MLTSA midterm: project proposal

AI's Changing Impact on Society, Education, and the Workplace

Andrew Kallai, andrewkallai, andrewka@udel.edu, Data/Literature Manager

Talha Mahmood, talhaMah56, talha@udel.edu, Communication/Visualization Manager

Abstract

This study proposes the examination of the impact of Artificial Intelligence (AI) on careers, education, and society by analyzing temporal trends in article titles. Using NLP techniques, including n-gram frequency and BERTopic classification [2], textual data is converted into numerical features for time series modeling. ARIMA and Facebook's Prophet models are employed to predict trends, addressing data variability and missing values. Gaussian processes are used to identify the classifiers clusterings for computed embeddings. This systematic approach enables structured tracking of AI's transformative role in key life domains.

Introduction

Artificial Intelligence (AI) is a topic that has touched almost all aspects of life in recent times. AI is present in the workplace, in educational settings, and even in conventional parts of society. As such, a question to poise from this would be which aspects of life have been affected the most? This question can be answered more simply by defining three categories: careers, education, and society. Analyzing the content of the article titles for keywords which fit into these respective categories will allow the articles to be classified effectively. The integration of the time series aspect of the data will show how classification changes over time. That is, how the content in article titles changes with respect to time. Because time series analysis methods require numerical data for exogenous variables, it is necessary to extract numerical features from the data. This can be done using NLP techniques and concepts. Take for example, counting the frequency of sentences of n size (n-grams) and then counting the number of n-grams which are classified into a respective category. A more advanced approach would be to classify titles into topics using a BERT-based model such as BERTopic and then predicting the trend of these classifications using ML techniques. Ultimately, ARIMA and Facebook's Prophet (FP) modeling are the two initial choices for predicting the trends of classification. ARIMA provides for effective handling of variability and identifying trends. The FP model is better suited to handle missing data which will be a problem when merging the datasets with different sampling intervals. Furthermore, by using topic classification techniques on the titles, three topic categories can be inferred. The fitting of various articles into these categories can be compared overtime using the autoregressive techniques and analysis of seasonal variations. Moreover, the continued prediction of the trends for AI can be validated by using Gaussian processes to check the kernel density of the classifiers used [1]. More specifically, the textual embeddings (numeric representation of the text) can be compared with the frequency distribution of categorical classification to determine the classifiers sensitivity [3]. This can also be used to evaluate the trend prediction for the ARIMA and FP models [4], [5]. Thus, machine learning can be used in a novel approach to identify the context to which AI affects daily life the most.

Data

The primary dataset used is MIT AI News Published till 2023. Using Python, the data was scraped from MIT news, which has published about technologies since the 1990s. This dataset contains News specifically about the topic "AI". The dataset contains more than 1000 news articles from the years 1994 to 2023. The data is hosted on Kaggle.com, with the owner attributing ownership of the content to MIT. No license is specified, but the data is provided for "educational purposes".

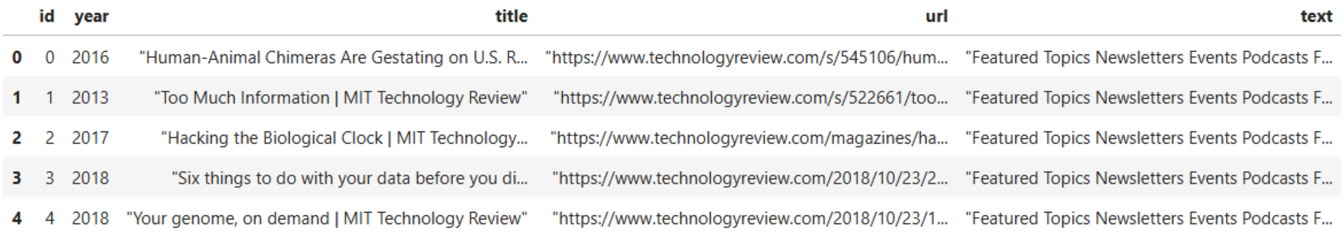
The secondary dataset is a collection of AI/tech articles scraped from the web. The data is provided with an MIT license. More data is present in this set from 2000-2023, however the data contains fewer sampling intervals.

