# Retrieving Data from Social Network Platforms: A State-of-Art Review

Harriet Sibitenda[1(✉)], Awa Diattara[1], Assitan Traore[2], and B. A. Cheikh[1]

[1] Laboratoire d'Analyse Numerique et Informatique, University of Gaston Berger, Saint-Louis, Senegal
{harriet.sibitenda,awa.diattara,cheikh2.ba}@ugb.edu.sn
[2] Business and Decision, Grenoble, France
assitan.traore@free.fr

**Abstract.** Analyzing public concerns gives feedback to organizations about goods and services. The means of collecting social concerns differ for each social media platform. There is a need to explore the literature about automated tools applied by researchers to collect social issues. The goal of this paper is to provide an overview of data collection methods from social media platforms. This is to guide the collection of public concerns through machine learning approaches. Following the preferred reporting items for systematic reviews and meta-analyses standards, we collected 2180 articles from the Google Scholar database based on keywords. We screened the reviews for relevancy based on abstract and title. Considering the exclusion criteria, we removed all full articles not related to social media platforms. We manually analyzed only 298 articles to identify classifications within methods of data collection. From the reviews, we retrieved five categories of data collection. These include; manual observations, self-report surveys, public repositories, existing licensed tools, public application programming interfaces, and web crawlers or scrapers. Considering sample networks of Facebook, Twitter, and YouTube, we explored the trends of tools. And we stated the pros and cons. In conclusion, to collect social concerns at no cost and in large amounts, we recommend using open libraries of public application programming interfaces and web scrapers. In the future, we plan to extract public data using the recommended trendy automated tools for each category and sample social networks.

**Keywords:** Data collection · social networks · social concerns

## 1 Introduction

In society, individuals usually experience challenges and hardships that hinder their daily effort at work and ways of living [1]. We call these personal troubles

PASET-RSIF, UGB.


K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 151–166, 2025.
https://doi.org/10.1007/978-3-031-86493-3_13

or problems. The occurrence of similar challenges to a large group of people within the same society implies the emergence of social issues. There is a need for an urgent response to reduce the magnitude of the issue's damage and spread. Consistent social issues affect the development of societies [2]. The collection of many social issues triggers the emergence of another. Social concerns have a historical cycle to their occurrence.

A study, by [3] highlights the occurrence of concerns in society. Initially, the occurrence of specific problems to many people creates common issues. People report their concerns to authorities like police, media platforms, government activists, and more. This stage is called legislation for action. After the legislative stage, the authorities develop laws and regulations to handle the situation. There is a probability that different sectors recover from the common social problem or adopt change. Finally, there is a need to assess the transformation action plan to preserve society for sustainable development and public governance [4].

There are four major methods of *Understanding social concerns* [5]. These include; (i) Surveys that gather data from a sample study population using questions that respondents interact by mobile phone, web, and face-to-face tools like paper. (ii) Experiments that are conducted in the natural and physical sciences based on cause-and-effect relationships. (iii) Observations that are field research sessions that involve watching the situation on participants to make reports. (iv) Existing dataset is data that someone else has already gathered and used in another study, for example, data from the US Census Bureau. Globally, the *Statista report 2020* highlights that the use of SNPs is increasing, and Facebook leads other networks like YouTube and WhatsApp [6]. Communication over SNPs involves user perceptions, posts, comments, reactions, emotions, and many more [7]. The social comments form big unstructured data, and this includes text, images, audio, and video clips. The big data attracts researchers to apply Artificial Intelligence (AI) tools to explore hidden knowledge insights. In Africa, social media usage penetration is growing higher with the Northern region at 45%, Southern 41%, western 16%, Eastern 10%, and Central 8%. [8]. We will focus on common SNPs like Facebook, Twitter, and YouTube to explore the methods of data collection for social issues reported on these networks.

The goal of this study is to give a global overview of methods of data collection from SNPs to identify public concerns using Machine Learning (ML) approaches. To attain the goal, we consider the following specific objectives: (i) To review the existing literature about collecting data from social networks. (ii) To explore Machine Learning (ML) tools to extract the social comments. (iii) To propose suggestions of trends for data collection from sample social media platforms.

This study includes three sections: The first section introduces the need for social feedback to policy-makers from comments reported on the SNPs. The second section describes findings from the State-of-art review with subsections like related literature, the methodology to use, exploration of the data types of comments, and the methods of collection used in reviews. Finally, the third section summarizes the findings of the study, discusses its limitations, and recommends future research directions.

## 2 State-of-Art Review for Data Collection from SNPs

### 2.1 Related Literature

Authors in [9] conduct a systematic review of ML text mining techniques on previous research from Twitter. They explored trends of topics between the years of 2006 to 2019. Collected data (18,000 articles) from IEEE, the Web of Science, and EBSCO. They used search keywords of "Twitter AND Survey" and "Twitter AND Review". The research displays a high-level analysis of topics from articles. The study has a limit of eliminating the study of subcategories from the full review of the papers. We note to use an index database and search keywords to collect review papers relevant to a study's purpose.

The study in [10] involved a systematic review of the use of Twitter for higher professional education. They used keyword search strings like "Twitter in higher education", and "higher education academics", to collect 615 paper reviews from databases like Scopus, and Google Scholar. Authors repeatedly collected some articles using Google Scholar and Scopus, and this increased the time for screening. They also used the snowball method to add some articles in the references to reviews and manually analyzed 28 reviews to classify the content. The authors excluded many articles to consider only articles of interest. This narrowed their findings to the themes, ethical issues, and theoretical methodologies used. We note to identify exclusion criteria to obtain relevant findings from full articles.

