

BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Cryptocurrency merupakan salah satu bentuk dobrakan baru pada sektor ekonomi. Namun masih banyak orang yang memandang *cryptocurrency* sebagai suatu aset yang tidak bernilai, apalagi banyak isu negatif yang membuat citra *cryptocurrency* menjadi buruk. Seiring dengan perkembangan zaman, mulai terdapat beberapa orang yang membuka dirinya terkait *cryptocurrency* untuk dijadikan pilihan sebagai sarana investasi. Walaupun begitu *cryptocurrency* dinilai sebagai suatu aset yang *high risk* karena memiliki volatilitas yang sangat tinggi, bahkan suatu isu yang beredar pada media sosial dapat menyebabkan keseluruhan *cryptocurrency* mengalami kenaikan atau penurunan yang signifikan. Objek penelitian ini merupakan pandangan masyarakat mengenai *cryptocurrency* yang berasal dari opini atau *tweets* pada *platform X* sebagai media sosial yang sering membicarakan *cryptocurrency* khususnya pada Januari 2024. Pada periode tersebut juga bertepatan dengan persetujuan Bitcoin ETF oleh SEC tepatnya 10 Januari 2024, sehingga sangat memungkinkan terjadi perubahan persepsi masyarakat mengenai *cryptocurrency* secara keseluruhan.

3.2 Metode Penelitian

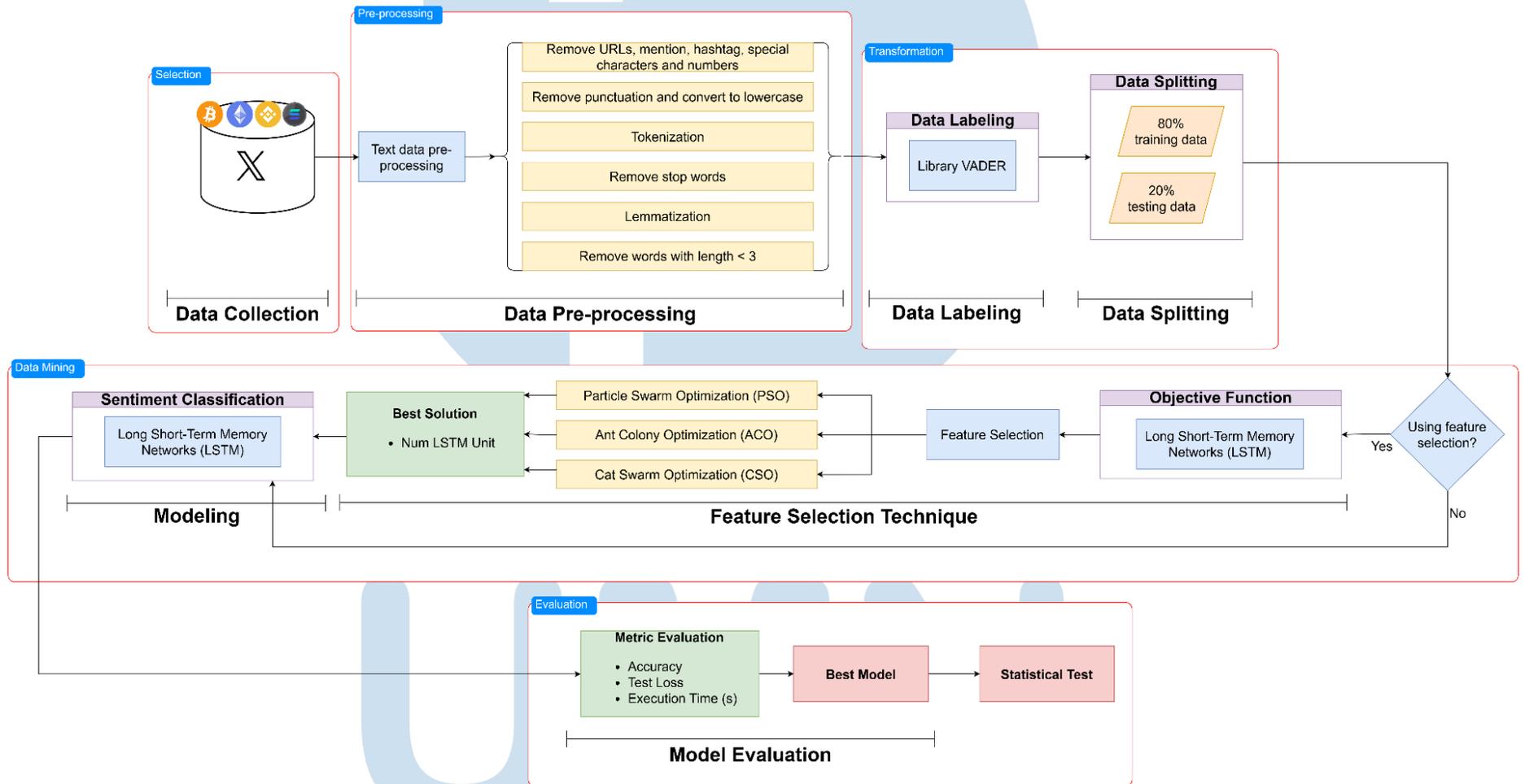
Penelitian ini termasuk ke dalam penelitian dengan pendekatan kuantitatif yang menggambarkan efektivitas prosedur pada *sentiment analysis* yang diusulkan [118]. Pada proses yang dilakukan, penelitian ini menguji model klasifikasi algoritma LSTM dengan dan tanpa *feature selection* berbasis *swarm intelligence*. Adapun proses evaluasi dengan *statistical test* menggunakan *Shapiro-Wilk Test* dan *ANOVA Test* untuk membuktikan efektivitas dari model yang diusulkan.

