Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(6), 191-198
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Social Media Is a Juggernaut: Lagged Correlation Analysis Using Ngram Data on "Internet" and "Social Media" with Amplification by the Advent of the "iPhone"

William Zywiak[1], Gao Niu[1], Nicolas Petrell[1], Victoria Nichele[1], Kirsten Hokeness[2]

[1]Department of Mathematics and Economics, College of Humanities and Social Sciences, Bryant University, Smithfield, USA
[2]College of Health and Behavioral Sciences, Bryant University, Smithfield, USA
Email: wzywiak@bryant.edu

How to cite this paper: Zywiak, W., Niu, G., Petrell, N., Nichele, V., & Hokeness, K. (2026). Social Media Is a Juggernaut: Lagged Correlation Analysis Using Ngram Data on "Internet" and "Social Media" with Amplification by the Advent of the "iPhone". *Open Journal of Social Sciences, 14,* 191-198.
https://doi.org/10.4236/jss.2026.146010

Received: May 8, 2026
Accepted: June 14, 2026
Published: June 17, 2026

Abstract

The amount of new information transmitted per day can be overwhelming. The data available through Ngram Viewer allows statistical examination to determine the most salient topics, as well as lagged correlation and lagged variance to support possible causal connections between concepts. Using this database, we determined that COVID, crypto, microplastics, and especially social media are important topics of the last few years. We also present results that suggest the development of the internet and the iPhone fueled the prominence of social media, and the access of the iPhone also increased access to the internet.

Keywords

Ngram, Lagged Variance, Crypto, Microplastics, Social Media

1. Introduction

In 1982, Naisbitt highlighted that the US was shifting from an industrial society to an information society. Indeed, anyone accessing the World Wide Web daily may feel overwhelmed by the new information encountered. The ngram database (Aiden & Michel, 2013; Michel et al., 2011) is publicly available data that can be statistically analyzed to see which themes are indeed salient and growing. This database was originally released with words up to the year 2009, and has since been updated with 2012, 2019, and 2022 versions (Solovyev, 2024). It is comprised of phrases from scanned books and phrases from one to five words may be searched. Google Books Ngram reflects word frequency in published books, not

direct public behavior or platform usage. Initially in this paper, we identify a number of "vanguard" terms; terms that have recently been increasing at a steep rate (crypto, microplastics, and covid). Following this, we found an extreme vanguard: "social media". Next, we conduct lagged correlations to suggest that the internet contributed to the rise of social media, and to suggest that the advent of the iPhone, provided a boost to "internet" and a more pronounced boost to "social media".

The first and second authors have been having Statistics 1 students use the Ngram database to conduct linear regressions, lagged correlations, and t-tests, since 2020. In our first article, we identified and discussed years that are more persistent in the Ngram database than the typical year, which drops off quickly. The years we identified were of historical importance and consisted of 1799 in the French and Russian corpus, 1865 in the American English and Hebrew corpus, 1917 in the Hebrew and Russian corpus, 1945 in the German and Hebrew corpus, and 1948 in the Hebrew corpus (Zywiak, Bobroff, & Niu, 2021). This paper used a t-test.

In our second article, we found the four most prominent character strengths in the American English corpus in 2019 (the most recent data point available when we were preparing this article) to be love, hope, perspective, and leadership (Zywiak & Niu, 2021). In this paper, we ranked frequencies, used four linear regressions, and used a pie-chart. These two papers are summarized by Solovyev (2024) of Russia, in her review of Ngram articles that assess societal change.

Our third paper was derived from an assignment turned in by a student for AA 205: Intro to Applied Analytics. As a student veteran he was interested in using the Ngram database to discern the extent to which money may be one of the causes and effects of war. In the Ngram plots for war in the American English corpus, the Revolutionary War, the War of 1812, the American Civil War, WW1, and WW2 were quite evident. Since a word may be salient because of good or bad press (valence is not tagged, see e.g, the plot for the insecticide DDT), we examined plots for "cost of war". This paper used linear regressions (McFadden, Zywiak, Bobroff, & Niu, 2022). The rationale and details for this line of research are further detailed in the McFadden et al. (2022) paper.

The present paper is meant to be an exemplar for providing support for causal relationships in historical data, though causality is not proved, since there is no random assignment to a manipulated independent variable. The inability to manipulate an independent variable (e.g., increased substance use to affect level of purpose in life) dictates that statistical approaches are used in naturalistic longitudinal designs to provide evidence of relationships between variables (Harlow, 2023: p. 19). In total, these four papers show examples of how regressions, lagged correlations, and t-tests can be used by Statistics 1 students to explore how different terms are related in published books.

