Evolutionary Psychological Science (2023) 9:385–394
https://doi.org/10.1007/s40806-023-00363-2

THEORETICAL ARTICLE

Exploring Large Digital Bodies for the Study of Human Behavior

Ulysses Paulino Albuquerque[1] · Anibal Silva Cantalice[1] · Edwine Soares Oliveira[1] · Joelson Moreno Brito de Moura[2] · Rayane Karoline Silva dos Santos[1] · Risoneide Henriques da Silva[1] · Valdir Moura Brito-Júnior[1] · Washington Soares Ferreira-Júnior[3]

Received: 27 February 2023 / Revised: 3 April 2023 / Accepted: 4 April 2023 / Published online: 23 May 2023

Abstract

Internet access has become a fundamental component of contemporary society, with major impacts in many areas that offer opportunities for new research insights. The search and deposition of information in digital media form large sets of data known as digital *corpora*, which can be used to generate structured data, representing repositories of knowledge and evidence of human culture. This information offers opportunities for scientific investigations that contribute to the understanding of human behavior on a large scale, reaching human populations/individuals that would normally be difficult to access. These tools can help access social and cultural varieties worldwide. In this article, we briefly review the potential of these *corpora* in the study of human behavior. Therefore, we propose Culturomics of Human Behavior as an approach to understand, explain, and predict human behavior using digital *corpora*.

Keywords Big data · Machine learning · Behavioral patterns · Personality traits · Digital media · Social media

Introduction

Approximately 5.3 billion people, or 66% of the world's population, have access to the Internet (ITU, 2023) and live in an era in which information has crossed temporal and spatial boundaries, allowing the world to remain connected with the aid of an increasingly intense flow of information (Valcanis, 2011). Therefore, we can infer that the Internet has become an important component of contemporary society, with great impacts in the most diverse areas that offer opportunities for new research insights (Jarić et al., 2020).

Widespread access to the Internet has allowed the search and storage of information in various digital media, forming large sets of data known as digital *corpora* (Correia et al., 2021; Leetaru, 2011; Michel et al., 2011). In turn, *corpora* are collections of items, including Internet pages, digitized books, and posts on social networks, that can be used to generate structured data, representing repositories of knowledge and/or evidence of different products of human culture (Correia et al., 2021).

This information offers opportunities for scientific investigations that contribute to the understanding of human behavior on a large scale, because it can reach individuals that research would normally have greater difficulty accessing (Gosling et al., 2010; Hargittai, 2018). Research involving digital data, which reach different societies, can help us reduce biases in human behavioral studies, since they are carried out mostly in Western, educated, industrialized, rich, and democratic (WEIRD) societies (Henrich et al., 2010).

WEIRD societies do not represent the cultural variety existing in the general population, which prevents large generalizations of human behavior (Henrich et al., 2010). One way around, this has been the creation of methods and tools

✉ Ulysses Paulino Albuquerque
upa677@hotmail.com

[1] Laboratório de Ecologia e Evolução de Sistemas Socioecológicos (LEA), Departamento de Botânica, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, Cidade Universitária, 123550670-901 Recife, Pernambuco, Brazil

[2] Instituto de Estudos do Xingu (IEX), Av. Norte Sul, Universidade Federal do Sul E Sudeste do Pará, Loteamento Cidade Nova, Lote N. 1, Qd 15, Setor 15, São Félix Do Xingu, Brazil

[3] Laboratório de Investigações Bioculturais no Semiárido, Universidade de Pernambuco, Campus Petrolina, BR203, Km 2, S/N, 56328-903 Petrolina, Pernambuco, Brazil

to measure the intercultural psychological distance[1] between different populations in WEIRD societies (Muthukrishna et al., 2020). This may allow samples from different cultures to make greater generalizations (Muthukrishna et al., 2020). However, until then, there has been no way to access all social and cultural varieties around the world truly. Culturomic tools can minimize this challenge (see, e.g., Bail, 2014) because this approach is dedicated to the collection and quantitative analysis of large *corpora* of digital data for the study of human culture (Michel et al., 2011).

From this perspective, we understand culture as any information that can be expressed through behavior and transmitted through teaching, language, and imitation, among other forms of cultural learning (Mesoudi, 2011; Richerson & Boyd, 2005). Virtual environments can be viewed as an extension of a user's social or *offline life* (Correia et al., 2017); therefore, it would be possible to explore trends and patterns that reflect human behavior. Some studies have been working within this approach, even without using the classic definition of culturomics (See Ding & Luo, 2022; Oliveira & Albuquerque, 2021); by looking at these works, we can observe the great potential that digital *corpora* have in helping us understand several phenomena related to human behavior.

For example, location data from cell phones helped in understanding urban mobility patterns and their relationship with social and health dynamics (Hassan Zadeh et al., 2019), and social networks such as Facebook, Twitter, Instagram, and Sina Weibo can be used to identify psychological characteristics and traits known as *the Big Five* (Azucar et al., 2018). Local press news from around the world can predict sociopolitical issues in different countries (Leetaru, 2011). Finally, dating sites can help identify human preferences concerning the search for romantic partners (Bergström, 2018). Thus, culturomic tools have great potential to increase the range of possibilities for the investigation of human behavior.

This opinion essay presents how culturomics have been growing as a study approach that can be used to understand human behavior and its main tools. Based on the presented scenarios and cited examples, we define culturomics of human behavior (CHB) as the approach that seeks to understand, explain, and predict human behavior from digital *corpora.*

Brief History of Culturomics

The term culturomics was first used by Michel et al. (2011) to describe cultural variation from sets of digitized texts written between 1800 and 2000, seeking to investigate lexicography, grammar evolution, and adoption of technologies, among other aspects. After this study, further efforts were made, and a new edition of this text *corpus* (*Google Books* Ngram Corpus) was conducted, with 6% of all books published (Lin et al., 2012). With the advancement of studies, the objectives have diversified, contemplating measures of cultural complexity based on linguistics in a *corpus* of texts published over the years, highlighting the cumulative aspect of human culture (Juola, 2013). On one hand, it was possible to identify cultural variation over time (see Gao et al., 2012; Petersen et al., 2012); conversely, it was possible to infer whether language suffers from political regimes (Caruana-Galizia, 2015) or social regimes (Bochkarev et al., 2014).

