Journal of Lorem
HTTPS://EJOURNALEJOURNAL.MEDIAKUNKUN.COM/INDEX.PHP/LOREM
LOREM : Computational Engineering and Computer Information Systems
E-ISSN : 3032-3924, P-ISSN : 3032-3894

Unveiling X/Twitter's Sentiment Landscape: A Python Crawler That Maps Opinion Using Advanced Search

Iqbal Sabilirrasyad[1], Agung Muliawan[2], Masud Hermansyah[3], Nur Andita Prasetyo[4], Abdul Wahid[5]

Software Engineering, Institut Teknologi dan Sains Mandala[1,2,5]
Information Systems and Technology, Institut Teknologi dan Sains Mandala[3,4]

ABSTRACT

Sentiment analysis is a method for determining attitudes towards a particular event or topic. Twitter or widely known as X now is a popular micro-blogging social media platform frequently used to express emotions. It is well-suited for sentiment analysis. However, some Twitter data retrieval applications have limited search capabilities. Sometimes, Twitter searches can lead to discussions that are unrelated to the intended topic, such as scams or frauds that exploit popular or common hashtags. In addressing this issue Twitter offers an advanced search function that enables detailed topic searches to meet specific information needs, such as sentiment analysis. This study presents a Python-based application for crawling Twitter data using a detailed search similar to advanced search on Twitter. The processed data is used to determine the average sentiment value of the searched topics. To calculate this sentiment value, the researcher utilises Stanza during the search process.

Keywords: *Sentimen Analysis, Twitter, Stanza, Social Media, Crawler*

Corresponding Author:
Iqbal Sabilirrasyad
- iqbalnorth@gmail.com

Received: September 11, 2023
Revised: October 11, 2023
Accepted: January 02, 2024
Published: January 11, 2024

1. INTRODUCTION

Social media is frequently used to express opinions and feelings. It can be a source of information on health, science or current events, such as the upcoming presidential election in Indonesia for the period 2024-2029. Twitter, which was formed in 2006 and is now owned by Tesla CEO Elon Musk, remains a popular social media platform due to its existing features and users. People from all over the world use Twitter for micro-blogging and to express their thoughts on various events. Twitter is well suited to sentiment analysis. Sentiment extraction, which involves studying people's responses to a particular topic, is a common research method. It has been widely used to gather information and data. Research conducted by Hermansyah analysed sentiment on the topic of MBKM using Twitter (Hermansyah et al., 2023). Catelli conducted research on sentiments towards COVID-19 vaccines in Italy (Catelli et al., 2023). Setiyawati and Cahyono conducted research on the sentiment generated towards smokers through Twitter (Setiyawati & Cahyono, 2023). The study focused on using tweet data from the platform. The use of subject-specific vocabulary is necessary to convey the meaning more precisely. The authors maintain a formal register and avoid biased language. The text adheres to style guides and citation consistency. The addition of further aspects has

been avoided to maintain the original content. Searching for data in third-party applications and the provided API can be suboptimal, especially when needing tweet data with specific words and topics. The text is grammatically correct and follows a clear and logical structure. This feature enables detailed search by keywords, hashtags and even language. On Twitter, users can use the advanced search feature to find specific tweets or topics. By entering the necessary search criteria, users can easily find the tweets they need. After conducting an advanced search on Twitter using specific keywords, a list of sentences containing relevant commands will be generated. The information displayed will correspond to the search pattern used. Several researchers have found that this process is rarely carried out, which can result in data that is mixed with ads or spam using the same hashtag or keywords.

This research aims to develop an application that crawls data from Twitter using the advanced search feature. The output will provide an average sentiment of the collected tweets. The data can also be saved in CSV format. The objective is to gain insights into the sentiment of the searched topic on Twitter.

2. METHODS

Several researchers have retrieved data through third-party applications, including Catelli et al. (2023), Hermansyah et al. (2023), and Setiyawati & Cahyono (2023). Additionally, Pratikakis (2018) developed a lightweight and high-performing Twitter crawler for research purposes (Pratikakis, 2018). In addition to Purwandari's research, this paper discusses Twitter applications that provide weather information for Indonesia (Purwandari et al., 2023). The application aims to retrieve tweet data from Twitter using Python and Selenium. The downloaded data is then analysed for sentiment, and the output of the application is the sentiment value of the searched topic. The system's flowchart is shown in the figure below.

