Unveiling the Post-Covid Economic Impact Using NLP Techniques

Kanishk Barhanpurkar
Thomas J. Watson College of Engineering and Applied Science, Binghamton University, USA
kbarhan1@binghamton.edu

Nikita Mandlik
Thomas J. Watson College of Engineering and Applied Science, Binghamton University, USA
nmandli1@binghamton.edu

Anand Singh Rajawat
School of Computer Science & Engineering, Sandip University, Nashik, India
anandrajawatds@gmail.com

S.B. Goyal
Faculty of Information Technology, City University, Petaling Jaya, Malaysia
sb.goyal@city.edu.my

Traian Candin Mihaltan
Faculty of Building Services, Technical University of Cluj-Napoca, 40033 Cluj-Napoca, Romania;
mihaltantraian83@gmail.com

Chaman Verma
Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University, 1053 Budapest, Hungary
chaman@inf.elte.hu

Maria Simona Raboaca
ICSI Energy Department National Research and Development Institute for Cryogenics and Isotopic Technologies
Ramnicu Valcea, Romania
Simona.raboaca@icsi.ro

2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI) | 979-8-3503-2138-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ECAI58194.2023.10194111

***Abstract*— The research paper presents a novel analysis of textual data based on Natural Language Processing (NLP) techniques to analyse New York Times articles from January 2019 to May 2023. The purpose of this paper is to gain an understanding of the economic impact that follows Covid-19 disease. New York Times (NYT), it was started in 1884 is one of the most prominent newspapers in the world. Additionally, we have used the New York Times Archive API to collect the data for the given timeframe. By analysing sentiment analysis, topic modelling, entity recognition, and keyword extraction, valuable insights can be gathered into market trends, industry shifts, and policy interventions. The authors have created a data science pipeline using Amazon Web Services (AWS), which enables data collection, storage, and visualization. It contributes to a better understanding of the pandemic's short-term and long-term economic effects. The results of this study demonstrate Natural Language Processing techniques' potential as a tool for financial analytics, assisting policymakers, economists, and businesses in formulating recovery strategies.**

Keywords*— *Post-Covid Economic, NLP, component, New York Times , Amazon Web Services (AWS).

I. INTRODUCTION

Economic activity generally declines during a recession as the business cycle contracts. Covid-19 spreads exponentially in the entire world by 2020, and many nations impose a lockdown to stop its spread. There is an impact on international trade as well as a decrease in currency flow among various countries because of this decision (Connaughton, J. E. et al., 2023). Several countries whose economies are heavily dependent on tourism suffered significant losses. This results in inflation and a lack of centralised money in 2022, which causes citizens to be unable to meet their basic needs (Bohl M. T. et al., 2023). In this study, textual data from news articles is used to shed light on post-COVID economic effects. The New York Times' extensive archives will be used to extract valuable economic trends, sentiments, and indicators (Taj, S. et al., 2019). The paper finds that uncertainty shocks had a greater impact on unemployment during the Great Recession due to factors such as the zero-linked interest rate and financial frictions in the economy (Eksi, O. et al., 2022). (Barbaglia, L. et al., 2022) has forecasted the recession over 26 major newspapers in 5 different languages. The related work study demonstrates that media sentiment, as measured by the Media Sentiment Index, is a strong predictor of stock market returns, highlighting the influence of media on investor behaviour (Baz, S. et al., 2022). The primary advantage of using the text data from the newspaper articles authored by professional journalists, scientists and industry experts. Thus, the data cleaning is more convenient for newspaper articles as compared to public opinion-based datasets (Example: Twitter, Reddit, Facebook comments etc.). Additionally, the authors have formulated research questions based are as follows-

RQ1: How do the articles from newspapers (NYTimes) associated with recession topics change over time?

RQ2: How do NY Times platforms influence the recession conditions?.