3.2.1 Alur Penelitian

Alur penelitian digunakan sebagai panduan agar suatu penelitian dapat dilakukan secara sistematis, dimana Gambar 3. 1 merupakan alur penelitian yang digunakan. Penelitian dimulai dari *data collection*, data dikumpulkan

menggunakan *tools* dalam melakukan *web scrapping* pada *platform X* secara otomatis. Namun data yang diambil dari media sosial seperti *platform X*, umumnya memiliki karakteristik seperti data tidak terstruktur, berdimensi tinggi, dan banyak *noise*. Tahapan selanjutnya perlu untuk dilakukan *data pre-processing* untuk memastikan data yang digunakan sudah bersih dan siap dikonsumsi pada algoritma-algoritma *machine learning* atau *deep learning*, karena kebersihan data dapat mempengaruhi hasil analisis dan kinerja model. Data yang sudah bersih, perlu dilakukan *data labeling* agar suatu model dapat memprediksi berdasarkan sentimen. Adapun *data splitting* yang digunakan untuk melakukan *training* pada model menggunakan algoritma-algoritma yang ditentukan, serta *data testing* digunakan untuk melakukan validasi pada model yang dilatih. Pada proses pelatihan model melalui tahapan *feature selection* berbasis *swarm intelligence* dengan algoritma PSO, ACO dan CSO mendapatkan hasil berupa LSTM *Unit*/fitur terpilih. LSTM Unit digunakan saat *modeling* untuk melakukan optimasi pada parameter algoritma LSTM, sehingga kinerja model LSTM dapat ditingkatkan. Adapun proses membandingkan model-model antara lain LSTM konvensional, LSTM yang dioptimasi menggunakan PSO, LSTM yang dioptimasi menggunakan ACO, dan LSTM yang dioptimasi menggunakan CSO berdasarkan *metric evaluation* seperti *accuracy*, *loss*, dan *execution time*. Hasil dari proses tersebut didapatkan model terbaik khususnya untuk *sentiment analysis* mengenai *cryptocurrency*. Terakhir untuk membuktikan apakah hasil yang didapat merupakan model terbaik, maka dilakukan *statistical test* untuk membuktikan hal tersebut. Keseluruhan proses penelitian yang dilakukan dapat diilustrasikan pada Gambar 3.1.

