Unveiling the AI Frontier

Insights for Predicting Industry Disruption and Job Transformation

MSCA 32018 Natural Language Processing and Cognitive Computing
Final Project

SHIJIA HUANG

Agenda

Executive Summary
High-level summary of the background and project objectives

Methodology & Source Data Overview
NLP techniques utilized and overview of the data source

Text Cleaning & Article Filtering
Data preprocessing, cleaning, and filtering relevant articles

Topic Detection
Identify major topics discussed around AI in the news corpus

Sentiment Analysis
Explicitly customized sentiment analysis to identity AI success/fail reasons

Entity Identification
Identify major AI solutions, companies, people, and locations in the news articles

Timeline Analysis
Timelines of sentiment changing and introduction of new AI technologies

Targeted Entity Sentiment Identification
Targeted sentiment analysis on major organizations and AI initiatives

Conclusions & Recommendations
Analysis summary and making actionable recommendations

Executive Summary

Background and Context

- By 2025, the global AI market is projected to reach a value of $126 billion, with a compound annual growth rate (CAGR) of 37.5% from 2021 to 2025. The impact of AI growth on job displacement and transformation is now a topic of significant discussion.
- Goldman Sachs' report and Facebook's research on Moravec's paradox predict that tasks requiring sensorimotor skills are less likely to be replaced by AI compared to those involving abstract thoughts or reasoning.
- However, these statements were strongly impacted by recent advances in Large Language Models (LLMs) like GPT-3, automating tasks that once required complex human thinking.

Problems and Questions

- Given the rapid growth of AI capabilities, it is uncertain whether job roles heavily reliant on sensorimotor skills like construction and installation will remain unaffected. Furthermore, the extent to which AI can replicate human reasoning remains questionable.
- Which industries and job roles will experience the greatest impact from AI in the coming years? How can AI be utilized to automate jobs and enhance employee productivity? How can organizations effectively leverage AI to adapt to the changing job landscape and maximize its benefits?

Resolution and Next Steps

- To answer those questions, I will analyze and extract meaningful insights from a large collection of news articles based on major topics, sentiments, entities, and timelines to discover opportunities for job automation and enhanced productivity.

Methodology & Source Data Overview

Processing Large Text Corpus with GCP Vertex AI

Large Volume of Unstructured Data

- Around 200K raw news articles were collected by web crawling online sources, covering topics such as data science, AI, and machine learning
- The data included the source URL, publish date, language, title, and text for each article
- The texts exhibit diverse sources, styles, and perspectives, resulting in variations in language, tone, and quality
- Not all articles directly align with the specific topic of interest (i.e. artificial intelligence in industries)
- The presence of noise, including web crawl remnants, links, abbreviations, and errors, adds complexity to the text

Methodology to Extract Insights

- **Topic Detection**
  - Utilized Latent Dirichlet Allocation (LDA) modeling with Gensim on lemmatized news text tokens and n-grams with hyperparameter tuning the number of topics to identify major topics discussed in the news corpus
  - Applied Zero-Shot (NLI) Modelling to positive and negative sentiment news articles separately using candidate labels from tuned LDA topics to identify top reasons for successful and failing data science and AI initiatives
- **Sentiment Analysis**
  - Pre-trained customized SVM model on open-source AI news perception data with sentiment labels (1-5) to predict the AI sentiment of each news article
  - Applied TF-IDF vectorizer to the open-source texts and SMOTE oversampling on imbalanced sentiment labels to increase the model accuracy
- **Entity Identification**
  - Processed news text with sentence segmentation using spaCy pipelines with multiprocessing to conduct Named-Entity Recognition (NER)
  - Manually cleaned the NER results to find the most frequently appeared AI technologies, organizations, people, and locations in the new articles
- **Timeline Analysis**
  - Plot sentiment-based article counts and average sentiment over time
  - Identify the date with the highest article count for each AI technology from NER to pinpoint when its introduction influenced the data science landscape
- **Targeted Entity Sentiment Identification**
  - Assigned the sentence sentiment to each AI and organization entity in the text during NER and calculated the average sentiment of each entity

