Celal Emre Hurma – Mert Süral *20.11.2025*

Progress report: A Comparative Analysis of Emerging Trends in Economic Literature: 2020–2025

Introduction: This project aims to analyze the evolution of academic focus within economic literature over the last six years (2020-2025). Based on the analysis of these economic articles we can infer a significant shift in the direction of economic research from one dominant topic to another. By utilizing N-gram data and analyzing keyword frequencies, we investigate three major global themes that have had a profound impact on the global economy: Artificial Intelligence, Climate Change, and the COVID-19 pandemic.

The primary objective is to apply statistical concepts to understand distribution of data by years and quarters, identifying trends and anomalies in economic literature.

Dataset: The dataset employed in this project is sourced from the arXiv.org repository. Specifically, the analysis incorporates all research articles published within the "Economics" category between the years 2020 and 2025. For further researches we plan to increase our datasets between 2000 and 2025 and access articles in most categories in these years.

Theme/Term Selection: Our data selection strategy focused on capturing a diverse range of trend patterns. Including data sets with rising (AI), decaying (Pandemic), and steady (Climate) trends allows for a robust comparative analysis of how academic focus evolves.

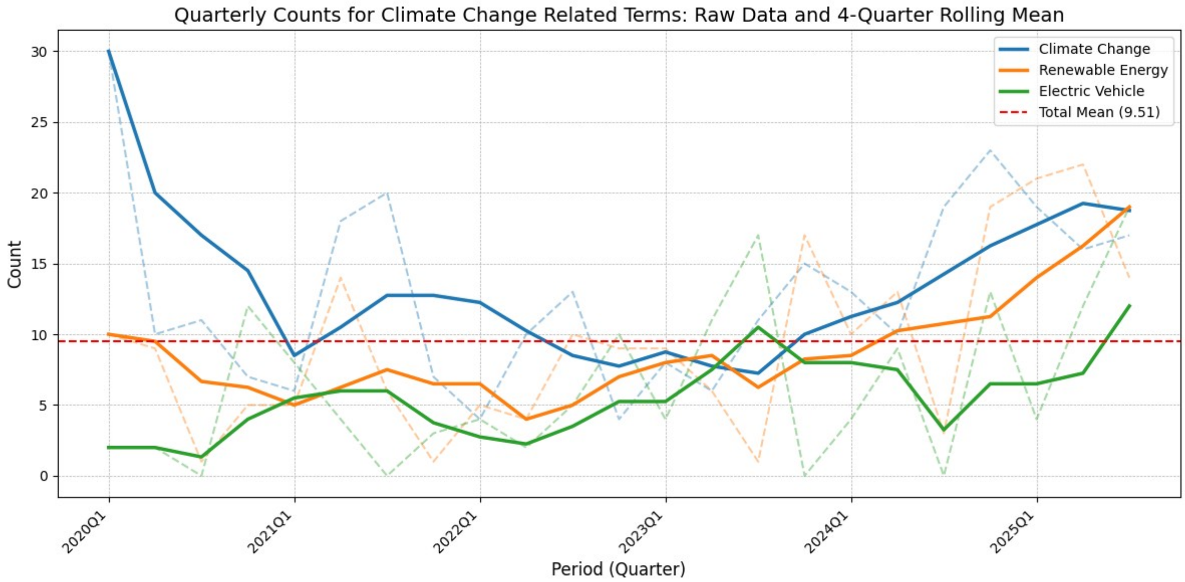
Methodology: Our methodology employs established statistical metrics, such as mean absolute error and standard deviation to analyze the data and interpret the events and trends of the period

-Mean Absolute Value: Mean Absolute Error (MAE) is a basic statistical metric used to measure the average magnitude of variability in a data set. MAE calculates the arithmetic average of the absolute differences between each data point. We used MAE to get a general review of the graphic because standard deviation is a better option when it comes to analyzing the critical data points

-Standard Deviation: Standard Deviation (SD) is a critical statistical measure used to quantify the amount of variation or dispersion in a dataset relative to its mean. The main difference between MAE and SD lies in their calculation: while MAE treats all deviations linearly, SD squares the differences. This squaring process is crucial because it assigns a higher weight to larger deviations, effectively penalizing outliers more severely than smaller variances**.** Therefore, SD is a better option when it comes to measuring structural instability and volatility. Furthermore, it allows us to apply the "⅔ Rule," which states that approximately 68 percent of the data should fall within the range of one standard deviation from the mean ($\mu \pm \sigma$).

