TUGAS AKHIR

# DETEKSI DAN ANALISIS KARAKTERISTIK AKUN BUZZER MEDIA SOSIAL TWITTER PADA SISTEM TERSEBAR SPARK


Axel Joseph Yang

NPM: 6181901025



PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN


2024

FINAL PROJECT

# DETECTION AND ANALYSIS OF TWITTER SOCIAL MEDIA BUZZER ACCOUNT CHARACTERISTICS BASED ON SPARK DISTRIBUTED SYSTEM


**Axel Joseph Yang**

**NPM: 6181901025**

**DEPARTMENT OF INFORMATICS**
**FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES**
**PARAHYANGAN CATHOLIC UNIVERSITY**
**2024**

# LEMBAR PENGESAHAN

# DETEKSI DAN ANALISIS KARAKTERISTIK AKUN BUZZER MEDIA SOSIAL TWITTER PADA SISTEM TERSEBAR SPARK

**Axel Joseph Yang**

**NPM: 6181901025**

**Bandung, 11 Januari 2024**

**Menyetujui,**

**Pembimbing**

Digitally signed by Veronica Sri Moertini

**Prof. Dr. Veronica Sri Moertini**

**Ketua Tim Penguji**

Digitally signed by Maria V. Claudia M.

**Maria Veronica, M.T.**

**Anggota Tim Penguji**

Digitally signed by Gede Karya

**Gede Karya, M.T.**

**Mengetahui,**

**Ketua Program Studi**

Digitally signed by Lionov

**Lionov, Ph.D.**

# PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa tugas akhir dengan judul:

**DETEKSI DAN ANALISIS KARAKTERISTIK AKUN BUZZER MEDIA SOSIAL TWITTER PADA SISTEM TERSEBAR SPARK**

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 11 Januari 2024


Axel Joseph Yang
NPM: 6181901025

## ABSTRAK

Terdapat bermacam-macam akun dalam media sosial, salah satunya adalah akun *buzzer*. Akun *buzzer* adalah akun pengguna media sosial yang secara sengaja menyebarkan informasi, opini, atau pesan tertentu dengan tujuan memengaruhi opini publik. Pada penelitian ini digunakan media sosial Twitter dengan domain Pemilihan umum Presiden Indonesia tahun 2024, karena Twitter merupakan salah satu media sosial dimana masyarakat bebas berpendapat dan kegiatan ini banyak dinanti oleh masyarakat Indonesia. Oleh karena itu ingin dideteksi akun *buzzer* untuk membedakan mana akun *buzzer* yang cenderung menggunakan kata-kata bernada positif (bersifat mendukung) dan yang bernada negatif (bersifat menyerang atau tidak sopan).

Data media sosial Twitter dikumpulkan dengan menggunakan teknologi *big data* bernama Apache Kafka. Apache Kafka adalah sebuah sistem pengumpul data dalam skala besar yang juga menjadi sarana dalam penyederhanaan *data stream*, nantinya hasil dari pengumpulan data akan disimpan pada Hadoop Distributed File System (HDFS). Lalu digunakan Apache Spark sebagai alat untuk mendeteksi dan menganalisis akun *buzzer* menggunakan data yang sudah tersimpan pada HDFS. Apache Spark sendiri merupakan sebuah *tools* yang dapat digunakan untuk memproses data secara paralel.

Sebelum dilakukan deteksi, datanya akan disiapkan terlebih dahulu sesuai kebutuhan. Deteksi pada penelitian ini terdiri dari 5 tahap dengan menggunakan fitur-fitur berdasarkan dari hasil studi literatur dan eksplorasi secara nyata, contoh salah satu fiturnya adalah *FollowersCount* atau jumlah pengikut. Diambil akun dengan jumlah pengikut di atas 500 pengikut, fitur ini didapatkan dari hasil studi literatur yang dilakukan. Selanjutnya hasil deteksi dianalisis untuk memberi contoh akun-akun yang tergolong sebagai *buzzer*, dari 59 akun yang didapatkan dari hasil deteksi, terdapat 51 akun yang merupakan akun *buzzer*. Oleh karena itu, ketika dihitung tingkat keberhasilan deteksinya, didapatkan angka sebesar 86.4%, dapat dikatakan bahwa deteksi yang dilakukan sudah berhasil.

