PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

AUTOMATING TWITTER DATA COLLECTION: A RAPIDMINER-BASED CRAWLING SOLUTION

Diemas Arya Komara[1*] Angga Hadiapurwa[2]
***email : diemas@upi.edu**
[1,2]Perpustakaan dan Sains Informasi Universitas Pendidikan Indonesia

Keyword:
Twitter
Rapidminer
Crawling

Article Info :
Submitted date
14/12/2022
Revised date
04/04/2023
Accepted date
06/04/2023

Abstract
Twitter is one of the social media that is favored by various groups, Twitter itself can display information in the form of textual information which can be supplemented with videos or images. This study will discuss how to crawl data on Twitter social media, using the RapidMiner application, 10000 data can be retrieved in one operation. This study aims to provide knowledge about how to use RapidMiner programming to obtain data that can later be processed for purposes related to information science. The results of this study are to show how to collect data on Twitter social media quickly using text mining.

Abstrak
Twitter merupakan salah satu media sosial yang digemari oleh berbagai macam kalangan, twitter sendiri dapat menampilkan informasi dengan kebanyakan berbentuk informasi tekstual yang dapat dilengkapi dengan video atau gambar, dalam kajian keilmuan perpustakan dan sains informasi hal tersebut dapat disebut sebagai data yang tersebar dalam dunia maya. Penelitian ini akan membahas tentang bagaimana cara untuk melakukan crawling terhadap data di media sosial twitter, dengan menggunakan aplikasi Rapidminer dapat diambil sebanyak 10000 data dalam satu kali operasi. Penelitian ini bertujuan untuk memberikan pengetahuan tentang bagaimana cara menggunakan pemograman Rapidminer untuk mendapatkan data yang nantinya dapat diolah untuk keperluan-keperluan yang berkaitan dengan ilmu informasi. Hasil dari penelitian ini adalah menunjukan bagaiman cara untuk mengumpulkan data dalam media sosial twitter secara cepat menggunakan text mining.

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

INTRODUCTION

Social media plays a crucial role in numerous aspects of human life. It is a medium for users to share information, communicate with each other, and share various types of content such as photos and videos. Social media is a digital platform that facilitates these activities (Auxier & Anderson, 2021; Putri, 2021; Rachmawaty, 2021). There are currently 2.8 billion active social media users, with Indonesia alone contributing more than 100 million active social media users (Fadly & Wantoro,2019; Yudhana, Riadi & Zuhriyanto, 2019). Depending on the type of social media they use, users fall into several categories. Social media is typically divided into five categories: social networks, such as Facebook and LinkedIn; microblogging, including Reddit and Twitter; photo sharing, such as Pinterest, Instagram, and Snapchat; video sharing, including TikTok, Twitch, Snackvideo, and Youtube; and instant messaging platforms, such as WhatsApp and Line. Currently, Twitter has 300 million social media users (Juditha, 2019; Al Asad et al., 2019; Mustafa et al., 2020).

Interactions occur on Twitter for approximately 700 million tweets per day, with an average of 7,500 tweets per second (Furqon & Setiawan 2020). Twitter serves as a massive data warehouse, storing billions of data from around the world. If processed and analyzed effectively, the data collected from Twitter can provide valuable insights and information. However, before the data can be transformed into useful information, it must first be collected and extracted using techniques such as crawling. Users can leverage crawling techniques from the Internet to efficiently collect data from Twitter and prepare it for further analysis (Dakhi & Aji, 2022; Sulastomo et al., 2022; Pakpahan & Simanullang, 2022). In this study, the author has utilized a crawling technique to gather data from Twitter. The outcome of this study can be further utilized to analyze and extract valuable information. Specifically, the crawling technique utilized in this study involves the use of the RapidMiner application.

This research has a main focus on how the implementation of data crawling techniques on Twitter social media, post-processing techniques or data processing that has been collected using crawling techniques is not discussed in this study. At the end of the study, testing will be carried out by crawling several keywords. Next, the results of crawling data will be displayed on several keywords that are currently the topic of chat or popular hashtags on Twitter on a certain day.