Back then, the *corpora* used were studied through a textual *corpus* gathered from the efforts of Michel et al. (2011) and Lin et al. (2012). However, before defining what we know today as culturomics, some studies have already been dedicated to analyzing large sets of data from the geolocation of cell phones to observe patterns of human mobility (González et al., 2008). Geolocation data from cell phones make it possible to identify patterns of movement and trips made by people and predict how human behavior affects the dynamics of epidemics. This is because mobility is a crucial factor in the spread of diseases, favoring the confrontation of these public health crises (Balcan et al., 2009; Song et al., 2010).

In the midst of this horizon of possibilities, conservation culturomics have emerged, which consist of analyzing digital data generated to provide new insights into human-nature interactions aimed at biodiversity conservation (Ladle et al., 2016). Conservation culturomics differ from iEcology because the latter studies ecological processes through online data (Jarić et al., 2020), while the former studies aspects of human culture and the human/nature relationship (Jarić et al., 2021). Conservation culturomics have been increasingly notable since their proposition, with the aim of investigating how the public interest can contribute to conservation through different approaches. These approaches include research on perceptions of national parks (Bhatt & Pickering, 2021) and people's thinking about specific animal species (Pickering & Norman, 2020). As more approaches have emerged beyond the books gathered on Google Ngram Viewer, other data *corpora* have become the focus of interest. This includes posts on social networks such as Instagram (Kroetz et al., 2021), Facebook (Altay et al., 2022), Twitter (Bhatt & Pickering, 2021), news available on digital platforms (Cooper et al., 2019; Francis et al., 2019), and even the association of data from different platforms, such as social networks and searches on research sites such as Wikipedia (Fernández-Bellon & Kane, 2020).

Digital Bodies and Big Data

At the origin of culturomics, themes such as big data, web scraping, machine learning, and artificial intelligence were not very evident, except for areas dedicated to information

[1] Intercultural psychological distance is understood as the size of the difference in psychology between different societies. For example, on how to perform such a measurement, see Muthukrishna et al. (2020).

technology and computing. The vast majority of digital bodies are formed by voluminous datasets that grow rapidly and that cannot be processed in the traditional way (Chen et al., 2014). Given the huge volume of data generated, for collection to be fast, accurate, and efficient, powerful tools are needed, such as web scraping. This comprises the procedure of extracting data from the web in an automated way without the need to manually copy or download them to a hard disk (Singrodia et al., 2019).

The culturomic approach can be further improved using machine learning (ML), which consists of sets of protocols that allow computers to automatically solve a class of tasks and continuously improve problem-solving based on performance measures (Janiesch et al., 2021; LeCun et al., 2015). Oliveira and Albuquerque (2021) used web scraping and machine learning to understand the dynamics behind the dissemination of messages with false information (fake news) on Twitter in the context of the COVID-19 pandemic. Heras-Pedrosa et al. (2020) utilized web scraping technique to analyze communication in the field of public health during the COVID-19 pandemic and recorded emotions generated in the population through data from Twitter, YouTube, Instagram, official press sites, and Internet forums in real time. This highlighted the potential of using multiple corpora for the same study.

In addition, advances in machine learning are important for the advancement of scientific practice in many areas. For example, in a study by Bae et al. (2021), ML was used to detect possible traces of schizophrenia in the posts on the Reddit forum aggregator. Chiong et al. (2021) used ML to track posts with depressive tendencies on social networks such as Facebook and Twitter.

One of the most recent and prominent techniques among culturomic methodologies is the natural language learning processing (NLP), which consists of machine learning that uses artificial intelligence to allow computers to read and interpret information from texts (Arbieu et al., 2021; Thessen et al., 2012). This technique is increasingly being used to process, analyze, and monitor trends in large volumes of digital data, generating deep insights and reducing human work time. In a study by Arbieu et al. (2021), for example, this technique was applied to perform automatic analysis of emotions in the textual content of news publications about the reinsertion of wolves (*Canis lupus*) in the region of Saxony, eastern Germany. From the expansion of the *corpora* used (social networks and online newspapers, among others), the use of these tools was optimized, as they allowed the exploration of these new sets of digital data, such as those arising from social networks and search engines.

Thus, these studies can be divided into two dimensions. The first concerns the content present in the corpus, examining changes in writing patterns, the frequency of specific terms, and the identification of motivations underlying their usage in a specific space–time context. This allowed for the detection of human cultural changes and trends through the quantitative analysis of words.

The second dimension seeks to understand people's engagement with elements of digital *corpora*, such as searches for a particular term on the Internet, views in videos and images, comments, likes, and shares. This dimension has been widely used in conservation culturomics. For example, Ladle et al. (2016) found that data from social network posts and searches for certain terms, which are two-dimensional data, can help identify unexplored conservation

Table 1 Books on tools, applications, and practices for using digital *corpora*

Title	Summary	Author
Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results	The book addresses the importance of using Big Data in the contemporary world, bringing methods and information on how large corporations (Netflix, Airbnb, Facebook, Microsoft, among others) use this information in their practices	(Marr, 2016)
Mining the social web	The book works as a practical guide on how to analyze social graphs, explore social network data, implement metadata, among other guidelines that can help the use of digital *corpora*	(Russell, 2014)
Data mining: practical machine learning tools and techniques	In this book, the reader will have access to a thorough grounding in machine learning concepts and practical advice on applying machine learning tools and techniques in real-world data mining situations	(Witten et al., 2011)
Analytics for big data	This e-book addresses in a practical way some of the main techniques and tools used in the analysis provided by the integrated use of analytics and big data	(Padilha et al., 2021)
Cultural evolution in the digital age	This fantastic book explores the application of cultural evolution studies to digital media	(Acerbi, 2020)

emblems and assess the cultural impact of conservation actions, such as the selection of an endemic animal as a mascot for a sporting event. The authors argue that through these data, it is possible to assess the quality of cultural ecosystem services and monitor how these services reach people (CICES, 2023).

