ANALISIS SENTIMEN OPINI PUBLIK TERHADAP PERISTIWA BITCOIN HALVING PADA DATA TEKS TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN PEMBOBOTAN FITUR TF-IDF

SKRIPSI

Diajukan oleh:

Andi Nur Halim

2011102441038

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR

JULI 2024

ANALISIS SENTIMEN OPINI PUBLIK TERHADAP PERISTIWA BITCOIN HALVING PADA DATA TEKS TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN PEMBOBOTAN FITUR TF-IDF

SKRIPSI

Diajukan Sebagai Salah Satu Persyaratan Untuk Memperoleh Gelar Sarjana Teknik Informatika Fakultas Sains Dan Teknologi Universitas Muhammadiyah Kalimantan Timur

Diajukan oleh:

Andi Nur Halim

2011102441038

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR

JULI 2024

LEMBAR PERSETUJUAN

ANALISIS SENTIMEN OPINI PUBLIK TERHADAP PERISTIWA BITCOIN HALVING PADA DATA TEKS TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN PEMBOBOTAN FITUR TF-IDF

SKRIPSI

Diajukan Oleh :
Andi Nur Halim
2011102441038

Disetujui untuk diujikan

Pada tanggal 29 Juni 2024

Pembimbing

Rudiman, S.Kom, M.Sc
NIDN. 1105068202

Mengetahui,

Koordinator Skripsi

Abdul Rahim S.Kom M.cs
NIDN. 0009047901

LEMBAR PENGESAHAN

ANALISIS SENTIMEN OPINI PUBLIK TERHADAP PERISTIWA BITCOIN HALVING PADA DATA TEKS TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN PEMBOBOTAN FITUR TF-IDF

SKRIPSI

Diajukan oleh:

Andi Nur Halim

2011102441038

Diseminarkan dan Diujikan

Pada Tanggal 8 Juli 2024

Penguji I	Penguji II
Naufal Azmi Verdikha, S.Kom., M.Eng **NIDN 1114048801**	**Rudiman, S.Kom., M.Sc** **NIDN 1105068202**

Mengetahui,

Ketua

Program Studi Teknik Informatika

Arbansyah, S.Kom., M.TI

NIDN 1118019203

PERNYATAAN KEASLIAN PENELITIAN

Saya yang bertanda tangan di bawah ini:

Nama	: Andi Nur Halim
NIM	2011102441038
Program Studi	: S1 Teknik Informatika
Judul Penelitian	: Analisis Sentimen Opini Publik terhadap Peristiwa Bitcoin Halving pada Data Teks Twitter Menggunakan Metode Naïve Bayes dan Pembobotan Fitur TF-IDF

Menyatakan bahwa Skripsi yang saya tulis ini benar-benar hasil karya saya sendiri, dan bukan merupakan hasil plagiasi/falsifikasi/fabrikasi baik sebagian atau seluruhnya.

Atas pernyataan ini, saya siap menanggung resiko atau sanksi yang dijatuhkan kepada saya apa bila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam skripsi saya ini, atau klaim dari pihak lain terhadap keaslian karya saya ini.

Samarinda 30 Juni 2024
Yang membuatan pernyataan

Andi Nur Halim
NIM: 2011102441038

ABSTRAK

Perkembangan teknologi telah mengubah cara orang berinteraksi dan melakukan transaksi. Salah satu inovasi penting adalah mata uang digital, yang biasa dikenal sebagai cryptocurrency. Baru-baru ini, topik Bitcoin Halving telah menarik perhatian besar di Twitter, bahkan menjadi trending topic di seluruh dunia. Peristiwa ini memicu banyak opini dan komentar dari pengguna Twitter. Mengingat banyaknya tweet yang terkait dengan Bitcoin Halving, sangat sulit untuk menentukan secara manual apakah sentimennya positif atau negatif. Oleh karena itu, diperlukan text mining untuk mengklasifikasikan sentimen tersebut, baik positif maupun negatif. Penelitian ini bertujuan memanfaatkan algoritma Naïve Bayes Classifier dan pembobotan fitur menggunakan TF-IDF (Term Frequency – Inverse Document Frequency). Dari total 538 data tweet yang diperoleh dari proses crawling di media sosial Twitter, dilakukan preprocessing dan pembobotan kata menggunakan TF-IDF, serta pembagian data untuk pelatihan dan pengujian model. Dengan beberapa rasio data latih dan data uji yaitu 90:10, 80:20, dan 80:20, hasil penelitian ini menunjukkan bahwa model Naïve Bayes dengan rasio 70:30 mendapatkan hasil akurasi terbaik yaitu 74%.