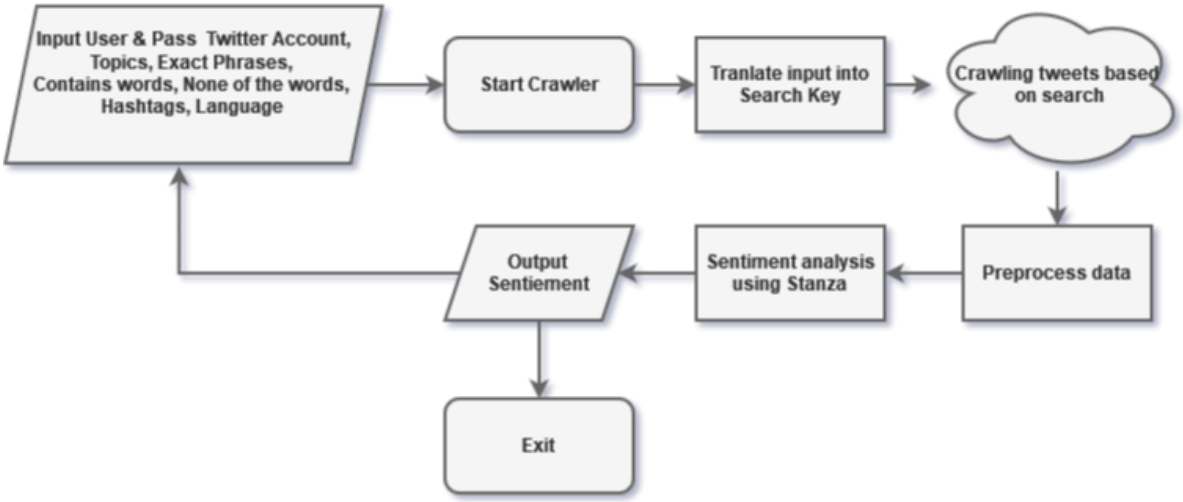

Figure 1. Application flowchart

Selenium and Python programming language are used due to the limited data that can be retrieved from the Twitter API. The retrieved data is not private but rather tweet data that exists or appears in the search based on the search criteria. The application is built on a Windows platform using a simple GUI in Python. Figure 1 shows the initial step of entering the user ID or email and the account to be used. Afterwards, several inputs will be entered into the advanced search to search for tweets in more detail. An explanation of each input is provided in the following table.

Table 1. Input Description

Input	Description
User ID	Userd ID yang digunakan untuk login kedalam situs twitter atau X
Password ID	Password yang digunakan untuk login kedalam situs twitter atau X
All of these words	Seluruh kata yang dimasukkan akan menjadi key search dalam satu tweets
This exact phrase	Kata yang spesifik yang harus ada dalam tweet
Any of these words	Kata yang harus terdapat salah satu kata yang tertera di dalam tweet
None of this	Tidak boleh ada kata yang sama dalam tweet dengan kata yang dimasukkan
Hashtag	Hastag yang ingin dicari di dalam tweet
Language	Bahasa spesifik tweet yang ingin dicari

Apart from the Twitter or X ID and password, all other data can be filled in according to the search needs. Once filled in, press the search button to proceed to the next step. The subsequent step involves converting the existing input into the key search used in Twitter. Advanced search in Twitter is similar to regular search, but with the addition of symbols and characters that indicate more specific search needs.

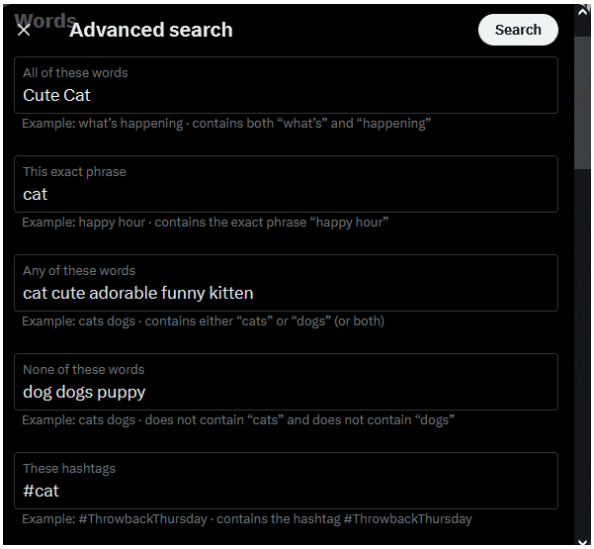

Figure 2. Twitter/X Advance Search
Source: https://twitter.com/search-advanced?lang=en

When entering words in the 'All of these words' section, such as 'Cute Cat', or using the exact phrase 'cat', or any of the words 'cat', 'cute', 'adorable', 'funny', or 'kitten', and excluding any of the words 'dog', 'dogs', or 'puppy', as well as using the hashtag '#cat', the search results will be updated accordingly as demonstrated in Figure 3.

