Natural Language Processing
&
Cognitive Computing

ADSP 32018 ON01 (Winter 2024)

Final Project Data Science

YuWei, Hsu
Winter, 2024

Introduction

The objective of our final project is to identify what types of tasks and jobs are most likely to see the biggest impact from AI by extracting **meaningful insights** from unstructured text.

Our goal is to provide **actionable recommendations** on what can be done with AI to automate the jobs and/or improve employee productivity.

Table of contents

01

Executive Summary

What news articles tell us?

➤ **The method we used**

In this analysis of articles related to AI innovation, we employed various techniques including N-gram analysis, BERTopic modeling, preselected keywords, customized sentiment analysis, VADER lexicon, named entity recognition (NER), and word cloud visualization.

➤ **Our findings**

Through our analysis, we have successfully identified emerging technologies such as Large Language Models (LLM) and generative AI within the given timeframe. Additionally, we have uncovered distinct trends in the dialogue surrounding AI technology. Our findings indicate that sentiments towards AI innovation vary significantly across different industries. Specifically, technology-related sectors express predominantly positive sentiments, crediting AI integration for driving development and innovation. In contrast, industries such as government, education, and media demonstrate a more negative sentiment towards AI, primarily due to ethical concerns and potential societal implications associated with its adoption.

➤ **Actionable recommendations**

We recommend that individuals across all industries leverage AI to enhance productivity, while also emphasizing the importance of cultivating soft skills such as cooperation, emotional intelligence, critical thinking, and ethical reasoning. These skills are highlighted in the articles as essential attributes that complement AI technology and contribute to individual and organizational success.
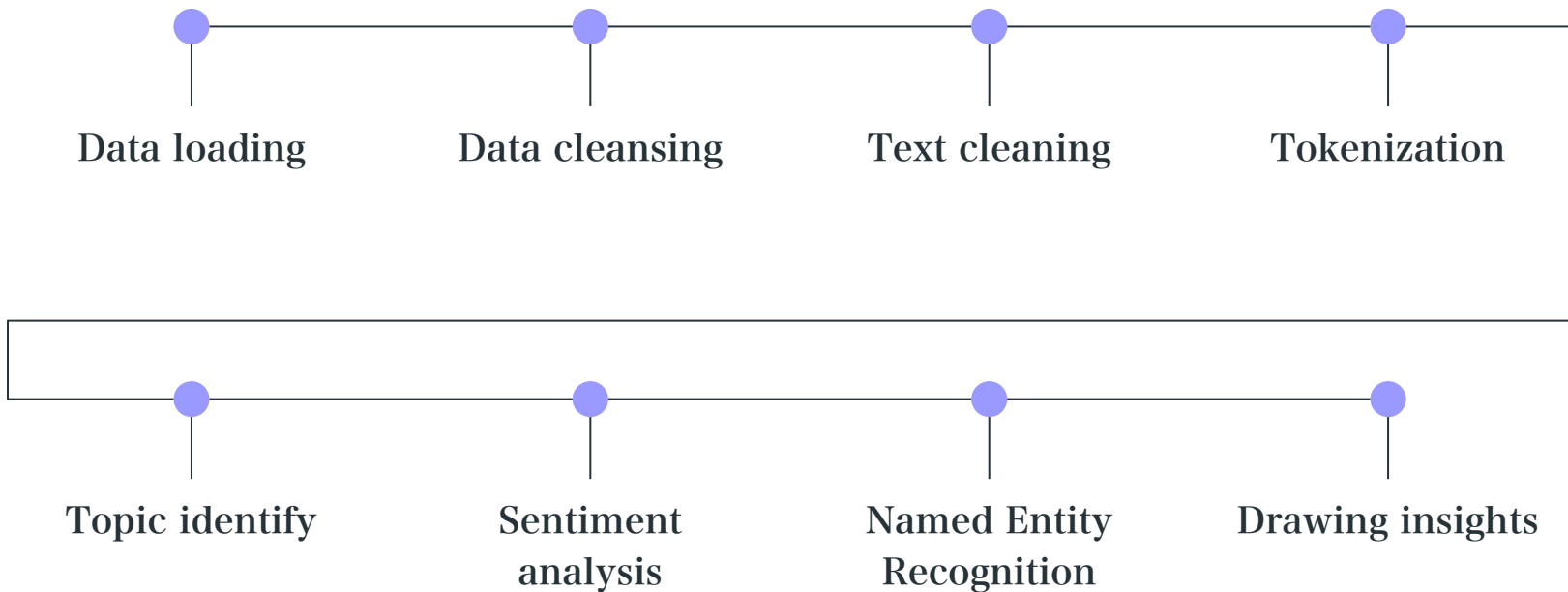
02