The authors in [11] explored challenges for data discovery, collection, preparation, and analysis on social media. They collected 260 papers using keyword search strings and databases like ACM, AIS, IEE, and Science Direct. They conducted a backward search using citations in the references to create a citation network of relevant papers. With Text mining techniques, they also extracted themes, opinions, and sentiments based on the text of titles and abstracts. The authors identified challenges for each stage and proposed solutions from reviews. We note that they didn't explore deeper subcategories for solutions. Analysis of each stage from a broader perspective would give profound findings about challenges, solutions, and methods used.

A study by [12] explored how digital technologies support urban and regional Agro-food purchasing and its characteristics. They used Scopus and Web of Science to collect 370 and 398 articles, respectively. Using the PRISMA (Preferred reporting items for systematic reviews and meta-analyses) protocol guidelines. The PRISMA guide protocol used was logical for the exclusion of articles. They manually analyzed inclusion articles and represented the findings of classification using graphs. This aided the understanding of the findings. We note the usefulness of the PRISMA protocol to select papers for inclusion.

Therefore, this study proposes to collect previous literature related to data collection on social media for a period of years, from 2018 to 2022. To attain this purpose, we explore three questions: (i) What is the type of data from social comments? (ii) What are the ML methods applied in reviews to collect the data, and how do they work? (iii) What are the trends, strengths, and limitations of ML methods for data collection? In this study, we adopted the use of one index

database named Google Scholar to reduce duplication of articles. We searched for articles using keywords. The PRISMA protocol was a guide for the exclusion and inclusion of articles. We manually analyzed inclusion articles to discover subcategories within methods of data collection for sample networks. Thus, our study adds an exploration of the state-of-art review about data collection and the classification of subcategories needed by researchers to select methods to extract social comments from SNPs.

### 2.2 Methodology

To begin with, we followed the PRISMA protocol guidelines suggested by [13] to construct a protocol to use in this study. The PRISMA standard for systematic review includes four major steps: identification of papers, screening, setting eligibility criteria, and selecting inclusion articles. We adopted the steps as shown in Fig. 1. Using the "Publish and Perish" open software, we selected the Google Scholar index database to collect articles for review. With the Google Scholar search, we entered keyword strings like "data collection and social networks", "data collection and Twitter", and "data collection and Facebook", for a specified period from 2018 to 2022. As a result, we collected 2180 articles and saved a copy of the output as "csv" and "ris" files. The csv file included the citation index details for each article. The ris file included referencing details for articles stored by reference software like Mendeley.

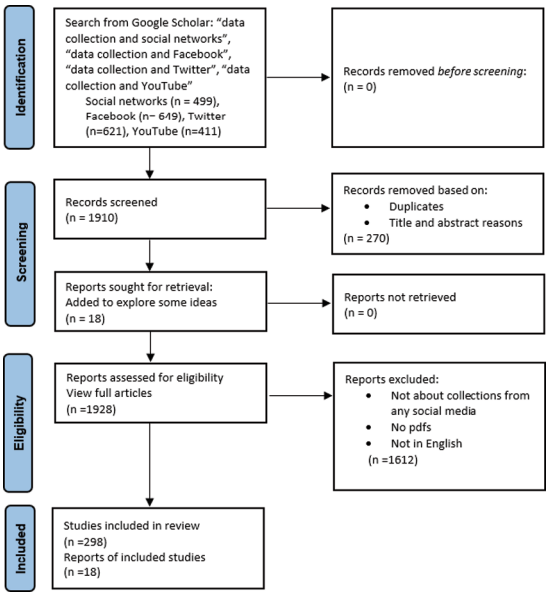


**Fig. 1.** The PRISMA flow protocol we adopted from Moher et al., 2009

After obtaining the articles, we began the screening process. We removed 270 irrelevant articles based on duplicates, the title, and the abstract. Using an eligibility criterion, we excluded 1612 articles. The criteria for exclusion and inclusion involved reading full articles. The csv file included a column of links to access the available full articles for each record of review. We eliminated records of articles without full PDFs, missing collections from any social media, and written in languages other than English. Finally, we had a csv file including 298 articles to consider for manual analysis. We added 18 papers to describe specific terms and ideas revealed from the reviews using the snowball method. We stored all these files in the Mendeley reference software to enable easy citation of articles.

**Exploration of Papers Collected**

Considering the csv dataset, the data contains data types such as text (with columns like Authors, Title, Type of document, Abstract, Age of paper). Number (with columns like Cites, GSrank, CitesPerYear, CitesPerAuthor). Date (with columns like Year, QueryDate). And links (with columns like ArticleURL, FullTextURL, CitesURL). The articles collected included years from 2018 to 2022. We collected articles in totals of 403, 355, 440, 382, and 380, from years, 2018, 2019, 2020, 2021, and 2022, respectively. After the screening and exclusion processes, we analyzed only 78, 38, 76, 56, and 45 inclusion articles respectively to the sequence of years.