II. RELATED WORK

In this study, newspaper articles from The Times and New York Times are analysed to provide an ecological analysis of the news coverage of the Covid-19 pandemic (Xue Y., & Xu Q., 2021). For analysing the political impact of pandemics over the years over a set of topics, the New York Times and The Times newspaper were used (Abbas A.H., 2022). In this study, topic modelling and keyword analysis are the major techniques used for generating insights. The real-time event mechanism can also be determined using social media analysis. Similarly, the spread of Covid-19 and the use of masks can be analysed in New York City and hence the mapping of both parameters can be done (Ma X. et al., 2022;

Adjei-Fremah S. et al., 2023). Additionally, the major newspaper's article data is analysed on a particular topic where the data streaming can be real-time or data-driven over time (Connaughton, J.E. et al., 2023). The social network analysis provides details about generating public opinions majorly for Twitter, and Reddit data (Kearney, M. S. et al., 2022) where the unformatted data is based on different users writing styles, and uniformity of the data is the major challenge.

Table 1: Comparative analysis for the different studies associated with newspaper data.

S.No	*Study*	*Data Source*	*Highlights*
1.	Malladi, R. K. (2022)	FRED Database	Supervised ML techniques accurately predicted the Covid-19 recession and stock market crash in advance.
2.	Bernaola Serrano, I. (2022)	Three major Spanish newspapers – El País, ABC, and El Mundo	Media agenda-building during the 2008 financial crisis and the Internet's influence, revealed loss of control
3.	Rios-Rodríguez, R. et al., 2023	98 Spanish print newspaper publishers over the period 2009–2018	Newspaper crisis persists despite economic recovery
4.	Zafri, N. M. et al., 2021	7,209 newspaper articles are assembled and analyzed from three popular local newspapers named "bdnews24.com", "New Age", and "Prothom Alo English"	Coverage of COVID-19 in Bangladesh informs a four-stage pandemic management framework
5.	Laudati, D. et al., 2023	News articles collected between 1989-2019	Sanctions on the Iranian economy from 1989-2019
6.	Wang, D.et al., 2023	230 news articles to examine the tone of each source	Media framing of GM foods differs between US and China, influencing public perception.
7.	Basak, G. K. et al., 2023	Brexit News 2016 Dataset	Improved lexicon enhances predictive power.

III. Proposed Methodology

The data is collected from the New York Times Archive API that returns the details for all the articles ranging from January 2019 to May 2023. In total, 230,178 articles have been collected in the given duration. The articles collected are stored using Amazon Web Services (AWS) S3 service for Data Storage. The S3 bucket web service is particularly used for creating a pipeline for data storage and organization. S3 is the most affordable service that provides all the services of data storage. Stored data is then filtered for the keywords related to recession to get the required data for the analysis. The Filtered headline data contains around 9,971 articles in the data that can then be used for further data processing. The next step in the pipeline constitutes Data Cleaning for NLP data. The data cleaning consists of several steps - (1) Punctuation - as the heading has a lot of punctuation, they do not contribute anything to the NLP analysis. (2) Tokenization - converts the string into a list of words. (3) Stop Words - it removes the irrelevant words from the list. (4) Lemmatizing/Stem - This reduces the words to their root form. And other methods like removing numbers from the headline. The data can be analysed using visualization techniques performing Natural Language Processing techniques like Sentimental analysis, Topic modeling, Entity recognition, and Keyword extraction to get key insights about the collected data. The entire proposed system is designed in Python language and we have used libraries like NLTK, seaborn, matplotlib, pandas, numpy, wordcloud, sklearn, textblob, textstat.

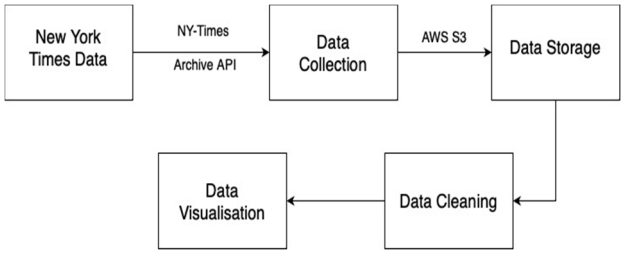

Figure 1: Data-flow diagram of entire process.

