E-ISSN : 2988-1986
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 10 No 8 Tahun 2026

DATA CRAWLING TECHNIQUES IN SOCIAL MEDIA AS AN INITIAL STAGE IN UNDERSTANDING BASIC MACHINE LEARNING TECHNIQUES (CASE STUDY: RAM PRICE INCREASE 2025)

Dicky Ardiansyah [1], Nadya Amalia [2], Mhd Bimanda Syafputra [3] , Eka Rahmadanta Sitepu [4]
[1,2,3,4] Information Systems Study Program , Faculty of Computer Science, STMIK Kaputama Binjai
E-mail: habibzamzami099@gmail.com [1], nadyaamaliaa1008@gmail.com [2], bimandasyafputra5338@gmail.com [3], dhanta07@gmail.com [4]

ABSTRACT

Social media is an important data source for understanding public responses to technological issues. This study aims to apply data crawling techniques as an initial stage of basic machine learning with a case study of the increase in Random Access Memory (RAM) prices in 2025. Data was collected from the Twitter (X) platform using the Twitter API, then went through pre-processing, tokenization, and token frequency analysis stages. The results of the study show that pre-processing effectively improves the quality of text data, while token frequency analysis is able to identify dominant words that represent the focus of public discussion. This study confirms that data crawling and token frequency analysis are important foundations in the initial pipeline of text-based machine learning.

Keywords: Crawling Data, Social Media, Twitter API, Text Data, Token Frequency, Machine Learning

Article History
Received: *January* 2026
Reviewed: *January* 2026
Published: *January* 2026

Plagiarism Checker No. 235
DOI Prefix:10.8734/Cohesion.v1i2.365
Copyright : Author
Published by: Kohesi

This work is licensed under aCreative Commons Attribution-NonCommercial 4.0 International License

1. INTRODUCTION

The development of digital technology has transformed the way people express their opinions and respond to various events, particularly through social media. Platforms like Twitter and X have become active public spaces for expressing opinions on social, economic, and technological issues in real time. This activity generates large amounts of text data that spontaneously and naturally represent public views, making it valuable for scientific analysis.[1].

In the information technology sector, changes in hardware component prices often spark widespread discussion on social media. For example, rising prices of Random Access Memory

(RAM), impact not only individual consumers but also the computer industry and digital businesses. Public reactions to this situation come in various forms, including complaints, criticism, and neutral opinions, spread through user posts and comments. This makes social media a relevant data source for understanding public perceptions of technology price dynamics.

To systematically collect this data, data crawling or web scraping techniques are widely used in data-driven research. Crawling allows for the automated retrieval of large amounts of data from online platforms in a relatively short time. This method is considered effective because it produces more consistent data than manual collection and supports advanced analysis in data science and machine learning.[2].

Twitter provides an Application Programming Interface (API) that allows researchers to access public data with certain restrictions. This API is commonly used because the data obtained reflects real user behavior and can be analyzed quantitatively and qualitatively. However, the use of social media data also requires attention to ethical aspects and privacy protection, so the crawling process must be conducted in accordance with applicable regulations.[3].

Crawled text data is generally unstructured, requiring further processing before analysis. One frequently used approach is sentiment analysis, a technique for classifying opinions into positive, negative, or neutral categories. In the context of learning basic machine learning techniques, sentiment analysis is a relevant initial step because it encompasses core processes such as data collection, preprocessing, and classification. Previous research has shown that this approach is effective for understanding public opinion on various social media-based issues.[4].

2. RESEARCH METHODOLOGY

This section outlines the systematic framework used to achieve the research objectives. The applied methodology integrates procedural steps with the data analysis life cycle to ensure accurate visualization results.

2.1 Types and Approaches of Research

Social media is a digital platform that allows users to share information, opinions, and experiences openly and in real time[5]. Twitter is particularly widely used in academic research due to its short text-based, public nature, and its ability to reflect people's spontaneous responses to events. Data generated from Twitter user activity is considered representative in describing public perceptions of economic and technological issues, including fluctuations in computer hardware prices. The main advantages of social media as a data source lie in its large volume, diversity, and speed of information updates.