### 2.3 Data Types of Social Comments from Reviews

Our first research question requires identifying types of data collected from social comments. From the reviews, researchers collected comments with data types like; text, URL links, images, video clips of ads, and stories. Text stores any kind of text data. It can contain both single-byte and multibyte characters. A URL (Uniform Resource Locator) is "a unique identifier used to locate a resource on the Internet". The video data type is "a type of file format for storing digital video data on a computer system". Image data type stores or references any type of image files in binary format like jpg, BMP, png, and more. From the reviews, the common data collected from SNPs is of text data type. The text was collected from sources like posts or comments, groups and pages, participant responses from surveys, news articles, publication papers, video comments and captions, and image text. Some reviews extract text from the metadata of images and transcribed text or captions from videos.

### 2.4 Methods of Data Collection

The second research question requires analyzing the ML methods used for data collection on SNPs. From the findings, we have retrieved five methods of data collection over the sequential years of 2018 to 2022. These included 23 articles for manual observations, 24 for self-reports, 15 for public repositories, 35 for existing

licensed tools, 190 for public APIs, and 23 for web scrapers or crawlers. The use of public APIs was most common. We also note that manual observations are still in use. We further explored each method of data collection to identify the pros and cons to suggest trends of adoption for sample social networks.

**Data Collection by Manual Observations**

Manual observation involves the visual analysis of data and writing findings by oneself. This is the manual coding of findings. Manual observation requires the user to navigate the website using keywords/hashtags and URL links. From our reviews, some researchers used manual observation to record findings. One pro of using manual observations is the ability to collect data from private accounts without sending notifications to owners [14]. The cons include: (i) More than one person may be required to collect data to quicken the process, [15]. (ii) They collect a small portion of results to generalize findings [15]. (iii) Inability to collect some meta-data features like subscriptions, and links to other websites [16]. After exploring these suggestions, we ignored the use of manual methods. We opt to recommend the use of automatic methods that use open-source tools to collect usage data.

**Data Collection by Self-report Surveys**

Self-reports give responses about a discussion. A self-report is a test or survey that relies on one's own interest/response. To do this, we use tools like questionnaires, interviews, and observations. Questionnaires involve questions with a choice to answer. Interviews involve structured conversation where one participant responds to asked questions by the other party. From the reviews, we analyzed three methods of self-reports, these include traditional, online, and mixed surveys.

Traditional surveys: These involve physical interaction with tools like questionnaires, and observations. The pros of using traditional surveys include; no requirement for technical and programming skills, and provide user interests and emotions. Some cons include: (i) Required participant compensation fees. (ii) Extra expenses to distribute user interaction tools like questionnaires. (iii) Biased responses like time spent on social media [17]. (iv) A need to automate the collection of responses from participants [18]. We note that traditional surveys use manual means of data collection.

Online Surveys: This involves sending electronic responses on the internet [19]. The collection of data is automatic or by manual observations. The tools used with online surveys include questionnaires, interviews, and observation. Most reviews about online surveys used questionnaires sent to participants over online platforms like Google Forms, phone apps, web portals, SurveyMonkey, and video ads. The interviews from online surveys involved videos and audio clips. Observation of responses on an online platform was also used to collect data. The pros of using online surveys include; the conduction of surveys at any preferable time, anywhere, and providing user interests and emotions. The cons include: (i) Some online platforms are non-user-friendly to participants [20]. (ii)

Inability to keep track of changes of content after closing the survey [21]. (iii) It is time-consuming for participants to use interfaces like ads. We thus disregard the use of online surveys because of increased expenses on the many participants.

Mixed surveys: These combine data from self-reports and usage data from the websites. The usage data is a collection of a visitor's actions on a website. The common sources of usage data reviewed include keywords, hashtags, and URL links. A keyword is a word or concept to identify digital content on a specific topic. A hashtag is a keyword preceded by the # symbol, thus we refer to keywords and hashtags interchangeably in our study. This mixed survey method requires participants to give responses online, and then collect usage data about user profile details on the social media account. From the reviews, the common tools used with the mixed method include questionnaires, interviews, and user usage data about profile accounts. The pros of using mixed surveys include the ability to gain depth in research findings [22]. And an increased number of participants [23]. The cons include: (i) Inability to restrict responses from non-target participants [23]. (ii) Requires the consent of participants to share their data about the account profile, [14,20]. (iii) A need to offer incentives to participants [24]. We thus recommend the use of mixed surveys for cases of combining user interests or emotions with usage data but at minimum or no participant compensation fees. Table 1 shows the pros and cons of self-report surveys.

**Table 1.** Comparing the methods of self-report surveys

| Method | Pros | Cons |
|---|---|---|
| Traditional survey | – Do not require programming skills | – Require participant compensation fees |
| | – Provide user interests and emotions | – Extra expenses to distribute user interaction tools like questionnaires |
| | | – Biased responses like time spent on social media |
| | | – Need to automate collection of responses from participants |
| Online Survey | – Conducted at any time and anywhere | – Some online platforms are not user-friendly to submit responses |
| | – Provide user interests and emotions | – Inability to keep track of changes for new content after the survey |
| | | – It is time-consuming for participants to use interfaces like ads |
| Mixed methods | – The ability to gain in-depth findings | – It is easy for non-target participants to give responses |
| | – Increases number of participants | – Requires consent of participants to share their account profile data |
| | – Provide user interests and emotions | – Complex method of data collection |
| | | – Need to offer incentives to participants |

### Data Collection by Public Repositories

Public repositories are accessible to everyone on the internet to collect existing datasets from both self-report and usage data. Downloading the datasets from public repositories may require using keywords related to a topic or a link to the repository. From the reviews, we also observed that previous studies provide public existing datasets on organizations or individual websites. Table 2 shows samples of existing datasets from studies of other authors and organizations.