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A



Gambar 3. 1 Alur Penelitian

3.2.2 Metode Data Mining

Fokus penelitian ini yaitu melakukan optimasi pada model LSTM menggunakan *feature selection* berbasis *swarm intelligence* yang dapat diterapkan menggunakan metode *data mining*. Terdapat beragam metode *data mining*, namun untuk yang populer terdapat tiga macam antara lain KDD, *Cross-Industry Standard Process for Data Mining* (CRISP-DM), dan *Sample, Explore, Modify, Model, and Access* (SEMMA). Berikut perbandingan antara setiap *framework* tersebut.

Tabel 3. 1 Perbandingan *Framework Data Mining* Populer

Sumber: [119]

Karakteristik	KDD	CRISP-DM	SEMMA
Fokus	Proses penemuan pengetahuan dalam <i>database</i>	Proses standar lintas industri untuk <i>data mining</i>	Metodologi untuk proyek pengembangan <i>data mining</i>
Tahapan	<i>Selection, Pre-processing, Transformation, Data Mining, Evaluation/Interpretation</i>	<i>Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment</i>	<i>Sample, Explore, Modify, Model, Assess</i>
Penerapan	Multidisiplin, tidak spesifik pada domain tertentu	Domain-agonistik, fleksibel, dan dapat diadaptasi ke berbagai industri	Khusus <i>software SAS</i>
Dokumentasi dan Contoh	Kurang mendetail	Dokumentasi lengkap dengan contoh konkret	Dokumentasi dan contoh melalui <i>software SAS</i>
Keunggulan	Fleksibel dan dapat diadaptasi	Proses terstruktur dengan langkah-langkah jelas, mendukung kolaborasi antar tim	Proses yang mudah diikuti, cocok untuk pengguna SAS
Kekurangan	Kurangnya struktur dan panduan langkah demi langkah	Mebutuhkannya penyesuaian untuk proyek tertentu	Terbatas pada lingkungan SAS, kurang fleksibel

Tabel 3. 1 menunjukkan perbedaan antara setiap *framework data mining* populer, dimana setiap *framework* memiliki fungsi, kelebihan, dan kekurangannya masing-masing. *Framework* KDD memungkinkan dalam menemukan informasi tersembunyi pada data [119]. Hal tersebut sejalan dengan tujuan penelitian ini yaitu ingin mengetahui pandangan masyarakat

dunia mengenai *cryptocurrency* dengan menggunakan pendekatan *sentiment analysis*, sehingga KDD dijadikan sebagai *framework* yang digunakan pada penelitian ini. Berikut penyesuaian antara alur penelitian dan *framework* KDD.

1) *Selection*

Selection merupakan tahapan pertama dalam *framework* KDD, dalam penelitian ini proses *selection* meliputi *data collection* pada internet dengan pengumpulan data menggunakan teknik *web scrapping*. Data diambil dari *tweets* atau opini pengguna *platform X* dengan pemilihan data yang memiliki *keywords* yang berkaitan dengan *cryptocurrency*. Selain itu data yang diambil pada rentang waktu 31 Desember 2023 - 31 Januari 2024 karena bertepatan dengan persetujuan Bitcoin ETF oleh SEC.

2) *Pre-processing*

Pre-processing merupakan tahapan kedua dalam *framework* KDD, dalam penelitian ini dilakukan *data pre-processing* menggunakan kumpulan data yang didapatkan dari tahapan *selection*. Seluruh data yang digunakan melewati beberapa teknik dalam memberihkan data khususnya pada *platform X* seperti 1) *remove URLs, mention, hashtag, special characters, and numbers*; 2) *remove punctuation and covert to lowercase*; 3) *tokenization*, 4) *remove stop words*, 5) *lemmatization*, 6) *remove words with length < 3*. Setelah melalui beberapa langkah dalam *data pre-processing*, semua teks pada data yang tidak relevan dihapus sehingga tidak menimbulkan ambiguitas.

3) *Transformation*

Transformation merupakan tahapan ketiga dalam *framework* KDD, dalam penelitian ini meliputi proses *data labeling* dan *data splitting*. *Data labeling* yang dilakukan yaitu melakukan klasifikasi data berdasarkan polaritasnya sehingga pada setiap data ditambahkan kolom baru untuk menampung label atau klasifikasi sentimennya apakah termasuk dalam sentimen positif atau negatif. Proses *transformation* tidak berhenti sampai

di sini, data yang sudah diberikan label akan melalui proses *data splitting* yaitu membagi data menjadi dua bagian yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih model menggunakan algoritma, sedangkan *data testing* digunakan untuk melakukan validasi atas model yang sudah dilatih.