2.2 Data Crawling Techniques

Data crawling, or web scraping, is the process of automatically extracting data from websites or online platforms using specific software or scripts. In the context of data science research, crawling serves as the initial stage in obtaining raw datasets for further analysis. This technique enables researchers to collect large-scale data with high time efficiency and consistent data structures. Data crawling on social media is generally carried out through two approaches: direct scraping from web pages and utilizing official Application Programming

Interfaces (APIs). APIs are preferred because they provide structured, legal, and platform-compliant data access[6]. With the right crawling technique, public opinion data regarding RAM price increases can be systematically collected based on specific keywords, time periods, and attributes.

2.3 Twitter API as a Data Collection Tool

Twitter provides an API that allows developers and researchers to access public data such as tweets, retweets, likes, and user metadata. The Twitter API supports tweet searches based on keywords, hashtags, and specific timeframe[7], making it highly suitable for a case study of RAM price increases in 2025. Utilizing the Twitter API in research offers the advantage of standardized data and easy integration with programming languages like Python. However, the API has limitations such as rate limits and historical data access policies. Therefore, the design of the crawling process needs to consider query efficiency and compliance with platform usage rules.

2.4 Text Data and Its Processing Challenges

Text data from social media crawling is unstructured and contains various irrelevant elements, such as URLs, symbols, emoticons, and non-standard language. This makes the data impossible to analyze directly without preprocessing[8].

The main challenges in processing text data lie in the diversity of user language styles and the presence of noise, which can influence analysis results. Therefore, systematic steps are required to clean and normalize the data to make it more representative of the research topic.

2.5 Text Data Pre-Processing

Text data preprocessing is a crucial step in data processing before further analysis. This stage aims to improve data quality and reduce noise. Common preprocessing processes include:

1. Case folding, which is changing all text to lowercase letters.
2. Text cleaning, by removing URLs, punctuation, numbers and special symbols.
3. Word normalization, to standardize non-standard word forms.
4. Stopword removal, to remove common words that have no analytical meaning.

This stage produces cleaner text data that is ready to be analyzed using token-based methods[9].

2.6 Tokenization

Tokenization is the process of breaking down text into smaller units called tokens, typically words. Tokenization is a fundamental step in text data processing as it forms the basis for various subsequent analysis techniques[10]. In the context of this research, tokenization was used to identify words that appeared in discussions regarding the increase in RAM prices.

The results of tokenization allow researchers to understand the basic structure of text data and map the dominant vocabulary used by social media users.

2.7 Token Frequency Analysis

Token frequency analysis is a method for counting the number of occurrences of each token in a text dataset. This technique is used to identify the most frequently occurring words and to identify the main topics discussed by users.

In early machine learning research, token frequency analysis serves as a simple form of text feature extraction. This stage helps researchers understand how text data is represented numerically before applying more complex analysis techniques[11].

3. RESULTS AND DISCUSSION

This research uses a descriptive quantitative approach with a text data exploration method. This approach was chosen because the research aims to describe the patterns of word occurrence in crawled social media data, without performing classification or prediction processes. The primary focus of the research is on the initial stages of the machine learning pipeline: data collection, pre-processing, and token frequency analysis as a basis for understanding text data processing.

3.1 Crawling Social Media Data

The data crawling process was conducted on the Twitter platform using keywords related to the increase in Random Access Memory (RAM) prices in 2025, from May to October. The collected data consisted of public tweets in Indonesian, reflecting users' spontaneous responses to the issue.

The crawling results show that social media has become an active discussion space regarding RAM price increases, as evidenced by numerous posts containing complaints, price comparisons, and mentions of RAM brands or capacities. The raw data obtained still contains various non-text elements such as URLs, symbols, and non-standard words, requiring further processing.