Kata kunci: Bitcoin Halving, Twitter, Analisis Sentimen, Naïve Bayes, TF-IDF

ABSTRACT

Technological advancements have revolutionized how people interact and conduct transactions. One significant innovation is digital currency, commonly known as cryptocurrency. Recently, the topic of Bitcoin Halving has garnered substantial attention on Twitter, even trending worldwide. This event has sparked numerous opinions and comments from Twitter users. Given the large volume of tweets related to Bitcoin Halving, manually determining whether the sentiment is positive or negative is extremely challenging. Therefore, text mining is necessary to classify these sentiments as either positive or negative. This study aims to utilize the Naïve Bayes Classifier algorithm and feature weighting using TF-IDF (Term Frequency – Inverse Document Frequency). From a total of 538 tweet data collected through crawling on social media Twitter, preprocessing and word weighting using TF-IDF were conducted, followed by data splitting for model training and testing. With various training and testing data ratios of 90:10, 80:20, and 80:20, the results of this study indicate that the Naïve Bayes model with a 90:30 ratio achieved the best accuracy of 74%.

Keywords: *Bitcoin Halving, Twitter, Sentiment Analysis, Naïve Bayes, TF-IDF*

PRAKATA

Assalamu'alaikum warahmatullahi wabarkatuh.

Alhamdulillah, dengan mengucap puji syukur kehadirat Allah SWT, atas segala nikmat yang di karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan tepat waktu, dengan judul " Analisis Sentimen Opini Publik Terhadap Peristiwa Bitcoin Halving Pada Data Teks Twitter Menggunakan Metode Naïve Bayes Dan Pembobotan Fitur TF-IDF " yang merupakan tugas akhir selama menempuh pendidikan di Universitas Muhammadiyah Kalimantan Timur dan merupakan salah satu syarat untuk kelulusan dan mendapatkan gelar sarjana yang harus di penuhi.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan kepada penulis dengan secara langsung maupun tidak secara langsung, penulis ingin menyampaikan rasa terima kasih kepada :

1. Puji syukur saya ingin ucapkan kepada Allah SWT atas segala limpahan rahmat dan berkah-Nya yang telah membimbing langkah-langkah saya dalam menyelesaikan skripsi ini.
2. Bapak Rudiman, S.Kom., M.Sc selaku dosen pembimbing yang selalu memberikan masukan, dukungan, semangat dan arahan kepada penulis untuk menyelesaikan skripsi ini.
3. Bapak Naufal Azmi Verdikha, S.Kom., M.Eng., sebagai dosen penguji pertama fakultas Sains dan Teknologi Universitas Muhammadiyah Kalimantan Timur.
4. Bapak Arbansyah, S.Kom., M.TI sebagai ketua Program Studi Teknik Informatika yang telah memberikan dukungan selama masa perkuliahan di Teknik Informatika.
5. Bapak Abdul Rahim S.Kom., M.Cs selaku Koordinator Skripsi yang telah memberikan dukungan dan arahan dalam proses penyelesaian skripsi.
6. Bapak Taghfirul Yoga Azhima Yoga Siswa selaku dosen pembimbing Akademik yang telah membimbing dari masa perkuliahan.
7. Dr. Muhammad Musiyam, M.T, selaku Rektor Universitas Muhammadiyah Kalimantan Timur yang telah memberikan dukungan, inspirasi, dan motivasi yang sangat berarti selama penulisan skripsi ini.
8. Seluruh dosen Teknik Informatika yang telah memberikan banyak ilmu dan pengetahuan kepada penulis, agar penulis bisa menyelesaikan skripsi ini dengan baik.
9. Kepada pahlawan dan panutanku, Ayahanda Abrani. Terima kasih selalu berjuang untuk kehidupan penulis, beliau memang tidak sempat menyelesaikan pendidikan sampai bangku perkuliahan, namun beliau mampu mendidik penulis, memotivasi dan memberikan dukungan hingga mampu menyelesaikan studinya sampai sarjana.
10. Kepada ibunda tercinta, Fitriah yang tidak henti hentinya memberikan kasih sayang dengan penuh cinta dan selalu memberikan motivasi serta do'a hingga penulis mampu menyelesaikan studinya sampai sarjana.