Hasil dari deteksi ini kemudian akan dilanjutkan dengan klasifikasi untuk memprediksi sentimen *buzzer* untuk memahami sentimen yang dihasilkan oleh akun *buzzer* tersebut. Pada prediksi sentimen ini dilakukan 2 eksperimen dengan perbedaan pada jumlah datanya. Terdapat 2 metode yang digunakan untuk prediksi sentimen, yaitu metode *lexicon based* dan metode *machine learning* (*decision tree* dan *random forest*). Model prediksi akan dievaluasi menggunakan *confusion matrix* untuk mengetahui seberapa bagus hasil yang dihasilkan oleh model. Hasil prediksi dianalisis untuk memberi contoh akun hasil prediksi yang didapatkan beserta dengan contoh *tweet* yang dilakukannya.

Kesimpulannya, berdasarkan eksperimen yang sudah dilakukan, metode tersebut dapat mendeteksi akun *buzzer* menggunakan fitur-fitur yang didapatkan dari hasil studi literatur dan eksplorasi secara nyata. Hasil deteksinya dapat dimanfaatkan untuk melakukan analisis sentimen agar mendapatkan model yang dapat memprediksi dengan baik. Sentimen yang sudah didapatkan dapat digunakan untuk membedakan mana akun *buzzer* yang bernada positif dan negatif. Untuk menampilkan hasil dari penelitian ini, dibangun sebuah perangkat lunak berbasis *website* yang dapat menerima masukkan sebuah *file* csv untuk dideteksi dan diprediksi lalu menampilkan hasilnya.



**Kata-kata kunci:** *Buzzer*, Twitter, *Data Stream*, Apache Kafka, Apache Spark, Klasifikasi, Sentimen

## ABSTRACT

There are various types of accounts on social media, one of which is the buzzer account. A buzzer account is a social media user account that intentionally spreads information, opinions, or specific messages with the aim of influencing public opinion. In this research, Twitter social media platform is used with the domain of the 2024 Indonesian Presidential Election, because Twitter is one of the social media platforms where people can freely express their opinions and this activity is eagerly awaited by the Indonesian people. Therefore, buzzer accounts will be detected to distinguish between buzzer accounts that tend to use positively toned words (supportive) and those with negatively toned words (aggressive or impolite).

Twitter social media data is collected using a big data technology called Apache Kafka. Apache Kafka is a large-scale data collection system that also serves as a means to simplify data streams. The collected data will be stored on the Hadoop Distributed File System (HDFS). Then Apache Spark is used as a tool to detect and analyze buzzer accounts using the data stored on HDFS. Apache Spark itself is a unified computing engine that can be used to process data in parallel.

Before the detection process, the data will be prepared according to the needs. The detection in this research consists of 5 stages using features based on literature studies and real-world exploration, such as one of the features being the FollowersCount or the number of followers. Accounts with more than 500 followers are selected, this feature is obtained from the literature study conducted. Next, the detection results are analyzed to provide examples of accounts classified as buzzer. Out of 59 accounts obtained from the detection, 51 of them are identified as buzzer accounts. Therefore, when calculating the detection success rate, it is found to be 86.4%. It can be said that the detection performed has been successful.

The results of this detection will then be followed by classification to predict the sentiment of buzzer accounts to understand the sentiment generated by these accounts. In this sentiment prediction, two experiments will be conducted with a difference in the amount of data. Two methods will be employed for sentiment prediction, that is lexicon based and machine learning (decision tree and random forest algorithms). The prediction models will be evaluated using a confusion matrix to assess the performance of the models. The prediction results will be analyzed to provide examples of predicted accounts along with sample tweets they have made.

In conclusion, based on the conducted experiments, the methods can successfully detect buzzer accounts using features obtained from data collection and exploration. The detection results can be utilized for sentiment analysis to build a model capable of making accurate predictions. The obtained sentiment can be used to distinguish between buzzer accounts with positive and negative tone. To showcase the results of this research, a web-based software has been developed that can accept input of a csv file to be detected and predicted, and then display the results.