Table 1 shows some texts about tools that can help researchers use *corpora*. *Corpora* are rich reservoirs of human culture, which can help us understand various scientific questions.

Investigations on Human Behavior from Digital Corpora

Every day, people worldwide use the Internet to search, shop, and share part of their lives through social networks, making the Internet a significant component in various aspects of contemporary society (Mora-Rivera & García-Mora, 2021). For example, if the content generated on the Internet is a reflection of everyday life (Correia et al., 2017), the data that comes from this content can help understand more complex phenomena. Although many studies that use culturomic methodologies do not have their themes focus on understanding human behavior, they provide clues about how using culturomics can be useful for several areas of knowledge that are dedicated to understanding it. In this section, we organize a synthesis of published works involving large digital corpora, which can offer insights for this theme (Table 2).

Therefore, we argue that several fields of knowledge seeking to investigate human behavior can take advantage of the potential demonstrated by culturomics. For example, studies have shown that from a dataset built on the basis of world news, it would have been possible to predict various political events, such as revolutions, stabilities, and even decisions

Table 2 Examples of research that used digital corpora to assess human behavior

Source of the corpora	Goals	References
Google Ngram	Check the variation of cultural complexity over time; identify trends in conscious and unconscious behavior	(Juola, 2013) (Dönmez, 2020)
Google Trends	Forecast oil consumption trends and values. Identify patterns in the search for words that reflect mental suffering (depression and suicide)	(Yu et al., 2019) (Stańdo et al., 2023)
Facebook	Identify personality variations and sociodemographic characterization from Facebook data	(Schwartz et al., 2013; Celli et al., 2014)
	Analyze online news consumption and engagement on Facebook during the COVID-19 pandemic	(Altay et al., 2022)
Instagram	Understanding tourist preferences for nature-based experiences in protected areas	(Hausmann et al., 2017)
	Identify suspected, counterfeit, and unapproved health products related to COVID-19	(Mackey et al., 2020)
Reports (Newspapers and Digital News)	To analyze whether the size of marine fish perceived in popular media corresponds to the actual size	(Francis et al., 2019)
	Predict revolutions and political conflicts	(Leetaru, 2011)
Twitter	Draw a disease occurrence map	(Young et al., 2014)
	Investigate the role toxicity plays in online discourse about face mask use	(Pascual- Ferrá et al., 2022)
	Check whether digital media reflect users' environmental awareness	(Fernández-Bellon & Kane, 2020)
YouTube	Check if videos in digital media could be used as a learning tool	(Rahman et al., 2021)
	Observe the occurrence of sport hunting in Brazil, its impacts, which game species are being affected, and their occurrence biomes	(El Bizri et al., 2015)
Flickr	Sort preferred travel destinations within a region	(Önder, 2017)
	Assess the vulnerability of places frequented by tourists	(Hale, 2018)
Relationship websites	Analyze preferences in partner choices on online search engines	(Bergström, 2018)
		(van Berlo & Ranzini, 2018) (Ting & McLachlan, 2022)
Wikipedia	Predict movie success based on digital media searches	(Mestyán et al., 2013)
	Observe the occurrence and intensity of diseases based on searching databases	(Sciascia & Radin, 2017)
	Check if natural history films are more interesting than conservation messages	(Nolan et al., 2022)

by state leaders (Leetaru, 2011). Moreover, evidence shows that such datasets can be an important foundation for understanding human preferences, based on cultural salience (the frequency of a given population characteristic), as a metric of visibility or interest (Correia et al., 2016), for example, assessing whether the body size and charisma of groups of species (amphibians, birds, mammals, and reptiles) influences their conservation (Berti et al., 2020).

Inferences regarding human mobility have also been made. Gonzalez et al. (2008), for example, analyzed data referring to the trajectory of 100,000 (anonymous) cell phone users, and observed that human trajectories have great temporal and spatial regularity. Studies such as this one help, for example, in urban planning and creation of strategies to address the spread of diseases. In this sense, the pattern of human mobility can be highly predictable, that is, people tend to frequently go to the same places and follow the same routes (Song et al., 2010). These results emphasize that the use of predictive models to understand urban mobility phenomena is not only possible, but also accurate (Song et al., 2010).

Efforts have also been made to understand how human behavior affects the dynamics of epidemics, mainly because human mobility is a crucial factor in the spread of diseases. Data from 29 countries worldwide were used for computational modeling of infectious diseases, opening the way for the development of necessary and accurate models for describing and, consequently, coping with epidemics (Balcan et al., 2009). Another aspect is that online behavior may present a tendency that is analogous to herd behavior becoming more collective, for example, in scenarios of risk to public health (Bentley et al., 2014). This makes online data an important tool for understanding human attitudes and actions during disease outbreaks.

Additionally, data from dating sites can be used as tools to investigate certain aspects of choosing romantic partners, as most relationships that start online do not differ much from those formed in other contexts (Bergström, 2018). For example, men have more initiative in initiating contact than women, and preferences regarding the age of partners can vary between genders in different age groups (Bergström, 2018). Furthermore, dating sites can be valuable sources of data on how people behave in situations of infectious disease outbreaks, where social isolation is recommended.

