Artificial Intelligence

Related News

Analysis

-Yichen Peng

Executive Summary

Objective:

- Identify what types of tasks and jobs are most likely to see the biggest impact from AI.
- Find actionable recommendations on what can be done with AI

Method:

- Sentiment analysis and named entity recognition to perform general analysis
- Utilize topic modeling and keywords to extract industries and technologies.
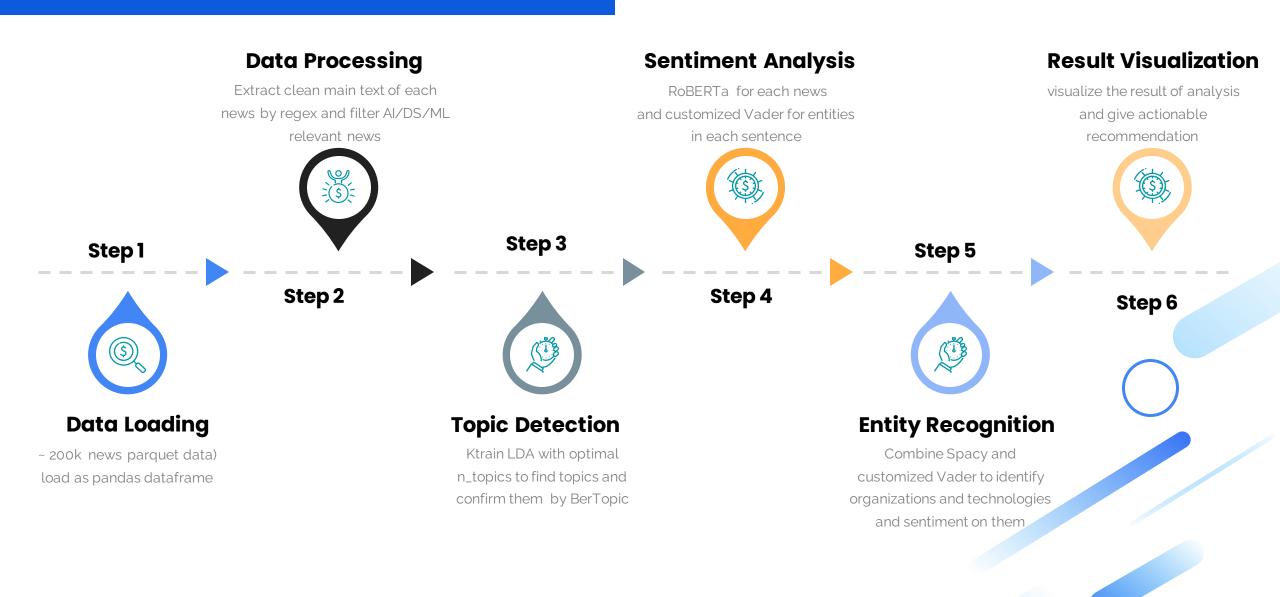- Then same techniques to perform target sentiment analysis.

Conclusion

- Healthcare benefits significantly from AI/DS/ML advancements
- Finance, particularly cryptocurrencies, face increased uncertainty due to AI/ML/DS and thus is negatively affected.
- Chatbots and AI content creation are most popular and could automate and improve tasks in customer service and design.
- Public apprehensions towards AI/ML/DS often stem from subpar user experience, worsened inequality, and privacy breaches.

Recommendation

- Healthcare entities should adopt AI and invest in AI-powered diagnostic solutions
- Finance firms might consider awaiting AI's maturity and regulatory clarity before fully embracing it.
- People in AI-impacted roles should stay adaptable to AI advancements
- Companies ought to gather representative data and avoid hasty product releases
- Academic bodies should bolster research into core technologies
- Governments need to swiftly set regulations to protect public data privacy and security.