Text Cleaning

Data Preprocessing

- Read Parquet file into Pandas dataframe and assigned an id for each news article

Clean News Titles

- Identified and eliminated news provider names at the end of news titles by matching splitting characters like '|' and '-'

Clean News Texts

- Finding main sentences (*discard web crawl remnants*):
  - Split text into sentences by newlines, tabs, or 3+ spaces
  - Then remove sentences with less than 10 or no words except for the cleaned title
- Finding main paragraphs (*discard irrelevant text*):
  - Divide main sentences into sections by the cleaned title
  - Then select the text section with the longest length as the main paragraph of the news article since the main text is usually followed by the cleaned title
  - Add the cleaned title back to the start of the main text
  - Dropped duplicate sentences which usually be Ads
- Removed links and special characters from the text

Article Filtering

Preparation: Text Normalization

- Tokenize and remove stop words from the cleaned news titles and texts with Genism
- Make bigrams and trigrams from the tokenized text with Genism and combine tokens and n-grams into a list
- Lemmatize the tokens and n-grams and keep only nouns, adjectives, verbs, adverbs, and proper nouns using the spaCy language processing pipeline

News Article Filtering with TF-IDF Keywords

- Extract the top 10 keywords from each normalized news title and text by applying the TF-IDF vectorizer
- Combine keywords from all news articles and identify the top 30 most frequent keywords for both titles and texts
- Manually select a subset of keywords related to artificial intelligence and industries
- Discard articles without any top 10 keywords in the subset
- Article Filtering Results: 200,332 -> 154,283 (~77%)