Recent studies have shown a change in sexual behavior during outbreaks of infectious diseases such as COVID-19, in which Chinese men and women aged between 18 and 45 years showed a decrease in the number of romantic partners and sexual frequency during the pandemic (Li et al., 2020). Conversely, virtual contact with potential sexual partners can be increased in frequency during this period (Seitz et al., 2020). Additionally, during the COVID-19 pandemic, data from dating apps gained more than 1.5 million daily users (Ting & McLachlan, 2022). Data from these apps also helped to outline the main profile of users, showing that being young, being single, and having higher levels of stress were predictors of greater app use (Ting & McLachlan, 2022).

Several studies have focused on understanding the characteristics that define individuals or groups within a sociocultural context based on spheres of human behavior linked to gender (Seewann et al., 2022), age (Agbo-Ajala et al., 2022), and personality traits (Azucar et al., 2018). For example, Schwartz et al. (2013) used big data tools to recover messages posted on Facebook and verified whether the language used in the posts reflected the personality, gender, and age of interlocutors. Women tend to be more affectionate, and men were more objective and possessive; language changes with advancing age, such as changing from a more singular "I" communication to plural "we" questions, and the propensity to use certain words is modified depending on the personality of the analyzed groups. For example, people with more outgoing personalities mentioned words related to greater sociability, such as "party," "love you," and "boys," while more introverted people mentioned words related to more solitary activities, such as "computer," "reading," and "Internet" (Schwartz et al., 2013).

Political positioning has also been investigated. For example, Twitter data were used to show that the flow of political information on this network is controlled by a limited number of influencers (Casero-Ripollés, 2021). Facebook data were also analyzed to see how different political candidates communicated with civil society (Caton et al., 2015). These aspects are important to analyze as one of the ways of aggregating people and/or groups today is through their affinity with different political parties.

Since these parties are formed by individuals to represent their beliefs and values in a political scenario, we can infer that they reflect the personality characteristics, thoughts, and ideologies of their members (Jost et al., 2014; Cohen, 2003). For example, partisan inclination influences the adoption of sanitary measures during public health crises (Gollwitzer et al., 2020). Although sex differences should be considered in these studies, as in the fight against COVID-19, evidence suggests that female leaders seek to minimize the impact of the virus, whereas male leaders implement risky short-term decisions to avoid harm to the economy (Luoto & Varella, 2021).

Several studies have investigated the relationship between digital media and human personality traits (Schwartz et al., 2013), as evidenced by Azucar et al. (2018). These authors showed that the way users interact on social media, such as profile privacy, language, age, gender, comments, and likes, can reflect many personality traits such as the positive link between extroversion and engagement on social media (Blackwell et al., 2017). Other studies also sought to understand through tweets that feelings are more prominent in environmental contexts hitherto unknown to the user, such as the COVID-19 pandemic, showing that fear was the most prominent feeling (Xue et al., 2020).

Another target of investigation was human morality, which is based on the salience of words related to moral (e.g., virtue, decency, and conscience) and virtuous (e.g., honesty, patience, and compassion) behaviors in digital corpora. With data from Google Ngram Viewer, Kesebir and Kesebir (2012) noticed a significant decline of these words in American books during the twentieth century, which for the authors would be linked to the disappearance of these concepts in public debate throughout the construction of modern history.

The way in which human beings relate to aspects of contagion and immunization against diseases can also be accessed and investigated through culturomics. For example, Young et al. (2014) used georeferencing of *tweets* related to HIV and the incidence maps of AIDS cases (https://aidsvu.org/), revealing a spatial correlation between publications and reported cases. Interactions in this sense (occurrence of tweets and occurrence of disease by the United States Centers for Disease Control and Prevention—CDC) have also been observed for other infectious diseases, such as *influenza* (Broniatowski et al., 2013) and flu (Hassan Zadeh et al., 2019).

Besides monitoring where the public interest is concentrated, efforts have been made to assess whether it is possible to change social attitudes toward environmental crises. From an association of data from Twitter and Wikipedia, to analyze engagement and searches on environmental crises, it was observed that people's involvement was greater after watching natural history films (Fernández-Bellon & Kane, 2020). Thus, certain digital resources can play an important role in creating connections with the natural world (Fernández-Bellon & Kane, 2020). Data from Google Trends were used to compare awareness of climate change in certain countries and the actual risk of impacts, which is necessary to identify countries where improving or adapting policies to face climate change are needed (Archibald & Butt, 2018).

The conservation culturomics approach, which is in increasing prominence and is discussed throughout this text, offers important perspectives for nature conservation, although it was not conceived as a specific discipline to study human behaviors. This approach recognizes the role of the public interest as an ally for nature conservation actions, as mentioned in previous studies (Ladle et al., 2016; Nghiem et al., 2016; Ladle et al., 2019). However, it is important to emphasize that the behavioral factors that drive the adoption of pro-conservation behaviors have not yet been adequately investigated.

Limitations of the Culturomic Approach

Although the use of digital corpora is a possibility for human behavior research, data collection, analysis, and interpretation of results need to be done with caution due to several sources of bias (Griffin et al., 2020; Tufekci, 2014). For example, information may be salient in digital media, even without a greater demand from the community. This can occur for two reasons: (1) artificial, when using programs and/or transmission lists, such as *bots* (Liu, 2019) and *spam* (Wang et al., 2012), and (2) natural, when a human manually inflates certain information, such as crowdturfing (Wang et al., 2012) and fake accounts (Shen et al., 2014). All of these options end up overvaluing information that is not of interest to a group or society, which can create social problems in the *offline world* (Bovet & Makse, 2019; Cantarella et al., 2023).