Methodology

Data Processing

Extract clean main text of each news by regex and filter AI/DS/ML relevant news

Sentiment Analysis

RoBERTa for each news and customized Vader for entities in each sentence

Result Visualization

visualize the result of analysis and give actionable recommendation

Step 1

Step 2

Step 3

Step 4

Step 5

Step 6

Data Loading

~ 200k news parquet data) load as pandas dataframe

Topic Detection

Ktrain LDA with optimal n_topics to find topics and confirm them by BerTopic

Entity Recognition

Combine Spacy and customized Vader to identify organizations and technologies and sentiment on them

Data Source Overview and Preprocessing

Data: Parquet

Size: 200332 rows, 5 columns

Column	Type	example
url	String	http://en.people.cn/n3/202…
date	date	2021-03-18
language	String	en
title	String	Artificial intelligence improves…
text	String	\n\nArtificial intelligence improves parking …

Cleaning to extract main text of news:

- Exclude non-English news
- Split by special symbol (\n\t…)
- Filter paragraph with over 10 words to exclude irrelevant content such website instructions
- Remove paragraphs after the recommendation phrase such as "Related Article"
- Remove paragraph containing url etc

Filtering to discard irrelevant news:

- Discard news irrelevant to Artificial Intelligence, Data Science and Machine Learning
- Number of news decrease from 200332 to 135927

Topic Detection

LDA with Hyperparameter Tuning （know optimal n_topics） → Ktrain (detection) → BerTopic (filtering)

Industry

- Finance
- Healthcare
- Education
- Automotive
- Studio Art
- AI-related Hardware

	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
Topic 3	best	insurance	credit	card	forbes
Topic 37	patient	health	cancer	care	hospital
Topic 4	science	student	education	university	course
Topic 5	image	art	photo	artist	like
Topic 40	chip	nvidia	computing	hardware	intel

Concerns:

- Privacy and Cybersecurity
- Ethical issues, bias, and risk

Application & Technology:

Keywords
- ChatBot
- Auto-writing
- Autonomous driving
- Prediction & Forecast
- Bard
- Generative
- ...

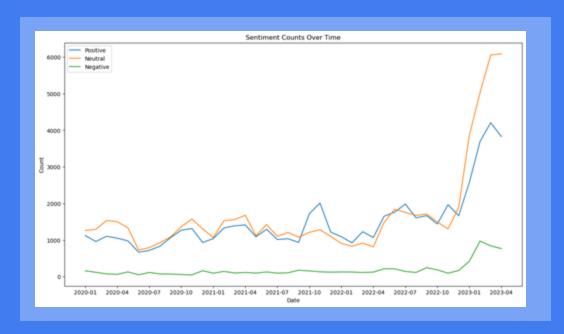|	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
Topic 1	security	military	cyber	cybersecurity	threat
Topic 20	right	bias	risk	regulation	ethical

Sentiment Analysis

Customized Vader by Editing Lexicon on Entity

positive: breakthrough; scalability; personalization; inclusive; streamlining

negative: overreach; unaccountable; inequality; disparity; manipulation

RoBERTa on News

Fine-tuned with financial news

Accuracy check as follows

```
[{'summary_text': " China's capital city Beijing has seen significant "
                  'improvement in the efficiency of parking fee collection, '
                  'turnover of roadside parking spots, order in roadside '
                  'parking, as well as traffic congestion . With the '
                  'AI-empowered system, drivers can park their vehicles at the '
                  'parking spots on the roadside, and then pay the parking '
                  "charge via their mobile phones'}]
```

Above is the text summarization of one news, and it is evident that the sentiment of this AI-related news is positive. Our model has accurately identified it as well. In fact, I have manually summarized and confirmed around 50 news articles, achieving an accuracy rate of 78%. Errors mostly occurred when distinguishing between neutral and positive sentiments and those news are indeed in between neutral and positive.