Topic Detection

Latent Dirichlet Allocation (LDA) Modelling with Hyperparameter Tuning

Best LDA Model

- Number of Topics: 18
- Coherence Score: 0.41

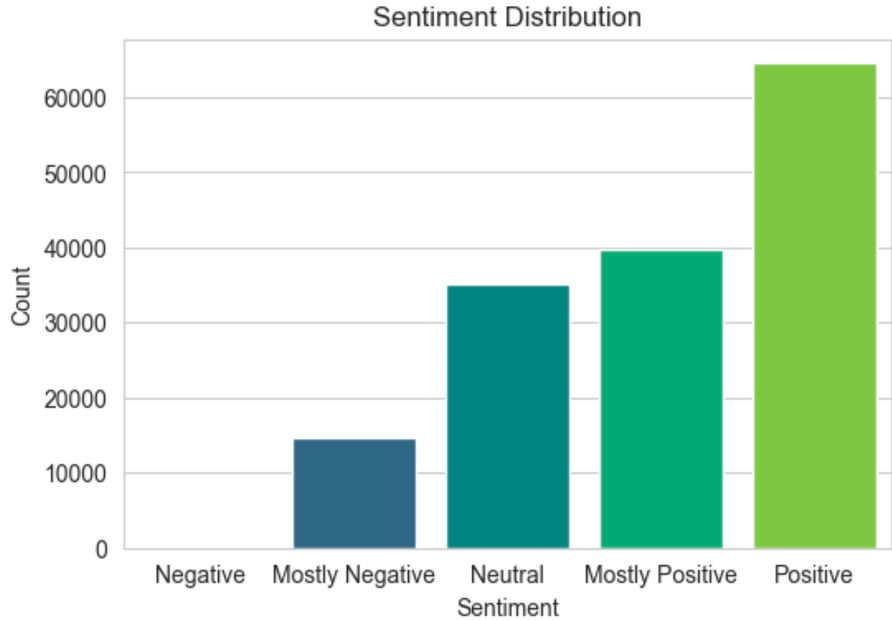
Top 5 Topics and Component Percentage

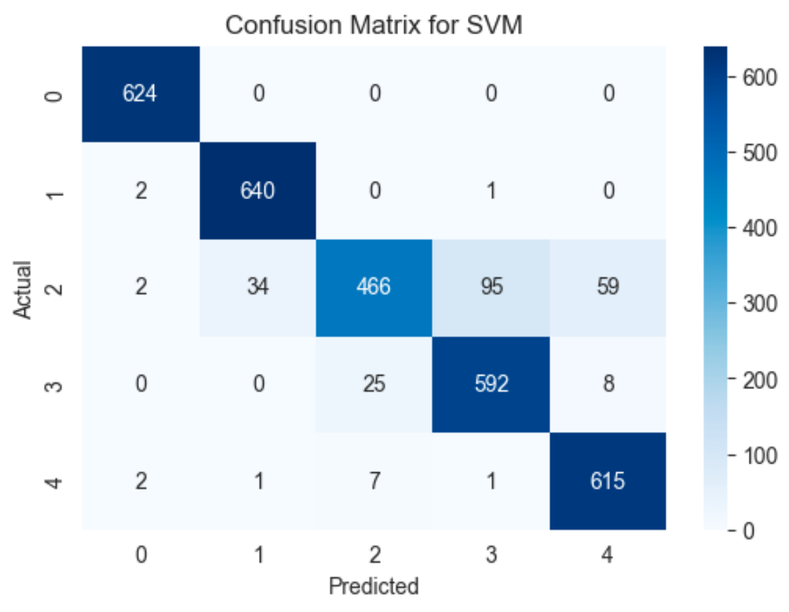
1. The global AI industry experiences growth, and research provides key insights on trends, players, and applications, driving development opportunities (15.7%)
2. AI improves systems, helps people, and revolutionizes work and interactions with systems (13.8%)
3. ChatGPT, an AI chatbot, uses new technology to assist people with work, answering questions, providing text help, and analyzing images (13.8%)