**Table 2.** Public repositories with existing datasets

| Source | Public repository | Description of dataset |
|---|---|---|
| Other authors | Ramos et al., 2018 | Has demographics from 1,000 Facebook status updates |
| | Wang et al. 2016 | About influenza study with Geo-Tagged Twitter data |
| | Basu et al. 2019 | Containing tweet IDs about Nepal and Italy earthquakes |
| | Dimitrov, D., et al. 2020 | Containing Tweets about the COVID-19 Pandemic |
| | Kaczmirek et al., 2014 | About German Bundestag elections 2013 for Facebook and Twitter |
| Organizations | Georgia State Uni. Lab | Can only provide tweet IDs |
| | Kaggle | Provide Covid-19 Tweets for late April Tweets using hashtags |
| | | Spam detection comments from Twitter and Email |
| | British Geological Survey | Has tweets related to landslide events |
| | National Science Foundation | Has posts about spam detection from Facebook |
| | Our World in Data | Provides daily world statistics about Covid-19 |
| | US state-level health data | Provide details for America's Health Rankings Annual Report |
| | Honeypot | useful for studying spam activity on Twitter |

The pros of using existing datasets include (i) The ability to download data, including worldwide posts in different languages [25]. (ii) Some public repositories enable the use of keywords to return datasets relevant to a specific topic [26]. The cons of using public existing datasets include: (i) Inability to track continuous changes in comments after the occurrence of the event, [27]. We ignore the use of public repositories.

### Data Collection by Existing Licensed Tools

An existing licensed tool is software that provides legally binding guidelines for data collection and requires a fee after the trial period. From the reviews, some existing licensed tools include Net viz, Netlytic, Gnip, NVivo, Sifter, SocialBlade, Brandwatch, Tuber, and more. The pros of using existing licensed tools include; (i) They provide other functionalities, such as text analysis and visualization [28]. (ii) The ability to access public APIs to extract meta-data in real-time and past

historical times, [29,30]. (iii) Offer trial periods and basic packages at no cost [24]. Some cons for using the existing licensed tools include (i) Searching one hashtag per request is time-consuming [28]. (ii) A need to filter columns with irrelevant features for a goal of study [31]. (iii) The use of search keywords eliminates relevant related data without the key terms [31]. (iv) A need to consider words based on the user sentiment found in the data traffic at a specific time [32]. (v) The basic versions usually have limitations to data volume [33]. We ignore using this method and explore other means that support free collections of data.

**Data Collection by Public APIs**

An Application Programming Interface (API) is a set of functions and procedures followed to create applications needed to access data from the network operating system. Open/Public APIs refer to APIs made publicly available to software developers. Public APIs require one to apply for developer access authentication rights to extract social network usage data without violating the public privacy laws of users. Users interact with comments/posts of text, audio, videos, and images. From the reviews, the common sources of usage data include hashtags/keywords, IDs or URL links, and user logs.

To begin with, Facebook Graph API "is an HTTP-based API that allows developers to extract data from the Facebook platform". Marketing API is "an HTTP-based API used to query data, create and manage ads, and perform a wide variety of other tasks". Marketing APIs are a collection of Graph API endpoints used to advertise on Facebook. There are open libraries used to call the Facebook Graph API like Facepager and CrowdTangle. Facepager "is for fetching public available data from YouTube, Twitter and other websites using APIs and web scraping". CrowdTangle uses its API to provide access to Facebook Graph API to collect posts from public groups and pages. CrowdTangle is most commonly used over sequential years.

For Twitter, the standard level of developers provides three APIs. These include streaming, search, and sampling APIs. From the reviews, we retrieved two commonly used Public APIs, namely; Streaming and Search APIs. Twitter API allows you to stream public Tweets from the platform in real-time. Streaming API has a limit of 15 tweets per request. The Twitter Search API "is an HTTP-based RESTful API that returns responses encoded in JSON format". Twitter's REST (Representational States Transfer) API "allows you to search terms based on specific parameters". The Search API includes a limit of 500 tweets per request in the past seven days. The Sampling API "delivers a random sample of publicly available Tweets in real-time, but supports selecting which fields return in the payload" [34]. Some reviews used enterprise-level Twitter APIs like; Power Track API [35], Historical PowerTrack API [36], and GetOldTweets API [37]. Commonly used open libraries for Twitter APIs, include; Tweepy, Twitter 4J, TwitterMySQL, Apache, Crowdbreaks, Rtweet, and Twarc. We also identified the Botometer API and Wayback Machine API. The Botometer API "checks the activity of a Twitter account and gives it a score

based on the extent to which it matches accounts that use automation". The Wayback Machine API "is a historical database that captures web pages using the internet archive". This API attains web data from different social media with some missing features.

For YouTube, the YouTube API was used to collect data. YouTube "provides the ability to retrieve feeds related to videos, users, and playlists". The YouTube Data tools library was commonly used to access data via the YouTube API. Table 3 demonstrates the use of these public APIs over the sequential years.