The above graph shows how the sentiment of news on AI/DS/ML change over time in general. The introduction of ChatGPT in January 2023 has clearly sparked a significant surge in interest and discussion surrounding this topic, encompassing positive, neutral, and negative sentiments.
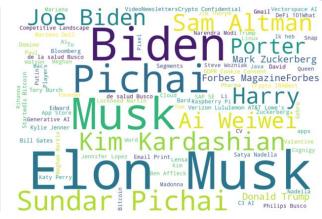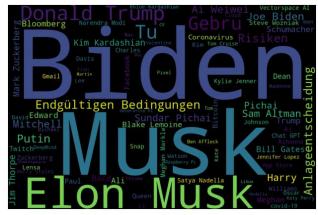
Entity Identification

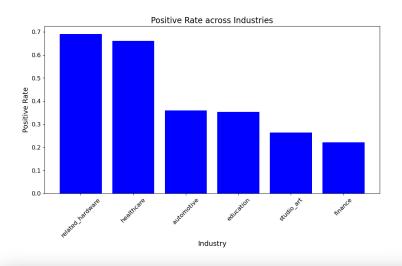Organization

People

Location

In these 6 graphs, white represents the top positive sentiment entities, while black represents the top negative sentiment entities.
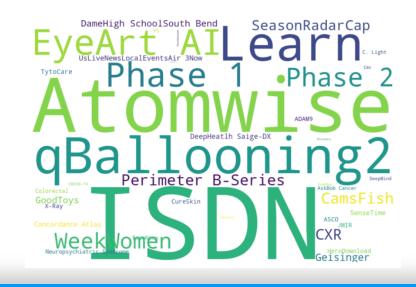
Whether it's organizations, individuals, or locations, we find that the most positive entities overlap heavily with the most negative ones.

This suggests that the field of AI/DS/ML is still largely dominated and led by large corporations such as Google, major countries such as US, and influential figures including Musk, and overall, it carries a considerable degree of controversy.

Therefore, we will dive into specific industries and applications.

Targeted Sentiment Identification–Industry Analysis (Success)

Among the six industries potentially influenced by the development of AI/DS/ML, healthcare is most positively impacted, if we exclude related hardware. This is because related hardware supplies necessary equipment for these technologies, rather than being a direct recipient of their impacts as we are interested in.
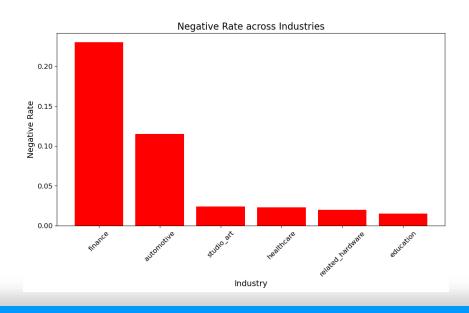
Based on the above word cloud, ATCOR, Brainnomix and Aidoc are healthcare companies providing AI solutions to support decision-making in diagnosis and treatment and all of them are viewed positive and promising. Therefore, healthcare companies should invest in AI and DS
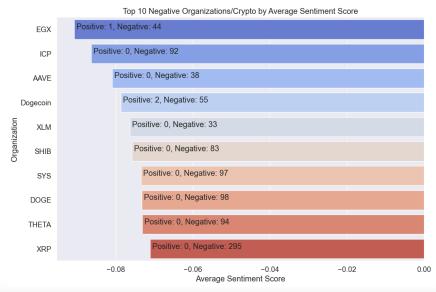
Based on the above word cloud, ISDN is Integrated Services Digital Network. EyeArt AI and qBallooning2 are both AI techonologies that support diagnosis. Therefore, healthcare companies should invest digitalization of medical data and AI technologies helping disease detection

Targeted Sentiment Identification–Industry Analysis （Failure）

Negative Rate across Industries

Top 10 Negative Organizations/Crypto by Average Sentiment Score

Organization	Sentiment counts
EGX	Positive: 1, Negative: 44
ICP	Positive: 0, Negative: 92
AAVE	Positive: 0, Negative: 38
Dogecoin	Positive: 2, Negative: 55
XLM	Positive: 0, Negative: 33
SHIB	Positive: 0, Negative: 83
SYS	Positive: 0, Negative: 97
DOGE	Positive: 0, Negative: 98
THETA	Positive: 0, Negative: 94
XRP	Positive: 0, Negative: 295

