IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12712

A High-Performance Approach to Real-Time Big Data Collection, Storage, and Analysis

Arjun B Prasad[1], Prof. K Sharat[2]

MCA Graduate (2nd Year), Dept. of MCA, Bangalore Institute of Technology, Bangalore, India[1]

Guide**,** Asst. Prof / Dept. of MCA, Bangalore Institute of Technology, Bangalore, India[2]

Abstract: Twitter is a well-known social media platform that has stood a test of time. a large majority of people chose Twitter as their social media channel of choice for trustworthy scientific news and information in many global communities. However, the constraints of The development of affordable data science is hindered by the Twitter app software interface (API). solutions for academic institutions. To fully utilise the data analytics offered by Researchers must pay considerable expenses in order to use Twitter with a free API account. We introduce our big data analytics tool This piece, it was created at Lakehead the University's DaTALab in Canada and lets users to quickly access vast amounts of Twitter data while focus on their search criteria on Twitter. The platform makes it easy to gather net data, which is subsequently cleared up using A series of filters before artificial intellect (AI) is used and ML systems. (AI). Our main area of concentration has been healthcare-related research, demonstrating the platform's potency. The platform itself, however, is adaptable to any intriguing subject. The data. was gathered and processed can be used for additional AI/ML analysis. To highlight the effectiveness of our system for upcoming healthcare research projects, we demonstrate our platform utilising a specific search topic.

Index Terms— Big data, social intelligence, analytics, web-based, robotic intelligence (AI), and data mining.

I. INTRODUCTION

TWITTER has proven to be an effective portal for research on social networks that enables the testing of new algorithms, investigation of belief dynamics, and collection of substantial amounts of data for study. The enormous amounts of facts that has been gathered and extensively filtered are now exposed to strong computing techniques, including classification, they can be hard for scholars to finish. For instance, a sizable data collection may require an exclusive server for a long time to gather data. or surpass the storage limits of a research team's accessible storage.

The choice and application of a variety of preprocessing and analytic approaches can take some time. Additionally, manually annotating Twitter data becomes a time-consuming operation when a researcher needs labelled data for a model's training, which usually limits the size the data gathering. any one of these problems is addressed by our scalable web-based platform, which also offers scalability so that researchers can gather data simultaneously. We provide an easy way to gather and host enormous data sets, apply cutting-edge preprocessing methods, and utilise large tweet sets may be labelled using machine learningA organisation can first set up a Twitter account. flow by indicating its search terms and desired amount of tweets. The quantity of tweets that can be gathered concurrently is a current Twitter- specific issue. As a result, we provide study teams with convenient queue entry, where we designate a server with few assets the greatest number of data collecting requests.

After data collection is complete, a study team can decide which preprocessing choices to use and can do procedures like removing duplicate tweets or converting emoticons, acronyms, and abbreviations into words.

Our platform also offers a thorough description of each option to aid Because alternatives must be used, research teams are assisted in learning their choices and effortlessly handling the ordering amongst them. in a specified order.

Not noting that it offers a cutting-edge technology solution to the challenge of marking huge numbers of tweets. Using Amazon Mechanical Turk categorise Twitter data, we specifically combine crowd-sourcing with Social network information gathering. When a Our system works with Amazon Mechanical Turk to establish a study that asks for volunteers and gives them access to an online poll for the annotation. The research team wishes to annotate a huge collection of tweets. The survey is also generated naturally, accounting for survey length (the number of tweets) and crossover (the number of respondents). participants needed to annotate each tweet). required for each participant to annotate) into account.

II. RELATED WORK

The collection, analysing, and evaluation of social data are all topics covered in this section. Significant uses in the following fields were identified by Cambria et al. [1] in their recent study of computer approaches for assessing huge social data. health research and education.

Alotaibi et al. [2] provided an in-depth review that targeted big data approaches to care supply chains. They covered important concepts, such as big data (analytics), notably big data in healthcare. The they examined management of supply chains in the context of healthcare.

Twitter was utilised by To recruit volunteers for a test of health, Wasilewski and many. The authors built a set of criteria that might be used to assess whether Twitter users belong for particular types of medical research relying on their tweet activity. endeavours. by watching how they use Twitter online.