2. Method

The 2022 version of the Ngram data was accessed, by googling "ngram viewer".

American English corpus was selected since that is the culture best known by most of the authors. Smoothing was set to zero (default is 3) so that the raw data could be accurately viewed. The Case Insensitive setting was left in the default position. Data was retrieved by having the mouse on a given year, and transcribing the value from the screen with pen and paper, and entering the data into an Excel file. Statistical analyses were conducted in Excel. The data for every year is available from the first author, or by repeating this procedure. The exact search strings and year ranges were "covid", "microplastics", and "crypto" from 2010 to 2022, and "social media", "internet", and "iPhone" from 1990 to 2022.

3. Results

After examining several terms, we noticed that three "vanguards" were rapidly increasing over time; crypto, microplastics, and covid are plotted in Figure 1 for the period from 1990 to 2022. These terms show rapid increases, and will be discussed further in the Discussion. When we add the 2-gram "social media", this overpowers the previous terms. See Figure 2. Note that because of scaling, the

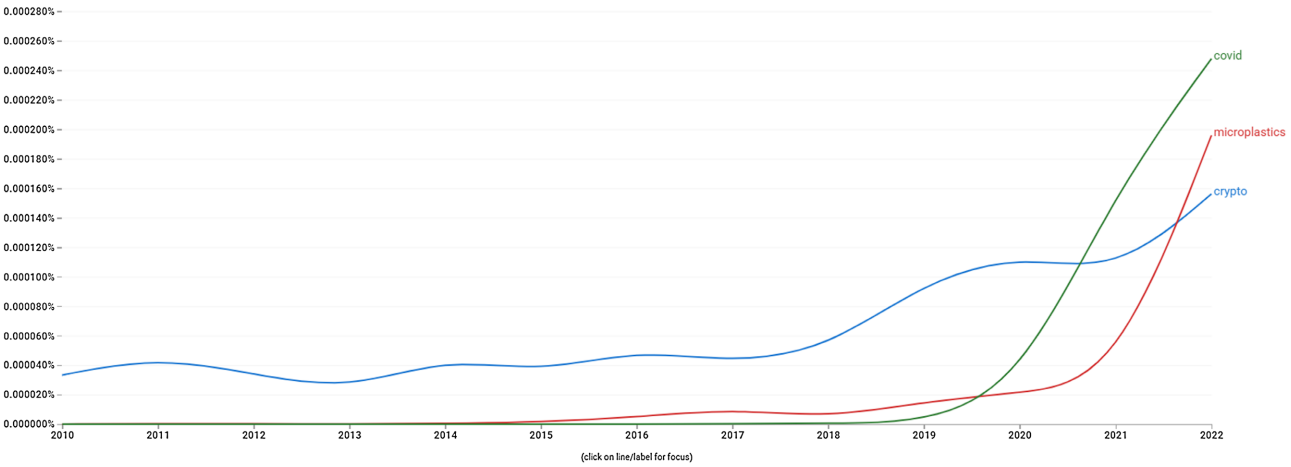

Figure 1. COVID, microplastics, and crypto, 2010 through 2022.