Among those 6 industries, Finance has much higher negative rate on news which means it may be most negatively impacted by AI/DS/ML.
Actually, I merge traditional finance, fintech, and crypto into one. So, let's dive into entities to see where negativity come from.
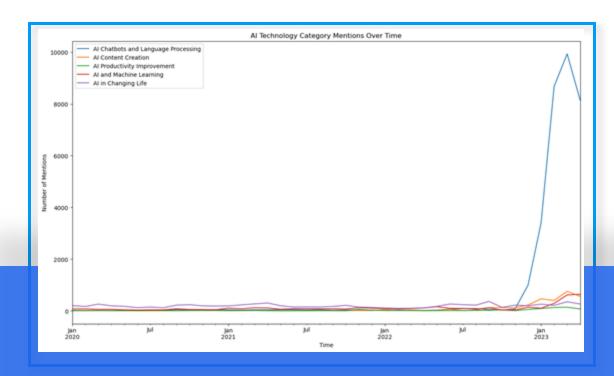
Based on the above, the entities with the lowest sentiment score are mostly cryptos.
In the finance fields, especially cryptos, it's hard to handle market volatility and uncertainty.
The introduction of AI in these fields may exacerbate investor anxiety, thereby increasing uncertainty.
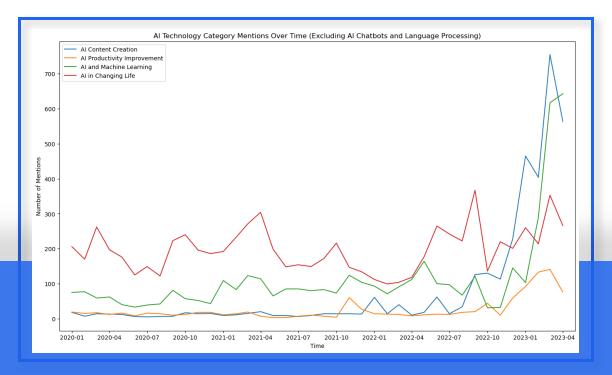Therefore, I do not think applications in finance (crypto) should be transformed by AI when it is not mature and well-regulated.

Targeted Sentiment Identification –Technology Analysis

AI Technology Category Mentions Over Time

AI Technology Category Mentions Over Time (Excluding AI Chatbots and Language Processing)

Apparently, natural language based technologies and chatbot based applications are much more intensively discussed and successful. It may because:

- meets people's biggest need – answer any questions
- comprehensive and fundamental functions mostly require basic knowledge (rich data availability)
- OpenAI and ChatGPT pushes the technology maturity

Among other categories, content creation and machine learning (DS) related technologies and applications are more popular probably because:

- Content creation fully harnesses the strength/potential of AI
- ML/DS technologies are fundamental support and therefore paid attention to

Productivity improvement applications do not spark intense discussion and AI in changing life such as autonomous driving has not seen a surge in topic popularity. It may because:

- Simply improving basic productivity cannot fully utilize the power of AI
- Human productivity cannot be fully substituted, or unemployment ensues
- AI in changing life has been frequently mentioned since 2020, but its popularity has been lukewarm, possibly due to the lack of significant breakthroughs
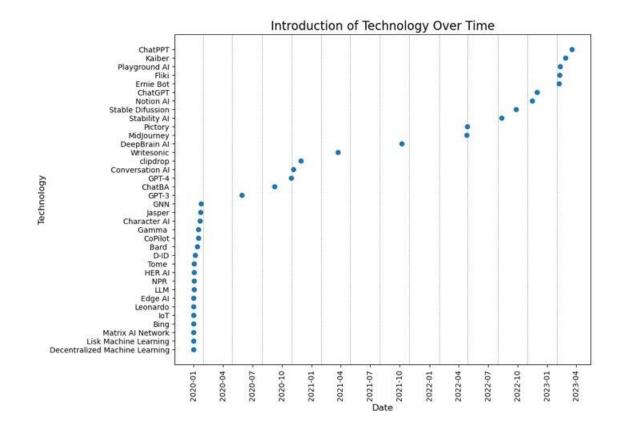
Concerns Analysis