4. AI technology powers the global business with intelligent solutions, enhancing capabilities, support, and industry leadership (9.8%)
5. AI and machine learning models use diverse data, including images, to develop new applications and involve researchers in the training process (6.6%)

Sentiment Analysis

Customized SVM Model Pre-trained on Open Source AI News with Sentiment Labels (1 - 5)

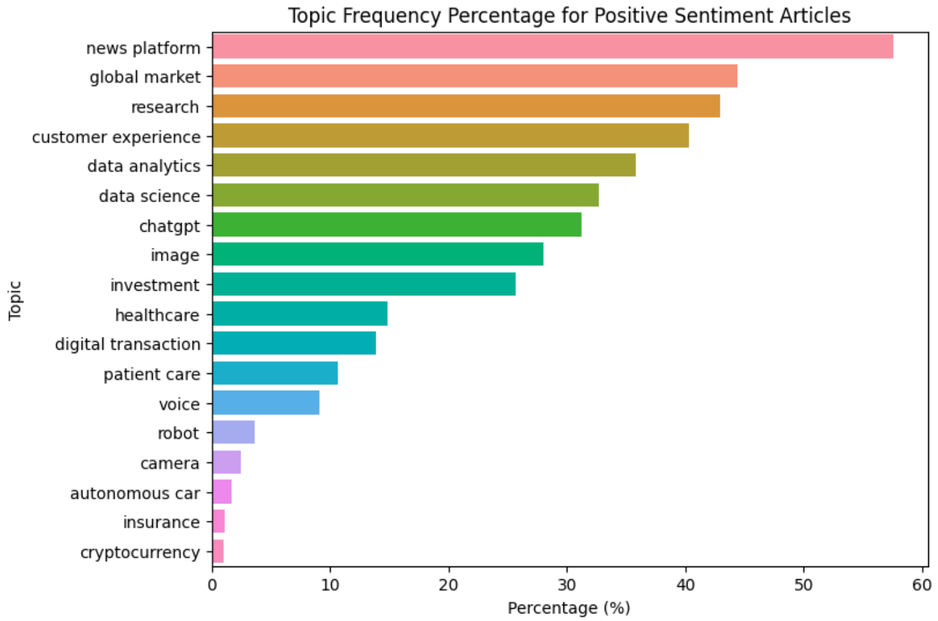
- Test Accuracy on predicting AI sentiment of open source data: 93%
- Precision & Recall: 0.93
- Accuracy on 50 hand-labeled news articles' sentiments: 87%