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

The development of technology and social media, especially Twitter, has had a major impact on various aspects of people's lives. In the academic field, the data generated by Twitter can be a valuable source of information for research in different fields, such as sociology, psychology, political science, medical science, and others. Twitter is a massively popular social media platform with an enormous user base and an astonishing number of tweets produced every second. Therefore, researchers need to have efficient and effective way to collect Twitter data. One way to collect Twitter data is by doing data crawling. However, the process of collecting data manually can take a lot of time and effort. Therefore, research on automating Twitter data collection can provide an efficient and effective solution for collecting large amounts of Twitter data.

Twitter is a commonly utilized social media platform where individuals can obtain information. Twitter has the advantage of displaying data in the form of sentences accompanied by pictures (Kurniawan & Maharani 2020; Pattiiha & Hendry, 2022). Twitter itself is a social media that has information in the form of sentences that can then be easily traced back by readers using the hashtag (#). Twitter, owned by the US-based company Twitter, Inc., is a social networking platform that relies on microblogging. Twitter users can interact with one another by posting tweets on their home page, and they can also create and join tweet communities. Additionally, Twitter allows users to exchange personal messages through chat (Murthy, 2018). Twitter users can post, like, and retweet tweets, Twitter can also be accessed without registering for an account, but users can only read public tweets without interacting with others. Users can interact with Twitter through a browser or mobile interface, or programmatically via its APIs (Chadijah, 2022; Seprina, 2019; Edwanto, 2019).

Tweets are by default public and viewable, but users can limit who sees their tweets by making their account private, which restricts messages to their followers only. This can be done by enabling the account's privacy settings. Users have the option to mute other users they don't want to interact with, hide their tweets from the account, and remove the account from their follower list.

Twitter is what is going on withinside the globally and what humans are speaking approximately proper now. Twitter can be accessed via applications (on mobile devices) and also through internet browsers. With the purposes of sharing information on Twitter each time viable, Programmatic access to Twitter is offered by Twitter to businesses, developers, and users through a Twitter Application Programming Interface

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

(API).

an API, which stands for application programming interface, facilitates communication between software programs by enabling them to exchange requests and data. APIs facilitate the interaction between two or more applications over the internet, functioning as a concept that enables software developers to leverage their programming skills in building new and innovative products (Karami & Collins, 2018; Lee & Kwon, 2018; Muri, Utomo & Sayyidati, 2019). API linking is done by allowing software applications to call an endpoint called: API endpoints that match a particular type of information and later provide an access point (endpoints are usually unique, like phone numbers). Twitter provides access to certain parts of Twitter services through APIs to allow users to create software that integrates with Twitter (Darmawan, 2020; Raharjo, 2021; Sanusi, 2020), such as solutions that help companies reply and feedback to customer comments on Twitter.

The information available on Twitter stands out from data shared on other social media platforms as it portrays details that users have voluntarily made public. By using the Twitter API Platform, individuals can gain entry to a broad range of public Twitter data that is shared worldwide. Twitter APIs also support users in managing their privacy by allowing them to control information sources on Twitter, such as direct messages, and make it available to developers who are approved to access it.

Web crawling begins with a list of URLs to access. The first URL on the list is known as the seed. When the crawler visits this URL, it communicates with the server and retrieves all the hyperlinks available on the web page. Afterward, it generates a collection of website links that need to be explored, known as the "crawl frontier," utilizing a set of criteria, based on a set of guidelines. The URLs that are visited are determined exclusively by the guidelines. If a user wants to archive a website, they can copy and save information during the crawling process. Archives are usually saved for viewing, reading, navigating, and browsing as if they were on the live web (Khder, 2021).

Archives or repositories are specialized platforms created to manage and store collections of website pages. These repositories store individual HTML pages as separate files. Although they function similarly to modern databases in terms of storing and managing data, repositories do not always require the full range of features that a database system provides. Repositories store the latest versions of web pages retrieved by crawlers (Ang, et al., 2021; Pratama, 2018; Sooai & Laniwati, 2021).