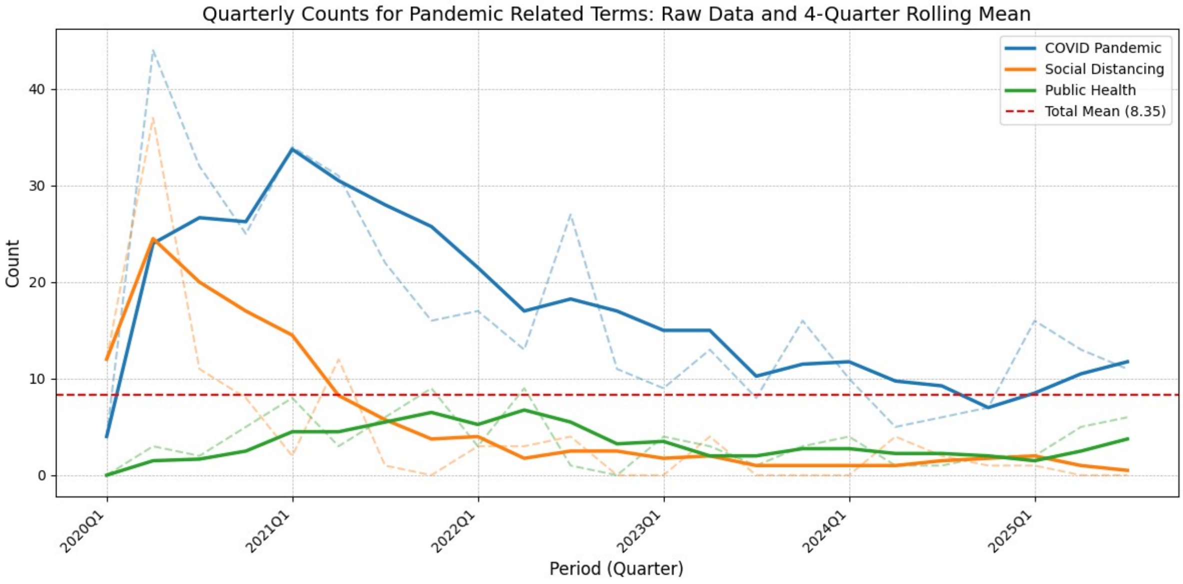
Consequently, data points falling outside this range are generally considered anomalies or indicators of rapid trend shifts.

Findings of the N-gram analysis
Climate Change and Energy: The Linear Topic



Analysis:
-By analyzing the standard deviation of this graphic we can see that these words are following a linear path
- Also in the late 2025 there is an increase on the line beyond the ⅔ Rule.
Evaluation:
- This graph illustrates that certain global challenges maintain a remarkable stability in economic literature.
- We observe that it is not affected by period trends such as artificial intelligence or corona and shows fluctuations independent of them.
- We can also infer a clear shift in the literature from problem identification to solution implementation. The diverging trends suggest that while general discussions of Climate Change are stabilizing, specific solutions like Renewable Energy are increasing in the economic discourse.
Mean Values of Quarterly Counts:
Climate Change: 12.91
Renewable Energy: 9.30
Electric Vehicle: 6.30

Findings of the N-gram analysis:
COVID Pandemic: The Decreasing Topic



Analysis:

- We can see a dramatic increase in the middle of the 2020 with the help of the standard deviation
- And also this ascension has a degradation part between 2021 and 2022
- We can also see one N-gram doesn’t leave the safe linear zone

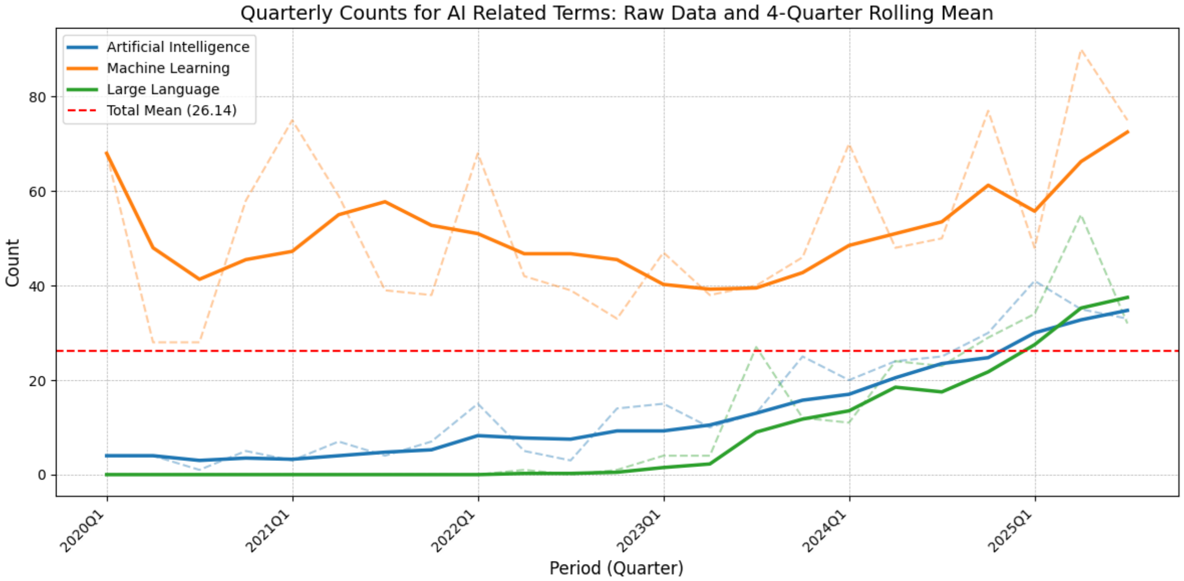
Evaluation:

- COVID pandemic acts as a pulse in economic articles it happened and then it’s popularity decreased incrementally.
- The term Social Distancing reached to zero in the latest quarters of 2025 that shows us the solution of the pandemic is no needed longer.
- Data trends indicate a divergence: The problem is over but the results are still affecting economic literature because of the impact.

Mean Values of Quarterly Counts:

Covid Pandemic: 16.96
Social Distancing: 4.57
Public Health: 3.52

Findings of the N-gram analysis:
Artificial Intelligence: The Increasing Topic



Analysis:

- There is a sudden increase in all of the N-gram starting in mid-2023 and continuing until the end of 2025
- Although the 2 of the N-grams are below the mean the other one is always an important word for the economic literature

Evaluation:

- The comparative analysis of “Artificial Intelligence” and “Language Models” signals a fundamental paradigm shift within the field. This transformation is evidenced by the exponential growth in “Language Models” usages starting in the first quarter of 2023 (2023Q1)
- From the dominance of the “Machine Learning” we can make an inference that the well known “Artificial Intelligence” is not that popular among the economic literature.

Mean Values of Quarterly Counts:

Machine Learning: 53.35
Artificial Intelligence: 14.91
Large Language: 11.17

WHAT WE HAVE DONE?

1- We have decided to connect API to arXiv.org to search Economy articles. The data are organically taken from 2020 to 2025. To ensure data consistency, we extracted metadata including titles, abstracts, and publication dates, filtering specifically for the 'Economics' (`econ`) subject category.

2- **Binning:** The year is divided into 4 equal pieces. Each Quarter represents the three months. Additionally the binning phase is applied to rolling mean at the first three graphs.

3- **Scrub:** The normal stopwords and academic stopwords removed to avoid noise. The “nltk” library helped us to find regular stopwords such as “the, a, and etc.”. However this does not contain full stopwords which we are looking for. So the economy based stopwords are created with ai tools and added to a config.json formatted file. This is the beginning of it:

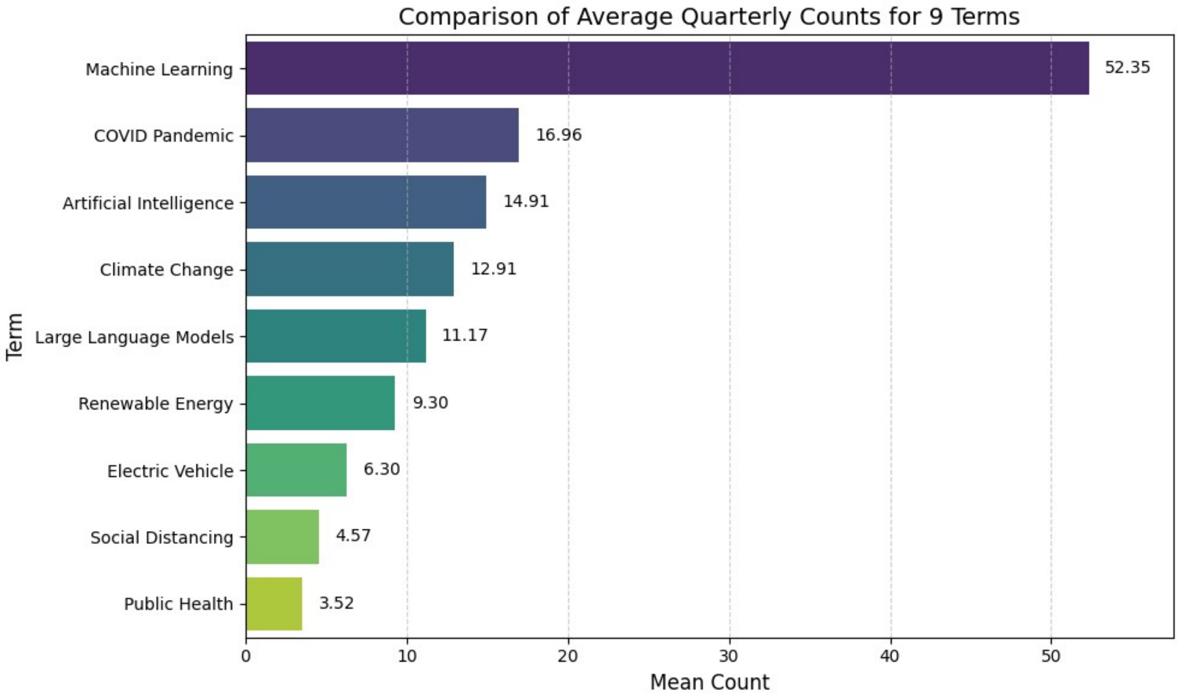
```
    "academic_stopwords": [
      "paper", "research", "study", "studies", "analysis", "analyze",
      "model", "models", "modeling", "data", "dataset", "datasets",]
```