In order to look into and characterise the dissemination of obtaining knowledge of medicinal drug abuse (DA) via internet social networks, A detailed examination Sequeira et al.'s analysis of the data for 0.42 million users of Twitter was done [4]. Twitter data was received for a set of words. that included the brand and generic names of popular drugs of abuse, and they utilised a variety of Classifying the data using data mining (ML) and deep learn (DL) techniques into classes for either drug abuse (DA) or nonabuse (NA).

This made it possible to look into the twitter stream advertising DA that spread over the platform in more detail.

For us to recognise them and determine whether their pharmacological treatments contributed to good viral sentiment, Adrover et al.'s [5] examination of the sentiment of HIV-positive Twitter users. They used both automatic and human techniques to remove noise from a batch of about 50 data points. million tweets, including crowdsourcing, computational techniques, keyword filtering, and machine learning (ML).

Angiani et al.'s [6] detailed examination of Basic prep techniques for attitude analysis of weblog data have shown how important it is to adopt Several strategies to increase system precision. Relevant information cleaning procedures for Twit are described, taking into account the usage of URLs, mentions, and tags., signs, and other widely used Internet fonts. This study on sentiment analysis employing filtering is significant and well-written.

III. TWITTER DATA IN RESEARCH

Since Twitter's launch in 2006, the extensive data sets as tweets may provide lead to the release of several research studies. Using seo and article alliance tools like Google Scholar, we were able to find articles that contained the word "Twitter" in the headline. To show the broad research expansion, the number of papers including "Twitter" were then categorised by the creating year. In 2009, there were 313 works with the term in the title; by 2019, there were 3760.

It would lead one to believe that as Twitter gets more well-known and widely used, more media outlets are also using the site. Work made in these papers frequently makes use of user data and features taken from tweets. The researchers must be able to gather or get sets of data that are focused on a certain topic, a group of keywords, or an appropriate region. Such Data sets may be acquired via either the Twitter API that is available. or curated data. When it's the case, the first example is preferable since it is uncertain if there data accessible in the latter case that meets the goals of the research.

The creation of a tool for the collecting and preparation of Twitter data is a labor-intensive but essential stage that the researchers must achieve before continuing with their work. This situation serves as the impetus for us to create a programme that would make it easier for us to gather, preprocess, and curate Twitter data. We go over how our application streamlines and expedites the repetitive data gathering processes while also providing a System for crowd-mediated analysis of sentiment and tagging of the gathered data.

IV. METHODOLOGY

A Singleton Tweet Harvester

Python script only called "tweepy," which includes the Twitter RESTful client API and gives access to its features, is the initial implementation of the tweet harvester.

The developer's Twitter API details are hard-coded to The tool's authentication Python source code. "pika," a library, uses its Advanced Messaging Waiting lists Protocol (AMQP) protocol to interact with the MongoDB instance. Twitter stream data is kept as JSON in a MongoDB database.

Using a streaming session, the singular tweet collector transmits collected tweets in serial format to an instance of the Tweety Stream Listener.

Each tweet that comes in is processed It is processed after being filtered with the cleaning library, before being sent to a single input address. As long as a live session is in progress, the script will continue to run. hasn't been interrupted by the user on purpose. Close the stream's broadcasting session so that you can change the keywords, and then manually edit the keywords within the code before initiating another session.

Although this harvesting technique is easy to deploy and use, it has several drawbacks. For For example, owing to limits imposed by Twitter, this procedure can only accept a single set of filters per tweet stream. API's Basic Engineer account. Additionally, the stream must be manually updated and launched, and there exists not an interface to control the service.