**Keywords:** Buzzer, Twitter, Data Stream, Apache Kafka, Apache Spark, Classification, Sentiment

*Dipersembahkan untuk kedua orang tua...*

# KATA PENGANTAR

Skripsi ini merupakan hasil dari perjalanan panjang yang penuh dengan tantangan dan pengalaman yang tak terlupakan. Maka dari itu dengan rasa syukur Penulis ingin menyampaikan terima kasih ke hadirat Tuhan yang Maha Esa, atas berkat-Nya dan karunia-Nya yang melimpah sehingga Penulis dapat menyelesaikan skripsi ini. Pada kesempatan kali ini, Penulis juga ingin mengungkapkan rasa terima kasih yang mendalam kepada semua pihak yang telah membantu dan mendukung selama menjalani kuliah dan selama proses penulisan skripsi ini, yaitu sebagai berikut:

1. Kepada kedua orang tua dan adik Penulis yang selalu mendukung, menyemangati dan memberikan doa.

2. Kepada Ibu Prof. Dr. Veronica Sri Moertini, Ir., M.T. selaku dosen pembimbing, yang telah banyak memberikan ilmu, saran, dan bantuan selama proses pengerjaan skripsi ini.

3. Kepada Ibu Maria Veronica Claudia, ST, MT dan Bapak Gede Karya, S.T., M.T., CISA, IPM, selaku dosen penguji yang telah memberikan masukan dan saran untuk skripsi ini.

4. Kepada teman-teman yang menemani Penulis selama masa kuliah, Jeremy, Febri, Rio, Gavin, Ferell, dan Indra.

5. Kepada teman-teman yang berjuang bersama selama skripsi, Filipus, Michael, Daryl, Brian, Kinan, dan Alma.

6. Kepada seluruh pihak yang mendukung Penulis yang tidak dapat disebutkan satu-satu.

Sebagai penutup, Penulis menyadari bahwa skripsi ini masih belum sempurna, oleh karena itu Penulis memohon maaf jika terdapat kesalahan penulisan atau metode yang digunakan pada skripsi ini. Penulis juga ingin berterima kasih kepada pembaca yang telah membaca skripsi ini, Penulis berharap skripsi ini dapat bermanfaat dan membantu untuk kemajuan ilmu pengetahuan, dan dapat dijadikan panduan untuk penelitian selanjutnya.

Bandung, Januari 2024

Penulis

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# DAFTAR KODE PROGRAM

# BAB 1

# PENDAHULUAN

Bab ini membahas mengenai latar belakang dilakukannya penelitian ini dengan mendeskripsikan gambaran besar permasalahan yang ada. Lalu dibahas rumusan masalah yang berisi mengenai inti permasalahan dari penelitian ini, dilanjutkan dengan tujuan yang berisi mengenai tujuan akhir yang ingin dicapai dalam penelitian ini sebagai solusi masalah. Kemudian terdapat batasan masalah yang membahas mengenai asumsi yang digunakan untuk membatasi ruang lingkup penelitian ini. Selanjutnya dibahas mengenai metodologi yang berisi mengenai tahapan pekerjaan dan eksperimen yang dilakukan, lalu diakhiri dengan pembahasan mengenai sistematika pembahasan yang berisi ringkasan apa saja yang dikerjakan pada setiap bab di buku ini.