4) *Data Mining*

Data mining merupakan tahapan keempat dalam *framework* KDD, dalam penelitian ini meliputi proses *feature selection* menggunakan algoritma *swarm intelligence* seperti PSO, ACO, dan CSO. Serta melakukan *modeling* menggunakan algoritma *deep learning* yaitu LSTM. Pada proses *feature selection* menggunakan algoritma *swarm intelligence*, *data training* digunakan untuk memilih fitur-fitur yang relevan, hal ini disebabkan karena data yang diambil khususnya dari media sosial memiliki karakteristik berdimensi tinggi dan memiliki banyak *noise*. Diterapkan *feature selection* berbasis *swarm intelligence* diharapkan dapat menangani permasalahan tersebut agar dapat meningkatkan kinerja pada model LSTM. Proses *data mining* selanjutnya yaitu *modeling* menggunakan algoritma LSTM. Algoritma LSTM merupakan algoritma *deep learning* yang canggih khususnya dalam mengklasifikasikan teks, namun tidak menutup kemungkinan kinerja dari model LSTM dapat ditingkatnya dengan bantuan *feature selection* berbasis *swarm intelligence*. Optimasi yang dilakukan yaitu optimasi parameter LSTM yang nilai parameternya didapatkan dari *feature selection*.

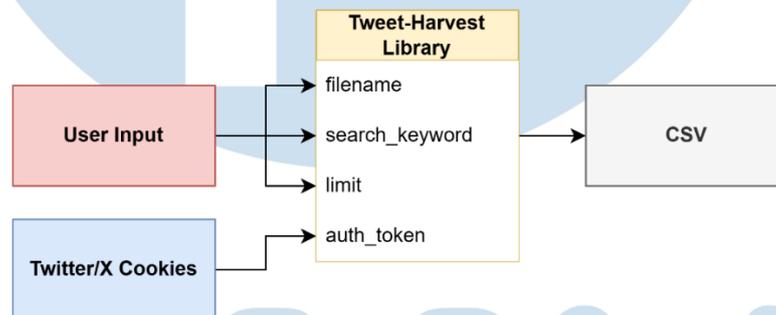
5) *Evaluation*

Evaluation atau *interpretation* merupakan tahapan terakhir dalam *framework* KDD, dalam penelitian ini meliputi proses membandingkan model-model yang dilatih menggunakan *data testing*. Model yang dibandingkan antara lain LSTM konvensional, LSTM yang dioptimasi dengan PSO, LSTM yang dioptimasi dengan ACO, dan LSTM yang dioptimasi menggunakan CSO berdasarkan *metric evaluation* seperti

accuracy, *loss*, dan *execution time*. Model terbaik adalah model yang memiliki akurasi tertinggi, *loss* terendah, dan *execution time* tercepat. Namun dari model terbaik saja tidak cukup untuk meyakinkan pembaca, maka *statistical test* dapat dilakukan untuk membuktikan bahwa dalam membandingkan model adil dan dapat menyatakan bahwa model tersebut memiliki kinerja yang berbeda secara signifikan dengan model yang lain.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan data primer yang dikumpulkan secara langsung menggunakan teknik *web scrapping* pada *platform X*. *Web scrapping* dilakukan secara otomatis menggunakan *tools* yang bernama *tweet-harvest*, yang dapat melanjutkan *scrapping* otomatis apabila sudah mencapai *limit* pada *platform X*. Data yang dikumpulkan berupa *tweets* atau opini masyarakat yang membahas *cryptocurrency*.