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

When dealing with a large amount of data, data crawlers must be selective in downloading web pages due to the limitation of resources. This means that they need to prioritize which pages to download first. Additionally, due to the high frequency of updates and deletions of web pages, it is important to ensure that the downloaded pages are current and accurate.

Due to the vast number of URLs that can be generated by server-side software, web crawlers often struggle to avoid retrieving duplicate content. This is because there are countless combinations of HTTP GET parameters that may lead to identical content. For instance, an online photo archive may allow users to select from several options via the HTTP GET parameter in the URL, resulting in 48 different URLs that lead to the same set of images. As a result, crawlers must navigate through an enormous number of possible script changes to extract unique content, presenting a challenge (Rennhard et al., 2022; Yadav, 2013).

RapidMiner is a software used for data science that was created by the company bearing the same name as the software, integrated and providing for environment data preparation, machine learning, deep learning, text mining, and predictive analytics. RapidMiner is a software platform that can be used for data processing activities until the data is finally processed into information (Hanif et al., 2018; Nofitri & Irawati, 2019; Sudarsono et al., 2021). RapidMiner is a versatile data science software that can serve multiple functions, including business and commerce, research, education, training, rapid prototyping, and application development. It covers all aspects of the machine learning process, such as data preparation, result visualization, model validation, and optimization (Hofmann & Klinkenberg, 2016). RapidMiner is developed on an open-core model. The RapidMiner Studio Free Edition is limited to one logical processor and 10,000 rows of data.

RapidMiner uses a client/server model with servers provisioned on-premises or on public or private cloud infrastructure. RapidMiner provides advanced analytics solutions through a model-driven framework that speeds deployment and reduces errors by virtually eliminating the need for coding. RapidMiner is a versatile data science software created by a company that shares its name. It has a multitude of potential applications, including business, research, education, training, rapid prototyping, and application development. RapidMiner can handle all stages of the machine learning process, such as preparing data, validating models, visualizing results, and optimizing the final output.

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

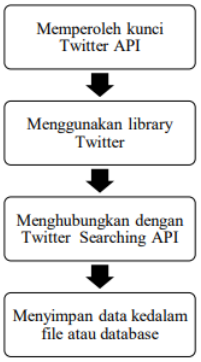
Alternatively, the tool can be called from another program or used as an API (Norris, 2013). RapidMiner offers command-line access to its functions, enabling users to call specific functions individually. Additionally, the software includes various learning schemas, models, and algorithms that can be leveraged, and it can also be extended with R and Python scripting for further functionality.

Research on Twitter data collection automation with RapidMiner can make a significant contribution to the development of data collection methodologies in the academic field. This could open up opportunities for researchers to use similar methods to collect data from other social media platforms, as well as develop more efficient and effective tools for future data collection.

METHODS

The research carried out to implement the crawling technique on Twitter social media uses the literature study method of library science and information science, utilizing related research as a reference in this study. This study also employs the RapidMiner application operating guidelines for Twitter as follows.

Text mining aims to generate information from a set of documents. Text mining is capable of generating information through processing, grouping, and analyzing large amounts of unstructured data (Ruhyana, 2019; Samsir et al., 2021). Text mining is used to obtain useful information from a series of documents with data sources in the text that have an unstructured format.