Furthermore, the motivation for choosing the *corpus* is often neglected during the investigation, as some research has shown that socioeconomic factors are highly discrepant between different social networks. For example, most users of networks such as Snapchat, Instagram, and TikTok are young people aged between 18 and 29 years (Pew Research Center, 2022) and have higher levels of education (Hargittai, 2018). That is, when using these networks as digital *corpora* of studies, caution is needed, especially when making large generalizations. For example, Mislove et al. (2021), when comparing a sample of US and Twitter audiences based on socioeconomic factors (geographical, race, and gender), the study observed that Twitter audiences did not represent the region's population. In addition to socioeconomic issues, it is important to consider the affinity of each platform with a certain type of content, because although many platforms allow the posting of text and photos, the public tends to prefer a specific type of media as a model (Di Minin et al., 2013, 2015).

Additionally, some researchers have noted that the use of big data must be associated with other methodologies, such as data incorporation or analysis. *Corpus* association can better predict some outcomes, data validation (e.g., interviews) (Azucar et al., 2018), and the presence of outliers within the sample (Griffin et al., 2020). Another way pointed out is the observation of the structure of the collected data, which sometimes does not allow for conventional analyses (for more details, see Dodds et al., 2011; Xue et al., 2020). For example, Koplenig (2017) pointed out statistical errors in the results in several articles that disregard the temporal characteristics of the data when testing their hypotheses. That is, observations that are close in time tend to be more similar than distant observations. Although it seems that these biases can make research with culturomics unfeasible, observing the biases already indicated can greatly minimize the risks of misinterpretation (Ruths & Pfeffer, 2014).

Conclusion

In short, the Internet has become a fundamental element of contemporary society, allowing the creation of large datasets that can be used to study and understand human behavior on a large scale. This information enables

scientific investigations that reach audiences who are normally difficult to reach and provides research opportunities in several areas. The culturomic approach to human behavior seeks to understand, explain, and predict human behavior using these digital corpora. With the constant increase in the volume of data, powerful tools, such as web scraping, are needed to collect and process this information. Therefore, the use of digital corpora is a rapidly developing area of research offering opportunities for new insights in several fields.

CHB is an innovative approach aimed at analyzing large cultural datasets, particularly social media data, to understand human behavior on a global scale. This approach is broader and more quantitative, emphasizing large-scale data analysis. While cross-cultural psychology explores the mind and behavior of individuals across different cultures, the data collected is primarily individual through interviews (Broesch et al., 2020).

It is important to note that CHB is more akin to historical psychology than to cross-cultural psychology, as historical psychology also conducts large-scale textual analyses (Muthukrishna et al., 2021). However, we argue that CHB should be considered a distinct field that dialogues with other areas mentioned earlier, given the specific nature of the analyses and theories involved in data collection and analysis.

Therefore, we can conclude that CHB is a promising approach for understanding human behavior on a global scale. While it may share some similarities with other disciplines, it is a field with its own characteristics and methodologies that deserve to be studied independently.

Author Contribution UPA: conceptualization, formal analysis, project administration, writing (original draft, review and editing). ASC, ESO, VMBJ, WSFJ: formal analysis, writing (original draft, review and editing). JMB, RKSS, RHS: writing (review and editing).

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

Acerbi, A. (2020). *Cultural Evolution in the Digital Age*. Oxford University Press.

Agbo-Ajala, O., Viriri, S., Oloko-Oba, M., Ekundayo, O., & Heymann, R. (2022). Apparent age prediction from faces: A survey of modern approaches. *Frontiers in Big Data, 5*, 1025806. https://doi.org/10.3389/fdata.2022.1025806

Altay, S., Nielsen, R. K., & Fletcher, R. (2022). Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media*, *2*. https://doi.org/10.51685/jqd.2022.020

Arbieu, U., Helsper, K., Dadvar, M., Mueller, T., & Niamir, A. (2021). Natural Language Processing as a tool to evaluate emotions in conservation conflicts. *Biological Conservation*, *256*, 109030. https://doi.org/10.1016/j.biocon.2021.109030

Archibald, C. L., & Butt, N. (2018). Using Google search data to inform global climate change adaptation policy. *Climatic Change, 150*(3), 447–456. https://doi.org/10.1007/s10584-018-2289-9

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Bae, Y. J., Shim, M., & Lee, W. H. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors*, *21*(17), Art. 17. https://doi.org/10.3390/s21175924

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society, 43*(3), 465–482. https://doi.org/10.1007/s11186-014-9216-5

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, *106*(51), Art. 51. https://doi.org/10.1073/pnas.0906910106

Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences, 37*(1), 63–76. https://doi.org/10.1017/S0140525X13000289

Bergström, M. (2018). De quoi l'écart d'âge est-il le nombre? L'apport des big data à l'étude de la différence d'âge au sein des couples. *Revue Française De Sociologie, 59*(3), 395–422. https://doi.org/10.3917/rfs.593.0395

Berti, E., Monsarrat, S., Munk, M., Jarvie, S., & Svenning, J.-C. (2020). Body size is a good proxy for vertebrate charisma. *Biological Conservation*, *251*, 108790. https://doi.org/10.1016/j.biocon.2020.108790

Bhatt, P., & Pickering, C. M. (2021). Public perceptions about Nepalese National Parks: A global Twitter discourse analysis. *Society & Natural Resources, 34*(6), 685–702. https://doi.org/10.1080/08941920.2021.1876193

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences, 116*, 69–72. https://doi.org/10.1016/j.paid.2017.04.039

Bochkarev, V., Solovyev, V., & Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, *11*(101), Art. 101. https://doi.org/10.1098/rsif.2014.0841

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications, 10*(1), 7. https://doi.org/10.1038/s41467-018-07761-2