Methodology & insights

Analysis workflow

Data loading   Data cleansing   Text cleaning   Tokenization

Topic identify   Sentiment analysis   Named Entity Recognition   Drawing insights

We used BERTopic and pre selected key words to cluster articles to the domains they land in

We first used BERTopic to find out what the potential topics are among the articles, and we used keywords in the topics to cluster articles. For example, the keywords "music", "song" could be in the domain of entertainment.

Intertopic Distance Map

D2

Topic 37
music | song | oneplus | artists | songs
Size: 120

D1
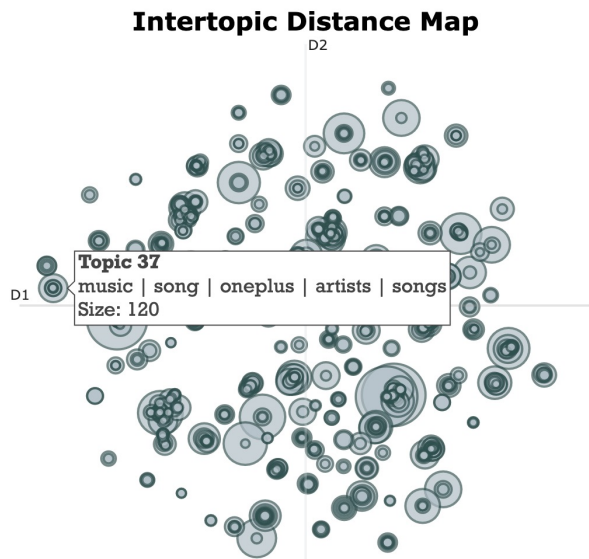
We finally clustered the articles to the specific domain areas, the domain areas are listed as below:

```
categories_with_keywords = {
    "Finance": ['stock', 'stocks', 'investors', 'finance', 'etf', 'shares',
                'banking', 'economy', 'trading', 'investment', 'portfolio'],
    "HealthCare": ['healthcare', 'clinical', 'medical', 'patient', 'care',
                'hospital', 'cancer', 'diabetes', 'pharmaceutical', 'treatment', 'health', 'medicine'],
    "Technology": ['technology', 'tech', 'software', 'hardware', 'internet',
                'innovation', 'startup', 'IT', 'computing', 'digital', 'cyber'],
    "Entertainment": ['music', 'spotify', 'sony', 'song', 'dj', 'artist',
                'radio', 'listening', 'film', 'movie', 'cinema', 'entertainment'],
    "Retail": ['retail', 'retailers', 'stores', 'shopping', 'costco', 'ecommerce',
                'online shopping', 'brick-and-mortar', 'mall', 'consumer', 'sales'],
    "Automotive": ['cars', 'traffic', 'vehicles', 'road', 'bikes', 'autonomous',
                'transport', 'automobile', 'car manufacturer', 'electric vehicle', 'EV'],
    "Education": ['students', 'teachers', 'education', 'learning', 'teaching',
                'assignments', 'professor', 'school', 'university', 'classroom', 'curriculum'],
    "Manufacturing": ['manufacturing', 'production', 'manufacturers', 'industrial',
                'maintenance', 'factory', 'assembly line', 'machinery', 'engineering'],
    "Supply Chain and Logistics": ['chain', 'supply', 'logistics', 'distribution',
                'warehouse', 'shipping', 'freight', 'inventory', 'fulfillment'],
    "Energy": ['oil', 'gas', 'energy', 'solar', 'renewable energy',
                'power', 'electricity', 'nuclear', 'utility', 'fuel'],
    "Media": ['news', 'media', 'television', 'journalism', 'broadcast', 'reporting',
                'press', 'publication', 'newspaper', 'journalist'],
    "Research and Development": ['research', 'researchers', 'researcher', 'chemistry',
                'chemist', 'science', 'scientist', 'lab', 'experiment', 'discovery'],
    "Food and Beverage": ['food', 'beverage', 'restaurant', 'cuisine', 'dining',
                'food industry', 'hospitality', 'culinary', 'menu', 'gastronomy'],
    "Real Estate": ['real estate', 'property', 'housing', 'housing market', 'realtor',
                'mortgage', 'homeownership', 'commercial real estate', 'residential'],
    "Travel and Tourism": ['travel', 'tourism', 'destination', 'hotel', 'airline',
                'vacation', 'holiday', 'travel agency', 'hospitality industry', 'tourist'],
    "Government and Public Sector": ['government', 'public sector', 'policy', 'politics',
                'public service', 'administration', 'legislation', 'government agency'],
    "Environmental": ['environment', 'climate', 'sustainability', 'green', 'ecology',
                'conservation', 'environmentalism', 'carbon footprint', 'renewable'],
    "Nonprofit and Social Services": ['nonprofit', 'charity', 'social services',
                'community', 'volunteer', 'philanthropy', 'NGO', 'non-governmental organization']
}
```