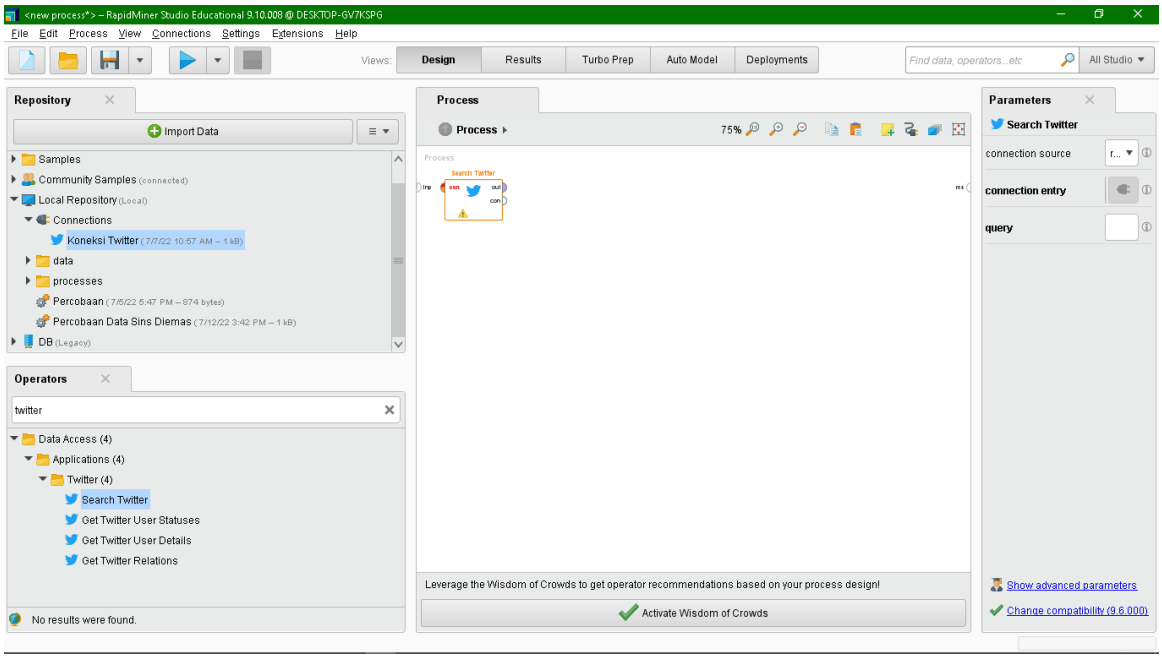
Picture. 1 Research Guidelines

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

In the picture, the stages of the process carried out by researchers to implement the crawling technique are explained. The process starts with obtaining the Twitter API key and continues with the use of the Twitter library. Then, the API is integrated with RapidMiner, followed by connecting the program code with the Twitter Searching API. The last stage involves finding and saving the data that has been collected into a file or database.