11. Kepada rekan-rekan sahabat dan teman atas dukungan dan kerjasamanya selama menempuh pendidikan hingga menyelesaikan penyusunan skripsi ini.

Akhir kata, penulis berharap agar skripsi ini dapat memberikan kontribusi yang bermanfaat bagi perkembangan ilmu pengetahuan, khususnya dalam bidang analisis sentimen dan text mining. Semoga penelitian ini dapat menjadi referensi bagi penelitian selanjutnya. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, oleh karena itu penulis sangat mengharapkan saran dan kritik yang membangun dari berbagai pihak. Terima kasih.

Samarinda 30 Juni 2024
Penyusun

Andi Nur Halim

DAFTAR ISI

Halaman

LEMBAR PERSETUJUAN ii
LEMBAR PENGESAHAN iii
PERNYATAAN KEASLIAN PENELITIAN iv
ABSTRAK v
ABSTRACT vi
PRAKATA vii
DAFTAR TABEL xi
DAFTAR GAMBAR xii
DAFTAR LAMPIRAN xiii
BAB I 1
PENDAHULUAN 1
1. Latar belakang Masalah 1
1.1. Rumusan masalah 3
1.2. Tujuan penelitian 4
1.3. Manfaat Penelitian 4
BAB II 5
METODE PENELITIAN 5
2.1 Objek penelitian 5
2.2 Alat dan bahan 5
2.3 Prosedur Penelitian 6
2.3.1 Pengumpulan data 7
2.3.2 Labeling Data 8
2.3.3 *Pre-Processing* 9
2.3.4 TF-IDF (*Term Frequency - Inverse Document Frequency*) 11
2.3.5 Split Data 13
2.3.6 *Naïve Bayes Classifier* 14
2.3.7 Evaluasi 16
BAB III 17
HASIL ANALISIS DAN PEMBAHASAN 17
3.1 Dataset 17
3.2 *Labeling* 18
3.3 *Pre-processing* 19

a. *Case folding* 20
b. *Cleansing* 21
c. *Tokenizing* 21
d. *Stopword Removal* 22
e. *Stemming* 23
f. *Delete Duplicates* 23
3.4 *Wordcloud* 24
3.5 Pembobotan Kata TF-IDF 25
3.6 Split Data 26
3.7 Klasifikasi 27
3.8 Evaluasi 30
BAB 4 33
PENUTUP 33
4.1 Kesimpulan 33
4.2 Saran 33
DAFTAR PUSTAKA 35
RIWAYAT HIDUP PENULIS 38
LAMPIRAN 39

DAFTAR TABEL

Halaman

Tabel 2. 1 Confusion Matrix 16
Tabel 3. 1 Dataset 17
Tabel 3. 2 Labeling Dataset 18
Tabel 3. 3 Hasil case folding 20
Tabel 3. 4 Hasil Cleansing 21
Tabel 3. 5 Hasil Tokenizing 21
Tabel 3. 6 Hasil Stopword Removal 22
Tabel 3. 7 Hasil Stemming 23
Tabel 3. 8 Hasil TF-IDF 25

DAFTAR GAMBAR

Halaman

Gambar 2. 1 Kerangka Penelitian 7
Gambar 2. 2 Tahapan Preprocessing 10
Gambar 2. 3 Split Data 13
Gambar 3. 1 Hasil Crawling 17
Gambar 3. 2 Visualisasi Persentase Sentimen 19
Gambar 3. 3 Delete Duplicates 24
Gambar 3. 4 Visualisasi Wordcloud 25
Gambar 3. 6 Split Dataset 27
Gambar 3. 7 Confusion Matrix Rasio 90:10 28
Gambar 3. 8 Confusion Matrix 80:20 29
Gambar 3. 9 Confusion Matrix Rasio 70:30 30
Gambar 3. 10 Perbandingan Accuracy Naïve Bayyes 31