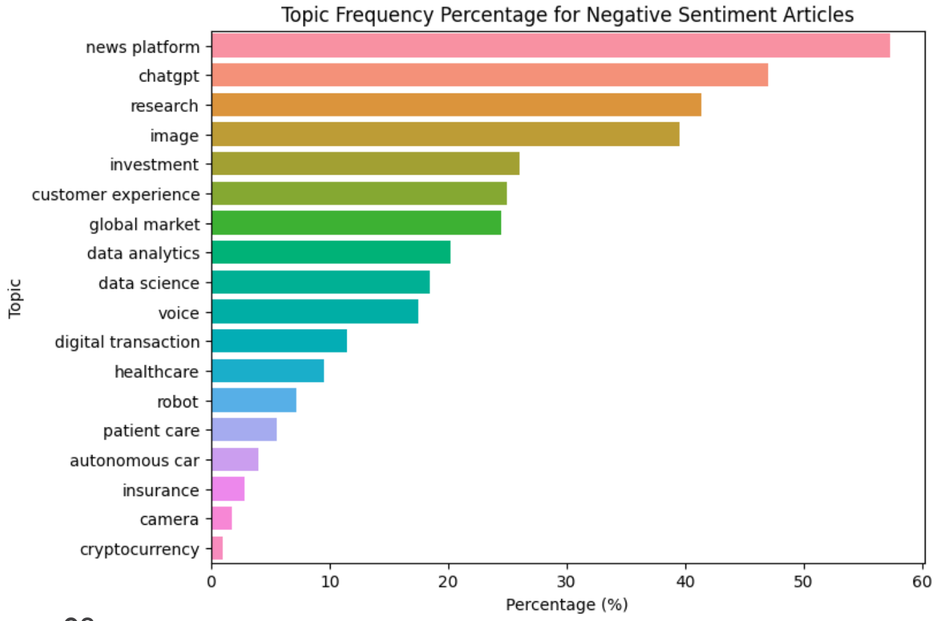




The Majority of Identified Sentiments Toward AI tend to be Positive

- Positive sentiment was strongly prevalent among the respondents, indicating a predominantly favorable view
- A notable proportion of respondents held a neutral stance, showing neither a strong positive nor negative sentiment toward AI
- While negative sentiment existed, it was relatively less pronounced compared to the prevailing positive and neutral sentiments





Topic Detection

Sentiment-Based Zero-Shot (NLI) Modelling Using Candidate Labels from 18 LDA Topics

Promising AI Initiatives

- Global market potential and growth of AI technology innovations
- Advancements and breakthroughs in AI research and algorithms
- Enhanced customer experience through personalized interactions and tailored recommendations
- Utilization of data science and analytics for better decision-making

Suspectable AI Practices

- Concerns and limitations related to AI language models like ChatGPT such as biased outputs and ethical considerations
- Controversial aspects and risks associated with AI research
- Criticism of AI Initiatives' impact, investment outcomes, and negative customer experiences

Entity Identification

AI Technologies and Solutions

- ChatGPT, Bing, and Generated AI are prominent AI solutions, advancing natural language processing, search engines, and generated content
- TensorFlow, Cloud Platforms, and Machine Learning played key roles in advancing efficient model development and deployment

People

- Tech company CEOs frequently mentioned in AI news showcase their influential role in shaping the AI landscape and driving industry-wide innovation
- The frequent mention of US presidents highlights the significant impact of AI on government policies, national strategies, and societal implications