IV. Results & Discussion

In this research paper, the authors have performed different types of text-based analysis to understand the data in a more efficient manner. The word frequency graph helps us to understand the number of occurrences of the most prominent words available in the dataset. Firstly, we have performed word frequency analysis where the Y-axis contains the number of occurrences for the word in the dataset and X-axis contains the most occurred word or symbols. In Figure 2, the most occurring keywords are "Market", "Business", "Economy", "Tax" and government" where the count range is in the range of 500-1000.

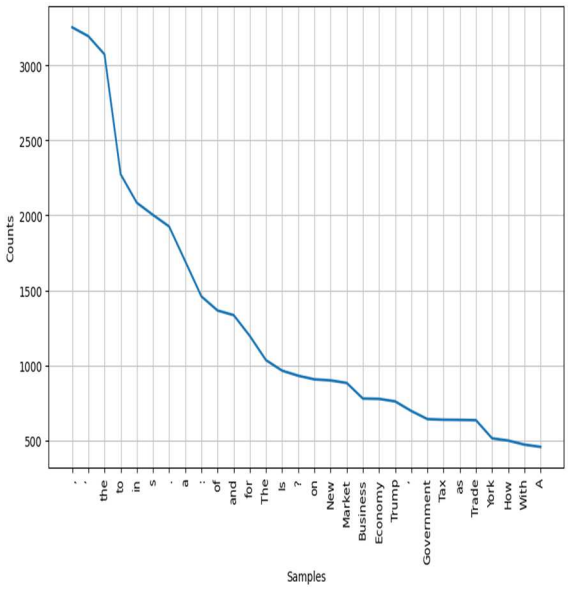

Figure 2: Word Frequency Analysis graph

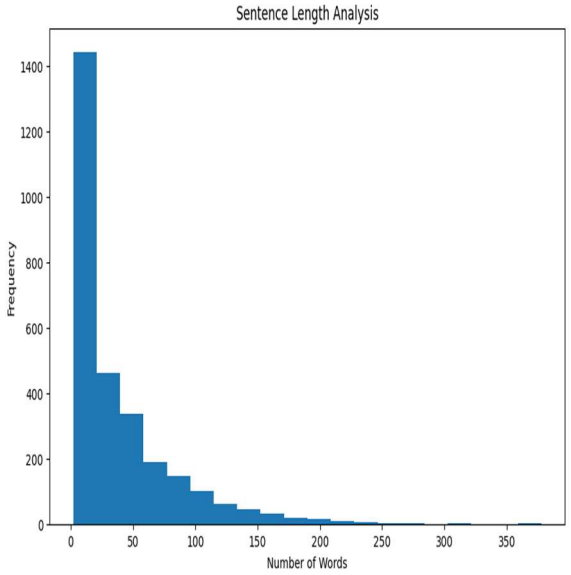

Figure 3. Sentence Length analysis.

In the sentence length analysis, the primary goal is to compare the frequency of words occurring in the headline in the dataset associated with the New York Times newspaper. The frequency and number of words are available on Y-axis and X-axis respectively. The average sentence length is maximum in the range of 0-25 words and sentence length is minimum in the range of 360-380 words. In the Figure 3, it can be observed that the frequency decreases with increasing the frequency of the words. In the word length analysis, we used a boxplot for analysing the word length (number of characters) for every word present in the dataset. In Figure 4, Y-axis represents the word length ranging from 0 to 20 characters per word. The Interquartile range (IQR) is available between word lengths of 2.0 to 5.0 per word. The word length for Q1 (25th percentile) range is 2.5, Q3 (75th percentile) range is 5.0 and the median word length is 3.7. The outliers are present for the word length of 11.0 to 20.0 characters. In Figure 5, we have performed the keyword analysis technique to extract the words which are associated with the New York Times dataset. We have specifically used the wordcloud library associated with Python to perform keyword analysis. The most used keywords are “business”, “market”, “inflation”, “government”, “trump”, “coronavirus” and “economy”. Word density depends on various factors like the number of occurrences of the word and the length of the word.