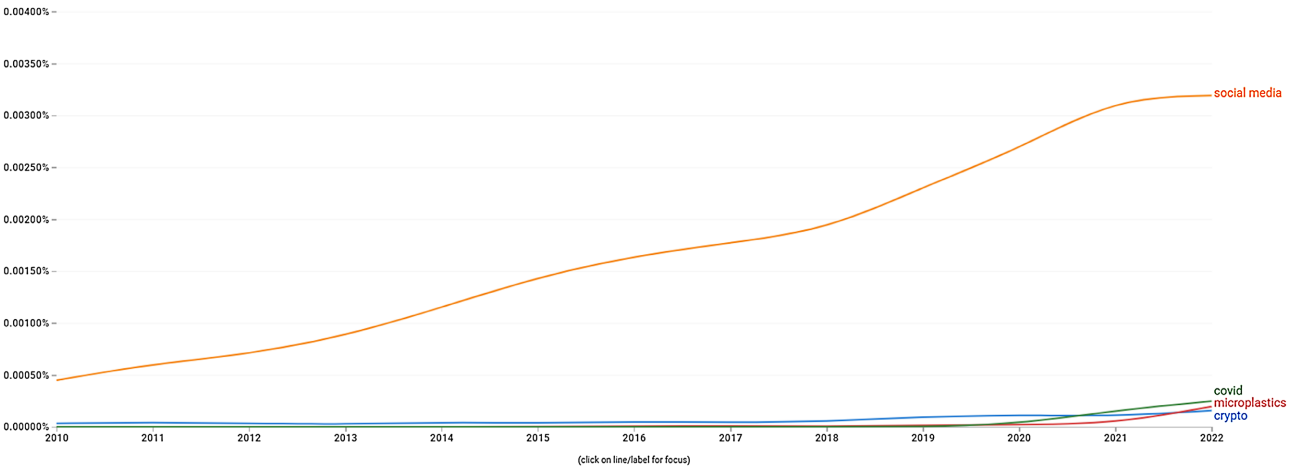

Figure 2. Social media, COVID, microplastics, and crypto, 2010 through 2022.

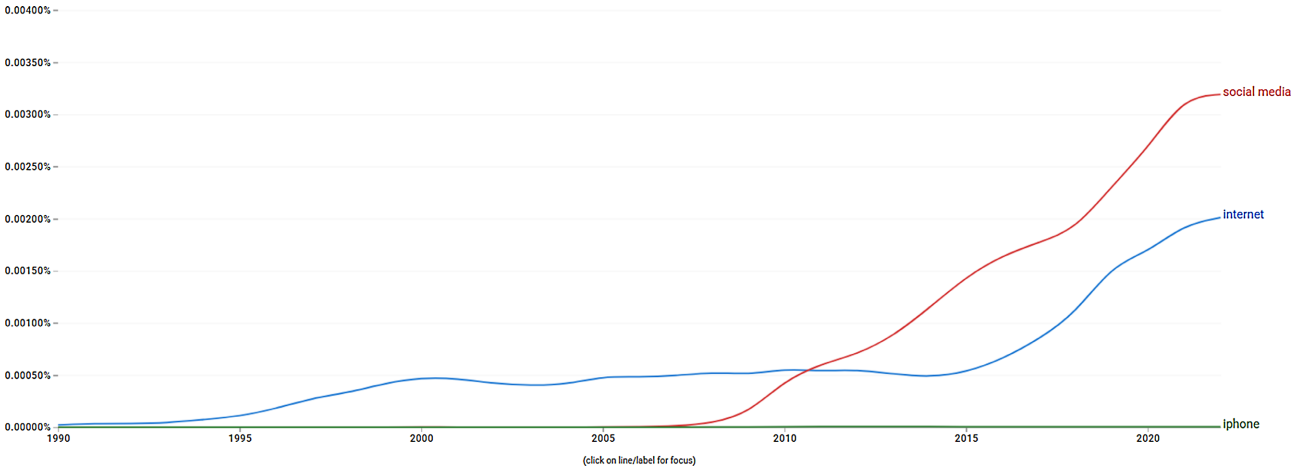

Figure 3. Social media, internet, and iPhone, 1990 through 2022.

three lines in **Figure 1**, look relatively flat in **Figure 2**. The internet began becoming prominent in the 1990s and a plot for social media and internet is depicted in **Figure 3** from 1990 to 2022. These two terms show a classic pattern that can be used to examine the extent of the lagged correlation.

With annual data for the frequency of "internet" from 1990 to 2000 and annual data for the frequency of "social media" from 1990 to 2022, we conducted lagged correlations for these two phrases, starting with a simultaneous correlation and ending with a lag of 17 years. These lagged correlations are depicted in **Figure 4**. The most pronounced correlation was equal to a near perfect 0.99235 associated with a lag of 14 years. Since correlations exaggerate the association between two variables, we also computed lagged variance, since this indicates the variance shared between two variables. (For example, a correlation of 0.7 is equivalent to only 49% shared variance between two variables.) Lagged variance is pictured in **Figure 5**, and is still very pronounced with 98% of the variance shared between the two phrases at a lag of 14 years.