Locations

- The United States, China, and India are prominent players in the field of AI

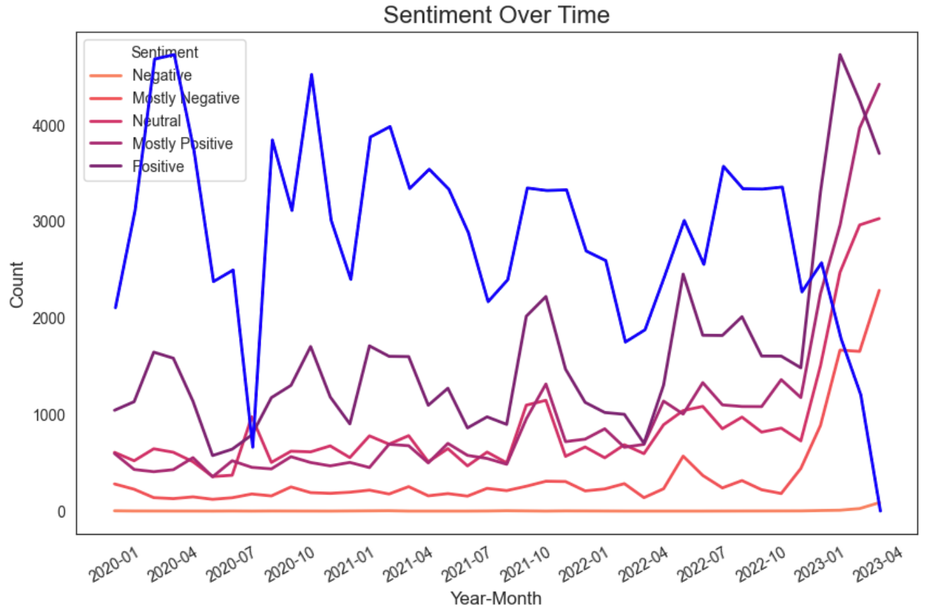
Organizations

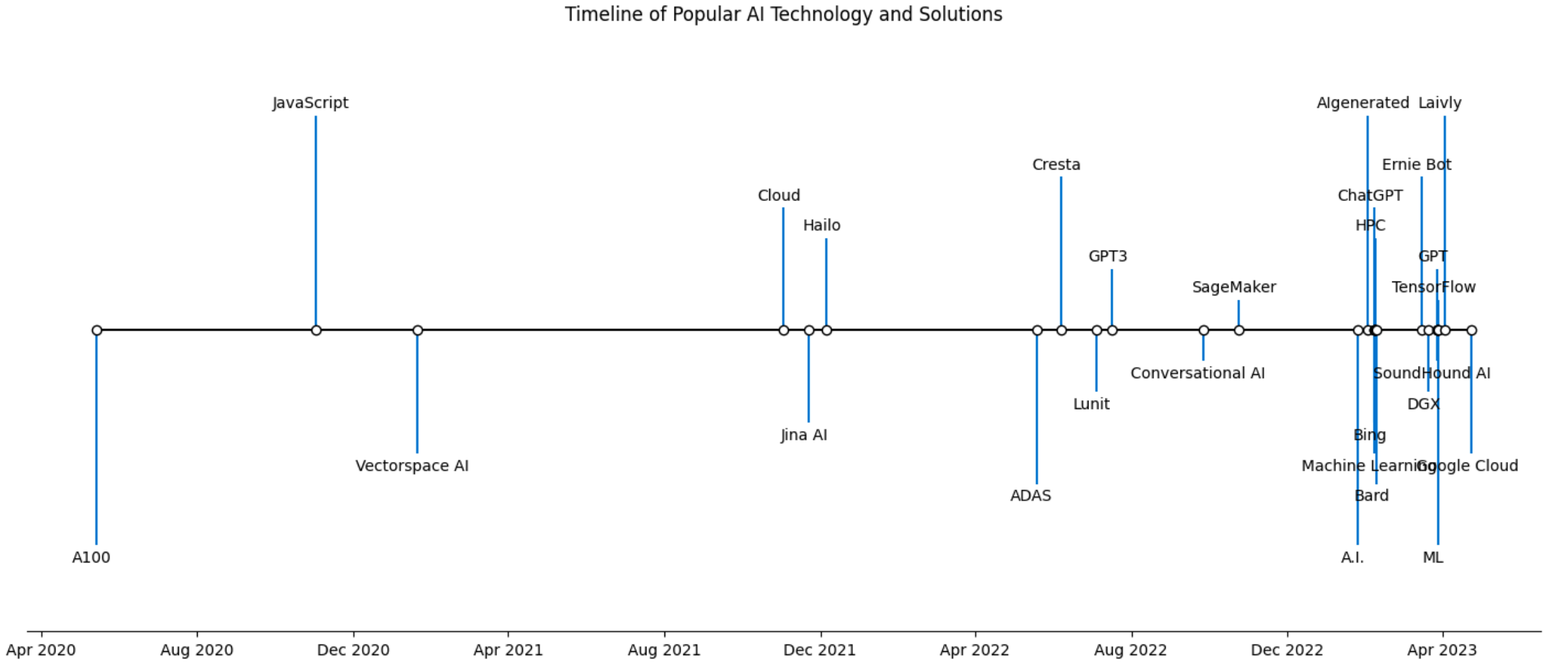
- Companies: Google, Microsoft, OpenAI, and many tech companies are leading the advancement of AI applications
- Academic Institutions: Joint Artificial Intelligence Center (JAIC) can accelerate transformative AI development through coordination, research, and ethical guidelines
- Governments: European Union (EU) can advance AI through regulatory frameworks, research and innovation, data sharing, education and skills support, and international collaboration