Dataset name	**URL**	**Number of rows**	**Number of columns**	**Number of relevant columns**	**Number of valid rows (not NaN on relevant columns)**	**Data type for each relevant column**
ai-tech-articles (train-00000-of-00001-4f269291413b781f.parquet)	https://huggingface.co/datasets/siavava/ai-tech-articles	**17,092**	**5**	**2**	**17,092**	**int64, string**
MIT AI News Published till 2023 (articles.csv)	https://www.kaggle.com/datasets/deepanshudalal09/mit-ai-news-published-till-2023/data	**1,018**	**8**	**2**	**1,018**	**DateTime, string**

ai-tech-articles

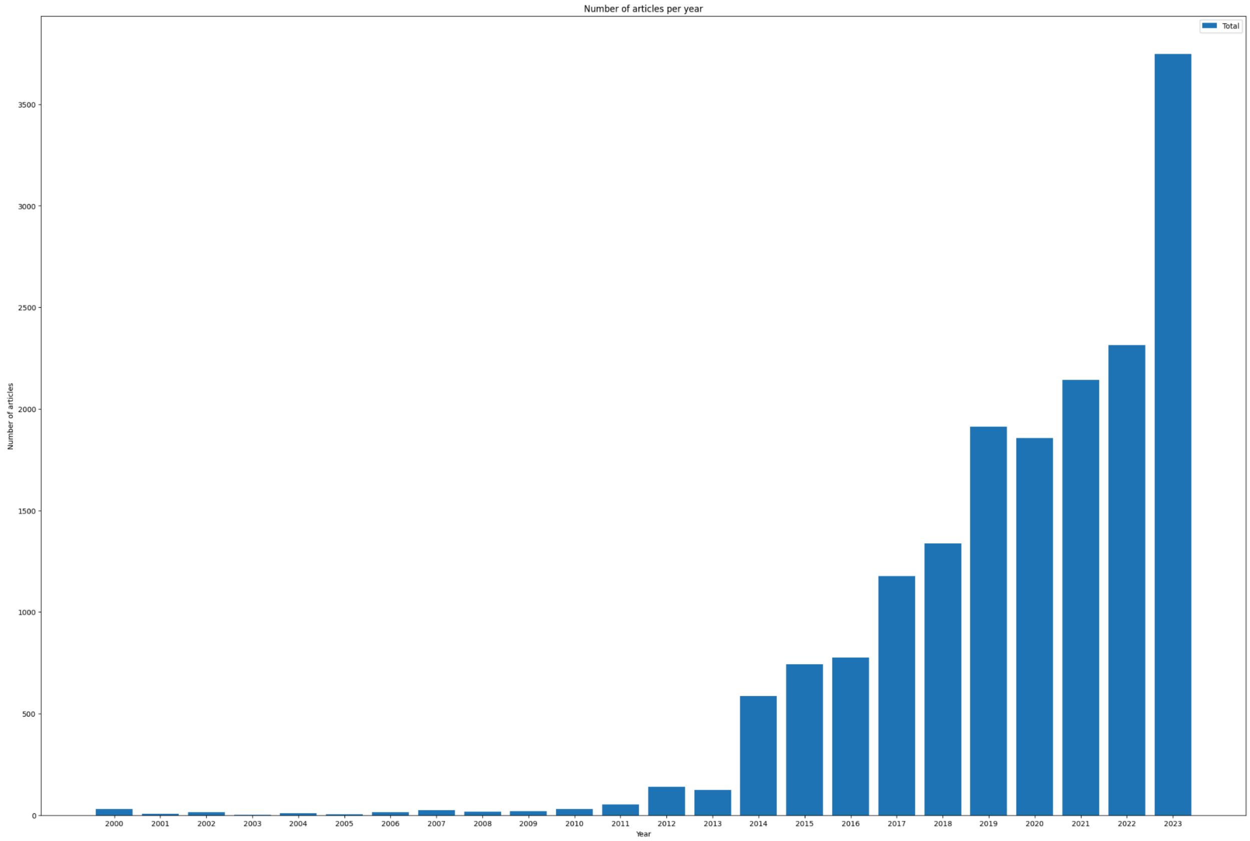
df.head()

