Utilizing Web Scraping Technology to Monitor Public Opinion on Environmental Policies

*Yu*e Cui*

College Of Arts & Science, New York University, 10003, New York, the United States

Abstract. Web scraping, an automated tool for extracting and analyzing large amounts of data from websites and social media platforms, has become an essential practice for researchers and policymakers. This paper examines the application of web scraping technology to monitor public sentiment towards environmental policies. In an age where digital platforms are crucial to public discourse, governments and researchers have increasingly utilised web scraping for real-time sentiment monitoring. This approach presents considerable ethical dilemmas, especially with user privacy and data security. This paper rigorously analyses these obstacles, emphasising concerns such as data integrity, user consent, and adherence to legislation. Contemporary academic methodologies illustrate the technical advantages of web scraping while simultaneously emphasising its ethical quandaries. This study proposes various methods for reconciling the necessity of data access with the safeguarding of user privacy, including the implementation of anonymisation techniques and the formulation of explicit ethical frameworks. The findings highlight the imperative for responsible web scraping to guarantee ethical adherence while optimising its effectiveness in analysing environmental debate.

1 Introduction

The digital age has transformed individual expression of thoughts, with social media and online forums providing abundant data for analysing popular sentiment. Novel and developing data types, particularly tweets extracted from social media platforms like Twitter, have become integral to the sociologist's data repertoire [1]. Public opinion can create laws, influence political leaders, and guide global environmental initiatives within the framework of environmental regulations. Web scraping, an automated tool for extracting and analysing extensive data from websites and social media platforms, has become an essential practice for researchers and policymakers. Web scraping employs automated bots and algorithms to methodically traverse web pages, gather publicly accessible data, and convert unstructured information into structured datasets suitable for analysis through data science methodologies, including natural language processing (NLP), sentiment analysis, and machine learning models.

In the context of environmental policies, public opinion can affect legislation, political leaders, and global environmental initiatives. Web scraping has become an essential tool for

* Corresponding author: yc6062@nyu.edu

researchers and policymakers in this field, allowing for real-time monitoring of public sentiment on urgent environmental concerns. However, the ethical implications of online scraping, notably the balance between data access and individual privacy, are still debated. Issues of user consent, data ownership, and compliance with privacy rules such as General Data Protection Regulation (GDPR) hinder the ethical application of this technology.

Scholars have approached web scraping from multiple perspectives. For example, psychologists can use web scraping to create modestly sized, more manageable datasets with tens of variables but hundreds of thousands of cases [2]. Web scraping can also be used to analyze craigslist rental listings to provide the situation of rental housing markets across the United States [3].

While some study its legal and ethical implications, others concentrate on its technical uses. Web scraping is a valuable tool for measuring public opinion on environmental concerns, such as climate change and the efficacy of policies, according to recent studies. On the other hand, issues with user permission, data integrity, and privacy regulations like the GDPR give rise to serious worries.

The purpose of this article is to investigate the ethical and practical implications of utilising online scraping to monitor public opinion on environmental policies. The paper will begin by providing an introduction of web scraping's uses in this industry, then examine ethical considerations, and lastly suggest principles for ethically balancing data access with privacy.

2 Web Scraping and Public Opinion on Environmental Policies

Web scraping serves as an effective method for analysing public sentiment and behaviour, especially in areas such as environmental policy, where public opinion significantly influences legislative and regulatory results. Web scraping facilitates the systematic extraction of data from social media platforms, blogs, forums, and news websites, enabling researchers to monitor real-time responses to environmental policies and global events concerning climate change, pollution, and resource management.

In the Carbon emissions cluster, a significant increase in the monthly proportion of discussions on Reddit was observed in 2009, exceeding +1 standard deviation. The increase can be partially attributed to the passage of the Alternative and Renewable Energy Portfolio Act in West Virginia in June, which aims to decrease reliance on coal and encourage the adoption of renewable energy sources. The launch of the US-India Partnership to Advance Clean Energy (PACE) in November further contributed to discussions by promoting inclusive low-carbon growth and advancements in clean energy technologies [4].

Web scraping has been utilised in environmental studies to examine public perceptions of specific policies, including carbon taxes and renewable energy subsidies. A national survey conducted by the Pew Research Centre from April 29 to May 5, 2024, which included 10,957 adults surveyed through the online American Trends panel, revealed that a majority of the U.S. public endorses the government implementing more assertive measures to combat climate change. This aligns with trends in social media discourse identified through web crawling methodologies. Approximately 65% of Americans believe that the federal government is insufficiently addressing the impacts of climate change, a perspective that remains consistent with views from the previous autumn [5].