**Table 3.** Use of trending Public APIs over sequential years

| Social Platform | Year | Public API(No. of papers)- Added packages |
|---|---|---|
| YouTube | 2018 | YouTube AP(I4)- YouTube Data Tool |
| | 2019 | YouTube API(1) |
| | 2020 | YouTube API(1)- YouTube Data Tool |
| | 2021 | YouTube API(4)- YouTube Data Tool |
| | 2022 | YouTube API(8)- YouTube Data Tool |
| Facebook | 2018 | Facebook Graph API(5) |
| | 2019 | Facebook Graph API (4)- pySocialWatcher, Facebook Marketing API(2)- pySocialWatcher |
| | 2020 | Facebook Graph API (8)- Facepager, CrowdTangle |
| | 2021 | Facebook Graph API (5)- CrowdTangle, Facebook Marketing API (1)- Facebook ads manager |
| | 2022 | Facebook Graph API (5)- CrowdTangle, Facebook Marketing API(1) |
| Twitter | 2018 | 1. Streaming API (26)- Tweepy, Twitter4j, Apache, TwitterMySQL |
| | | 2. Search API (11)- Twiter4j, Historical PowerTrack API (2), Botometer API (2) |
| | 2019 | 1. Streaming API (17)- Tweepy, Apache |
| | | 2. Search API (3)- Rtweet, Botometer API (2) |
| | 2020 | 1. Streaming API (13)- Tweepy, Apache |
| | | 2. Search API (3)- Rtweet |
| | 2021 | 1. Streaming API (13)- Tweepy, twitteR, Crowdbreaks |
| | | 2. Search API (6)- Rtweet, Twarc, Twitter4J, Sampling API(1) |
| | 2022 | 1. Streaming API (9)- Tweepy, TwitterMySQL |
| | | 2. Search API (7)- Rtweet, Twarc, Sampling API (1), Botometer API (1) |

### Sample Trends of Public APIs Used

We identified four trendy open libraries used with the Twitter API. These include Tweepy, Twitter4J, Rtweet, and Twarc. Tweepy is a Python-based package that gives one access to Twitter Streaming and Search APIs [38]. Tweepy has a limit of providing fewer streams of data (1% of total tweets) and has a bias in returning most tweets in the English language. Twiter4J is an open-source Java library, which provides access to both Twitter Streaming and Searching APIs [39]. Twitter4J has a limit to access private accounts. Rtweet is the "R package that provides users a range of functions designed to extract data from Twitter REST/Search and streaming APIs" [40]. Rtweet has a limit on the use

of R programming skills. Twarc "is a command-line tool and Python library for collecting and archiving Twitter JSON data via the Twitter API". This Python library allows the use of a premium version of Search API at a fee [41]. Twarc gives an entire conversation, including its direct replies and nested replies. Using Twarc has a limit of the restrictive 7-day window of data collection by Search API.

For Facebook, we explored that the most commonly used open library is CrowdTangle. CrowdTangle uses its API to access Facebook Graph API to retrieve posts from public groups and pages. It provides historical data on posts shared by public pages or groups, and the user can add new account IDs to collect posts. Some limitations of CrowdTangle include: (i) the inability to track every public account, private profile, and group, [42]. (ii) Some public pages have removed content [43]. (iii) And it ignores public pages whose likes and followers are more than 25K [25].

We observed that the reviews commonly used YouTube API with the open library of YouTube Data tools. Some pros of using YouTube Data tools include: (i) Provides access to modules like video list info and video network [44]. (ii) It enables the retrieving of new channels that satisfy a search query [45]. The cons include: (i) Some feature variables of meta-data were missing, like watch time [44]. (ii) Some captions of transcribed text were missing [46]. (iii) It has a limit to access 1000 requests per day.

**Data Collection by Web Crawlers or Scrappers**

From a general perspective, [47] explains that we use the terms "web scraping" and "web crawling" interchangeably to imply data extraction from web pages. Web scraping is "a procedure of automatic extraction of data from websites using software". Web crawling is "about finding or discovering URLs or links on the web". For this study, we refer to these two terms to mean the same. Web scrapers access public data without the limitations that exist with APIs. The user searches the web pages of interest using tools like hashtags, and IDs/URL links. And thereafter analyses the HTML/XML inspect elements (CSS or XPath selectors) [48]. The CSS Selector "combines an element selector and a selector value to identify particular elements on a web page". XPath is "used to navigate through elements and attributes in an XML document". Collecting a list of web CSS or XPath elements requires web crawlers that use coding languages such as Python and R [49]. We have three modes for scraping or crawling data [50]. These include web extensions, existing licensed software, and open-source tools. A web browser extension is "a small software application that adds a capacity or functionality to a web browser". Table 4 demonstrates the modes of web crawlers/scrapers used in the reviews.