DAFTAR LAMPIRAN

Halaman

Lampiran 1. 1 CV Expert Labelling 39
Lampiran 2. 1 Code Crawling Twitter 40
Lampiran 2. 2 Import Library & Pip Install 41
Lampiran 2. 3 Read Dataset & Count Sentiment 42
Lampiran 2. 4 Preprocessing 42
Lampiran 2. 5 Code Untuk Menyimpan Hasil Teks Preprocessing 43
Lampiran 2. 6 Code Delete Duplicate 43
Lampiran 2. 7 Code Visualisasi Persentase Sentimen 44
Lampiran 2. 8 Code Wordcloud Sebelum Teks Preprocessing 44
Lampiran 2. 9 Code Wordcloud Setelah Teks Preprocessing 44
Lampiran 2. 10 TF-IDF (Term Frequency – Inverse Document Frequency) 45
Lampiran 2. 11 Naive Bayes Classification 45
Lampiran 2. 12 Code Confusion Matrix 46
Lampiran 3. 1 Kartu Kendali Bimbingan 47

BAB I
PENDAHULUAN

1. Latar belakang Masalah

Seiring dengan kemajuan teknologi informasi yang terus meningkat, media sosial juga mengalami pertumbuhan yang sangat cepat. Fenomena tersebut mengakibatkan timbulnya istilah 'banjir data' yang sebagian besar sumbernya berasal dari platform-platform media sosial yang ada (Julianto *et al*., 2022). Media sosial merupakan salah satu tempat yang sering digunakan untuk memberikan opini serta pendapat, opini dan pendapat tersebut bisa berbeda beda, dapat berupa positif, negatif maupun netral (Humam & Laksito, 2023). Dari berbagai platform media sosial yang tersedia, salah satu yang paling diminati dan digunakan adalah *twitter*. (Asmara *et al*., 2020).

Twitter merupakan platform sosial media yang memungkinkan penggunanya untuk membuat unggahan dan berinteraksi melalui tulisan yang disebut dengan *tweet*. Pada awalnya *tweet* dibatasi maksimal 140 karakter, akan tetapi kemudian diperluas menjadi 280 karakter. Penyebaran informasi melalui *twitter* bersifat *real-time* dan memiliki fitur *trending topic* apabila topik-topik yang muncul banyak dibahas penggunanya (Fauzianto *et al*., 2023), salah satu perbincangan yang cukup banyak di bicarakan di *twitter* adalah peristiwa bitcoin halving yang terjadi pada bulan April tahun 2024.

Terkait dengan hal tersebut, bitcoin adalah mata uang digital atau kripto yang diciptakan pada tahun 2009 oleh seseorang atau kelompok yang dikenal dengan nama samaran Satoshi Nakamoto (Septiarini *et al*., 2020). Bitcoin merupakan mata uang digital desentralisasi pertama yang tidak dikendalikan oleh lembaga atau otoritas keuangan manapun (Ramos *et al*., 2020). Transaksi bitcoin diverifikasi dan dicatat dalam sebuah buku besar publik yang disebut blockchain.

Halving adalah proses pengurangan *reward* atau imbalan yang diterima oleh para penambang *(miners)* setelah berhasil memverifikasi sebuah blok transaksi baru di dalam

blockchain (Meynkhard, 2019). Peristiwa halving terjadi setiap empat tahun sekali dengan tujuan untuk mengurangi jumlah pasokan bitcoin yang akan diedarkan di pasar (Meynkhard, 2019). Ketika halving harga bitcoin secara historis menunjukkan tren kenaikan. Fenomena ini dapat dijelaskan oleh konsep sederhana tentang permintaan dan penawaran, ketika jumlah bitcoin yang dihasilkan berkurang, maka kelangkaan akan meningkat sehingga nilai dari bitcoin itu sendiri akan menjadi lebih tinggi (M. H. Z. K. Ramadhani, 2022). Hal yang membuatnya sangat banyak diperbincangkan di *twitter* adalah dikarenakan bitcoin merupakan instrumen investasi dengan volatilitas yang sangat tinggi atau tidak stabil, serta fluktuasinya yang sangat cepat (M. H. Z. K. Ramadhani, 2022).

Dengan menggunakan kata kunci "Bitcoin Halving *lang:id*" banyak *tweet* yang membahas tentang peristiwa ini, mulai dari prediksi harga, analisis pasar, sampai dengan opini dan spekulasi dari para pengguna *twitter* mengenai dampak halving terhadap harga bitcoin ke depannya. Dengan demikian opini-opini terkait peristiwa halving yang tersebar di *twitter* dapat analisis menggunakan metode klasifikasi sentimen (Fauzianto *et al*., 2023). Klasifikasi sentimen merupakan sebuah proses pengolahan data teks secara otomatis untuk memperoleh informasi terkait kecenderungan penilaian terhadap suatu objek, baik itu penilaian yang bersifat positif, negatif maupun netral (Julianto *et al*., 2022).