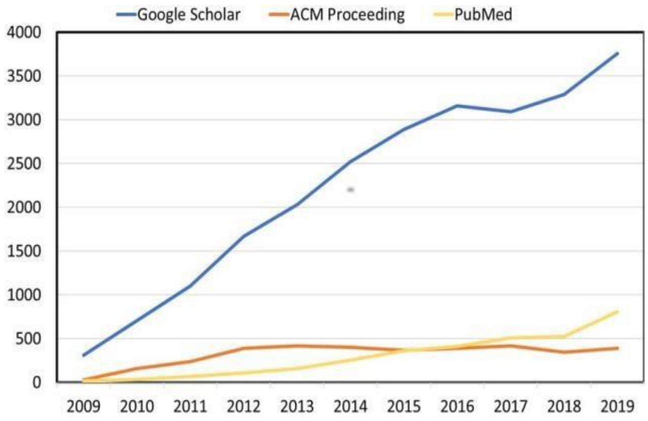

Fig. 1. Publication trend.

Parallel Stream Support for the Singleton Tweet Harvester

The sporadic tweet harvester is updated to handle a few parallel streams in order to get around these restrictions. The consumer-key, consumer_secret, access_token, and access_token_secret need to set up streams with the Twitter API are stored in the application's setup file. A code block for a tag manager as a code block for tweeting segregation have been added to the Python script already in place. By scanning several streams and creating a word pattern to be used for post-segregation of new tweets, the keyword manager block talks with the setting file. Each tweet is put through a series of regular expression tests, and if they pass, they cause the replies to be grouped into streams that match the regular expressions. The implication of this is that one tweet may replace numerous normal

Tweet Cleaning Library and Service

To do analysis properly, tweets must be cleansed following their provision by the buyer or collection. Traditional data mining methods in addition to more modern technologies like computer vision (AI), algorithmic learning (ML), and fuzzy learning (DL) may all be used to analyse data. The tweet filter is a Python programme that may sequentially apply filters that the user selects to the gathered data set.

A separate file called filters.py contains the useful filters for actions like taking away hashtags, taking away emojis, and fixing mistakes. The basic TweetFilter class, which has a sole filter that produces a simpler form of a supplied text, is the ancestor of all filter classes in the file.

This solution generates increasingly cleaner tweets by looping the original text of the tweet after each filter is applied. The tweet scrubber is a simple Flask site that exposes tweet cleaning capability to API endpoints, allowing interaction with the web site and its algorithms.

DOI: 10.17148/IJARCCE.2023.12712

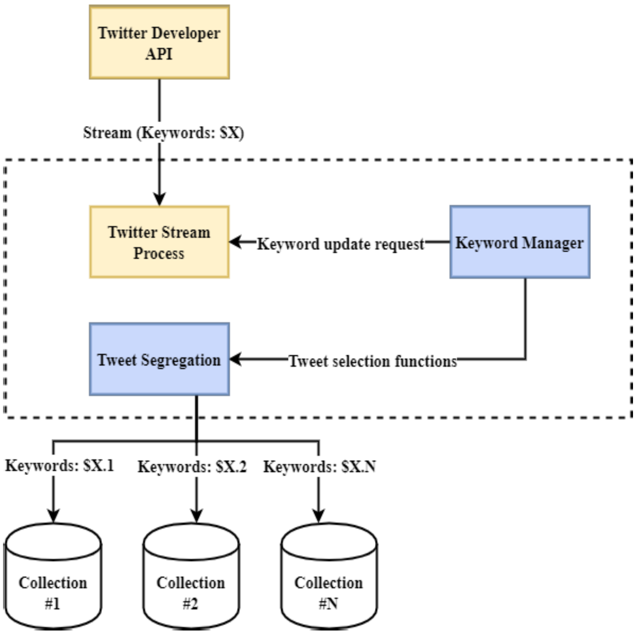

Fig. 2. Concurrent Stream.

SCALING

A technique, network, or system&#39;s capacity to growis referred to as its &quot;scalability&quot; accommodate a growing load. For instance, as more assets come in and the total result rises, a system may be termed scalable.

Since We may expand the variety of nodes in the set because our solution is highly scalable. To exponentially increase system output. Horizontal scaling enables several hardware entities to increase capabilities by operating as a single entity.

The nodes indicate the computers that are capable of supporting the tasks and can each have separate RabbitMQ worker sharing the same network. In the case that one node in the cluster fails, the RabbitMQ broker handles balancing, sharing, and requeuing the task to the left servers. fails.

