1: # OF ARTICLES OVER TIME



APPENDIX 2: TOPIC RESULT

topic 0 | stock fund investment investor distribution capital price return network trading
topic 1 | government state india security risk law right privacy policy china
topic 2 | drug day discovery covid19 researcher disease medicine scientist lab cell
topic 3 | student university education program school skill institute project course innovation
topic 4 | energy financial management risk process sector operation supply change increase
topic 5 | say image example thing think algorithm problem question different computer
topic 6 | medical clinical care healthcare cancer imaging hospital disease treatment image
topic 7 | cloud edge computing nvidia network device chip computer vision smart
topic 8 | forecast player trend region size opportunity provides segment factor revenue
topic 9 | enterprise organization partner insight capability analytics today leading process ceo
topic 10 | gray release press statement view http content reflect television original
topic 11 | google microsoft openai search chatbot tech generative musk bing engine
topic 12 | video news day app game music home image hour best

APPENDIX 3: NER RESULT

entity	index	label	date
chatgpt	14708	14708	14708
microsoft	13413	13413	13413
covid-19	9530	9530	9530
google	7901	7901	7901
amazon	5682	5682	5682
ibm	5434	5434	5434
ml	4164	4164	4164
dr.	3364	3364	3364
fda	2903	2903	2903
don	2616	2616	2616
intel	2268	2268	2268
ap	2250	2250	2250
nyse	2200	2200	2200
reuters	2094	2094	2094
congress	1814	1814	1814
cto	1578	1578	1578
healthcare	1517	1517	1517
eu	1505	1505	1505
android	1451	1451	1451
nasa	1425	1425	1425

APPENDIX 4: INDUSTRIES