Timeline Analysis

Surging AI Discussions and Declining Average Sentiment

- AI discussions have surged since January 2023, reflecting heightened interest, awareness, and an evolving technological landscape
- The average sentiment (blue line) towards AI has declined since January 2023 which might be caused by various factors, such as ethical concerns, negative media coverage, and a lack of transparency

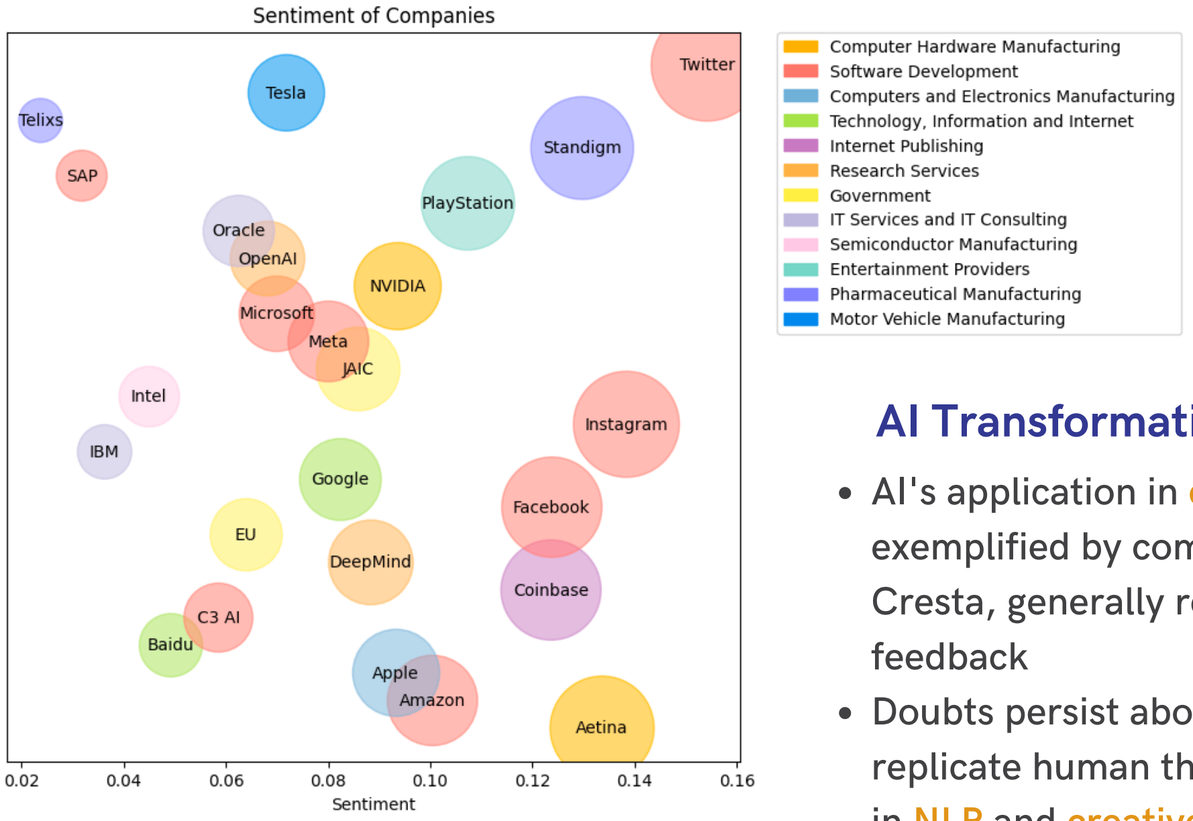




Introduction of New Tech is Reshaping Industries and Data Science Applications

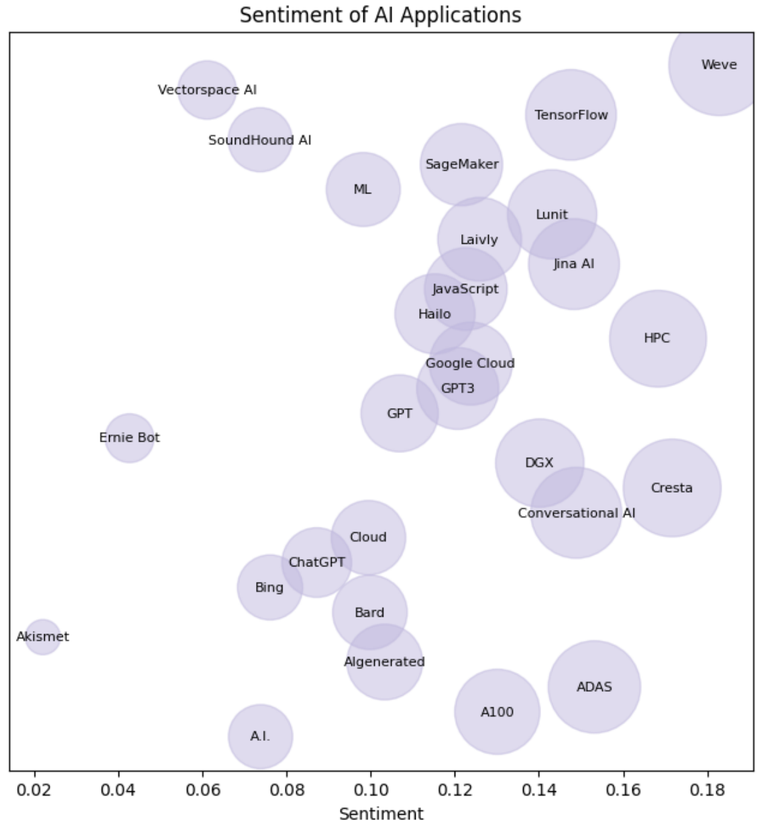
- Abundant new AI technologies and solutions since January 2023 such as GPT and Bing, revolutionizing various industries with advanced applications and frameworks
- Solutions like Sagemaker and Tensorflow transformed data science applications, streamlining development, enabling scalability, and fostering collaboration