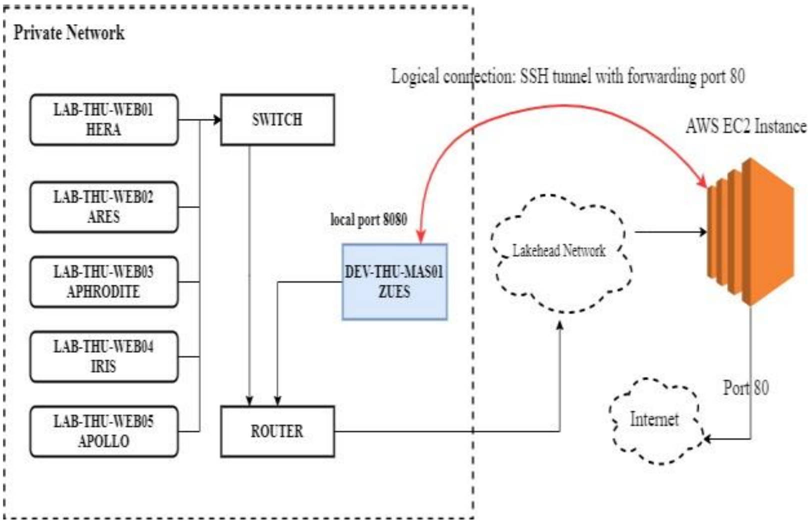

Fig. 3. Deployment Strategy.

DOI: 10.17148/IJARCCE.2023.12712

V. APPLICATIONS

Real-Time Analysis Application

The programme users can submit a list of topics to be watched in real-time for immediate Twitter monitoring. It runs a continuous Twitter feed listener using a collecting worker. The listener&#39;s streaming tweets are sent over the web port and display on a map alongside their positions.

Two columns give some of the tweets that have received the most likes and praise, and immediate analysis of the mood of the posts is also provided.

VI. CONCLUSION

When huge amounts of user-generated or user- driven When gathering data for research projects, Twitter has cemented its place as the go-to source. Due to the current economic value of big data, the free API only makes a tiny portion of all the data that a paid user on the API could have access to. Nearly 1% of the real data is seldom sent by using the API. set, despite being a cost-effective method. We presented the scalable, domestically developed Twitter infrastructure we deployed at Lake Head College in Canada in this study. Our technology has demonstrated that it can increase the amount of Twitter API data that is accessible by levels judged enough for reliable scholarly research. When it comes to healthcare-related applications, our technology has shown considerable level of efficacy. challenges and research through implementation in healthcare, as demonstrated in this paper and our prior work. Numerous papers with and will use our tool serve as evidence of how effectively it prepares what is collected for use in ML-based research. Being localised would be really beneficial for this remote web site. able to run concurrently in North Europe and America since it would allow a greater range of data streams to enter the data collection process.

REFERENCES

[1] E. Cambria, N. Howard, Y. Xia, and T.-S. Chua, “Computational intelligence for big social data analysis,” IEEE Comput. Intell. Mag., vol. 11, no. 3, pp. 8–9, Aug. 2016.

[2] S. Alotaibi, R. Mehmood, and I. Katib, “The role of big data and Twitter data analytics in healthcare supply chain,” in Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies. Berlin, Germany: Springer, 2019, pp. 267–279.

[3] M. B. Wasilewski, J. N. Stinson, F. Webster, and J. I. Cameron, “Using Twitter to recruit participants for health research: An example from a caregiving study,” Health Inform. J., vol. 25, no. 4, pp. 1485–1497, Dec. 2019.

[4] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, “A large-scale study of the Twitter follower network to characterize the spread of prescription drug abuse tweets,” IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1232–1244, Dec. 2019.

[5] C. Adrover, T. Bodnar, Z. Huang, A. Telenti, and M. Salathé, “Identify- ing adverse effects of HIV drug treatment and associated sentiments using Twitter,” JMIR Public Health Surveill., vol. 1, no. 2, p. e7, Jul. 2015.

[6] G. Angiani et al., “A comparison between preprocessing techniques for sentiment analysis in Twitter,” in Proc. KDWeb, 2016, pp. 1–15.