Penelitian ini menggunakan metode klasifikasi sentimen dengan fitur TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk pembobotan kata, dan *Naive Bayes Classifier* (NBC) untuk mengkategorikan data teks ke dalam kelas sentimen positif atau negatif (Imelda & Arief Ramdhan Kurnianto, 2023). Metode *Naive Bayes Classifier* memiliki sejumlah kelebihan, di antaranya adalah kecepatan dalam melakukan komputasi, kesederhanaan algoritma yang digunakannya, serta kemampuan untuk menghasilkan klasifikasi dengan akurasi yang tinggi (Zhafira *et al*., 2021).

Dari penjabaran diatas metode tersebut banyak digunakan oleh para peneliti untuk melakukan klasifikasi sentimen, seperti penelitian yang dilakukan oleh Srividya *et al*. pada tahun 2019 penelitiannya berjudul *Aspect Based Sentiment Analysis using POS Tagging and TFIDF* mengadopsi analisis sentimen berbasis aspek dengan menggunakan metode *Naïve Bayes* dan *Support Vector Machine* (SVM). Dalam penelitian tersebut, terdapat dua model yang dikembangkan, yaitu model 1 yang menggunakan *POS tagging*, dan model 2 yang menggunakan TF-IDF. Hasil penelitian menunjukkan bahwa performansi model 2 lebih unggul daripada model 1 (Srividya & Mary Sowjanya, 2019). Penelitian lainnya oleh Elly Firasari *et al*. pada tahun 2020 mengenai *Comparation of K-Nearest Neighboor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance.* Hasil penelitian tersebut menunjukan kedua metode bekerja dengan baik dalam melakukan klasifikasi, namun metode *Naive Bayes* lebih unggul dari metode *K-Nearest Neighbor* (K-NN) dengan mencapai akurasi yang lebih tinggi (Firasari *et al*., 2020).

Berdasarkan uraian di atas, riset ini akan melakukan klasifikasi sentimen dengan memanfaatkan opini pengguna Twitter terhadap peristiwa halving. Opini-opini tersebut akan dibagi menjadi data sentimen positif maupun negatif. Pengklasifikasian dalam pengumpulan data tweet dilakukan dengan menggunakan metode *Naïve Bayes Classifier* dan TF-IDF. Selanjutnya, hasil dari klasifikasi data tersebut digunakan untuk memberikan informasi tentang bagaimana persepsi publik terkait peristiwa halving, apakah cenderung positif atau negatif. Tidak jarang ada beberapa peristiwa yang direspons positif oleh masyarakat tetapi dinilai negatif oleh para analis pasar, atau sebaliknya. Namun, dilakukannya analisis sentimen ini dapat memberikan gambaran yang lebih objektif mengenai persepsi publik secara umum, sehingga dapat menjadi pertimbangan penting bagi para pemangku kepentingan dalam mengevaluasi dampak dan implikasi dari peristiwa halving secara lebih komprehensif.

1.1. Rumusan masalah

Rumusan masalah dalam penelitian ini adalah bagaimana hasil akurasi yang akan di dapatkan dalam klasifikasi teks twitter mengenai peristiwa Bitcoin Halving dengan menggunakan pembobotan TF-IDF dan metode *Naïve Bayes*?

1.2. Tujuan penelitian

Tujuan penelitian ini adalah Mendapatkan akurasi klasifikasi sentimen pada peristiwa Bitcoin halving menggunakan pembobotan TF-IDF dengan Algoritma *Naïve Bayes*.

1.3. Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini meliputi. (i) Bagi penulis, penelitian ini merupakan sebuah eksplorasi teori-teori yang selama ini dipelajari, serta menambah wawasan dan ilmu pengetahuan serta pengalaman terhadap analisis sentimen *data mining.* (ii) Bagi Universitas, sebagai tolak ukur pengetahuan mahasiswa dalam menguasai ilmu yang sudah di pelajari dan sebagai referensi untuk penelitian selanjutnya. (iii) Bagi pembaca, memberikan informasi mengenai sentimen terhadap peristiwa Halving dan bermanfaat untuk referensi penelitian analisis sentimen di bidang Teknik Informatika.