We used unique Bigram and Trigram to find out what new techologies appeared in the articles

With the output of unique Bigram and Trigram, we found out there were some technologies mentioned in the articles, the results are as below

```
tech_words = [
    "artificial intelligence",
    "machine learning",
    "generative ai",
    "deep learning",
    "conversational ai",
    "big data",
    "ai chatbot",
    "language model",
    "language processing",
    "digital transformation",
    "ai applications",
    "facial recognition",
    "auto transportation",
    "communications cloud",
    "swot analysis",
    "wireless communications",
    "supply chain logistics",
    "large language models",
    "natural language processing",
    "amazon web services",
    "generative artificial intelligence",
    "ai drug discovery"
]
```

And we traced back to the date when the technologies were first mentioned in the articles, we found:

"large language models (LLM)" & "generative AI"

were relatively innovative technologies in the data science domain

Timeline of First Mention of Technologies
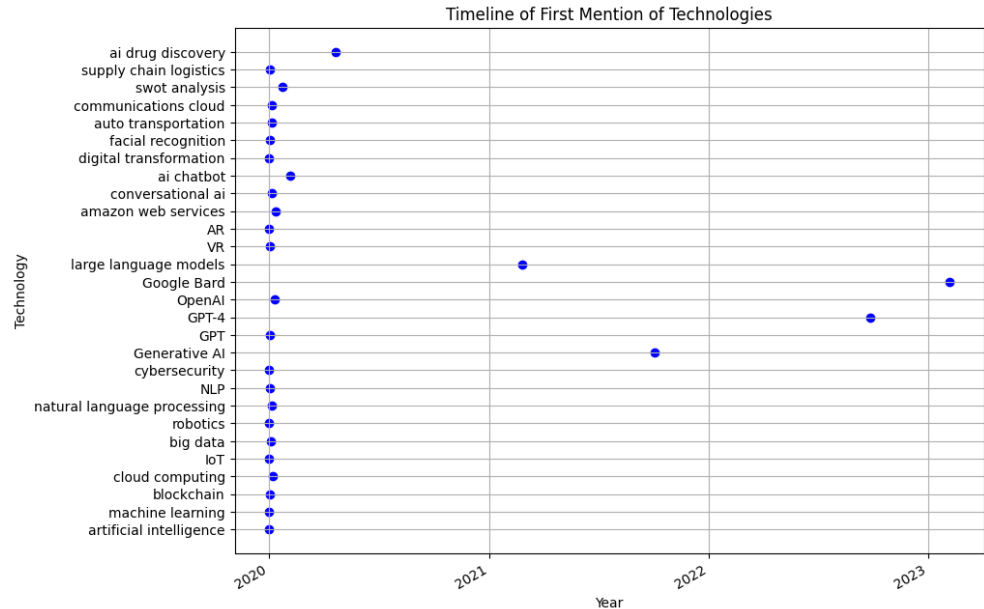
We used both customized sentiment and vader_lexicon for tagging articles' sentiments to find out sentiment trend