Additionally, we noted that "iPhone" was increasing in the ngram database from 2006 to 2011 and levelling off thereafter [with a value of 0.0000000665% in 2005, 0.0000001963% in 2006 (triple the 2005 value), and peaking at 0.0000072305%

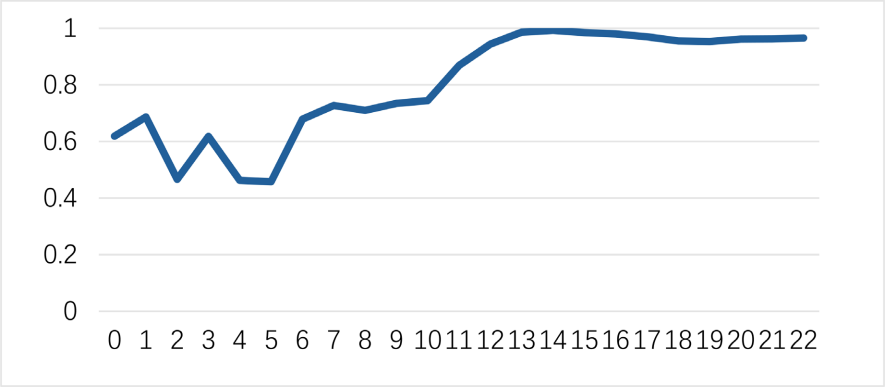

Figure 4. Correlation between "internet" and "social media" based on lag in years.

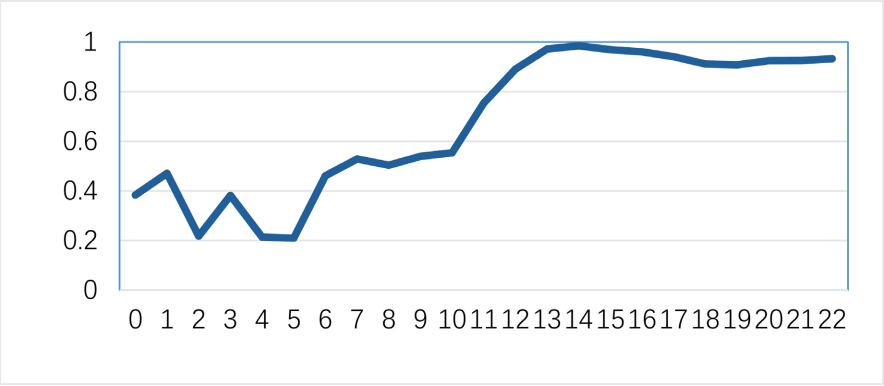

Figure 5. Lagged variance between "internet" and "social media".

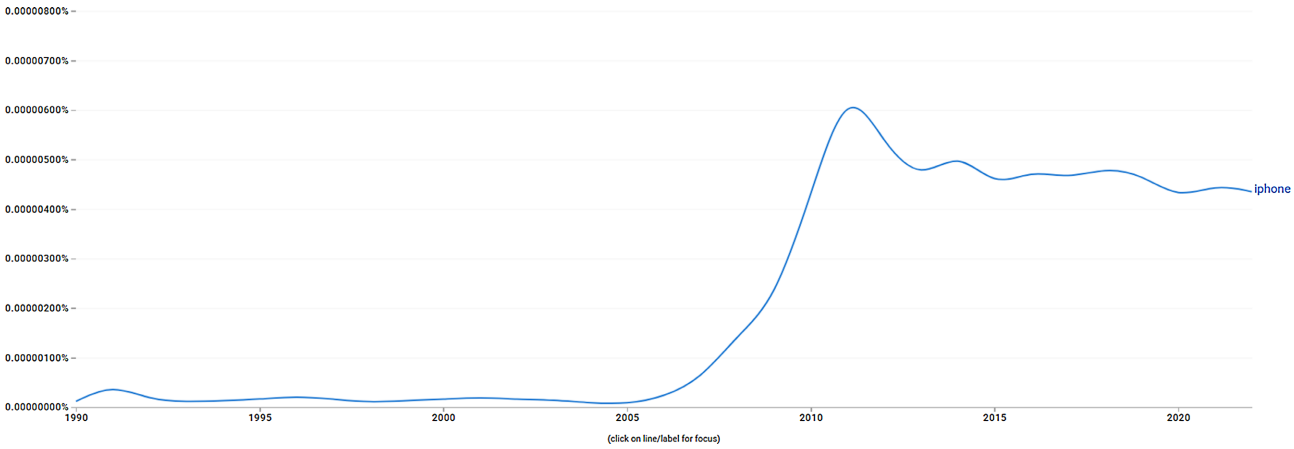

Figure 6. iPhone 1990 through 2022.

in 2011 (37 times the 2006 value) see Figure 6]. Since *iPhone* increased for a discrete period of time and to include the third bivariate analysis focused on in Statistics I (i.e., a t-test), we examine the frequency of internet and social media, from 2006 and earlier compared to 2007 and later using a t-test. There was a huge difference in the frequency of social media comparing 1990-2006 versus 2007-2022: respective M's (SD's): 0.00875 (0.00005) versus 13.79 (113.37) words per million, t (15) = 5.18, $p = 0.0001$. Additionally, there was an increase in the frequency of internet from 1990-2006 versus 2007-2022, respective M's (SD's): 2.741 (0.0003481) versus 9.0385 (0.0031673) words per million, t (18) = 4.26, $p = 0.0002$. Finally, while the ngram data is right censored at 2022, we note that COVID practically returns to zero in 2026 in the more up to date Google Trends.