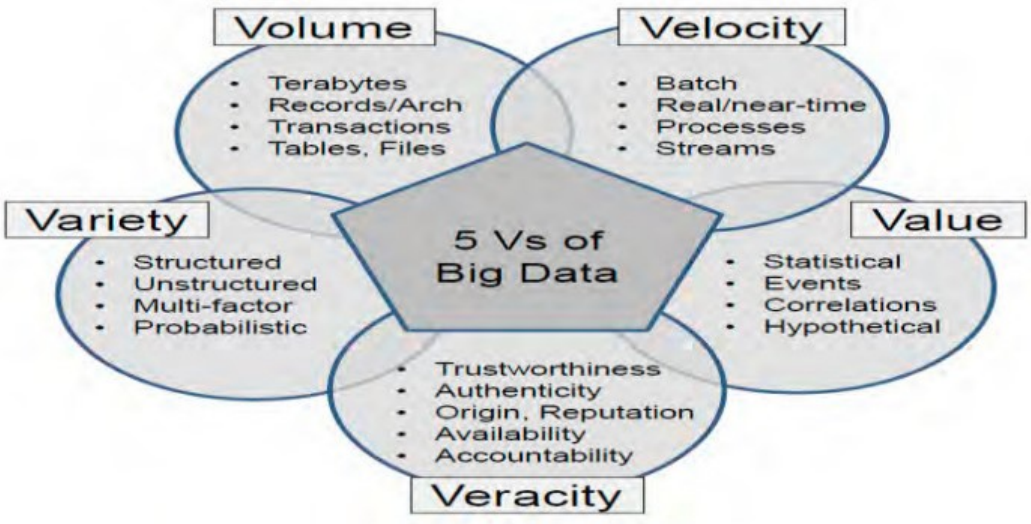
## 1.1 Latar Belakang

Media sosial adalah sebuah sarana digital yang memfasilitasi penggunanya untuk saling bersosialisasi satu sama lain atau membagikan konten seperti foto dan video yang dilakukan secara *online* yang memungkinkan manusia untuk saling berinteraksi tanpa dibatasi oleh ruang dan waktu. Beberapa contoh dari media sosial adalah Instagram, Tiktok, dan Twitter. Instagram adalah sebuah aplikasi yang memungkinkan penggunanya untuk mengunggah foto dan video ke dalam *post* dan *story* yang dapat diedit dengan berbagai filter, pengguna juga dapat menyukai foto dan video serta mengikuti pengguna lain untuk menambahkan konten pada *home* mereka. Tiktok adalah sebuah aplikasi yang memberi wadah bagi penggunanya untuk dapat berekspresi melalui konten video. Twitter adalah sebuah media sosial yang memungkinkan penggunanya untuk mengirimkan sebuah pesan singkat atau yang disebut dengan *tweet* kapanpun dan dimanapun mereka berada. *Tweet* sendiri bisa terdiri dari pesan teks maupun foto dan video, tetapi mayoritas isi dari *tweet* adalah pesan teks.

Pengguna Twitter terdiri dari berbagai macam tipe, seperti akun untuk berjualan, untuk berbagi tentang informasi suatu acara, maupun digunakan sebagai akun *buzzer*. Istilah *Buzzer* sendiri adalah orang atau sekumpulan orang yang pendapatnya memiliki pengaruh terhadap hal-hal tertentu. *Buzzer* berada di mana-mana, salah satunya pada media sosial Twitter. Akun *buzzer* pada Twitter adalah pengguna akun Twitter yang mayoritas mempunyai jumlah *followers* yang banyak dan memiliki pengaruh terhadap suatu hal atau acara. Akun *buzzer* biasanya mempunyai tugas untuk mempromosikan sesuatu, baik berupa produk, orang, dan lainnya secara berulang.

Banyaknya akun *buzzer* ini menyebabkan kebingungan kepada para pembacanya, seperti akun mana yang benar-benar memberi informasi yang informatif dan dapat dipercaya dan akun mana yang hanya melakukan *tweet* yang kurang informatif. Oleh karena itu dilakukan deteksi dan analisis karakteristik akun *buzzer* pada media sosial Twitter.

Data Twitter memenuhi karakteristik 5V dari *big data*. *Big Data* merupakan salah satu kemajuan di bidang teknologi yang berkembang dan dibutuhkan saat ini. *Big data* adalah sekumpulan data yang berukuran besar yang jumlahnya terus bertambah, yang terdiri dari berbagai macam jenis data dan terbentuk dengan terus menerus dengan kecepatan tertentu [6]. *Big data* memiliki karakteristik khusus 5V, yaitu *volume, velocity, variety, veracity,* dan *value* seperti pada Gambar 1.1.




Gambar 1.1: Karakteristik *big data* [1]

*Volume* mengacu pada sekumpulan data dalam jumlah dan ukuran yang sangat besar. *Velocity* adalah kecepatan pertambahan data dalam kurun waktu tertentu. *Variety* mendefinisikan berbagai jenis data yang ada, mulai dari data terstruktur seperti tabel basis data, data semi-terstruktur seperti *Extensible Markup Language* (XML) dan *JavaScript Object Notation* (JSON), dan data tidak terstruktur seperti citra, suara, dan teks. *Veracity* mengarah kepada seberapa akurat dan dapat dipercaya suatu data. *Value* merupakan seberapa bernilainya atau bermaknanya suatu data.