Targeted Entity Sentiment Identification



AI Transformation Suggestions

- AI's application in customer service, exemplified by companies like Weve and Cresta, generally receives positive feedback
- Doubts persist about AI's capacity to replicate human thinking and originality in NLP and creative generation, as seen in sentiment towards ChatGPT, Ernie Bot, Vectorspace AI, and similar models



AI Investment Recommendations

- AI investment recommended for software, pharmaceutical, entertainment, and e-commerce companies based on more optimistic sentiments
- Hardware companies may face challenges in AI investment due to complex integration, limited compatibility, and higher costs

Conclusions & Recommendations

Artificial Intelligence brings impactful changes to the industry landscape

The AI industry is experiencing significant growth, driven by research insights on emerging trends, key players, and diverse applications. This growth indicates a strong potential for automation and productivity improvements in various sectors

Positive sentiments prevail among news articles, indicating a generally favorable view of AI's impact on systems, work interactions, and people's lives. The neutral stance reflects a cautious but open attitude toward AI, while negative sentiments are relatively less prominent

Key AI initiatives, such as ChatGPT and AI-generated content, demonstrate the potential to revolutionize work tasks, assist with inquiries, and analyze data. However, there are concerns surrounding biased outputs, ethical considerations, and the need for researcher involvement in model training

Foster AI-driven Innovation

emerging explore
innovation development
automation opportunities
solutions research
potential advancements
productivity
industries
technology unlock
applications

Address Ethical Considerations

transparency inclusivity
guidelines fairness compliance
mitigate biases
responsible regulations
trustworthy measures
ethics diversity
discrimination
accountability

Enhance Academia and Industry Collaboration

cooperation breakthroughs exchange
resources innovation partnerships
academic
knowledge industry shared
organizations
collaboration synergy
joint challenges
researchers
advancements
institutions
efforts
research

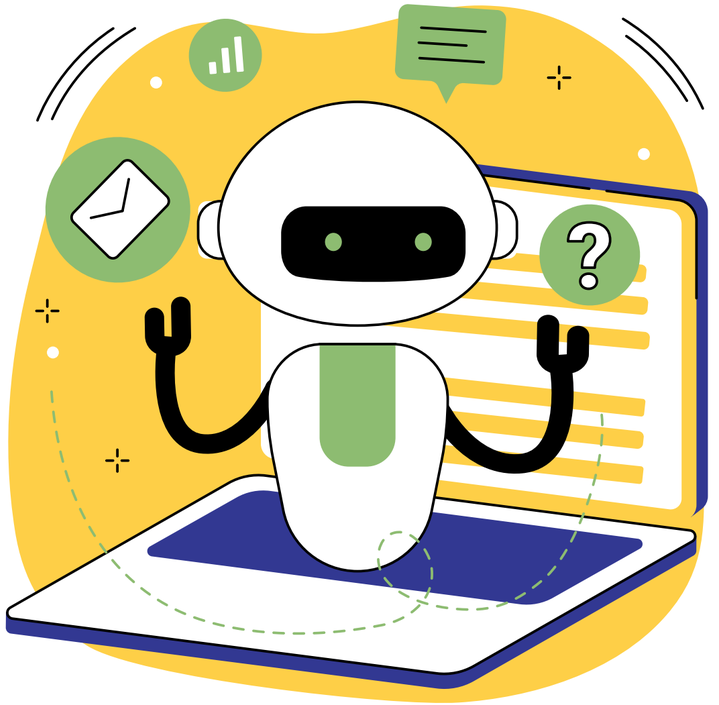
Thank you!

Info

Shijia Huang, University of Chicago

Email

shijia@uchicago.edu



References

1. GitHub Repository: https://github.com/slvhuang/NLP-AI-News-Insights
2. Open Souce Data: https://www.kaggle.com/datasets/saurabhshahane/public-perception-of-ai
3. https://medium.com/grabngoinfo/zero-shot-topic-modeling-with-deep-learning-using-python-a895d2d0c773
4. https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/