Broesch T., Crittenden, A.N., Beheim, B.A., Blackwell, A.D., Bunce, J.A., Colleran, H., Hagel, K., Kline, M., McElreath, R., Nelson, R.G., Pisor, A.C., Prall, S., Pretelli, I., Purzycki, B., Quinn, E.A., Ross, C., Scelza, B., Starkweather, K., Stieglitz, J., & Mulder, M.B. (2020). Navigating cross-cultural research: Methodological and ethical considerations. *Proceedings of the Royal Society* B.287202012452020124. https://doi.org/10.1098/rspb.2020.1245

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS One, 8*(12), e83672. https://doi.org/10.1371/journal.pone.0083672

Cantarella, M., Fraccaroli, N., & Volpe, R. (2023). Does fake news affect voting behaviour? *Research Policy, 52*(1), 104628. https://doi.org/10.1016/j.respol.2022.104628

Caruana-Galizia, P. (2015). Politics and the German language: Testing Orwell's hypothesis using the Google N-Gram corpus. *Digital Scholarship in the Humanities*, *31*(3), Art. 3. https://doi.org/10.1093/llc/fqv011

Casero-Ripollés, A. (2021). Influencers in the political conversation on Twitter: Identifying digital authority with big data. *Sustainability, 13*(5), 2851. https://doi.org/10.3390/su13052851

Caton, S., Hall, M., & Weinhardt, C. (2015). How do politicians use Facebook? An applied social observatory. *Big Data & Society, 2*(2), 2053951715612822. https://doi.org/10.1177/2053951715612822

Celli, F., Bruni, E., & Lepri, B. (2014). Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia*. https://doi.org/10.1145/2647868.2654977

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), Art. 2. https://doi.org/10.1007/s11036-013-0489-0

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, *135*, 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

CICES. (2023). *Common International Classification of Ecosystem Services*. https://cices.eu/

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Cooper, M. W., Di Minin, E., Hausmann, A., Qin, S., Schwartz, A. J., & Correia, R. A. (2019). Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement. *Biological Conservation, 230*, 29–36. https://doi.org/10.1016/j.biocon.2018.12.004

Correia, R. A., Jepson, P., Malhado, A. C. M., & Ladle, R. J. (2017). Internet scientific name frequency as an indicator of cultural salience of biodiversity. *Ecological Indicators, 78*, 549–555. https://doi.org/10.1016/j.ecolind.2017.03.052

Correia, R. A., Jepson, P. R., Malhado, A. C. M., & Ladle, R. J. (2016). Familiarity breeds content: Assessing bird species popularity with culturomics. *PeerJ*, *4*, e1728. https://doi.org/10.7717/peerj.1728

Correia, R. A., Ladle, R., Jarić, I., Malhado, A. C. M., Mittermeier, J. C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R., & Di Minin, E. (2021). Digital data sources and methods for conservation culturomics. *Conservation Biology*, *35*(2), Art. 2. https://doi.org/10.1111/cobi.13706

Di Minin, E., Fraser, I., Slotow, R., & MacMillan, D. C. (2013). Understanding heterogeneous preference of tourists for big game species: Implications for conservation and management. *Animal Conservation, 16*(3), 249–258. https://doi.org/10.1111/j.1469-1795.2012.00595.x

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, *3*. https://www.frontiersin.org/articles/https://doi.org/10.3389/fenvs.2015.00063

Ding, Q., & Luo, X. (2022). People with high perceived infectability are more likely to spread rumors in the context of COVID-19: A behavioral immune system perspective. *International Journal of Environmental Research and Public Health, 20*(1), 703. https://doi.org/10.3390/ijerph20010703

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One, 6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Dönmez, İ. (2020). Analyzing five conscious and unconscious behaviors using Google n-gram database generated from millions of books. In *2020 5th International Conference on Computer Science and Engineering (UBMK)* (pp. 19–24). Presented at the 2020 5th International Conference on Computer Science and Engineering (UBMK). https://doi.org/10.1109/UBMK50275.2020.9219540

El Bizri, H. R., Morcatty, T. Q., Lima, J. J. S., & Valsecchi, J. (2015). The thrill of the chase: uncovering illegal sport hunting in Brazil through YouTube posts. *Ecology and Society, 20*(3). https://doi.org/10.5751/es-07882-200330

Fernández-Bellon, D., & Kane, A. (2020). Natural history films raise species awareness—A big data approach. *Conservation Letters*, *13*(1), e12678. https://doi.org/10.1111/conl.12678

Francis, F. T., Howard, B. R., Berchtold, A. E., Branch, T. A., Chaves, L. C. T., Dunic, J. C., Favaro, B., Jeffrey, K. M., Malpica-Cruz, L., Maslowski, N., Schultz, J. A., Smith, N. S., & Côté, I. M. (2019). Shifting headlines? Size trends of newsworthy fishes. *PeerJ*, *7*, e6395. https://doi.org/10.7717/peerj.6395

Gao, J., Hu, J., Mao, X., & Perc, M. (2012). Culturomics meets random fractal theory: Insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface*, *9*(73), Art. 73. https://doi.org/10.1098/rsif.2011.0846

Gollwitzer, A., Martel, C., Brady, W. J., Pärnamets, P., Freedman, I. G., Knowles, E. D., & Van Bavel, J. J. (2020). Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour, 4*(11), 1186–1197. https://doi.org/10.1038/s41562-020-00977-7

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), Art. 7196. https://doi.org/10.1038/nature06958

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences*, *33*(2–3), Art. 2–3. https://doi.org/10.1017/s0140525x10000300

Griffin, G. P., Mulhall, M., Simek, C., & Riggs, W. W. (2020). Mitigating bias in big data for transportation. *Journal of Big Data Analytics in Transportation, 2*(1), 49–59. https://doi.org/10.1007/s42421-020-00013-0