4. Discussion

4.1. Vanguards

A dramatic increase for the term *microplastics* is seen in the ngram database starting in 2021. Microplastics are defined as being less than 5 mm in diameter. Microplastics may be poisonous based on the chemicals they are made of. They are also an excellent absorbent of other pollutants given their irregular surface area

(Albazoni, Al-Haidarey, & Nasir, 2024). Microplastics have contaminated water ways, oceans, soil, and the atmosphere. Microplastics can be found in drinking water. Microplastics can be directly ingested by animals and ingested secondarily through ingestion of organisms that ingested microplastics. In comparing microplastic concentrations in polychaetas, copepods, and shrimp, microplastics are particularly pronounced in shrimp. Microplastics affect invertebrates, fish, birds, and mammals (Albazoni, Al-Haidarey, & Nasir, 2024). In humans, microplastics cause oxidative stress conditions, trigger inflammation, cause hormonal disruption, and increase cancer risk (Sudaryanti & Joewono, 2025).

The most popular cryptocurrencies (crypto) are Bitcoin and Ethereum. Benefits of crypto are that it eliminates barriers in international trade and currency exchange rates (Swathi, 2023). Almeida and Gonçalves (2023) reviewed the literature on crypto and noticed herding behavior, driven by market sentiment, along with irrational investors, which leads to high trading and speculative bubbles. Bitcoin peaked at over 123K in October 2025. Both Bitcoin and Ethereum are products of blockchain. The mathematics that underly blockchain include Markov chains (Zhang, 2021).

Global excess deaths due to COVID have been estimated as high as 17.7 million (Jha, Brown, & Ansumana, 2022). COVID-19 affected many aspects of life including physical health, mental health, economics, society, and policy (Boutsioli, Bigelow, & Gkounta, 2022). The epidemic revealed structural disparities in access to and quality of healthcare (Wang & Naeem, 2025). As we noted previously, COVID practically returns to zero in 2026 in the more up to date Google Trends.

4.2. Social Media

Social media has become ubiquitous, originating on platforms such as MySpace and Hyves (Utz, 2011). More recent platforms include Facebook and LinkedIn. As is this case with social interactions, social media has benefits and costs. Social media affects people in different ways (Pouwels, Beyens, Keijsers, & Valkenburg, 2025; Van der Wal, Valkenburg, & van Driel, 2024). Social media can help people network at a rapid pace, and social media profiles can both benefit and hurt job seekers. Social media has amplified bullying, polarized public opinion, and may be used to promote products as well as ideas.

4.3. Limitations

Ngram data is biased in a number of ways. Given the rapid pace of technological developments, data anchored on whole years may not be sensitive enough to study fast changing processes. The word counts are for published books only, and do not take into account conversations, texts, or posts. The data is right censored at 2022, and updates occur every four years on average. There may be optical character recognition (OCR) errors present in even the newest data release. Scientific literature is overrepresented. OCR errors may also add error variance to the years that words are tagged to (Zhang, 2015). Finally, unless special phrases are used

like "cost of war" valence is difficult to discern: is a concept popular, unpopular, or both?

5. Conclusion

Overall, our results suggest that the ngram data can be used to identify hot topics, and we hypothesize that this is also true within disciplines. Our statistical analyses suggest that the internet fueled the development of social media, and that increases in the iPhone are associated with increases in the internet and social media. Other researchers have highlighted the strengths of lagged correlations. For example, Luftensteiner, Krikova, and Rainer (2026) emphasize the use of lagged correlations in industrial settings to optimize predictive maintenance before mechanical failures occur. We hope other researchers will apply lagged correlations (and the more accurate lagged variance analysis) to the ngram data in particular, and time series data in general.

Acknowledgements

This paper was supported by NSF grant award 2030584 "Enhancing Undergraduate Enrollment, Persistence, and Graduation in Science and Mathematics", PI: Kirsten Hokeness.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Aiden, E., & Michel, J. B. (2013). *Uncharted: Big Data as a Lens on Human Culture.* Penguin.