BAB II
METODE PENELITIAN

2.1 Objek penelitian

Objek pada penelitian ini adalah data sosial media *twitter* yang di dapatkan dari proses (*crawling*), dan peneliti melakukan beberapa tahapan dalam melakukan penelitian meliputi, peristiwa halving sebagai topik sentimen yang dianalisis, analisis sentimen sebagai metode klasifikasi sentimen yang akan diaplikasikan, serta pembobotan fitur yang diterapkan menggunakan TF-IDF, dan *Naïve Bayes Classifier* sebagai algoritma klasifikasi yang akan diimplementasikan untuk melakukan klasifikasi sentimen terhadap data *tweet* terkait peristiwa bitcoin halving.

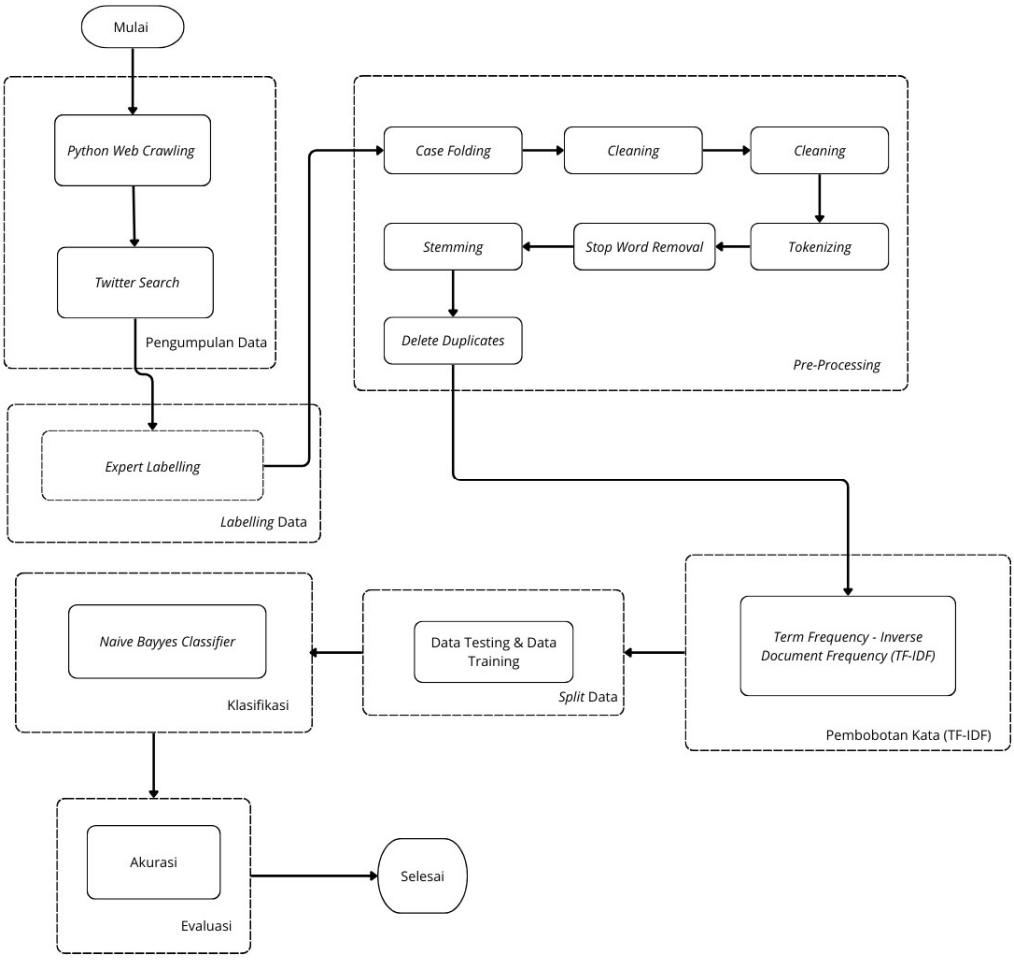
2.2 Alat dan bahan

Alat yang digunakan dalam penelitian ini meliputi perangkat keras dan perangkat lunak. Perangkat keras yang digunakan peneliti adalah laptop dengan spesifikasi AMD Ryzen 5 Series 5000H, RAM 16GB, dan penyimpanan SSD 512GB. Perangkat lunak yang digunakan mencakup *Google Collaboratory* versi 1.0.0 yang dapat diakses di https://colab.research.google.com/, Python versi 3.8.10, Node.js versi 14.16.0 digunakan untuk menjalankan *Tweet Harvest* versi 2.6.0 yang fungsinya *collect* data berdasarkan kata kunci yang di pakai. yang digunakan dalam penelitian ini berupa *library Python* seperti API *Twitter*, (i) *Pandas* versi 2.0.3 untuk manipulasi dan analisis data, (ii) *Numpy* versi 1.25.2 untuk operasi data numerik, (iii) *Re* versi 2.2.1 untuk operasi *Regex,* (iv) NLTK versi 3.8.1 untuk pemrosesan bahasa alami, (v) *Scikit-Learn* versi 1.2.2 untuk machine learning dan analisis data, (vi) *Matplotlib* versi 3.7.1 untuk visualisasi data, (vii) *Sastrawi* versi 1.0.1 untuk stemming bahasa indonesia.

Bahan yang digunakan dalam penelitian ini adalah dataset *tweet* terkait peristiwa bitcoin halving di sosial media *twitter*. Data tersebut menjadi sumber utama dalam untuk analisis percakapan dan opini publik mengenai peristiwa bitcoin halving di *twitter*.

2.3 Prosedur Penelitian

Penelitian dimulai dengan pengumpulan data dari *twitter* menggunakan *python*. Selanjutnya, proses labeling dilakukan oleh ahli bahasa (*expert*) untuk memberikan label sentimen pada data, apakah positif atau negatif. Setelah itu pra-pemrosesan data meliputi *case folding*, *cleansing*, *tokenizing*, *stopwords removal* dan *stemming.* Ekstraksi fitur dilakukan dengan metode TF-IDF. Data dibagi menjadi data latih dan data uji. Data latih digunakan untuk melatih model *Naive Bayes* dalam mengklasifikasikan sentimen menjadi positif dan negatif. Terakhir, kinerja model dievaluasi menggunakan *Confusion Matrix* untuk menilai kualitas klasifikasi sentimen pada data teks *twitter* seperti yang ditampilkan dalam Gambar 2.1.