MLTSA midterm: project proposal



	id	year	title	url	text
0	0	2016	"Human-Animal Chimeras Are Gestating on U.S. R...	"https://www.technologyreview.com/s/545106/hum...	"Featured Topics Newsletters Events Podcasts F...
1	1	2013	"Too Much Information | MIT Technology Review"	"https://www.technologyreview.com/s/522661/too...	"Featured Topics Newsletters Events Podcasts F...
2	2	2017	"Hacking the Biological Clock | MIT Technology...	"https://www.technologyreview.com/magazines/ha...	"Featured Topics Newsletters Events Podcasts F...
3	3	2018	"Six things to do with your data before you di...	"https://www.technologyreview.com/2018/10/23/2...	"Featured Topics Newsletters Events Podcasts F...
4	4	2018	"Your genome, on demand | MIT Technology Review"	"https://www.technologyreview.com/2018/10/23/1...	"Featured Topics Newsletters Events Podcasts F...

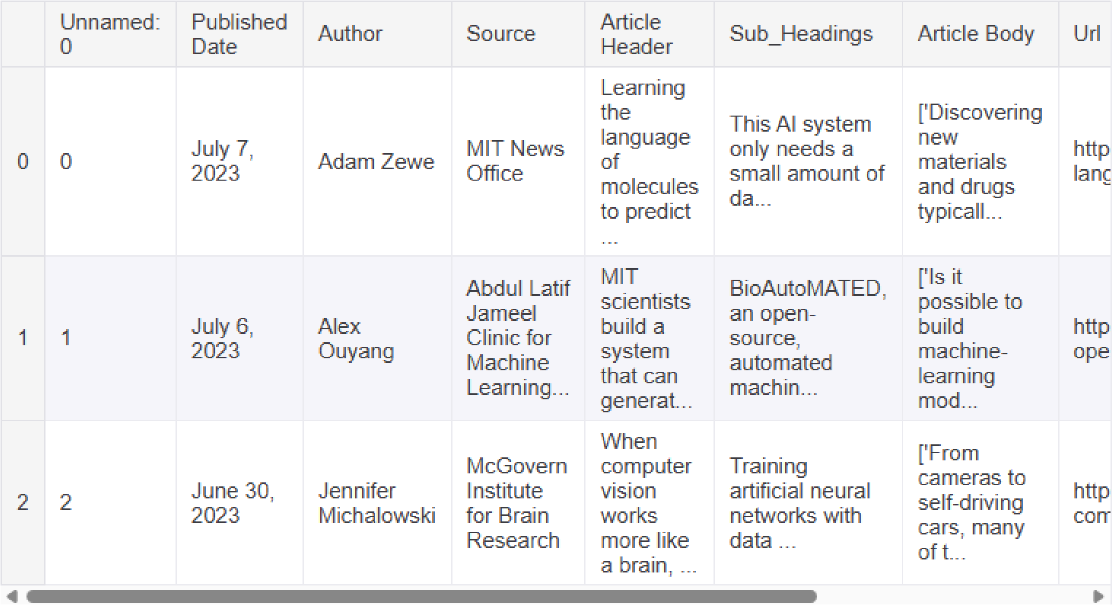
Frequency count plot



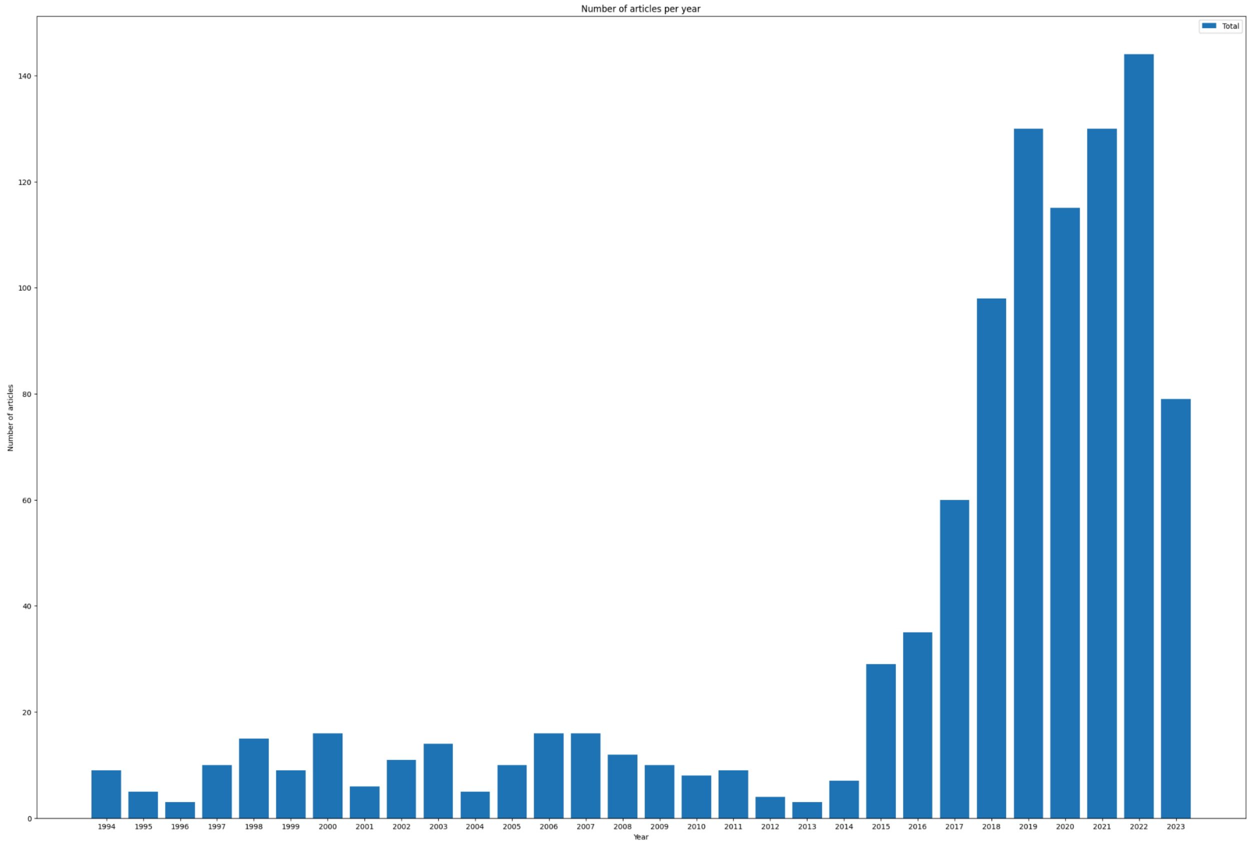
AI Articles Figure:
The data is very sparse until 2014, where an apparent exponential rise in the frequency occurs.