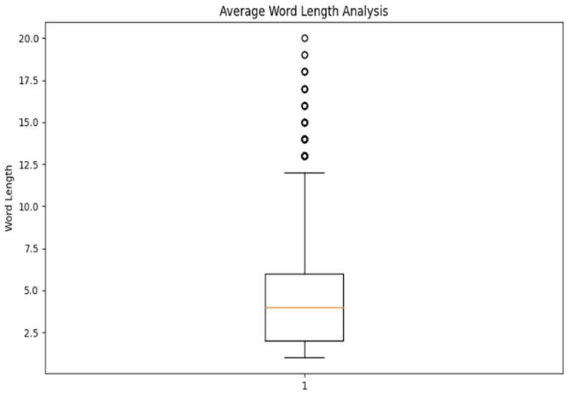

Figure 4. Average Word Length Analysis.

Figure 5. keyword analysis (wordcloud).

The Flesch Reading Base (FRE) is used to analyse the text complexity for the New York Times Newspaper headlines. The average score for the entire dataset ranges in the range of 50 - 100 which shows that text is highly readable. It also provides insights into the textual data that newspaper articles need less effort in the data cleaning process. The high score on the FRE scale is directly proportional to the easiness of reading the text. The mathematical equation to calculate the FRE scale is as follows-

FRE = 206.835 − 1.015 * (total words/total sentences)rn− 84.6 * (total syllables/ total words)

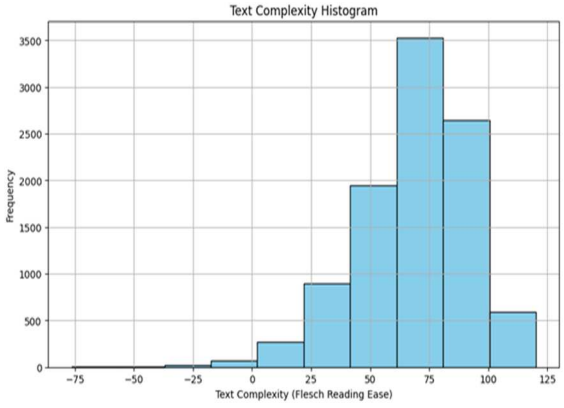

Figure 6:Text-complexity (Flesch Reading Base).

It is possible to discard stopwords in further textual analysis by identifying the words that repeat most often in a language. A common stopword is "the", "to", "in", "a", "of", and "and", which is an optimized technique to discard stopwords during textual analysis. In Figure 6, “the”, “to”, “in”, and “a” has around 3000, 2300, 1900, and 1800 occurrences respectively. Additionally, the y-axis represents the number of occurrences and the x-axis represents the recurring stopwords. After discarding the stop-words, we analysed the remaining non-stop-words in each headline associated with the dataset. As shown in Figure 7, the context

changed for the keywords "the", "is", "a" and "on" according to grammar rules and language syntax. The keywords are organized by the number of occurrences in each highlight, such as "market", "tax", "business", "Trump", "U.S.", and "inflation". Relational mapping can be performed using entity-topic modeling or social network analysis.

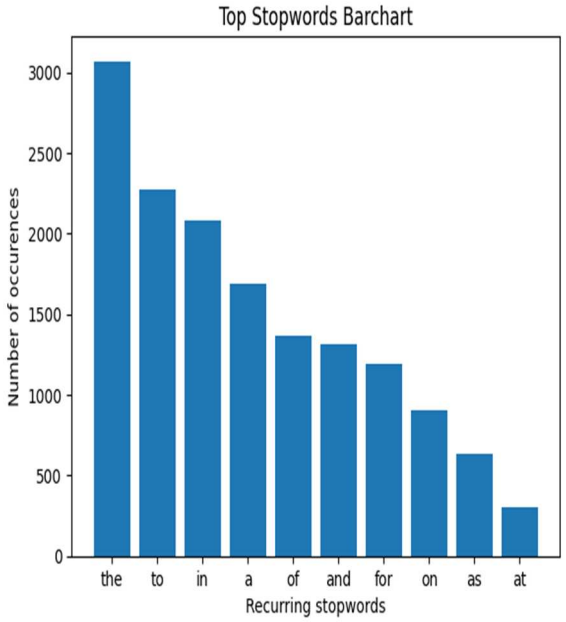

Figure 7. Stopwords analysis.