[{'summary_text': ' Corporates are using Artificial Intelligence (AI) '
                  'platforms in the interview process to simplify and speed up '
                  'hiring . But real world use of AI highlights concerns that '
                  'discrimination caused by flaws in the technology and '
                  'associated processes is causing suitable candidates to miss '
                  'the short list . The coronavirus pandemic has really '
                  'impacted the global job market, with nearly all'}]
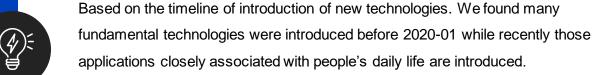
[{'summary_text': ' Capgemini report says 70% of customers expect fair AI '
                  'interactions from firms . Discriminatory bias of AI systems '
                  'and negative AI customer experiences come at a high cost '
                  'for organisations . Share of customers who believe '
                  'organisations are being fully transparent about how they '
                  'are using their personal data has fallen from 76% in 2019 '
                  'to 62% today'}]

I chose the 10 most negative sentiment news from the concerns topic and perform text summarization by transformer

The above results well shows people's concerns mainly from

- Flaws in technology and poor data representativeness may exacerbate discrimination and inequality
- The premature advent of immature technology, which may be stimulated by ChatGPT, instead led to a poor user experience
- The absence of appropriate regulations and supervisions could potentially result in the infringement of personal privacy data

Based on the timeline of introduction of new technologies. We found many fundamental technologies were introduced before 2020-01 while recently those applications closely associated with people's daily life are introduced.

This timeline effectively demonstrates the progression of a technology from the technical side to the practical application side and it takes time.

Conclusions

- ✓ AI/DS/ML landscape is driven by large entities, manifesting substantial controversy

- ✓ Healthcare sees the most positive impact from AI/DS/ML progress

- ✓ Finance, especially cryptocurrencies, experiences more uncertainty from AI/ML/DS advancement and thus AI applications in it seem to fail

- ✓ Chatbots and AI content creation may automate roles like customer service and designers, enhancing productivity and creativity