**Table 4.** Modes of web crawlers or scrapers

| Mode | Functionality | Crawler or Scraper | Social platform |
|---|---|---|---|
| Open software | Require user guidelines to create new codes | Scrapy | Twitter, YouTube, Facebook |
| | | Selenium and Beautifulsoup | YouTube |
| | | IMcrawler | Facebook |
| | | TwitterScraper | Twitter |
| | | Twint | Twitter |
| | | Snscrape | Twitter |
| Licensed software | Private and public existing tools that require a fee | Octoparse scraper | Twitter |
| | | OpinionScraper | News sites |
| Web extensions | Public web extensions that run on a browser | WebDataRA | Twitter |
| | | Ncapture | Twitter, YouTube |

The web extensions used in reviews include WebDataRA and NCapture. These extensions have some limitations. (i) Some web extensions are free, but others require a fee. (ii) The user may not access all the features that APIs offer. We ignore the use of web extensions because of the manual methods involved in accessing data. Existing licensed software is "proprietary software distributed under a licensed agreement to enable users to access data". These tools have terms of service required to follow. And after the trial period, one requires payment fees for premium packages. We retrieved some common licensed tools used, such as Octoparse scraper [51] and OpinionScraper [52]. These tools are public or private. The public ones like Octoparse provide both standard and premium packages to extract data from different websites. Existing licensed tools have a limit on costs incurred to access large amounts of data. The open-source tools are freely available to extract data without a commercial license or paying for standard packages in particular. Most open-source tools available require programming skills to follow user guidelines set to extract the data. The open software tools from reviews include Scrapy, Selenium, Beautifulsoup, IMcrawler, TwitterScraper, Twint, and Snscrape. We will further explore the pros and cons of top open software tools.

**Trends of Scrapers and Crawlers for Sample Tools**

We reviewed four common open-source tools for web scraping or crawling. The tools include Scrapy, TwitterScraper, Twint, and Beautifulsoup. Scrapy is "a free and open-source web crawling framework written in Python". Networks like Facebook, Twitter, and YouTube. Some pros of using Scrapy include; (i) The ability to attain historical data [53]. (ii) Provides some features of meta-data that public APIs omit, such as subscribers and thumbnails from YouTube [54]. The cons of using Scrapy include: (i) Requires the consent of user login credentials to access data from a website. (ii) And the user has to disable pop requests that would discontinue the crawling process [55]. TwitterScraper is for Twitter only. It is "a simple script to scrape Tweets into JSON raw format, using the Python package requests [56]". TwitterScrapper overpasses limitations of public APIs such as the length of tweets and the style of posts, [47]. TwitterScraper has a limit of internet speed/bandwidth, and how many instances started at each request [47]. Twint is another trending scraping method used for Twitter. Twint "is a scraping tool developed in Python to extract and scrape tweets

from specific users and tweets on specific topics". Twint is most reliable when a list of users of interest is to be extracted [57]. Because of the depreciation of the developer support end, Twint has a limit on the probability of reduced functions. Selenium is "an open-source tool that is used for automating the tests carried out on web browsers". Beautiful Soup "is a Python library for pulling data out of HTML and XML files". At first, Selenium allows access to a website (different websites), and then Beautifulsoup crawls the data into a list of data represented in JSON format or csv and more. Using Selenium and Beautifulsoup provides the ability to collect more features as viewed on the screen, [54]. Selenium requires automatic scroll through all files to capture as much data as possible, [58]. From the reviews, researchers rarely used web scrapers. They applied them to specific objectives, such as volume of data increase, and access to features that are not offered by the Public APIs.

## 3 Conclusion

### 3.1 Discussion of Findings

This study aims to provide solutions to three research objectives. First, we reviewed existing literature about methods of data collection from social networks. Using the Google Scholar database and search keywords, we collected previous articles from 2018 to 2022. Second, we explored the kinds of ML methods used for data collection. We reviewed five general methods, these are; self-report surveys, Public APIs, Web crawlers or scrapers, public repositories, and manual observations. We explored the pros and cons of each method. Finally, we identified trends in ML methods to collect data for sample networks like Facebook, Twitter, and YouTube. Table 5 demonstrates a summary of a comparison of the methods of data collection.

**Table 5.** A summary of comparison for methods of data collection

| Method | Pros. | Cons. |
|---|---|---|
| Manual observations | Ability to record all data displayed on screen | Tiresome, and requires another software to record data |
| Traditional surveys | The ability to collect user interests and emotions | Requires participants compensation expenses |
| Online surveys | Timely collection of user interests and emotions | Require participant fees and training before use |
| Mixed surveys | Increased access to user interests, emotions, and profiles | Requires participant compensation fees and their consent |
| Public repositories | Provide already compiled datasets with easy access | Inflexibility to adjust contents with trends in comments |
| Existing Licensed tools | Offer more functionalities like analysis of data | Requires a fee after the trial period |
| Public APIs | Provide access to most features of meta-data | Limited streaming and historical data |
| Crawlers or Scrapers | Provide increased volume of data and missing features | Require advanced programming skills |

The limitations of our study include; having fewer reviews for some categories of methods used for data collection. Most reviews did not highlight the challenges faced during the data collection step. The different sample social networks have different methods of data collection, there was a need to review each separately to attain better perceptions.

### 3.2 Recommendation and Future Work

This study provided answers to all our research questions ((i) What is the type of data from social comments? (ii) What are the ML methods applied in reviews

to collect the data, and how do they work? (iii) What are the trends, strengths, and limitations of the ML methods of data collection?). Based on the findings, to collect data at no cost and in large amounts, we recommend using open-source tools that support using ML algorithms. These are public APIs and web scrapers. In the future, we plan to extract public data using the recommended trendy ML tools for sample social networks.