Figure 2. Crawling Data

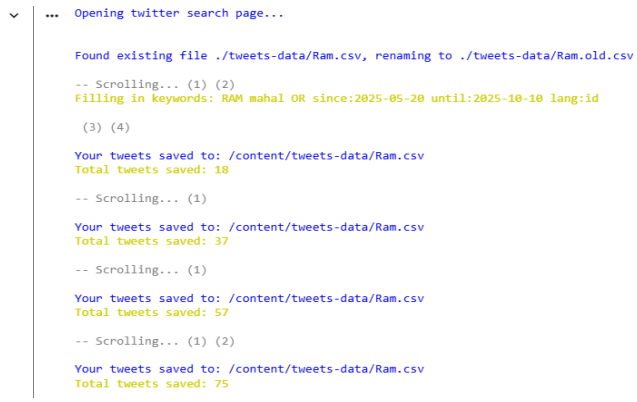
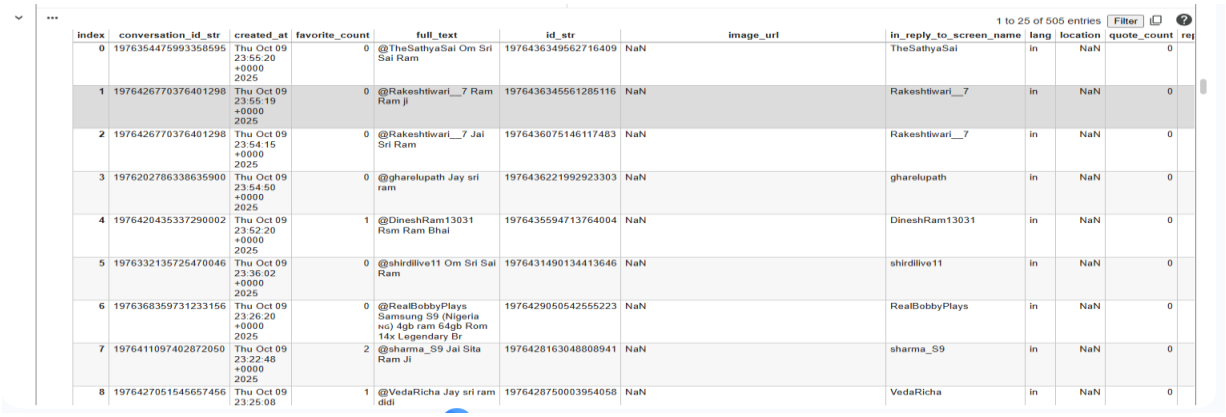


Figure 3. Data Analysis in Table Form



1 to 25 of 505 entries Filter

index	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	rep
0	1976354475993358595	Thu Oct 09 23:55:20 +0000 2025	0	@TheSathyaSai Om Sri Sai Ram	1976436349562716409	NaN	TheSathyaSai	in	NaN	0	
1	1976426770376401298	Thu Oct 09 23:55:19 +0000 2025	0	@Rakeshtiwari__7 Ram Ram ji	1976436345561285116	NaN	Rakeshtiwari__7	in	NaN	0	
2	1976426770376401298	Thu Oct 09 23:54:15 +0000 2025	0	@Rakeshtiwari__7 Jai Sri Ram	1976436075146117483	NaN	Rakeshtiwari__7	in	NaN	0	
3	1976202786338635900	Thu Oct 09 23:54:50 +0000 2025	0	@gharelupath Jay sri ram	1976436221992923303	NaN	gharelupath	in	NaN	0	
4	1976420435337290002	Thu Oct 09 23:52:20 +0000 2025	1	@DineshRam13031 Rsm Ram Bhai	1976435594713764004	NaN	DineshRam13031	in	NaN	0	
5	1976332135725470046	Thu Oct 09 23:36:02 +0000 2025	0	@shirdilive11 Om Sri Sai Ram	1976431490134413646	NaN	shirdilive11	in	NaN	0	
6	1976368359731233156	Thu Oct 09 23:26:20 +0000 2025	0	@RealBobbyPlays Samsung S9 (Nigeria NG) 4gb ram 64gb Rom 14x Legendary Br	1976429050542555223	NaN	RealBobbyPlays	in	NaN	0	
7	1976411097402872050	Thu Oct 09 23:22:48 +0000 2025	2	@sharma_S9 Jai Sita Ram Ji	1976428163048808941	NaN	sharma_S9	in	NaN	0	
8	1976427051545657456	Thu Oct 09 23:25:08	1	@VedaRicha Jay sri ram didi	1976428750003954058	NaN	VedaRicha	in	NaN	0	