We used combined methods to filter articles to make sure our keywords function aligns with VADER sentiment's function. For example, if our keyword's function tagged the article as Positive but VADER showed as Negative, we would filter out this kind of mismatched data.

```
positive_keywords = ['innovative', 'breakthrough', 'revolutionary', 'advancement', 'improve', 'enhance', 'boost',
                     'innovate', 'growth', 'success', 'advance', 'benefit', 'positive', 'increase', 'empower',
                     'create', 'potential', 'effective']
negative_keywords = ['concern', 'challenge', 'disruption', 'replace', 'challenges', 'concerns',
                     'decline', 'fall', 'negative', 'loss', 'problem', 'restrict', 'destroy', 'sue', 'scramble','block', 'fight', 'against', 'issue']
neutral_keywords = ['study', 'research', 'trend', 'analysis', 'survey', 'report']
```

The surge in AI-related topics, especially noticeable post-December 2022, signifies the introduction of revolutionary innovations. This observation aligns with our findings indicating the emergence of large language models (LLMs) and generative AI around this timeframe. While a majority of the population holds a positive outlook on AI's impact, it's noteworthy that up to one-third of the population harbors suspicions and expresses negative opinions.

Sentiment Trend (Monthly)
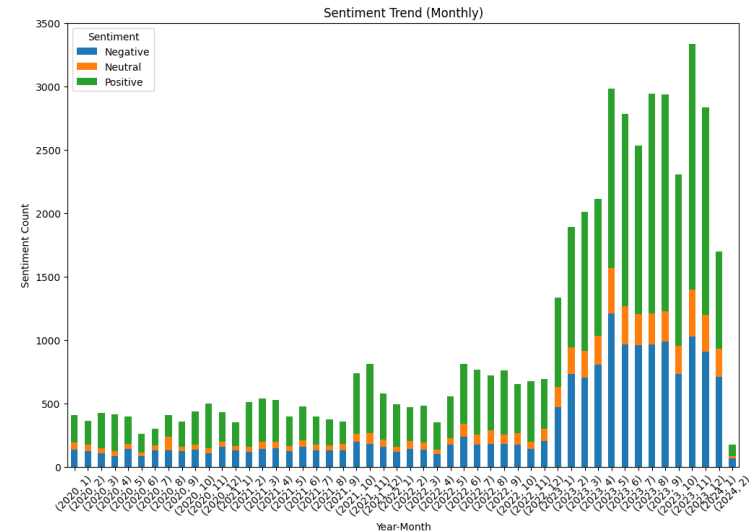
There is substantial discourse surrounding AI technologies across a spectrum of industries including Technology, Media, Finance, Healthcare, Education, and Research and Development.

However, sentiment analysis unveils contrasting attitudes towards AI adoption. Industries like Technology, Healthcare, and Research and Development, which stand to benefit from AI algorithms, image generation, and data analytics, demonstrate a favorable disposition towards integrating AI technologies into their operations.

Conversely, sectors such as Education, Media, and Finance, which harbor concerns about being replaced by large language models (LLM), tend to harbor more negative sentiments regarding AI. This dichotomy highlights the divergent perceptions and readiness levels across industries regarding the adoption of AI solutions.