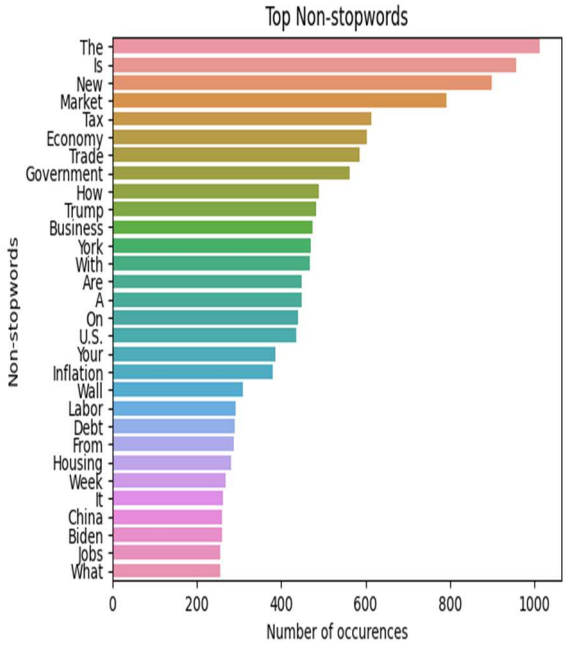

Figure 8. Non-stopwords analysis.

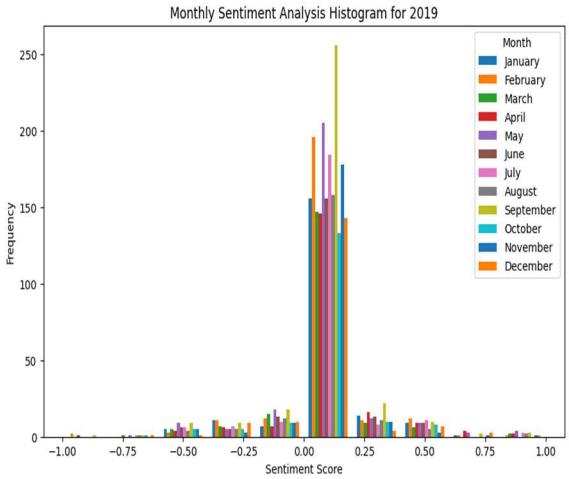

Figure 9. (a)

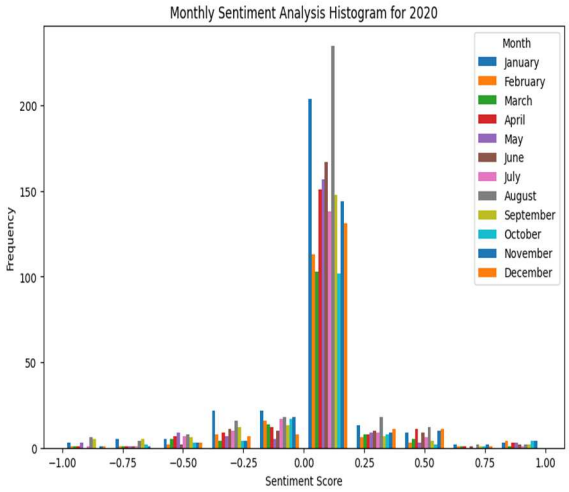

Figure 9. (b)

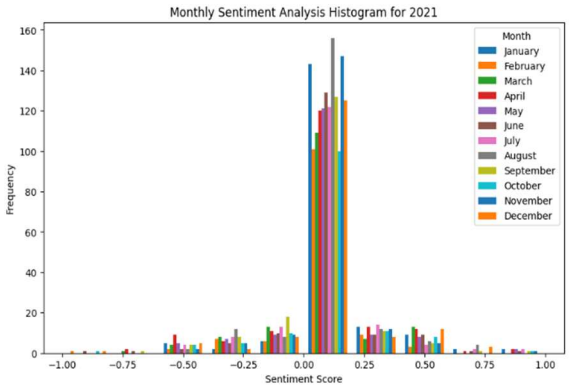

Figure 9. (c)