Nonetheless, the characteristics of web scraping present difficulties related to the variety of opinions obtained. The demographics of Twitter users tend to be younger and more urban in nature. A study by McKinsey indicates that millennials represent the most active demographic on social media, with an estimated 68.8% projected to be using these platforms in 2024 [6]. Young individuals may introduce biases in sentiment analysis regarding environmental policy, particularly as older demographics tend to be less active online. Public

discourse on social media frequently exhibits polarisation, and scraping methodologies may unintentionally amplify vocal minorities or opinions that are geographically concentrated. The ethical and practical challenges of data collection are thus heightened, particularly regarding the accurate representation of diverse viewpoints and the preservation of data integrity.

3 Ethical Considerations and Challenges of Web Scraping

3.1 Legal and Ethical Frameworks

The growth and prevalence of research on platforms like Facebook and Twitter have raised significant concerns regarding investigator conduct and the ethics of social media use. The availability of large data sets has drawn researchers, including those not typically linked to health data, such as computer and data scientists, to engage with its ethical considerations. The dependence on oversight by ethics review boards is insufficient, and the accessibility of social media data frequently leads to ambiguity regarding the distinction between public and private spaces. Furthermore, social media users and researchers may overlook conventional terms of use [7]. Therefore, the legal and ethical implications of web scraping are complex, particularly when it comes to privacy and consent.

Notable incidents, exemplified by Facebook's 2018 data privacy scandal with Cambridge Analytica, highlight the risks associated with unregulated data collection and its potential for misuse. This instance involved the unauthorised collection of personal data from millions of Facebook profiles, resulting in significant privacy violations. This incident ignited public outrage and underscored the pressing necessity for enhanced data protection regulations, given that Facebook's inadequate oversight facilitated the misuse of web scraping techniques to sway political campaigns [8].

These incidents illustrate the ethical risks associated with data scraping and highlight the need for robust legal frameworks. The GDPR in the European Union establishes stringent guidelines for the collection, storage, and processing of personal data. Under GDPR, personal data encompasses any information capable of identifying an individual, such as social media handles, IP addresses, or opinions expressed online. Comparable legislation is present in other areas, including the California Consumer Privacy Act (CCPA) in the United States.

Web scraping frequently exists in a "grey area" within legal frameworks [9]. The act of scraping publicly available data, despite its apparent accessibility, may contravene platform terms of service or, in specific jurisdictions, infringe upon data protection laws without user consent. For instance, a researcher who collects public posts from a social media platform regarding a government environmental initiative without obtaining explicit consent from users may violate privacy laws, particularly if the data is utilised in unforeseen ways by the users.

Additionally, ethical guidelines from research organisations emphasise the necessity of obtaining informed consent when engaging with human subjects, including in digital environments. Users typically do not anticipate that their online activities will be monitored for research purposes unless this is clearly communicated, highlighting concerns regarding the ethical transparency of web scraping practices.

3.2 Practical Challenges

In addition to legal considerations, web scraping entails various technical challenges that are closely linked to ethical dilemmas. In the social sciences, many researchers are still in the early stages of acquiring foundational skills in data scraping, which presents significant

technical challenges. Ensuring data accuracy and reliability presents the initial challenge. Numerous websites are increasingly adopting dynamic programming languages such as JavaScript, which present greater challenges for scraping and necessitate the application of advanced programmatic methods, including headless browser scraping [10].

Headless browsers like Puppeteer and Selenium replicate user actions by fully loading web pages, including JavaScript content, while omitting the graphical interface. This enables scrapers to engage with dynamic websites, access data that is only available post-page load, and automate tasks such as button clicks or scrolling. Headless browsers, by simulating standard browser behaviour, enhance data extraction capabilities. However, they necessitate advanced programming skills and may raise ethical concerns regarding the extent of data scraping.

Furthermore, public sentiment data frequently presents as unstructured, necessitating the use of advanced tools by researchers to parse and analyse text, images, and other interactive data. Natural language processing (NLP) tools can assist in analysis; however, they may misinterpret sentiment in posts, especially regarding nuanced subjects such as environmental policy, where sarcasm, irony, or contextual factors can substantially change meaning.

Bias represents a significant challenge. Web scraping may skew the representation of opinions from specific demographic groups that exhibit higher online activity. Young, tech-savvy users may predominate on social media platforms, whereas older or rural populations may be under-represented. This introduces a bias in the data, complicating policymakers' ability to obtain a comprehensive understanding of public opinion across various age groups, socio-economic classes, or geographic regions.

Consent represents a significant ethical challenge. Numerous platforms possess terms of service that forbid automated data extraction. Although this does not invariably deter researchers, it prompts enquiries regarding the ethical engagement with user-generated data. A viable approach involves developing web scraping techniques that anonymise and aggregate data prior to analysis, thus minimising potential harm and safeguarding individual privacy.