Industries that have most discourse of AI technologies

Category	custom_sentiment	count	total
Technology	Positive	6679	11044
Technology	Neutral	1088	11044
Technology	Negative	3277	11044
Media	Neutral	1056	9728
Media	Negative	3511	9728
Media	Positive	5161	9728
Finance	Positive	2561	4886
Finance	Neutral	612	4886
Finance	Negative	1713	4886
HealthCare	Negative	1080	4746
HealthCare	Positive	3340	4746
HealthCare	Neutral	326	4746
Education	Positive	1860	3607
Education	Neutral	313	3607
Education	Negative	1434	3607
Research and Development	Positive	2161	3409
Research and Development	Neutral	313	3409
Research and Development	Negative	935	3409

Negative sentiment proportion in each industries

Category	
Government and Public Sector	0.581312
Environmental	0.456929
Entertainment	0.454515
Travel and Tourism	0.448276
Education	0.397560
Media	0.360917
Finance	0.350594
Real Estate	0.343220
Energy	0.340952
Automotive	0.322063
Nonprofit and Social Services	0.313305
Technology	0.296722
Food and Beverage	0.289548
Research and Development	0.274274
Supply Chain and Logistics	0.248252
HealthCare	0.227560
Retail	0.218563
Manufacturing	0.133140

We used NER technique to find out entities laying in the articles

Organization	Location	Product	People
•Microsoft	•Chine	•GPT-4	•Sam Altman
•Gray Media Group, Inc.	•India	•DeppMind	•Joe Biden
•Google	•US	•Windows	•Elon Musk
•PR Newswire	•UK	•Excel	•Donald Trump
•Amazon	•Russia	•Llama2	
•Associated Press	•Japan	•Google Cloud	
•IBM			
•Apple			
•Intel			
•Meta			
•Nvidia			

Organizations such as Microsoft, Google, Nvidia, and Amazon, alongside countries like China, India, and the US, as well as influential figures in the government sector such as Joe Biden and Elon Musk, are prominently highlighted in numerous articles. This underscores their pivotal role in driving the advancement of transformative AI capabilities. Their contributions extend across various domains, including legislation, regulation of AI usage, and investment in enhancing AI-related development such as wafer processing, cloud computing, GPU, and CPU technologies. These entities are widely recognized as key stakeholders shaping the future landscape of AI technology and its applications.

After pairing organization entities with sentiment analysis, we discovered that companies such as Gray Media Group, Inc., PR Newswire, Microsoft, Google, and Nvidia exhibit a positive attitude towards new AI inventions. This observation suggests that the technology and media industries are particularly enthusiastic about investing in, introducing, or integrating these AI technologies.

Applications related to people culture, heavy industry, entertainment media, journalism, and public policy face challenges in being fully transformed by AI due to their complexity, subjectivity, reliance on human interaction and emotion, need for creativity and innovation, ethical considerations, and context sensitivity. While AI can assist with certain tasks in these domains, it is unlikely to fully replace human expertise, creativity, and judgment in the near future.

03

Actionable
Recommendations

Recommendation derived from articles and our data analysis

From the data presented, it is evident that different industries exhibit varying attitudes towards AI technologies. In technology-related sectors, such as IT and innovation-driven industries, there is a notable emphasis on investing in AI research and innovation to accelerate development. However, in human-centric industries like government, education, and media, there is a prevalent sentiment of negativity and suspicion towards AI, with concerns about potential job displacement and the fear of being replaced by AI-driven systems. While AI can certainly assist with certain tasks in these domains, there remain essential skill sets that AI currently cannot replicate, including decision-making, creativity, problem-solving, and emotional intelligence. Therefore, it is imperative for individuals in our generation to cultivate essential soft skills such as cooperation, emotional learning, critical thinking, and ethical reasoning to complement AI technologies. While the articles demonstrate how AI can enhance productivity, it is undeniable that human qualities and capabilities distinguish us from machines. Thus, while we cannot avoid the integration of AI into our lives, it is our uniquely human traits that define our value and significance in society.