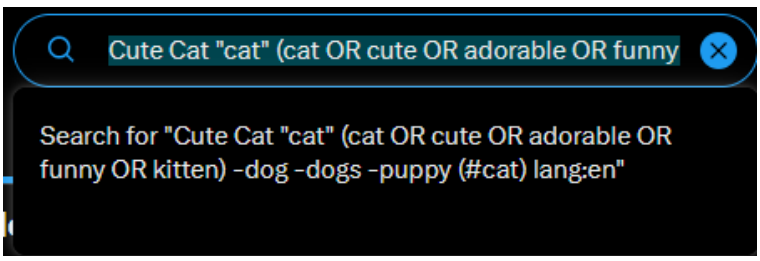

Figure 3. Result of advance search
Source: https://twitter.com/

The words are changed by splitting and adding appropriate characters to resemble the pattern of the advanced search on Twitter. Then, the crawling program searches for tweet data based on the specified search.

The obtained data must be preprocessed before extracting the sentiment from each tweet. To retrieve the sentiment, Stanza is used. Stanza is a collection of accurate and efficient tools for linguistic analysis in multiple human languages (Qi et al., 2020). Elkins conducted research on the shape of a story using machine learning, including the use of Stanza (Elkins, 2022) such as preprocessing stages using Machine Learning techniques such as data cleaning, data transformation, feature extraction, and data labeling (Wiranto, Sabilirrasyad, et al., 2023). The library allows for the analysis of sentiment with two or five different classes. The text describes the use of the Stanza library for sentiment analysis, which employs a CNN classifier trained on several data sources, including Stanford Sentiment Treebank, MELD, SLSD, Arguana, and Airline Twitter Sentiment. The text is clear and concise, with a logical flow of information and no grammatical errors or spelling mistakes. The program seeks the average sentiment value of all tweets obtained to determine whether the searched topic has a positive or negative sentiment. The sentiment assessment scale ranges from 0 to 1. To search for a different topic, refill the corresponding column or exit the program.

Existing applications undergo multiple tests with specific topics to evaluate performance and identify areas for development. Throughout this process, various tests were conducted on the following topics.

Table 2. Topics of Experiments

	All of these words	**This exact phrase**	**Any of These words**	**None of these words**	**Hashtag**
Topic 1	Soccer	soccer	soccer football player transfer club sports	baseball volleyball bowling handball hand shoes chlotes discord join official bet put lucky luck	#soccer
Topic 2	Video Game	game	awards game of the year genre cool	table dnd poker monopoly tabletop discord join sell buy sold	#game
Topic 3	Health	Healthy	sick healty hospital diet medicine	died parent relationship love	#healthy

Each trial will decrease the number of individually filled out forms. The presented topics are intentionally common to reduce bias. The parameters considered are the crawling process speed and the sentiment value generated for each topic.

3. RESULTS AND DISCUSSION

The figure displays the results of the sentiment analysis application developed using data crawlers.