RESULT

Twitter API Making

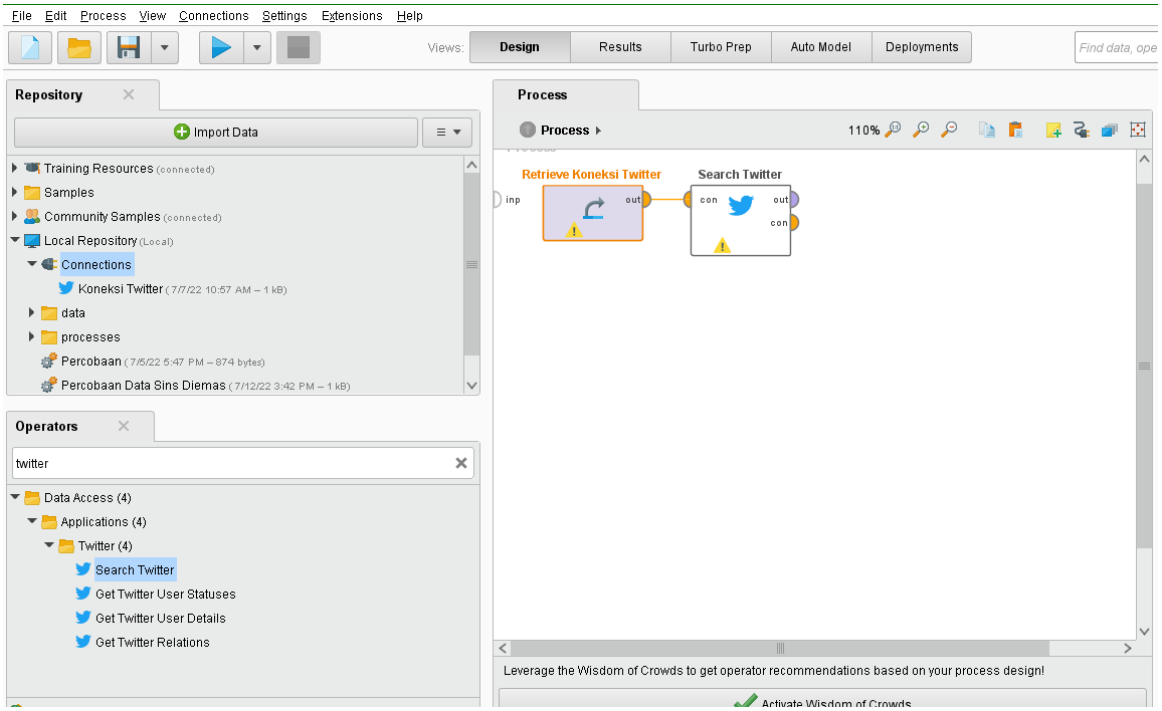


Picture. 2 Twitter API Making

The first step in crawling data from Twitter is to obtain the Twitter API source, which will be used to retrieve or crawl data from the social media platform. The API will then be connected to the RapidMiner application using a token obtained from the API confirmation

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

Twitter API Connecting

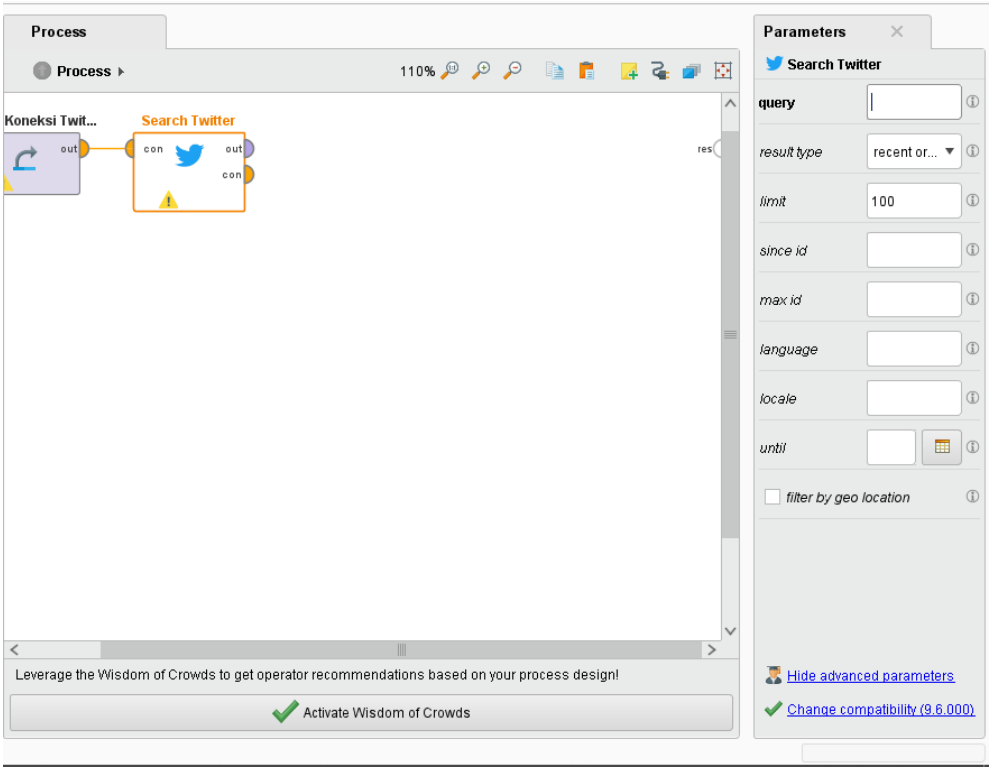


Picture. 3 Twitter API Connecting

Connecting the API with RapidMiner, at this stage the connection between the API and RapidMiner must be established, as it is crucial for the proper execution of the data crawling stage. The process of connecting the API with RapidMiner is the most important aspect to consider, as a series of data crawling processes depend on this step.