4 Suggestions and Future Prospects

The increasing significance of web scraping for data collection necessitates the establishment of explicit ethical guidelines and legal frameworks to safeguard both users and platforms. The LinkedIn vs. HiQ Labs case illustrates the increasing necessity for regulatory clarity within the industry. LinkedIn's decision to sue HiQ Labs under the Computer Fraud and Abuse Act (CFAA) for scraping public data highlights the challenges faced by platforms in protecting user privacy while balancing the public nature of online data [11]. The legal dispute has initiated a discussion regarding the scope of rights that platforms possess over publicly available data, marking a significant juncture for forthcoming regulations. Future guidelines must address user consent and the limitations associated with scraping public information to prevent similar legal conflicts.

A multi-faceted approach is essential to tackle the ethical and practical challenges related to web scraping. Transparency and consent must be fundamental components of web scraping practices. It is essential for researchers and institutions to clearly communicate the intended use of data and to obtain consent when possible. When consent is unattainable, particularly in large-scale data collection from public forums, ethical guidelines must prioritise the anonymisation of data to safeguard individual identities.

Another suggestion is the development of ethical scraping frameworks, which could include platform-specific guidelines that balance the benefits of data collection with the need for user privacy. These frameworks could draw on existing privacy laws like GDPR and CCPA, providing researchers with clear boundaries for ethical data collection. According to

Art. 5.1, c) of the GDPR, personal data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed [12]. This implies that, even for research purposes,data that are not relevant and necessary should be deleted or anonymized — hence the importance of anonymization or data omission for research activities.

Future advancements in artificial intelligence and machine learning may address various ethical and technical challenges related to web scraping. AI algorithms can be developed to identify and eliminate biassed or inaccurate data, concurrently enhancing the efficiency and scalability of data scraping techniques. These tools may assist in identifying ethically questionable practices, including the scraping of data from platforms that explicitly prohibit such actions.

Interdisciplinary collaboration among computer scientists, legal experts, and ethicists is essential for addressing the evolving ethical landscape of web scraping. With the advancement of data collection technologies, the establishment of new ethical standards and technical solutions will be essential for the responsible and effective use of web scraping.

5 Conclusion

The ethical issues and practical difficulties of employing online scraping technology to track public opinion on environmental policy have been examined in this research. The paper has demonstrated how online scraping may be a useful tool for academics and policymakers while also presenting serious ethical challenges by analysing its advantages and disadvantages. These concerns are particularly related to privacy, permission, and data veracity. When used in conjunction with environmental discourse, web scraping techniques provide real-time insights into public opinion; nevertheless, users' rights must be respected when using them. According to the research, enforcing legal requirements like GDPR, anonymizing data, and putting in place ethical frameworks for data scraping can help strike a balance between data access and privacy. Furthermore, developments in AI present viable answers to a few of the technological problems, like lowering bias and enhancing data dependability. Subsequent investigations have to concentrate on honing the moral guidelines pertaining to web scraping and investigating the ways in which new technologies can augment the precision and morality of data gathering. All things considered, ethical web scraping techniques can significantly improve our comprehension of public sentiment towards important topics like environmental regulations, but only if they are implemented in a manner that upholds both legal and personal rights.

Reference

1. M.L. Williams, P. Burnap, L. Sloan, Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. Sociology. **51**, 1149–1168 (2017)
2. R.N. Landers, et al., A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychological Methods. **21**, 475 (2016)
3. G. Boeing, P. Waddell, New insights into rental housing markets across the United States: Web scraping and analyzing Craigslist rental listings. Journal of Planning Education and Research. **37**, 457–476 (2017)
4. A. Mandal, A. Kaushal, A. Acharjee, Climate-related discussions on social media: Critical lessons for policymakers. Natl. Inst. Econ. Rev. 1–8 (2024)

5. A. Tyson, B. Kennedy, Two-thirds of Americans think government should do more on climate, Pew Research Center, **23** June (2020) https://www.pewresearch.org/science/2020/06/23/two-thirds-of-americans-think-government-should-do-more-on-climate/
6. J. Zote, Social media demographics to inform your 2024 strategy, Sprout Social, 14 February (2024) https://sproutsocial.com/insights/new-social-media-demographics/
7. E. Chiauzzi, P. Wicks, Digital trespass: Ethical and terms-of-use violations by researchers accessing data from an online patient community. J. Med. Internet Res. **21**, e11985 (2019)
8. R. McNamee, Zucked: Waking Up to the Facebook Catastrophe (Penguin Publishing Group, 2020), accessed 10 October (2024)
9. D. Possler, S. Bruns, J. Niemann-Lenz, Data is the new oil—but how do we drill it? Pathways to access and acquire large data sets in communication science. Int. J. Commun. **13**, 3894–3911 (2019)
10. A. Luscombe, K. Dick, K. Walby, Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Qual. Quant. **56**, 1023–1044 (2022)
11. G. Xiao, Bad bots: Regulating the scraping of public personal information. Harv. JL & Tech. **34**, 701 (2020)
12. I. Siegert, et al., Personal data protection and academia: GDPR issues and multi-modal data-collections. Online J. Appl. Knowl. Manag. (OJAKM) **8**, 16–31 (2020)