Figure 4. Data Amount Display

```
[50]
✓ 0s
    # Cek jumlah data yang didapatkan

    # Hitung jumlah baris (tweet) dalam DataFrame dan simpan ke variabel num_tweets.
    num_tweets = len(df)

    #Cetak jumlah tweet ke layar agar kita tahu berapa banyak data yang berhasil diambil.
    print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")

    Jumlah tweet dalam dataframe adalah 505.
```

3.2 Data Preprocessing

A preprocessing stage was applied to clean the text data from noise and normalize the language structure. This process included case folding, URL and symbol removal, and stopword removal. After this stage, the text data became more concise and focused on words that had contextual meaning related to the research topic.

Pre-processing results show a significant reduction in the number of words compared to the initial data, but the quality of the information is more representative. Common words with no analytical contribution were successfully removed, leaving only tokens relevant to the issue of RAM price increases.

a. Folding case

Figure 5. Folding Case

```
Case Folding Result :

0                      @thesathyasai om sri sai ram
1                        @rakeshtiwari__7 ram ram ji
2                        @rakeshtiwari__7 jai sri ram
3                            @gharelupath jay sri ram
4                        @dineshram13031 rsm ram bhai
5                       @shirdilive11 om sri sai ram
6    @realbobbyplays samsung s9 (nigeria ng) 4gb ra...
7                          @sharma_s9 jai sita ram ji
8                         @vedaricha jay sri ram didi
9    @flowerroad_52 dari segi performa redmi note 1...
Name: full_text, dtype: object
n
```

b. Text Cleanup

Figure 6. Text Cleaning

```
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist

def remove_tweet_special(text):
  text = text.replace('\\t'," ").replace('\\n'," ").replace('\\u'," ").replace('\\',"")
  text = text.encode('ascii', 'replace').decode('ascii')
  text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text).split())

  return text.replace("http://", " ").replace("https://", " ")

  TWEET_DATA['full_text'] = TWEET_DATA['full_text'].apply(remove_tweet_special)
```

```
<>:10: SyntaxWarning: invalid escape sequence '\w'
<>:10: SyntaxWarning: invalid escape sequence '\w'
/tmp/ipython-input-3841879406.py:10: SyntaxWarning: invalid escape sequence '\w'
  text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text).split())
```

```
def remove_number(text):
  return re.sub(r"\d+", "", text)

  TWEET_DATA['full_text'] = TWEET_DATA['full_text'].apply(remove_number)

  #remove punctuation
  def remove_punctuation(text):
    return text.translate(str.maketrans("","",string.punctuation))

    TWEET_DATA['full_text'] = TWEET_DATA['full_text'].apply(remove_punctuation)

  #remove whitespace leading & trailing
  def remove_whitespace_LT(text):
    return text.strip()

    TWEET_DATA['full_text'] = TWEET_DATA['full_text'].apply(remove_whitespace_LT)
```

3.3 Tokenization

Tokenization is performed on text data that has undergone pre-processing. This process produces a set of tokens consisting of individual words that users use to discuss RAM price increases.

The tokenization results show that the vocabulary that emerges tends to be repetitive and focuses on aspects of price, memory capacity, and market conditions. Tokenization plays a crucial role as the basis for constructing text data representations, as each token serves as the primary unit of analysis in the word frequency stage.

Figure 7. Tokenization

```
Tokenizing Result :

0    [om, sri, sai, ram]
1         [ram, ram, ji]
2        [jai, sri, ram]
3        [jay, sri, ram]
4       [rsm, ram, bhai]
Name: tweet_tokens, dtype: object
```