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

Twitter Data Crawling

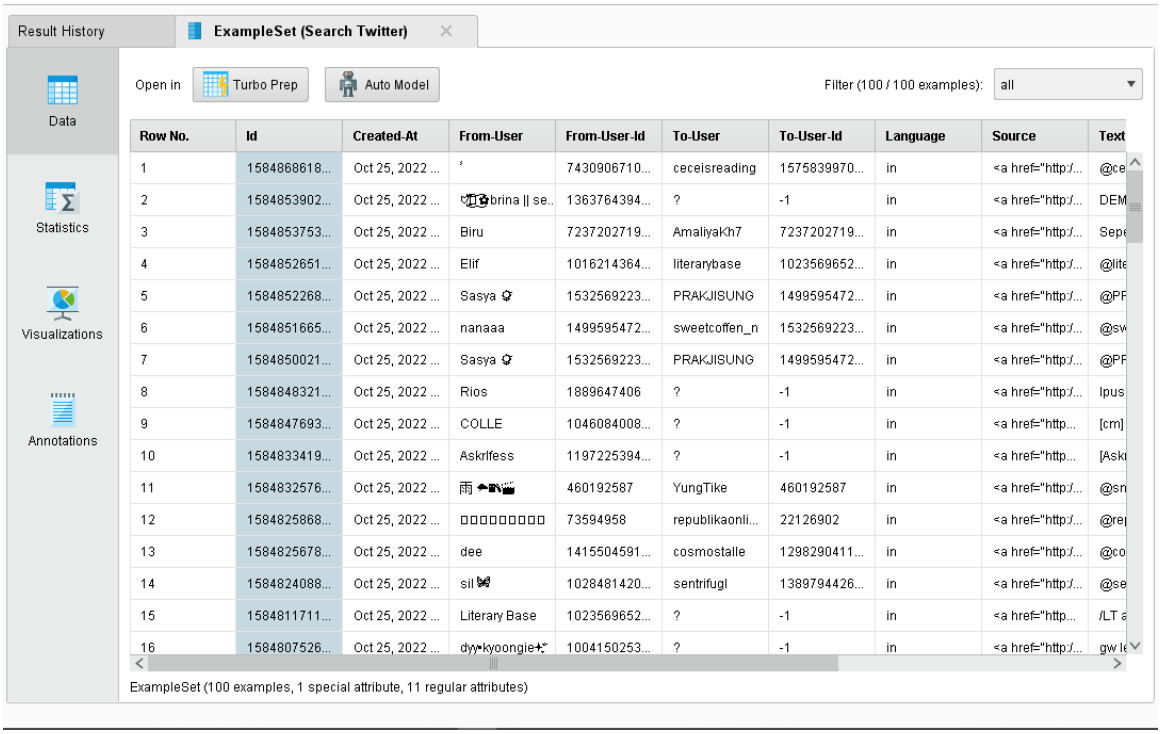


Picture. 4 Twitter Data Crawling

We can start data crawling by using the query feature. In this feature, we can enter search keywords for information or tweets on Twitter. Then, we can choose results based on the popularity of the tweet and/or the time the tweet was made. The limit is how much data we want to retrieve from Twitter. The maximum amount of data that Rapidminer can retrieve in one operation is 10,000. We can also search by language, location, and certain dates.

PUBLIS JOURNAL
Publication Library and Information Science
ISSN 2598-7852
Volume 6 Number 2 November 2022
http://journal.umpo.ac.id/index.php/PUBLIS

Twitter Data Result



Picture. 5 Twitter Data Result

After running the query with the keyword 'ipusnas', we obtained 100 tweets containing the word 'ipusnas' from the Twitter search engine. The obtained data can be processed to become useful information, which can be analyzed for sentiment analysis, marketing, information literacy, or other purposes that can benefit from the data analysis

CONCLUSIONS

Twitter is one of the largest social media platforms in the world with over 300 million users, and 700 million tweets are sent per day, making it a valuable source of data for various research disciplines. Data crawling is an efficient and effective way to collect Twitter data, and research on automating Twitter data collection can provide a solution for collecting large amounts of Twitter data. Twitter users can post, like, and retweet tweets, and tweets are publicly viewable by default. Users can interact with Twitter via a browser, mobile application, or programmatically via its APIs. Overall, Twitter is a

powerful social media platform that provides an extensive source of data for research and other purposes.