## References


1. Stepney, P.: Social Work in the 21st Century. Int. Soc. Work. **41**(3), 394–396 (1998)
2. Hart, H.: What is a social problem? Am. J. Sociol. **29**(3), 345–352 (1923)
3. Geels, F.W., Penna, C.: Societal problems and industry reorientation: elaborating the Dialectic Issue LifeCycle (DILC) model and a case study of car safety in the USA (1900–1995). Res. Policy **44**(1), 67–82 (2015)
4. Anthikad, J.: Social problems. In: Sociology for Graduate Nurses, p. 156 (2014)
5. Martin, B.: Sociological research methods, pp. 36–67 (2017)
6. Statcounter. Social Media Stats Africa—StatCounter Global Stats (2021)
7. Hartmann, J., Huppertz, J., Schamp, C., Heitmann, M.: Comparing automated text classification methods. Int. J. Res. Mark. **36**(1), 20–38 (2019)
8. Ortiz-Ospina, E.: The rise of social media - Our World in Data (2019)
9. Karami, A., Lundy, M., Webb, F., Dwivedi, Y.K.: Twitter and research: a systematic literature review through text mining. IEEE Access **8**, 67698–67717 (2020)
10. Singh, L.: A systematic review of higher education academics' use of microblogging for professional development: case of twitter. Open Educ. Stud. **2**(1), 66–81 (2020)
11. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics–Challenges in topic discovery, data collection, and data preparation. Int. J. Inf. Manag. **39**, 156–168 (2018)
12. Samoggia, A., Monticone, F., Bertazzoli, A.: Innovative digital technologies for purchasing and consumption in urban and regional agro-food systems: a systematic review. Foods **10**(2), 208 (2021)
13. Moher, S., Liberatl, A., Tetzlaff, J., Alttman, D.H., Parisma Group: So schaffst du deine Ausbildung. Ausbildungsbegleitende Hilfen (abH) **151**(4), 264–269 (2009)
14. Stellefson, M., Paige, S., Apperson, A., Spratt, S.: Social media content analysis of public diabetes facebook groups. J. Diab. Sci. Technol. **13**(3), 428–438 (2019)
15. Livas, C., Delli, K., Pandis, N.: "My Invisalign experience;;: content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. Prog. Orthodont. **19**, 1–8 (2018)
16. Lenczowski, E., Dahiya, M.: Psoriasis and the digital landscape: YouTube as an information source for patients and medical professionals. J. Clin. Aesthet. Dermatol. **11**(3), 36–38 (2018)
17. Seabrook, E.M., Kern, M.L., Fulcher, B.D., Rickard, N.S.: Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. J. Med. Internet Res. **20**(5), e168 (2018)
18. Busam, B., Solomon-Moore, E.: Public understanding of childhood obesity: qualitative analysis of news articles and comments on facebook. Health Commun. **00**(00), 1–14 (2021)
19. Eysenbach, G.: Improving the quality of web surveys: the checklist for reporting results of internet E-surveys (CHERRIES). J. Med. Internet Res. **6**(3), 1–6 (2004)
20. Nova, F.F., et al.: "Facebook promotes more harassment": social media ecosystem, skill and marginalized hijra identity in Bangladesh. In: Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW1, pp. 1–35 (2021)

21. Docimo, S., Jacob, B., Seras, K., Ghanem, O.: Closed Facebook groups and COVID-19: an evaluation of utilization prior to and during the pandemic. Surg. Endosc. **35**(9), 4986–4990 (2021)
22. Spiliotopoulos, T., Oakley, I.: Post or tweet: lessons from a study of facebook and twitter usage. In: Following User Pathways: Using Cross Platform and Mixed Methods Analysis in Social Media Studies Workshop at ACM CHI 2016 (2016)
23. Pötzschke, S., Weiß, B.: Realizing a global survey of emigrants through facebook and instagram (2021)
24. Gündüzalp, S., Şener, G.: The analysis of opinions about teaching profession on twitter through text mining. Res. Educ. Media **12**(1), 3–12 (2020)
25. Etta, G., et al.: COVID-19 infodemic on Facebook and containment measures in Italy, United Kingdom and New Zealand. PLoS ONE **17**(5 May), 1–14 (2022)
26. Inuwa-Dutse, I., Liptrott, M., Korkontzelos, I.: Detection of spam-posting accounts on Twitter. Neurocomputing **315**, 496–511 (2018)
27. Lyu, J.C., Luli, G.K.: Understanding the public discussion about the centers for disease control and prevention during the COVID-19 pandemic using twitter data: text mining analysis study. J. Med. Internet Res. **23**(2), e25108 (2021)
28. Anderson, M.: Social media and COVID-19 : Characterizing anti-quarantine comments on Twitter, pp. 2–5 (2020)
29. Al-Ramahi, M., Elnoshokaty, A., El-Gayar, O., Nasralah, T., Wahbeh, A.: Public discourse against masks in the COVID-19 Era: infodemiology study of twitter data. JMIR Public Health Surveill. **7**(4), 1–12 (2021)
30. Dashtian, H., Murthy, D.: Cml-Covid: a large-scale covid-19 twitter dataset with latent topics, sentiment and location information. Academia Lett. **1**, 1–6 (2021)
31. Karami, A., et al.: 2020 U.S. presidential election in swing states: gender differences in Twitter conversations. Int. J. Inf. Manag. Data Insights **2**(2) (2022)
32. Fahey, R.A., Boo, J., Ueda, M.: Covariance in diurnal patterns of suicide-related expressions on Twitter and recorded suicide deaths. Soc. Sci. Med. **253**(March), 112960 (2020)
33. Park, H.W., Park, S., Chong, M.: Conversations and medical news frames on twitter: infodemiological study on COVID-19 in South Korea. J. Med. Internet Res. **22**(5), e18897 (2020)
34. Mahaini, M.I., Li, S.: Detecting cyber security related Twitter accounts and different sub-groups: a multi-classifier approach. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2021, pp. 599–606 (2021)
35. Mohammadi, E., Thelwall, M., Kwasny, M., Holmes, K.L.: Academic information on twitter: a user survey. PLoS ONE **13**(5), 1–18 (2018)
36. Kim, Y., Emery, S.L., Vera, L., David, B., Huang, J.: At the speed of Juul: measuring the twitter conversation related to ENDS and Juul across space and time (2017–2018). Tob. Control **30**(2), 137–146 (2021)
37. Dev, J.: Spring 2020 discussing privacy and surveillance on twitter: a case study of COVID-19 Jayati Dev, pp. 1–10 (2020)
38. Reuter, K., et al.: Monitoring Twitter conversations for targeted recruitment in cancer trials in Los Angeles county: protocol for a mixed-methods pilot study. JMIR Res. Protoc. **7**(9), 1–17 (2018)
39. Abdulsattar, G., Alkubaisi, A.J., Kamaruddin, S.S., Husni, H.: Conceptual framework for stock market classification model using sentiment analysis on twitter based on Hybrid Naïve Bayes Classifiers. Int. J. Eng. Technol. (UAE) **7**(2), 57–61 (2018)