Hale, B. W. (2018). Mapping potential environmental impacts from tourists using data from social media: A case study in the Westfjords of Iceland. *Environmental Management, 62*(3), 446–457. https://doi.org/10.1007/s00267-018-1056-z

Hargittai, E. (2018). Potential Biases in Big Data: Omitted Voices on social media. *Social Science Computer Review*, *38*(1), Art. 1. https://doi.org/10.1177/0894439318788322

Hassan Zadeh, A., Zolbanin, H. M., Sharda, R., & Delen, D. (2019). Social media for nowcasting flu activity: Spatio-temporal big data analysis. *Information Systems Frontiers*, *21*(4), Art. 4. https://doi.org/10.1007/s10796-018-9893-0

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2017). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters, 11*(1), e12343. https://doi.org/10.1111/conl.12343

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83–135. https://doi.org/10.1017/S0140525X0999152X

Heras-Pedrosa, C., Sánchez-Núñez, P., & Peláez, J. I. (2020). Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. *International Journal of Environmental Research and Public Health*, *17*(15), Art. 15. https://doi.org/10.3390/ijerph17155542

ITU. (2023). *ITU-D ICT Statistics*. https://www.itu.int/itu-d/sites/statistics/

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), Art. 3. https://doi.org/10.1007/s12525-021-00475-2

Jarić, I., Bellard, C., Correia, R. A., Courchamp, F., Douda, K., Essl, F., Jeschke, J. M., Kalinkat, G., Kalous, L., Lennox, R. J., Novoa, A., Proulx, R., Pyšek, P., Soriano-Redondo, A., Souza, A. T., Vardi, R., Veríssimo, D., & Roll, U. (2021). Invasion culturomics and iEcology. *Conservation Biology, 35*(2), 447–451. https://doi.org/10.1111/cobi.13707

Jarić, I., Roll, U., Arlinghaus, R., Belmaker, J., Chen, Y., China, V., Douda, K., Essl, F., Jähnig, S. C., Jeschke, J. M., Kalinkat, G., Kalous, L., Ladle, R., Lennox, R. J., Rosa, R., Sbragaglia, V., Sherren, K., Šmejkal, M., Soriano-Redondo, A., … Correia, R. A. (2020). Expanding conservation culturomics and iEcology from terrestrial to aquatic realms. *PLOS Biology*, *18*(10), Art. 10. https://doi.org/10.1371/journal.pbio.3000935

Jost, J. T., Nam, H. H., Amodio, D. M., & Van Bavel, J. J. (2014). Political neuroscience: The beginning of a beautiful friendship. *Political Psychology, 35*, 3–42. https://doi.org/10.1111/pops.12162

Juola, P. (2013). Using the Google N-Gram corpus to measure cultural complexity. *Literary and Linguistic Computing*, *28*(4), Art. 4. https://doi.org/10.1093/llc/fqt017

Kesebir, P., & Kesebir, S. (2012). The cultural salience of moral character and virtue declined in twentieth century America. *The Journal of Positive Psychology, 7*(6), 471–480. https://doi.org/10.1080/17439760.2012.715182

Koplenig, A. (2017). Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digital Scholarship in the Humanities,* fqv030. https://doi.org/10.1093/llc/fqv030

Kroetz, A. M., Brame, A. B., Bernanke, M., McDavitt, M. T., & Wiley, T. R. (2021). Tracking public interest and perceptions about smalltooth sawfish conservation in the USA using Instagram. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *31*(10), Art. 10. https://doi.org/10.1002/aqc.3680

Ladle, R., Jepson, P., Correia, R., & Malhado, A. (2019). A culturomics approach to quantifying the salience of species on the global internet. *People and Nature, 1*, 1–9. https://doi.org/10.1002/pan3.10053

Ladle, R. J., Correia, R. A., Do, Y., Joo, G.-J., Malhado, A. C., Proulx, R., Roberge, J.-M., & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment, 14*(5), 269–275. https://doi.org/10.1002/fee.1260

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Art. 7553. https://doi.org/10.1038/nature14539

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*. https://doi.org/10.5210/fm.v16i9.3663

Li, W., Li, G., Xin, C., Wang, Y., & Yang, S. (2020). Challenges in the practice of sexual medicine in the time of COVID-19 in China. *The Journal of Sexual Medicine, 17*(7), 1225–1228. https://doi.org/10.1016/j.jsxm.2020.04.380

Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books NGram Corpus. *Proceedings of the ACL 2012 System Demonstrations*, 169–174. https://aclanthology.org/P12-3029

Liu, X. (2019). A big data approach to examining social bots on Twitter. *Journal of Services Marketing, 33*(4), 369–379. https://doi.org/10.1108/jsm-02-2018-0049

Luoto, S., & Varella, M. A. C. (2021). Pandemic leadership: Sex differences and their evolutionary–developmental origins. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.633862

Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., et al. (2020). Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram. *JMIR Public Health and Surveillance, 6*(3), e20794. https://doi.org/10.2196/20794

Marr, B. (2016). *Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results*. John Wiley & Sons.