In conclusion, the use of RapidMiner in crawling data on Twitter can provide researchers with a reliable and automated solution for collecting large amounts of structured data. With RapidMiner's efficient data collection and processing features, researchers can optimize their data analysis efforts and enhance their understanding of social media behavior. Additionally, this research could inspire further advancements in data collection methodologies, leading to new and innovative solutions for social media data collection and analysis.

ACKNOWLEDGMENTS

The author declares that this paper does not contain plagiarism and there is no other interest in writing this paper.

REFERENCE

Al Asad, N., Pranto, M. A. M., Afreen, S., & Islam, M. M. (2019, November). Depression detection by analyzing social media posts of user. In 2019 I*EEE International Conference on Signal Processing, Information, Communication & Systems* (SPICSCON) (pp. 13-17). IEEE.

Ang, P. S., Teo, D. C. H., Dorajoo, S. R., Prem Kumar, M., Chan, Y. H., Choong, C. T., ... & Poh, J. W. W. (2021). Augmenting product defect surveillance through web crawling and machine learning in Singapore. *Drug safety*, 44(9), 939-948.

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*, *1*, 1-4.

Chadijah, I. (2022). PERILAKU CELEBRITY WORSHIP DAN HUBUNGAN PARASOSIAL DI MEDIA SOSIAL (STUDI KASUS FOLLOWERS TWITTER@ jefrinichol) (Doctoral dissertation, Universitas Hasanuddin).

Dakhi, R. P. B., & Aji, A. M. B. (2022). Sentiment Analysis Of Public Opinions On The Effectiveness Of Online Learning Using Naïve Bayes Algorithm. *JISICOM (Journal of Information System, Informatics and Computing)*, *6*(1), 273-279.

Darmawan, R. (2020). Implementasi Metode K-Nearest Neighbor (K-Nn) Untuk Identifikasi Ujaran Kebencian Terhadap Tokoh Politik Pada Twitter (Doctoral Dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau).

Edwanto, I. K. (2019). Pembangunan Aplikasi Media Sosial Untuk Penyebaran Informasi Keadaan Sekitar Dengan Memanfaatkan Gps Dan Geofencing Berbasis Android (Doctoral dissertation, Universitas Komputer Indonesia).

Fadly, M., & Wantoro, A. (2019, August). Model Sistem Informasi Manajemen Hubungan Pelanggan Dengan Kombinasi Pengelolaan Digital Asset Untuk Meningkatkan Jumlah

Pelanggan. In *Prosiding Seminar Nasional Darmajaya* (Vol. 1, pp. 46-55).

Furqon, R. M., & Setiawan, E. B. (2020). Deteksi berita rumor pada sosial media twitter menggunakan metode naïve bayes multinomial dengan pembobotan tf-idf. *eProceedings of Engineering*, 7(2).

Hanif, M. H. M., Adewole, K. S., Anuar, N. B., & Kamsin, A. (2018). Performance evaluation of machine learning algorithms for spam profile detection on Twitter using WEKA and RapidMiner. *Advanced Science Letters*, 24(2), 1043-1046.

Hofmann, M., & Klinkenberg, R. (Eds.). (2016). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

Juditha, C. (2019). Dukungan Sosial Warganet Di Twitter Terhadap Gaya Komunikasi Pasangan Calon Presiden Pada Debat Pemilu 2019. *Jurnal Studi Komunikasi dan Media*, *23*(1), 87-100.

Karami, A., & Collins, M. (2018). What do the US West Coast public libraries post on Twitter?. *Proceedings of the Association for Information Science and Technology*, 55(1), 216-225.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3).

Kurniawan, F. W., & Maharani, W. (2020). Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec. *eProceedings of Engineering*, 7(2).

Lee, S. Y., & Kwon, Y. (2018). Twitter as a place where people meet to make suicide pacts. *Public health*, 159, 21-26.

Muri, M. F. A., Utomo, H. S., & Sayyidati, R. (2019). Search Engine Get Application Programming Interface. *Jurnal Sains dan Informatika*, 5(2), 88-97.