- ✓ People's concerns on AI/ML/DS mainly come from poor user experience, aggravated inequality, and invasion of privacy
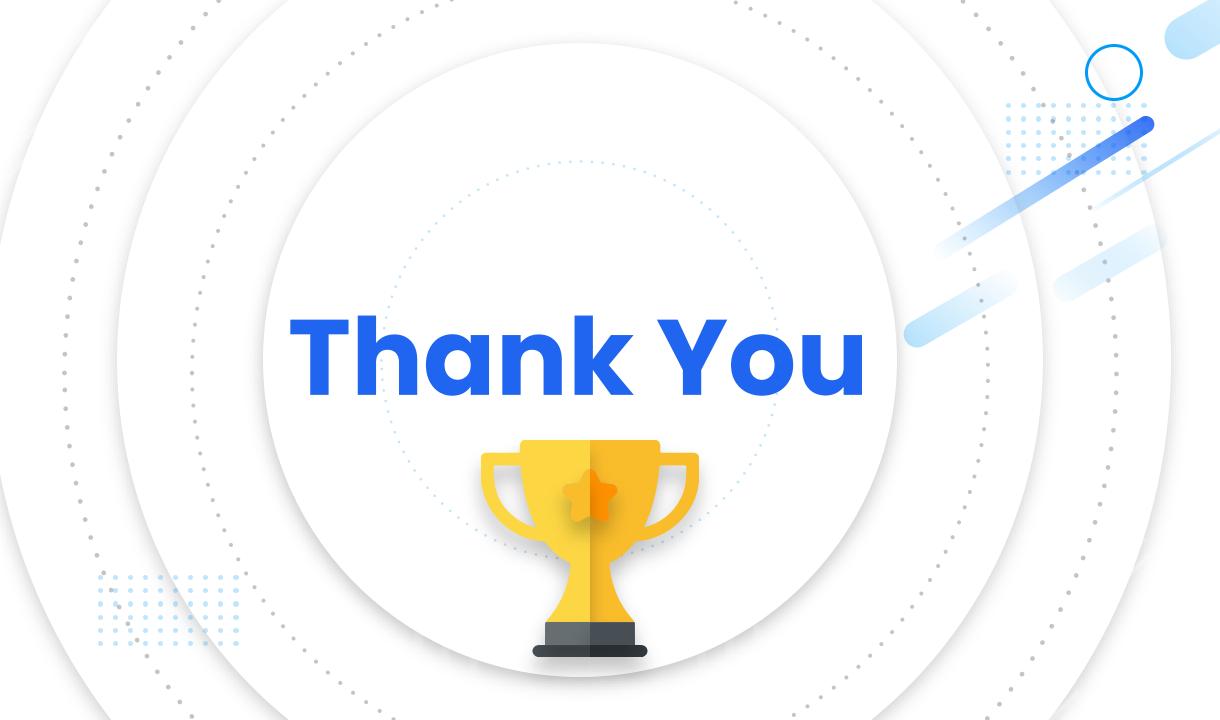
Recommendations

To deal with impact from AI/DS/ML:

- Healthcare firms, especially those with legal access to medical data should embrace AI transformation and invest in AI-assisted solutions for disease diagnosis
- Finance firms can hold off on the integration of AI technology until it becomes mature and properly regulated.
- People engaged in jobs that heavily impacted by AI should learn and adapt to the advancements in AI technology.

To accelerate the development of the transformative capabilities:

- Companies should strive to obtain sufficiently representative data while refraining from prematurely launching immature products.
- Academic institutions should keep intensifying research into foundational technologies.
- Governments should promptly establish relevant regulations to safeguard the data privacy and security of the public.

Thank You

Appendix -- Model Selection

Topic Modeling: LDA with hyperparameter tuning to find optimal n_topics, then use Ktrain to detect some clear topics, and finally use BerTopic to do twice filtering. All three models work together to give topics including industries, technologies and applications, and concerns

Sentiment Analysis:
- regular Vader / Customized Vader: by nature works for short texts (sentence level) but bad for long text (news level)
- SVM/LogReg fined tuned with Yelp: Yelp data (subjective and colloquial) is essentially different from news (objective and formal). So they work poorly
- BertClassifier fined tuned with Yelp: better than above two but still not suitable
- RoBERTA: best

NER: use spacy NER and customized Vader together to do target sentiment analysis since customized vader performs not bad and much faster than RoBERTA

Appendix – News Source Analysis

domain	positive_ratio	neutral_ratio	negative_ratio	count
www.businesswire.com	0.952618	0.047382	0.000000	401
itbusinessnet.com	0.872611	0.117834	0.009554	314
www.01net.it	0.867384	0.132616	0.000000	279
www.topnewsphil.com	0.863248	0.136752	0.000000	117
digitalmedianet.com	0.829060	0.170940	0.000000	117
www.prnewswire.co.uk	0.809609	0.185053	0.005338	562
www.design-reuse.com	0.806667	0.193333	0.000000	150
www.prnewswire.com	0.778869	0.213681	0.007450	2953
www.newswire.ca	0.776435	0.220544	0.003021	331
investorplace.com	0.761111	0.116667	0.122222	180

The news source has a little bit bias such as
businesswire.com.
It's not normal for a news source to have articles
whose positive sentiment ratio is over 95%