MIT AI News Published till 2023

df.head()



	Unnamed: 0	Published Date	Author	Source	Article Header	Sub_Headings	Article Body	Url
0	0	July 7, 2023	Adam Zewe	MIT News Office	Learning the language of molecules to predict ...	This AI system only needs a small amount of da...	['Discovering new materials and drugs typicall...	http lang
1	1	July 6, 2023	Alex Ouyang	Abdul Latif Jameel Clinic for Machine Learning...	MIT scientists build a system that can generat...	BioAutoMATED, an open-source, automated machin...	['Is it possible to build machine-learning mod...	http ope
2	2	June 30, 2023	Jennifer Michalowski	McGovern Institute for Brain Research	When computer vision works more like a brain, ...	Training artificial neural networks with data ...	['From cameras to self-driving cars, many of t...	http com



MIT_AI figure:
The data has slight upward and downward trends until 2014 when an apparent exponential rise is present.

Methodology

MLTSA midterm: project proposal

To analyze AI's evolving impact on careers, education, and society, this study proposes a multi-stage approach combining natural language processing, time series modeling, and machine learning. First, textual data from article titles are categorized into the three predefined domains using keyword-based classification. For granular analysis, NLP techniques such as n-gram frequency extraction quantify keyword prevalence, while BERTopic, a BERT-based topic modeling framework, clusters titles into semantically coherent topics. Numerical features derived from these methods (e.g., topic probabilities, n-gram counts) are aggregated temporally to form time series datasets.

Time series forecasting is conducted using ARIMA (AutoRegressive Integrated Moving Average) and Facebook's Prophet (FP) models. ARIMA captures trends, seasonality, and variability, while FP addresses missing data and irregular sampling intervals inherent in merged datasets. To validate classifier robustness and trend predictions, Gaussian processes evaluate the kernel density of textual embeddings (e.g., count vectors) against frequency distributions of categorical classifications, ensuring sensitivity and reliability. Seasonal variations and autoregressive patterns are analyzed to compare shifts in topic dominance over time.

Finally, model performance is assessed via residual analysis and cross-validation. Predictive trends from ARIMA and FP are compared to gauge consistency, while Gaussian process validation confirms alignment between textual semantics and classification outcomes. This integrated methodology enables dynamic tracking of AI's societal influence, offering insights into its prioritized domains and future trajectory.

Deliverable

1. Statistical Conclusions:
   - A comprehensive report detailing trends in AI's societal impact across careers, education, and society, derived from time series analysis (ARIMA and Prophet models). This will include:
     - Quantitative measures of how the prominence of each category (e.g., frequency of topic-related keywords or BERTopic clusters) has evolved over time.
     - Seasonal variations or autoregressive patterns in AI's perceived influence.
     - Forecasts predicting which domain(s) AI will most significantly impact in the near future.

2. Algorithmic Tools:
   - A custom NLP pipeline for classifying article titles into the three categories using keyword extraction and BERTopic-based topic modeling.
   - Validated time series forecasting scripts (ARIMA and Prophet) to replicate trend predictions.

3. Validation Framework:
   - A Gaussian process-based validation report assessing the sensitivity and reliability of classifiers and trend predictions by comparing textual embeddings with categorical frequency distributions.

4. Visualizations:
   - Time series plots showing trends in AI's impact across categories.

- Heatmaps or bar charts highlighting seasonal variations or shifts in topic dominance.

Link to GitHub repo:
https://github.com/talhaMah56/news_project.git

Bibliography

[1] J. Cho, G. Hwang, and C. Suh, ‘A Fair Classifier Using Kernel Density Estimation’, in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 15088–15099.

[2] M. Grootendorst, ‘BERTopic: Neural topic modeling with a class-based TF-IDF procedure’, arXiv preprint arXiv:2203. 05794, 2022.

[3] Samuel, Jim and Khanna, Tanya and Sundar, Srinivasaraghavan, Fear of Artificial Intelligence? NLP, ML and LLMs Based Discovery of AI-Phobia and Fear Sentiment Propagation by AI News (March 09, 2024). Available at SSRN: https://ssrn.com/abstract=4755964 or http://dx.doi.org/10.2139/ssrn.4755964

[4] S. J. Taylor and B. Letham, ‘Forecasting at scale’, *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[5] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2018.