Albazoni, H. J., Al-Haidarey, M. J. S., & Nasir, A. S. (2024). A Review of Microplastic Pollution: Harmful Effect on Environment and Animals, Remediation Strategies. *Journal of Ecological Engineering, 25,* 140-157. https://doi.org/10.12911/22998993/176802

Almeida, J., & Gonçalves, T. C. (2023). A Systematic Literature Review of Investor Behavior in the Cryptocurrency Markets. *Journal of Behavioral and Experimental Finance, 37,* Article 100785. https://doi.org/10.1016/j.jbef.2022.100785

Boutsioli, Z., Bigelow, V., & Gkounta, O. (2022). COVID-19: A Selective Short Literature Review. *Athens Journal of Health and Medical Sciences, 9,* 71-86. https://doi.org/10.30958/ajhms.9-2-1

Harlow, L. L. (2023). *The Essence of Multivariate Thinking: Basic Themes and Methods* (3rd ed.). Routledge.

Jha, P., Brown, P. E., & Ansumana, R. (2022). Counting the Global COVID-19 Dead. *The Lancet, 399,* 1937-1938. https://doi.org/10.1016/s0140-6736(22)00845-5

Luftensteiner, S., Krikova, K., & Rainer, R. (2026). OCCF—Leveraging Lagged Correlation Analysis for Enhanced Insights in Continuous Industrial Data. *Procedia Computer Science, 277,* 3655-3662. https://doi.org/10.1016/j.procs.2026.02.400

McFadden, R. H., Zywiak, W. H., Bobroff, R. P., & Niu, G. (2022). War and Money in Ngram Viewer. *Advances in Historical Studies, 11,* 188-195. https://doi.org/10.4236/ahs.2022.114016

Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P. et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331,* 176-182. https://doi.org/10.1126/science.1199644

Naisbitt, J. (1982). *Megatrends: Ten New Directions Transforming Our Lives.* Warner Books Inc.

Pouwels, J. L., Beyens, I., Keijsers, L., & Valkenburg, P. M. (2025). Changing or Stable? The Effects of Adolescents' Social Media Use on Psychosocial Functioning. *Child Development, 96,* 752-770. https://doi.org/10.1111/cdev.14207

Solovyev, V. (2024). Using the Google Books Ngram Corpus to Study Social Evolution. *Social Evolution & History, 23,* 144-164. https://doi.org/10.30884/seh/2024.02.06

Sudaryanti, L., & Joewono, H. T. (2025). Systematic Literature Review: The Presence of Microplastics in the Body and Their Impact on Human Health. *Journal of Ecohumanism, 4,* 1456-1470. https://doi.org/10.62754/joe.v4i2.6522

Swathi, B. (2023). Cryptocurrencies: A Review of Literature. *International Journal of Current Science, 13,* 974-979.

Utz, S. (2011). Social Network Site Use among Dutch Students: Effects of Time and Platform. In *Networked Sociability and Individualism: Technology for Personal and Professional Relationships* (pp. 23). Business Science Reference. https://www.igi-global.com/gateway/chapter/60494

Van der Wal, A., Valkenburg, P. M., & van Driel, I. I. (2024). In Their Own Words: How Adolescents Use Social Media and How It Affects Them. *Social Media+Society, 10,* 1-11. https://doi.org/10.1177/20563051241248591

Wang, R., & Naeem, M. A. (2025). COVID-19 Pandemic: A Comprehensive Meta-review of Global Impacts, Responses, and Future Preparedness. *The Clinical Respiratory Journal, 19,* 1-14. https://doi.org/10.1111/crj.70134

Zhang, S. (2015). The Pitfalls of Using Google Ngram to Study Language. *Wired.*

Zhang, Y. X. (2021). Blockchain Viewed from Mathematics. *Notices of the American Mathematical Society, 68,* 1740-1751. https://doi.org/10.1090/noti2365

Zywiak, W. H., & Niu, G. (2021). Love, Hope, Perspective, and Leadership in the Ngram Database: Solace for Modern Times. *Open Journal of Social Sciences, 9,* 159-166. https://doi.org/10.4236/jss.2021.911013

Zywiak, W. H., Bobroff, R. P., & Niu, G. (2021). Black Swan Years in American English, French, German, Hebrew, and Russian: Years That Reverberate in Ngram Viewer. *Advances in Historical Studies, 10,* 208-214. https://doi.org/10.4236/ahs.2021.103013