40. Rahman, M.M., Ali, G.G.M.N., Li, X.J., Paul, K.C., Chong, P.H.J.: Twitter and census data analytics to explore socioeconomic factors for post-COVID-19 reopening sentiment. medRxiv (2020)
41. Graham, T., Bruns, A., Angus, D., Hurcombe, E., Hames, S.: #IStandWithDan versus #DictatorDan: the polarised dynamics of Twitter discussions about Victoria's COVID-19 restrictions. Media Int. Aust. **179**(1), 127–148 (2021)
42. Celestini, A., Di Giovanni, M., Guarino, S., Pierri, F.: Information disorders on Italian Facebook during COVID-19 infodemic, pp. 1–16 (2020)
43. Broniatowski, D.A., et al.: Facebook pages, the "disneyland" measles outbreak, and promotion of vaccine refusal as a civil right, 2009–2019. Am. J. Public Health **110**, S312–S318 (2020)
44. Rieder, B., Matamoros-Fernández, A., Coromina, Ò.: From ranking algorithms to 'ranking cultures': investigating the modulation of visibility in YouTube search results. Convergence **24**(1), 50–68 (2018)
45. Vargas Meza, X., Yamanaka, T.: Food communication and its related sentiment in local and organic food videos on YouTube. J. Med. Internet Res. **22**(8), e16761 (2020)
46. Kim, T., Jo, H., Yhee, Y., Koo, C.: Robots, artificial intelligence, and service automation (RAISA) in hospitality: sentiment analysis of YouTube streaming data. Electron. Mark. **32**(1), 259–275 (2022)
47. Permatasari, R., Rakhmawati, N.A.: Features selection for entity resolution in prostitution on twitter. Int. J. Adv. Data Inf. Syst. **2**(1), 53–61 (2021)
48. Gunawan, R., Rahmatulloh, A., Darmawan, I., Firdaus, F.: Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. In: IcoIESE 2018, vol. 2, pp. 283–287 (2019)
49. Krotov, V., Silva, L.: Legality and ethics of web scraping. In: Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018 (2018)
50. Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M., Mbaye, S.N.: Web scraping: state-of-the-art and areas of application. In: Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, pp. 6040–6042 (2019)
51. Mittelmeier, J., Cockayne, H.: Global representations of international students in a time of crisis: a qualitative analysis of Twitter data during COVID-19. In: International Studies in Sociology of Education, pp. 1–18 (2022)
52. Faty, L., Ndiaye, M., Sarr, E.N., Sall, O.: OpinionScraper: a news comments extraction tool for opinion mining. In: 2020 7th International Conference on Social Network Analysis, Management and Security, SNAMS 2020, pp. 4–8 (2020)
53. Hinduja, S., Afrin, M., Mistry, S., Krishna, A.: Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. Int. J. Inf. Manag. Data Insights **2**(2), 100113 (2022)
54. Anand, V., Shukla, R., Gupta, A., Kumar, A.: Customized video filtering on YouTube, pp. 1–13 (2019)
55. Gray, L.: Gender Bias Detection Using Facebook Reactions (2020)
56. Franco-Riquelme, J.N., Bello-Garcia, A., Ordieres-Meré, J.: Indicator Proposal for measuring regional political support for the electoral process on twitter: the case of Spain's 2015 and 2016 general elections. IEEE Access **7**, 62545–62560 (2019)
57. Gutierrez, C.G., et al.: Analyzing and visualizing twitter conversations. In: Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering, pp. 4–13 (2021)
58. Jin, J., Lam, S., Savas, O., McCulloh, I.: Approaches for quantifying video prominence, narratives, discussion: engagement on COVID-19 related youtube videos. In: Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020, pp. 811–818 (2020)