Gambar 3. 2 Langkah Pengumpulan Data

Gambar 3. 2 merupakan langkah pengumpulan data dengan menggunakan *library tweet-harvest*. *Library tweet-harvest* yang di *import* pada JupyterLab dengan diintegrasikan dengan Node.JS belum dapat digunakan langsung, perlu dilakukannya pengaturan tambahan. Pengaturan tambahan yang dimaksud adalah pengaturan pada parameter *library tweet-harvest* dengan memasukkan parameter *filename* sebagai nama *file* CSV yang akan menjadi *output* setelah proses pengumpulan data selesai, parameter *search_keyword* merupakan *data tweets* dengan kata kunci apa yang ingin diambil pada *platform X*, parameter *limit* merupakan batasan data yang dikumpulkan pada setiap pemanggilan fungsi *tweet-harvest*, serta parameter *auth_token* merupakan identitas *token* pengguna *platform X* yang dapat diakses di pengaturan *browser* bagian *cookies* saat membuka halaman

X dengan akun yang ingin digunakan untuk pengumpulan data. Setelah proses pengumpulan data selesai, maka secara otomatis terdapat *file* CSV yang dapat di *download* untuk dijadikan *dataset* penelitian.

3.3.1 Populasi dan Sampel

Teknik pengumpulan data yang digunakan pada penelitian ini menggunakan teknik *non-probability sampling*, dimana sampel tidak diambil secara acak. *Non-probability sampling* yang digunakan merupakan gabungan antara *time-based sampling* dan *keyword sampling*, sehingga pengumpulan data dari *platform* X hanya mengambil data-data berdasarkan rentang waktu dan berdasarkan *keywords* yang dimasukkan. *Keywords* yang dimasukkan pada *tools web scrapping* yaitu '(*cryptocurrency OR crypto OR bitcoin OR \ \$btc OR ethereum OR \ \$eth OR binance OR \ \$bnb OR solana OR \ \$sol*) lang:en'. Pada *keywords* tersebut meliputi berbagai *cryptocurrency* populer seperti Bitcoin, Ethereum, Binance, dan Solana dengan data berbahasa Inggris. Alasan memilih 4 *cryptocurrency* tersebut sebagai *keywords* dikarenakan memiliki kapitalisasi pasar terbesar, dengan kata lain sudah banyak orang percaya dan menginvestasikan uang mereka pada 4 *cryptocurrency* tersebut [120]. Adapun penggunaan data Bahasa Inggris karena *cryptocurrency* sering disandingkan dengan mata uang dolar sebagai mata uang global, selain itu Bahasa Inggris juga merupakan bahasa internasional yang digunakan oleh masyarakat dunia. Data yang berhasil dikumpulkan dengan teknik *web scrapping* yaitu berjumlah 9884 data.

3.3.2 Periode Pengambilan Data

Data yang dikumpulkan dengan menggunakan teknik *web scrapping* pada *platform* X merupakan *tweets* atau komentar yang berada pada rentang waktu 31 Desember 2023 – 31 Januari 2024. Periode tersebut bertepatan dengan persetujuan Bitcoin ETF oleh SEC, sehingga memungkinkan masyarakat banyak yang membicarakan *cryptocurrency*. Namun pengambilan data per hari hanya dibatasi 1000 data dan data yang didapat tidak konsisten, sehingga pengumpulan data bisa lebih atau kurang dari yang ditetapkan. Permasalahan tersebut terjadi karena terdapat aturan baru terkait limitasi penggunaan

platform X, sehingga jika pengambilan data melebihi batas dapat terkena *warning* bahkan *suspend*.

3.4 Variabel Penelitian

Terdapat dua jenis variabel yang digunakan pada penelitian ini yaitu variabel independen dan variabel dependen. Setiap variabel mempunyai perannya masing-masing dalam memprediksi hasil sentimen. Berikut penjelasan penggunaan variabel independen dan variabel dependen pada penelitian ini.

3.4.1 Variabel Independen

Variabel independen merupakan variabel prediktor atau penyebab yang dapat mempengaruhi variabel dependen. Pada penelitian ini, data *tweets* atau opini masyarakat pada *platform X* yang membahas *cryptocurrency* merupakan variabel independen. Hal tersebut dikarenakan *tweets* berupa opini dapat mempengaruhi pandangan masyarakat mengenai *cryptocurrency* yang biasanya terkandung suatu sentimen di dalamnya.