Data Twitter merupakan data terstruktur yang memenuhi karakteristik 5V dari *big data* dari segi *volume* dan *velocity*, karena seperti yang disebutkan pada penelitian Gupta dan Hewett [7], diperkirakan setiap harinya ada 500 juta *tweets* yang ditulis di Twitter, yang dapat menyebabkan *data stream* yang tidak terbatas dan terus berkembang. Oleh karena itu, Twitter menghasilkan *volume* data yang sangat besar dengan *velocity* yang sangat tinggi.

*Big data* memiliki beberapa masalah seperti metode penyimpanan dan cara pemrosesannya, maka dari itu harus diproses menggunakan *tools* khusus untuk menggunakannya. Hal ini disebabkan karena data yang digunakan jumlahnya sangat banyak, sehingga kurang memadai jika diolah menggunakan *tools* biasa atau tradisional. *Big data* memerlukan *tools* atau teknologi yang memadai dengan kapasitas penyimpanan yang besar atau dengan nama lainnya adalah teknologi *big data*.

Teknologi *big data* adalah teknologi khusus untuk menangani masalah *big data*. Walaupun *big data* memiliki karakteristik 5V, tetapi masalah yang harus ditangani oleh teknologi *big data* hanya *volume*, *velocity*, dan *variety*. Pertama untuk masalah *volume*, digunakan teknik penyimpanan dan pemrosesan data terdistribusi. Kedua untuk masalah *velocity*, digunakan pemrosesan *stream* dan terdistribusi. Ketiga untuk masalah *variety*, digunakan teknik integrasi data dan penyimpanan data tidak terstruktur [8]. Beberapa teknologi *big data* yang memadai adalah Hadoop, Spark, Hive, Kafka dan masih banyak lagi, dalam tugas akhir ini digunakan Kafka untuk pengumpulan data media sosial Twitter dan digunakan Spark untuk menganalisis akun *buzzer*. Oleh karena itu akan dibahas lebih lanjut tentang Kafka dan Spark.

Kafka diciptakan untuk menangani aliran data secara *real-time* yang aman dan efisien. Seiring dengan berjalannya waktu kebutuhan untuk pemrosesan data secara *real-time* semakin meningkat, oleh karena itu Kafka menjadi semakin banyak digunakan dan populer. Kafka sendiri adalah sistem pengumpul data dan sistem penyederhanaan *data stream*, yang dapat menangani jumlah *volume* data yang sangat banyak secara *real-time*. Komponen dari Kafka terdiri dari *topics*, *partitions*, *cluster*, *broker*, *producer*, dan *consumer*.

Spark adalah sebuah mesin komputasi dan sekumpulan *library* yang dapat digunakan untuk memproses data secara paralel [9]. Spark memiliki keunggulan yang berupa *in-memory processing*, dimana komputasi dilakukan di dalam *memory*. Spark juga dapat digunakan pada berbagai bahasa

pemrograman seperti Python, Java, Scala, dan R, lalu Spark juga mendukung menggunakan kueri *Structured Query Language* (SQL). Spark memungkinkan penggunanya untuk memproses data dalam jumlah yang besar, dengan berbagai macam *library* yang relatif mudah digunakan, seperti Spark SQL, Spark Streaming, Spark MLlib, dan GraphX.

## 1.2 Rumusan Masalah

Rumusan masalah dari latar belakang yang telah dipaparkan adalah sebagai berikut:

1. Bagaimana mengumpulkan data Twitter menggunakan Kafka?
2. Bagaimana eksplorasi dan penyiapan data Twitter yang sudah dikumpulkan?
3. Bagaimana mendeteksi akun *buzzer* pada media sosial Twitter pada sistem tersebar Spark dengan pendekatan analisis *dataset batch*?
4. Bagaimana analisis akun *buzzer* dari data Twitter yang sudah dikumpulkan?
5. Bagaimana menampilkan informasi *buzzer* dan hasil analisisnya dari perangkat lunak dengan menarik dan informatif?