4- **Frequency:** The n-gram terms are detected to see the tests more clearly. We have decided the terms which are combined with each other. Then we compared themselves group by group.

5- **Graphs:** We have used the Matplotlib library to perform graphs. The analyse (mean, composition, frequency) is shown with three different graph types.

6- **Analysis:** With the help of the histograms, we could make clear analyses.

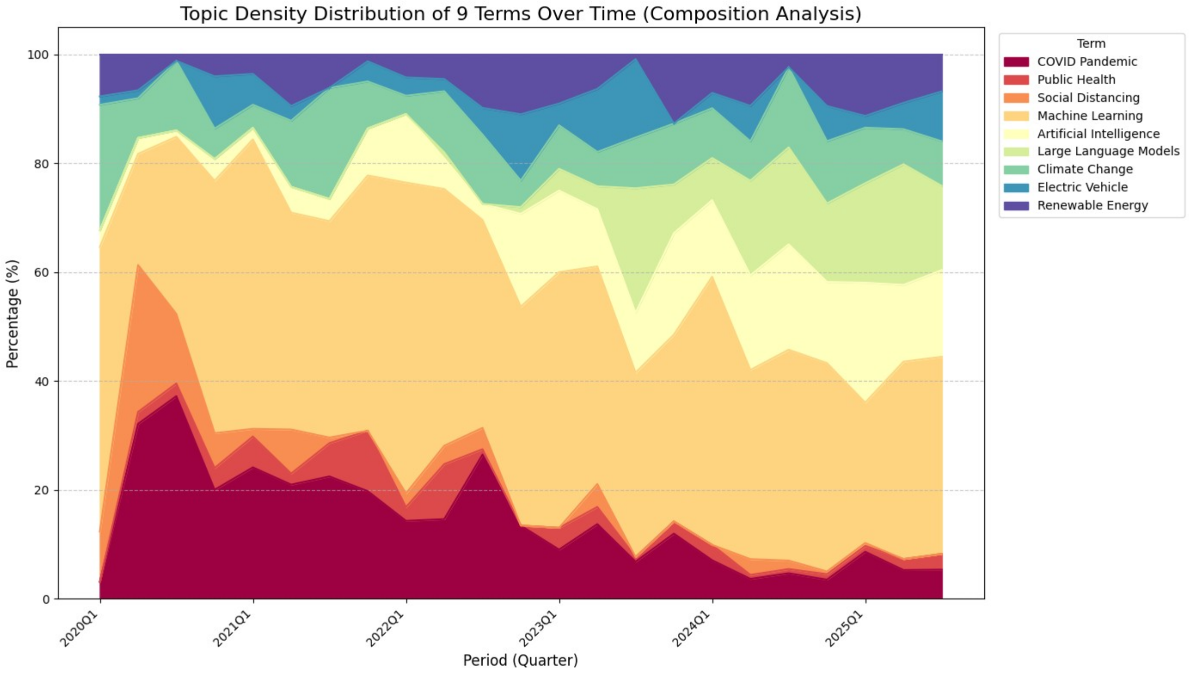
Average Quarterly Mean Comparison



Analysis:

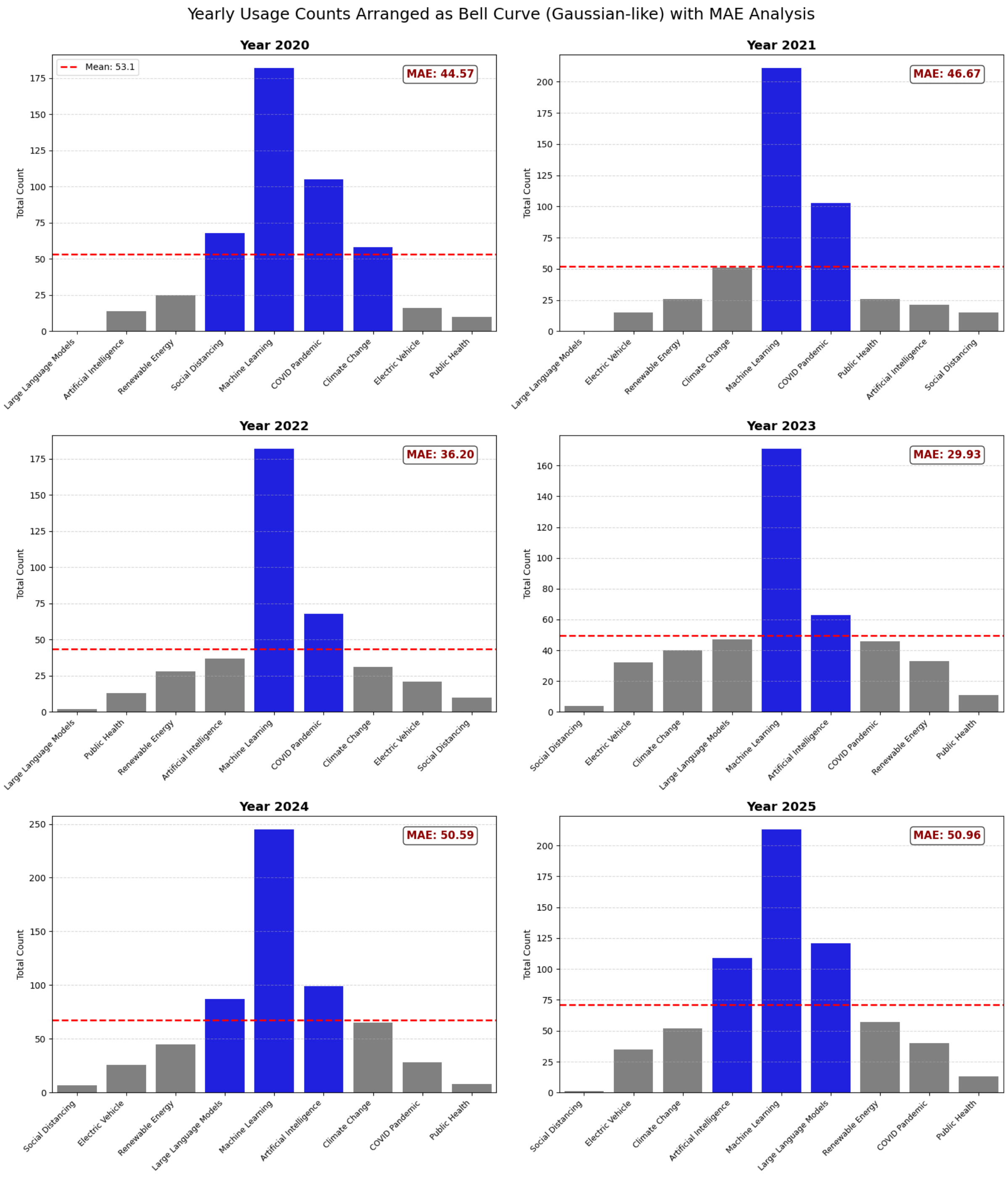
- Machine learning is still dominating the economic literature.
- We can see that a general topic like Public Health is placed at the last. We can grasp that general topics don't dominate economic literature.
- The more it is new the more it is mentioned. It is understandable because our datasets are between 2020 and 2025.

Topic Density: General Usage Analysis



Analysis:

- We added this graphic to show the percentages of all topics compared to each other
- We can understand that Machine Learning is dominating the economic literature
- Only at the peak of corona pandemic machine learning leaves the first place.



Analysis:

- In these distinct graphs covering a 6 year period, mean and MAE values are presented some values fall within the MAE limits, while others fall outside.
- Although different words fall within this MAE range each year, the one that stands out the most is Machine Learning.
- The topics falling within the MAE range in these graphs are those that have created a significant impact in the economic literature.
- While values under this line correspond to undervalued topics with limited research, the outliers above it highlight the year's most visible and important topic

REPO FOR CODES

https://github.com/cemrelistan/VBA_111.git