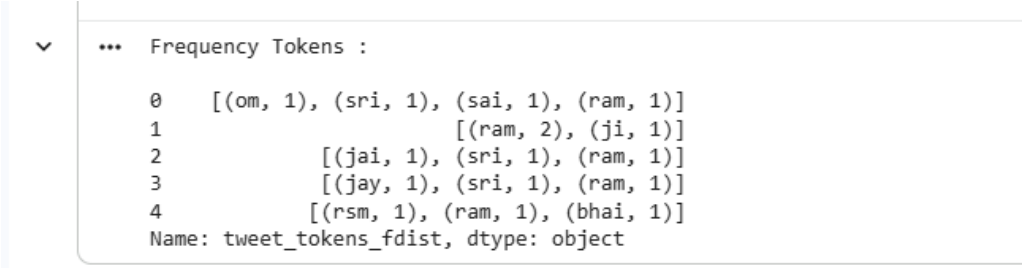
3.4 Token Frequency

Token frequency analysis was performed by counting the number of occurrences of each token in the entire dataset. The analysis revealed that some words had a significantly higher

frequency of occurrence than others. These words were generally directly related to price, market conditions, and the RAM product itself.

Tokens with high frequency indicate the primary focus of social media user discussions. Conversely, tokens with low frequency reflect more specific and individual opinions or experiences. This pattern suggests that public discussions tend to focus on the primary issue of price increases, with a variety of opinions emerging as complementary. These token frequency results can be presented in a table or simple visualization to demonstrate the dominance of certain words in discussions.

Figure 8. Token Frequency



4. CONCLUSION

This study demonstrates that social media data crawling techniques are effective as an initial step in text data processing for basic machine learning. Twitter (X) provides relevant data to represent public opinion regarding RAM price increases in 2025. The pre-processing stage plays a crucial role in improving data quality, while tokenization and token frequency analysis are able to describe dominant word patterns in public discussions. Thus, this study provides a systematic initial overview of the text data processing pipeline before applying more complex machine learning methods.

REFRENCES

[1] N. Van Robert Jhosefhin *et al.*, “Analisis Sentimen Crawling Data dari Sosial Media X tentang Gaza Menggunakan Metode SVM dan Decision Tree,” 2025. [Online]. Available: https://journal.stmiki.ac.id

[2] M. A. Khder, “Web scraping or web crawling: State of art, techniques, approaches and application,” *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[3] S. S. Sohail *et al.*, “Crawling Twitter data through API: A technical/legal perspective,” May 2021, [Online]. Available: http://arxiv.org/abs/2105.10724

[4] T. Agustiranti, A. Khalfani Izzati Kurdiana, B. Al Ghiffari, E. Dwi Juniar, and D. Gita Purnama, “Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh),” *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, vol. 7, no. 1, pp. 297–305, 2024.

[5] S. Suhendra and F. Selly Pratiwi, “Peran Komunikasi Digital dalam Pembentukan Opini Publik: Studi Kasus Media Sosial,” *Iapa Proceedings Conference*, p. 293, Oct. 2024, doi: 10.30589/proceedings.2024.1059.

[6] “Perancangan Aplikasi Web Crawler untuk Menghasilkan Dokumen Teks pada Domain Tertentu”.

[7] “ANALISIS SENTIMEN DATA TWITTER TERHADAP PELAKSANAAN PEMBELAJARAN ONLINE DI INDONESIA PADA MASA PANDEMI COVID-19 MENGGUNAKAN METODE NATURAL LANGUAGE PROCESSING”.

[8] P. Deski Manalu *et al.*, “Implementasi Algoritma Klasifikasi untuk Analisis Sentimen Media Sosial Tiktok Tahun 2025,” *Jurnal Teknik Informatika dan Teknologi Informasi*, doi: 10.55606/jutiti.v5i1.5644.

[9] U. Khairani, V. Mutiawani, and H. Ahmadian, “Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 887–894, Aug. 2024, doi: 10.25126/jtiik.1148315.

[10] “ANALISIS SENTIMEN PADA KOMENTAR INSTAGRAM MENGGUNAKAN GAUSSIAN NAÏVE BAYES”.

[11] A. Liken Anggoro, L. V. P. Ken, and M. G. Setiawan, “Analisis Media Text Clustering pada Twitter Akan Kasus Selebriti Menggunakan Orange Data Mining,” *remik*, vol. 7, no. 1, pp. 189–195, Jan. 2023, doi: 10.33395/remik.v7i1.12001.