## 1.3 Tujuan

Tujuan penelitian dari rumusan masalah yang telah dipaparkan adalah sebagai berikut:

1. Mengumpulkan data Twitter menggunakan Kafka.
2. Melakukan eksplorasi dan penyiapan data Twitter yang sudah dikumpulkan.
3. Mendeteksi akun *buzzer* pada media sosial Twitter pada sistem tersebar Spark dengan pendekatan analisis *batch dataset*.
4. Melakukan analisis terhadap akun *buzzer* dari data Twitter yang sudah dikumpulkan.
5. Menampilkan informasi *buzzer* dan hasil analisisnya dari perangkat lunak dengan menarik dan informatif.

## 1.4 Batasan Masalah

Batasan masalah untuk penelitian ini adalah:

1. *Tweet* yang dianalisis hanya *tweet* yang berupa teks saja.
2. Teks *tweet* yang dianalisis hanya berasal dari domain Pemilihan umum Presiden Indonesia tahun 2024.

## 1.5 Metodologi

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai akun *buzzer*.
2. Melakukan studi literatur mengenai teknik analisis data terstruktur, statistika dasar, dan dasar-dasar *text mining*.
3. Melakukan studi literatur mengenai Apache Spark dan Apache Kafka.
4. Melakukan studi literatur mengenai analisis sentimen dan algoritma klasifikasi.
5. Melakukan instalasi atau konfigurasi Kafka pada klaster Hadoop/Spark dan pengumpulan data media sosial Twitter menggunakan Kafka.
6. Melakukan eksplorasi terhadap fungsi-fungsi yang dapat digunakan untuk mengumpulkan data deteksi akun *buzzer*.
7. Melakukan eksplorasi *buzzer* secara nyata pada Twitter.
8. Melakukan eksplorasi dan penyiapan data media sosial Twitter.
9. Melakukan deteksi akun *buzzer* dan analisis terhadap *buzzer* yang didapatkan.
10. Melakukan prediksi sentimen akun *buzzer* dan analisis terhadap hasil prediksinya.

11. Mengevaluasi hasil analisis yang didapatkan.
12. Membangun perangkat lunak yang dapat mendeteksi *buzzer*, memprediksi sentimen *buzzer*, dan menampilkan hasilnya.

## 1.6 Sistematika Pembahasan

Sistematika penulisan tugas akhir ini adalah sebagai berikut:

1. Bab 1 Pendahuluan
   Pada Bab 1 membahas tentang pengantar tugas akhir yang berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.
2. Bab 2 Landasan Teori
   Pada Bab 2 membahas tentang studi literatur yang mencakup penjelasan dari konsep-konsep yang digunakan, seperti teori *buzzer*, teori teknik analisis data terstruktur, teori statistika dasar, teori dasar *text mining*, teori Apache Spark, teori Apache Kafka, teori analisis sentimen dan teori mengenai algoritma klasifikasi.
3. Bab 3 Studi Eksplorasi
   Pada Bab 3 membahas tentang langkah-langkah dan eksperimen untuk melakukan konfigurasi pada teknologi yang digunakan, serta hasil eksplorasi *buzzer* secara nyata dan contoh komputasi manual metode pada Python.
4. Bab 4 Pengumpulan, Eksplorasi, dan Penyiapan Data
   Pada Bab 4 membahas tentang bagaimana pengumpulan data dilakukan, lalu eksplorasi terhadap data yang didapatkan, dan diakhiri dengan penyiapan data agar dapat digunakan dengan baik.
5. Bab 5 Analisis Data
   Pada Bab 5 membahas tentang tahap-tahap untuk mendeteksi akun *buzzer* beserta dengan hasil analisis dan evaluasinya, lalu dilanjutkan dengan prediksi sentimen *buzzer* dan hasil analisis serta evaluasinya.
6. Bab 6 Perancangan Perangkat Lunak dan Implementasinya
   Pada Bab 6 membahas tentang fitur yang tersedia pada perangkat lunak dan juga memperlihatkan cara kerja dari perangkat lunak yang dibangun beserta hasilnya.
7. Bab 7 Kesimpulan dan Saran
   Pada Bab 7 membahas tentang kesimpulan dari keseluruhan penelitian yang sudah dilakukan dan saran yang membangun yang bisa diterapkan dipengembangan selanjutnya.