Gambar 2. 1 *Kerangka Penelitian*

2.3.1 Pengumpulan data

Pengumpulan data dilakukan melalui *crawling* data di platform media sosial *Twitter* menggunakan *library Tweet-Harvest* yang dikembangkan dengan *Node.js* dan dapat diakses dengan bahasa pemrograman *Python*. *Library Tweet-Harvest* digunakan sebagai alat untuk mengumpulkan dan menganalisis informasi dari *Twitter* dengan kata kunci "*Bitcoin Halving lang:id*". *Tweet-Harvest* adalah *tools* untuk mengumpulkan data dari *Twitter* dengan memanfaatkan *auth_token Twitter*(Yuniarossy *et al*., 2024). Dengan menggunakan *library Tweet-Harvest*, peneliti dapat mengekstrak *tweet* yang relevan dengan topik tertentu terkait

konteks penelitian. Berdasarkan lampiran (2.1) Tahapan-tahapan pengumpulan data yang dilakukan dalam penelitian adalah sebagai berikut:

a. Tahap pertama adalah token autentikasi dari twitter, token autentikasi tersebut disimpan di variabel *twitter_auth_token*. Token tersebut digunakan untuk melakukan akses API Twitter untuk mengambil data tweet yang diperlukan dalam penelitian.
b. Tahap kedua adalah instalasi *library pandas* dan *Node*.js. Library *pandas* digunakan untuk manipulasi dan analisis data yang telah dikumpulkan. dan *Node.js* digunakan untuk menjalankan *tweet-harvest* dan mengakses API *Twitter.*
c. Pada tahap ketiga, peneliti menentukan nama *file* dan format penyimpanan data yang akan digunakan untuk menyimpan *tweet* yang dikumpulkan. Data *tweet* yang telah di *crawling* disimpan dalam format CSV untuk memudahkan proses analisis selanjutnya.
d. Pada tahap keempat, yaitu tahap *read* dan visualisasi data yang telah di *crawling.* Data yang telah dikumpulkan dalam format CSV dibaca ke dalam dataframe menggunakan *library pandas*. Kemudian *dataframe* tersebut ditampilkan untuk memeriksa data yang berhasil diambil.
e. Tahap terakhir, peneliti memeriksa jumlah dataset yang ada di dalam *dataframe*. Proses ini dilakukan dengan menghitung panjang dataframe, langkah tersebut dilakukan untuk memastikan jumlah data yang telah dikumpulkan sesuai dengan kebutuhan penelitian.