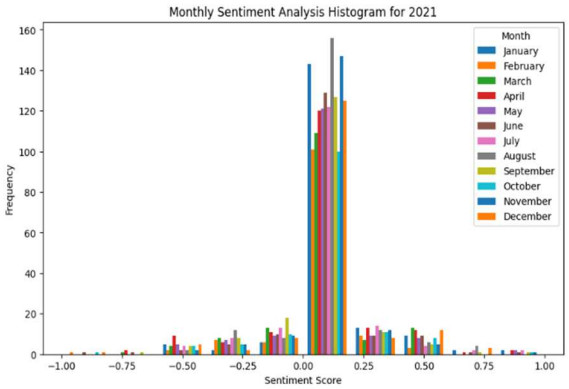

Figure 9. (d)

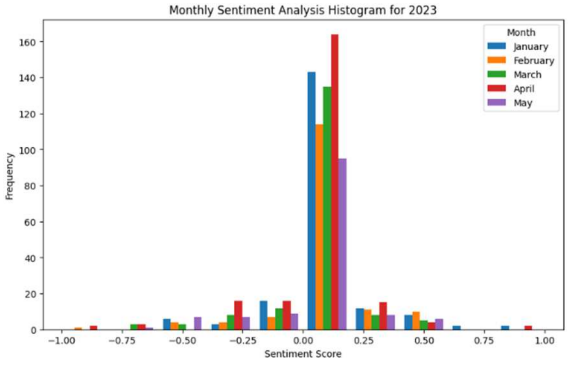

Figure 9. (e)

Figure 9. (a)(b)(c)(d)(e) Comparison of monthly sentimental analysis from 2019 to 2023.

To analyse the impact of Covid-19, we have used sentimental analysis for every headline every month in the years 2019, 2020, 2021, 2022, and 2023. The Sentimental Analysis scale is used to analyse the polarity of the given scale from -1.0 (extreme negative) to +1.0 (extreme positive) and 0 for neutral. In Figure 8, the negative scale is not observed from the month of January 2019 to May 2019. However, the negative scaling increases from June 2019 to December 2019 where the Covid-19 on its peak in the North America and European continent. At the start of 2020, Covid-19 peaks have been observed, with more negative results reported in the first eight months of the year. The year 2021 has also shown the highest positive scores and no signs of economic crisis. In the year 2022, positive articles have been observed for the duration of January 2022 to August 2022. However, the negative sentiment score for the economic crisis has increased from September 2022 to December 2022. In the year 2023, the extreme negative and extreme positive headlines are found in April 2023.

V. CONCLUSION & FUTURE WORK

In this research paper, the authors have developed a data pipeline to store the data, filter the data and visualize the data collected from New York Times Archive API. We have stored the data using services offered by AWS. AWS S3 is the one of the most convenient storages which is used for creating large data pipelines. We have performed various data cleaning techniques like stemming, tokenization, and removing punctuations. The main aim of this research paper is to assess the post Covid-19 economic impact using Natural Language Techniques. We have performed analytical techniques like keyword analysis, sentiment analysis, and entity recognition. Conclusively, the analysis conducted on New York Times headlines spanning from 2019 to 2023 illustrates the fluctuating sentiments associated with the Covid-19 impact. Negative sentiment experienced an upsurge during 2019-2020, followed by a prevalence of positive sentiment in 2021. However, negative sentiment resurfaced in 2022-2023, underscoring the persistent economic challenges that persist. In the future scope, we will implement predictive modeling techniques to forecast the upcoming recession or economic crisis conditions based on data associated with newspaper datasets.

REFERENCES

[1] Connaughton, J. E., Cebula, R. J., & Amato, L. H. (2023). The regional economic impact of the 2020 COVID-19 recession in the USA. Journal of Financial Economic Policy, (ahead-of-print).