3.4.2 Variabel Dependen

Variabel dependen adalah variabel respon atau hasil yang biasanya terpengaruh dari variabel independen. Pada penelitian ini, sentimen masyarakat mengenai *cryptocurrency* merupakan variabel dependen. Hal tersebut dikarenakan sentimen seseorang dapat terpengaruh dari opini masyarakat yang membahas mengenai *cryptocurrency*. Apabila banyak orang yang menganggap *cryptocurrency* sebagai suatu aset yang dapat dipercaya dapat melawan inflasi maka menyebabkan sentimen masyarakat terhadap *cryptocurrency* menjadi baik, begitu pula sebaliknya.

3.5 Teknik Analisis Data

Analisis data yang dilakukan pada penelitian ini yaitu *sentiment analysis* dengan pengujian beberapa model algoritma LSTM dengan dan tanpa *feature selection* berbasis *swarm intelligence* untuk optimasinya. Dalam membandingkan model-model tersebut dilihat berdasarkan *evaluation metrics* seperti *accuracy*, *loss*, dan *execution time*. Adapun *statistical test* yang digunakan untuk membuktikan hasil *modeling*. Keseluruhan analisis data dilakukan menggunakan bahasa

pemrograman Python, Tabel 3. 2 merupakan perbandingan bahasa pemrograman yang menjadi pertimbangan untuk digunakan pada penelitian ini.

Tabel 3. 2 Perbandingan Bahasa Pemrograman Populer untuk Analisis Data

Sumber: [116]

Faktor	Python	R	MatLab
Deskripsi	Python adalah bahasa pemrograman yang sering digunakan untuk membangun <i>website</i> dan <i>software</i> , otomatisasi tugas, dan melakukan analisis data.	R adalah bahasa pemrograman yang dioptimalkan untuk analisis statistik dan visualisasi data.	MatLab adalah bahasa pemrograman yang digunakan sebagian besar oleh para <i>engineers</i> dan <i>data analysts</i> untuk komputasi numerik.
Harga	Dapat digunakan secara gratis karena <i>open-source</i> .	Dapat digunakan secara gratis karena <i>open-source</i> .	Perlu melakukan pembayaran yang tidak sedikit dan tidak <i>open-source</i> .
Penggunaan	Mudah dipelajari dan dibaca dibandingkan bahasa pemrograman yang lain, karena struktur program sederhana dan bersih.	Mudah dipelajari dan dibaca, karena struktur cukup sederhana. Hal tersebut menyebabkan bahasa R banyak digunakan untuk <i>data mining</i> .	Mudah dipelajari oleh pemula karena dilengkapi oleh beberapa fasilitas yang mendukung dalam analisis data dan komputasi numerik.
Library	Dilengkapi <i>libraries</i> , <i>lists</i> , dan <i>dictionaries</i> yang dapat membantu dalam analisis. Jumlah <i>library</i> terus bertambah karena komunitas Python sangat besar.	Dilengkapi <i>libraries</i> , <i>lists</i> , dan <i>dictionaries</i> yang dapat membantu dalam analisis. Komunitas R masih sedikit dibandingkan dengan Python, sehingga variasi <i>library</i> masih kurang beragam.	Dilengkapi <i>package</i> dan <i>libraries</i> yang dapat membantu dalam analisis. Namun karena berbayar, komunitas dari MatLab sangat sedikit dibandingkan bahasa pemrograman yang lain.
Kinerja	Cepat dalam menjalankan program hingga selesai.	Cepat dalam menjalankan program hingga selesai.	Lebih lambat dalam menjalankan program hingga selesai.

Penelitian ini memilih Python sebagai bahasa pemrograman untuk melakukan *sentiment analysis cryptocurrency* karena kelebihan-kelebihan yang ditawarkan seperti yang dijelaskan pada Tabel 3. 2 yaitu bahasa yang *open-source* dan kemudahannya untuk dipelajari membuat komunitas Python sangat besar dibandingkan bahasa pemrograman lain. Hal tersebut membuat sebagian besar *developer* mengembangkan banyak inovasi, salah satunya pada pembuatan *library* yang dapat digunakan oleh semua orang termasuk penelitian ini yang melakukan *sentiment analysis cryptocurrency*.