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS One, 8*(8), e71226. https://doi.org/10.1371/journal.pone.0071226

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), Art. 6014. https://doi.org/10.1126/science.1199644

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2021). Understanding the demographics of Twitter users. *Proceedings of the International AAAI Conference on Web and Social Media, 5*(1), 554–557. https://doi.org/10.1609/icwsm.v5i1.14168

Mora-Rivera, J., & García-Mora, F. (2021). Internet access and poverty reduction: Evidence from rural and urban Mexico. *Telecommunications Policy*, *45*(2), 102076. https://doi.org/10.1016/j.telpol.2020.102076

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science, 31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology, 72*, 717–749. https://doi.org/10.1146/annurev-psych-082820-111436

Nghiem, L. T. P., Papworth, S. K., Lim, F. K. S., & Carrasco, L. R. (2016). Analysis of the capacity of Google trends to measure interest in conservation topics and the role of online news. *PLoS One, 11*(3), e0152802. https://doi.org/10.1371/journal.pone.0152802

Nolan, G., Kane, A., & Fernández-Bellon, D. (2022). Natural history films generate more online interest in depicted species than in conservation messages. *People and Nature, 4*(3), 816–825. https://doi.org/10.1002/pan3.10319

Oliveira, D. V. B., & Albuquerque, U. P. (2021). Cultural evolution and digital media: Diffusion of fake news about COVID-19 on Twitter. *SN Computer Science, 2*(6), 430. https://doi.org/10.1007/s42979-021-00836-w

Önder, I. (2017). Classifying multi-destination trips in Austria with big data. *Tourism Management Perspectives, 21*, 54–58. https://doi.org/10.1016/j.tmp.2016.11.002

Padilha, J., Soares, J. A., Alves, N. S. R., Abreu, E. M., Silva, F. R., Morais, M. S. de F., Lacerda, P. S. P., Neto, R. M., & Machado, V. de A. (2021). *Analytics para Big Data*. SAGAH.

Pascual-Ferrá, P., Alperstein, N., & Barnett, D. J. (2022). Social network analysis of COVID-19 public discourse on Twitter: Implications for risk communication. *Disaster Medicine and Public Health Preparedness, 16*(2), 561–569. https://doi.org/10.1017/dmp.2020.347

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*(1), Art. 1. https://doi.org/10.1038/srep00313

Pew Research Center. (2022). *Pew Research Center*. Pew Research Center. https://www.pewresearch.org/

Pickering, C., & Norman, P. (2020). Assessing discourses about controversial environmental management issues on social media: Tweeting about wild horses in a national park. *Journal of Environmental Management*, *275*, 111244. https://doi.org/10.1016/j.jenvman.2020.111244

Rahman, N. A., Ng, H. J. H., & Rajaratnam, V. (2021). Big data analysis of a dedicated YouTube channel as an open educational resource in hand surgery. *Frontiers in Applied Mathematics and Statistics, 7*. https://doi.org/10.3389/fams.2021.593205

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Russell, M. A. (2014). *Mining the Social Web* (2° ed). O'Reilly Media, Inc.,

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One, 8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Sciascia, S., & Radin, M. (2017). What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. *International Journal of Medical Informatics, 107*, 65–69. https://doi.org/10.1016/j.ijmedinf.2017.09.002

Seewann, L., Verwiebe, R., Buder, C., & Fritsch, N.-S. (2022). "Broadcast your gender." A comparison of four text-based classification methods of German YouTube channels. *Frontiers in Big Data, 5*, 908636. https://doi.org/10.3389/fdata.2022.908636

Seitz, B. M., Aktipis, A., Buss, D. M., Alcock, J., Bloom, P., Gelfand, M., Harris, S., Lieberman, D., Horowitz, B. N., Pinker, S., Wilson, D. S., & Haselton, M. G. (2020). The pandemic exposes human nature: 10 evolutionary insights. *Proceedings of the National Academy of Sciences, 117*(45), 27767–27776. https://doi.org/10.1073/pnas.2009787117

Shen, Y., Yu, J., Dong, K., & Nan, K. (2014). *Automatic fake followers detection in Chinese micro-blogging system* (pp. 596–607). Springer International Publishing. https://doi.org/10.1007/978-3-319-06605-9_49

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications. *International Conference on Computer Communication and Informatics (ICCCI), 2019*, 1–6. https://doi.org/10.1109/ICCCI.2019.8821809

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), Art. 5968. https://doi.org/10.1126/science.1177170

Stańdo, J., Fechner, Ż, Gmitrowicz, A., Andriessen, K., Krysinska, K., & Czabański, A. (2023). Increase in Search interest for "suicide" and "depression" for particular days of the week and times of day: Analysis based on Google trends. *Journal of Clinical Medicine, 12*(1), 191. https://doi.org/10.3390/jcm12010191

Thessen, A. E., Cui, H., & Mozzherin, D. (2012). Applications of natural language processing in biodiversity science. *Advances in Bioinformatics, 2012*, 1–17. https://doi.org/10.1155/2012/391574

Ting, A. E., & McLachlan, C. S. (2022). Intimate relationships during COVID-19 across the genders: An examination of the interactions of digital dating, sexual behavior, and mental health. *Social Sciences, 11*(7), 297. https://doi.org/10.3390/socsci11070297

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 505–514. https://doi.org/10.1609/icwsm.v8i1.14517

Valcanis, T. (2011). *An iPhone in every hand: Media ecology, communication structures, and the global village*. https://www.semanticscholar.org/paper/An-iPhone-in-Every-Hand%3A-Media-Ecology%2C-Structures%2C-Valcanis/55495f57c43e325ae681514c51e567fad6b6e4db

van Berlo, Z. M. C., & Ranzini, G. (2018). Big dating: A computational approach to examine gendered self-presentation on tinder. In *Proceedings of the 9th International Conference on Social Media and Society* (pp. 390–394). Presented at the SMSociety '18: International Conference on Social Media and Society, Copenhagen Denmark: ACM. https://doi.org/10.1145/3217804.3217951

Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., & Zhao, B. Y. (2012). Serf and turf. *Proceedings of the 21st international conference on World Wide Web*. https://doi.org/10.1145/2187836.2187928

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed). Morgan Kaufmann.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLoS One, 15*(9), e0239441. https://doi.org/10.1371/journal.pone.0239441

Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine, 63*, 112–115.https://doi.org/10.1016/j.ypmed.2014.01.024

Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting, 35*(1), 213–223. https://doi.org/10.1016/j.ijforecast.2017.11.005