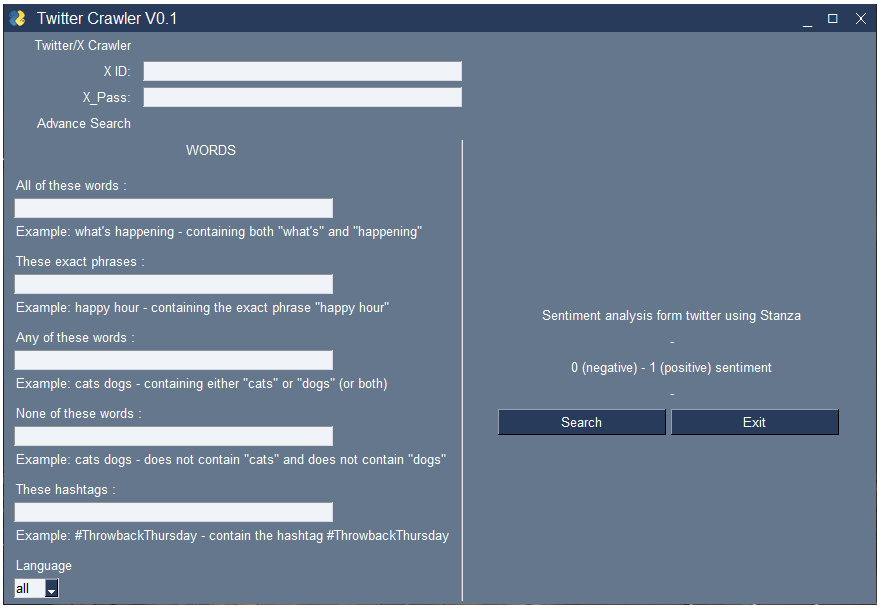

Figure 3. Application GUI

The application requires improvements as some processes produce errors when dealing with tweets that lack text. However, no issues arise when processing text-dominant tweets.

The table below displays the results of the experiment conducted on the application using various topic themes.

Table . Result of Experiments

	Sentiment			**Time Elapsed**		
	Full	No Hashtag	No Unincluded Word	Full	No Hashtag	No Unincluded Word
Topic 1	0.44	0.32	0.30	434.06	489.52	523.22
Topic 2	0.23	0.27	0.24	536.86	539.17	539.17
Topic 3	0.27	0.35	0.38	552.37	617.98	753.490

The experiment was conducted by reducing the input of each topic. It can be seen that each sentiment issued has a different variation. With the highest sentiment being Topic 1 with a value of 0.44 where all existing search parts are fulfilled. From the results of the length of time the programme runs also increases as the points are reduced, this explains that by reducing the points in the search column it is possible to search for tweets more broadly and generally.

So that it produces more data taken and takes longer to do the process of determining the sentiment of each topic.

4. CONCLUSION

The application requires improvement and additional features. During the experiment process, errors were frequently encountered due to tweet data being in image format. This issue can be resolved by implementing filters. The advanced search feature includes a filter for links, which the crawler application has yet to utilize. The advanced search feature includes a filter for links, which the crawler application has yet to utilize. The advanced search feature includes a filter for links, which the crawler application has yet to utilize. Additionally, some of the retrieved information is in the form of advertisements, which can be reduced by limiting tweets that do not contain links. This may affect the amount of data obtained. In the future, researchers may develop a more reliable system and an improved process.

REFERENCES

Catelli, R., Pelosi, S., Comito, C., Pizzuti, C., & Esposito, M. (2023). Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. *Computers in Biology and Medicine*, *158*, 106876. https://doi.org/10.1016/j.compbiomed.2023.106876

Elkins, K. (2022). The Shapes of Stories: Sentiment Analysis for Narrative. *Elements in Digital Literary Studies*. https://doi.org/10.1017/9781009270403

Hermansyah, M., Firdausi, F., Wahid, A., & Prasetyo, N. A. (2023). Twitter Sentiment Analysis for Exploring Public Opinion on the Merdeka Belajar-Kampus Merdeka (MBKM) 2023 with the Naïve Bayes Classifier Algorithm. *PROCEEDING INTERNATIONAL CONFERENCE ON ECONOMICS, BUSINESS AND INFORMATION TECHNOLOGY (ICEBIT)*, *4*, 852–860.

Pratikakis, P. (2018, April 20). *twAwler: A lightweight twitter crawler*. arXiv.Org. https://arxiv.org/abs/1804.07748v1

Purwandari, K., Perdana, R. B., Sigalingging, J. W. C., Rahutomo, R., & Pardamean, B. (2023). Automatic Smart Crawling on Twitter for Weather Information in Indonesia. *Procedia Computer Science*, *227*, 795–804. https://doi.org/10.1016/j.procs.2023.10.585

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages* (arXiv:2003.07082). arXiv. https://doi.org/10.48550/arXiv.2003.07082

Setiyawati, D., & Cahyono, N. (2023). Analisis Sentimen Pengguna Sosial Media Twitter Terhadap Perokok Di Indonesia. *Indonesian Journal of Computer Science*, *12*(1), Article 1. https://doi.org/10.33022/ijcs.v12i1.3154

Wiranto, F., Sabilirrasyad, I., Hermansyah, M., Mandala, S., Wiranto, F., & Mandala, S. (2023). *Optimizing Forecasting of Dow Jones Stock Index in New York amid Uncertain Global Conditions in 2023 : A Combined Approach of ARIMA and Machine Learning Models*. *1*, 73–88.