Murthy, D. (2018). *Twitter*. Cambridge: Polity Press.

Mustafa, M. Z., Kasmani, M. F., Ariffin, M. Y. M., & Noor, K. M. (2020). Perhubungan Awam, Media Sosial, Etika Kerja Islam Serta Trend Penggunaan Media Sosial di Sektor Awam Kementerian di Malaysia. *ABQARI: Journal of Islamic Social Sciences and Humanity*.

Nofitri, R., & Irawati, N. (2019). Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 5(2), 199-204.

Norris, D. (2013). Rapid Miner-a potential game changer. *IT-Director. com*.

Pakpahan, S., & Simanullang, A. (2022). Analisis Sentimen Integritas KPK Tahun 2021 Pencegahan Korupsi pada Twitter KPK Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes. *Citra Sains Teknologi*, *2*(1), 63-73.

Pattiiha, F. S., & Hendry, H. (2022). Perbandingan Metode K-NN, Naïve Bayes, Decision Tree untuk Analisis Sentimen Tweet Twitter Terkait Opini Terhadap PT PAL Indonesia. *JURIKOM (Jurnal Riset Komputer)*, 9(2), 506-514.

Pratama, R. P. (2018). Aplikasi deteksi plagiarisme menggunakan metode cosine similarity (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

Putri, A. D. (2021). Maksimalisasi Media Sosial untuk Meningkatkan Pendapatan dan Pengembangan Diri Generasi Z di MAN 1 Pesawaran. *Journal of Social Sciences and Technology for Community Service (JSSTCS)*, *2*(2), 37-42.

Rachmawaty, A. (2021). Optimasi Media Sosial Dalam Meningkatkan Penjualan di Masa

Pembatasan Sosial Berskala Besar. *Tematik: Jurnal Teknologi Informasi Komunikasi (e-Journal)*, *8*(1), 29-44.

Raharjo, M. (2021). Analisis Sentiment Twitter Terkait Kondisi Social Distancing Di Indonesia Dengan Model Voting Ensemble.

Rennhard, M., Kushnir, M., Favre, O., Esposito, D., & Zahnd, V. (2022). Automating the Detection of Access Control Vulnerabilities in Web Applications. *SN Computer Science*, 3(5), 1-24.

Ruhyana, N. (2019). Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil/Genap Pada Twitter Dengan Metode Klasifikasi Naive Bayes. *ikraith-informatika*, 3(1), 94-99.

Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, 5(1), 157-163.

Sanusi, R. (2021). Analisis Sentimen Pada Twitter Terhadap Program Kartu Prakerja Menggunakan Long Short Term Memory (Doctoral Dissertation, Stmik Akakom).

Seprina, I. (2019). Analisis Sentimen Masyarakat Terhadap Data Tweet Lazada Menggunakan Text Mining Dan Algoritma Naive Bayes Classifier. In *Bina Darma Conference on Computer Science* (BDCCS) (Vol. 1, No. 4, pp. 998-1004).

Sooai, A. G., & Laniwati, M. (2021). Ekstraksi Ulasan Sentimen Film dari Twitter dengan Naïve Bayes pada Situs Web Media Sosial Penggemar Film. *INSYST: Journal of Intelligent System and Computation*, 3(1), 49-54.

Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner. *JBASE-Journal of Business and Audit Information Systems*, 4(1).

Sulastomo, H., Ramadiansyah, R., Gibran, K., Maryansyah, E., & Tegar, A. (2022). Analisis Sentimen Pada Twitter@ Ovo_Id dengan Metode Support Vectore Machine (SVM). *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 6(2), 1050-1056.

Yadav, S. (2013). Protection from Crawler using. htaccess. IJLTET *International Journal of Latest Trends in Engineering and Technology.*

Yudhana, A., Riadi, I., & Zuhriyanto, I. (2019). Analisis Live Forensics Aplikasi Media Sosial Pada Browser Menggunakan Metode Digital Forensics Research Workshop (DFRWS). *Techno (Jurnal Fakultas Teknik, Universitas Muhammadiyah Purwokerto)*, *20*(2), 125-130.