[2] Bohl, M. T., Kanelis, D., & Siklos, P. L. (2023). Central bank mandates: How differences can influence the content and tone of central bank communication. Journal of International Money and Finance, 130, 102752.

[3] Taj, S., Shaikh, B. B., & Meghji, A. F. (2019, January). Sentiment analysis of news articles: a lexicon-based approach. In 2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET) (pp. 1-5). IEEE.

[4] Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. International Journal of Social Research Methodology, 23(3), 277-290.

[5] Xue, Y., & Xu, Q. (2021). An ecological discourse analysis of news coverage of COVID-19 in China in The Times and The New York Times. Journal of World Languages, 7(1), 80-103.

[6] Abbas, A. H. (2022). Politicizing the pandemic: A schemata analysis of COVID-19 news in two selected newspapers. International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 35(3), 883-902.

[7] Ma, X., Luo, X. F., Li, L., Li, Y., & Sun, G. Q. (2022). The influence of mask use on the spread of COVID-19 during pandemic in New York City. Results in Physics, 34, 105224.

[8] Adjei-Fremah, S., Lara, N., Anwar, A., Garcia, D. C., Hemaktiathar, S., Ifebirinachi, C. B., ... & Samuel, R. (2023). The effects of race/ethnicity, age, and area deprivation index (ADI) on COVID-19 disease early dynamics: Washington, DC case study. Journal of racial and ethnic health disparities, 10(2), 491-500.

[9] Connaughton, J.E., Cebula, R. J., & Amato, L. H. (2023). The regional economic impact of the 2020 COVID-19 recession in the USA. Journal of Financial Economic Policy, (ahead-of-print).

[10] Kearney, M. S., Levine, P. B., & Pardue, L. (2022). The puzzle of falling US birth rates since the Great Recession. Journal of Economic Perspectives, 36(1), 151-76.

[11] Eksi, O., & Tas, B. K. O. (2022). Time-varying effect of uncertainty shocks on unemployment. Economic Modelling, 110, 105810.

[12] Barbaglia, L., Consoli, S., Manzan, S., Forecasting GDP in Europe with Textual Data (June 28, 2022). Available at SSRN: https://ssrn.com/abstract=3898680 or http://dx.doi.org/10.2139/ssrn.3898680

[13] Bailey, D. (2022). Contention in Times of Crisis: Recession and Political Protest in Thirty European Countries. Contemporary Sociology, 51(4), 300–302. https://doi.org/10.1177/00943061221103312

[14] Baz, S., Cathcart, L., Michaelides, A., What is the Value of Financial News? (October 18, 2022). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4251414

[15] Malladi, R. K. (2022). Application of Supervised Machine Learning Techniques to Forecast the COVID-19 US Recession and Stock Market Crash. Computational Economics, 1-25.

[16] Bernaola Serrano, I. (2022). Effects of the 2008 crisis on agenda building: Internally originated content versus external dependence. Journalism Practice, 1-18.

[17] Rios-Rodríguez, R., Fernández-López, S., Dios-Vicente, A., & Rodeiro-Pazos, D. (2023). Reconversion in a declining market: the

return to profitability of the print newspaper industry. Journal of Media Business Studies, 20(2), 204-222.

[18] Zafri, N. M., Afroj, S., Nafi, I. M., & Hasan, M. M. U. (2021). A content analysis of newspaper coverage of the COVID-19 pandemic for developing a pandemic management framework. Heliyon, 7(3), e06544.

[19] Laudati, D., & Pesaran, M. H. (2023). Identifying the effects of sanctions on the Iranian economy using newspaper coverage. Journal of Applied Econometrics, 38(3), 271-294.

[20] Wang, D., Li, Y., Mao, Z., He, M., Hon, C., & Liu, Z. (2023). Risk definers and social discourse of GM foods–a comparative analysis of the People's Daily and the New York Times.

[21] Basak, G. K., Das, P. K., Marjit, S., Mukherjee, D., & Yang, L. (2023). The British Stock Market, currencies, brexit, and media sentiments: A big data analysis. The North American Journal of Economics and Finance, 64, 101861.