2.3.2 Labeling Data

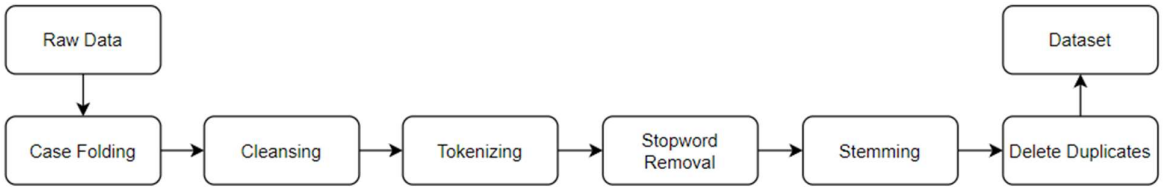
Pada proses klasifikasi teks pada dataset yang telah didapatkan sebagai bagian dari tugas akhir (skripsi), peneliti memerlukan bantuan dari seorang ahli bahasa yang memiliki pengalaman dalam pelabelan data. Oleh sebab itu peneliti mengirim permintaan melalui website *project.co.id* untuk mencari individu yang sesuai dengan kriteria tersebut. Dalam permintaan tersebut, peneliti menjelaskan bahwa peneliti mencari lulusan yang saat ini bekerja di bidang terkait seperti guru atau dosen, atau memiliki keahlian khusus dalam bahasa. Calon

ahli diminta untuk mengajukan penawaran yang mencantumkan pekerjaan saat ini, yaitu pengalaman relevan dalam pelabelan data, serta gelar akademik (Lampiran 1). Peneliti juga memberikan kesempatan bagi calon ahli untuk mengajukan pertanyaan lebih lanjut jika diperlukan. Pedoman untuk melakukan pelabelan oleh ahli bahasa dalam data ini adalah sebagai berikut:

a. Positif : Bitcoin merupakan aset masa depan, antusiasme dan optimisme pada peristiwa halving, serta harapan akan peningkatan nilai Bitcoin menjadi respons yang mendukung, motivasi untuk pertumbuhan, serta analisis positif tentang dampak halving menambah keyakinan dalam komunitas.
b. Negatif : penggunaan kata kata kasar dan kotor, skeptisme terhadap peristiwa halving, kekhawatiran terhadap volatilitas pasar yang tidak terduga, serta ketidakpastian mengenai stabilitas jangka panjang pada nilai bitcoin.

2.3.3 *Pre-Processing*

dataset yang sudah dilabeli, Selanjutnya adalah tahapan *Preprocessing*. Tahapan ini merupakan salah satu komponen utama yang penting di dalam *text mining* (S. Ramadhani *et al*., 2022). *Preprocessing* merupakan metode dalam penambangan data yang mengubah data asli menjadi format yang terstruktur dan lebih mudah dimengerti (Barus, 2022). Tahapan *preprocessing* mencakup enam langkah yaitu *case folding*, *cleansing*, *tokenizing*, *stopword removal*, *stemming*, dan penghapusan duplikat. Data yang dihasilkan dari proses crawling masih mengandung duplikat yang disebabkan oleh *tweet* yang diposting berulang kali, sehingga perlu dilakukan penghapusan data duplikat.(Fitriyah & Kartikasari, 2023) berdasarkan lampiran (2.4), (2.5) dan (2.6), tahapan preprocessing pada gambar 2.2 sebagai berikut :



Gambar 2. 2 *Tahapan Preprocessing*

a. *Case Folding*

Case folding merupakan proses mengonversi semua teks menjadi huruf kecil (*lowercase*) untuk menghilangkan variasi bentuk huruf besar dan kecil dalam analisis teks. Fungsinya adalah menyeragamkan representasi teks agar lebih konsisten dan mengurangi dimensi fitur yang perlu diproses saat analisis teks.

b. *Cleansing*

tahap pembersihan teks yang melibatkan eliminasi elemen-elemen yang tidak relevan atau tidak penting dalam teks, seperti tanda baca, karakter khusus, atau karakter yang tidak diinginkan lainnya. Hal ini dilakukan untuk membersihkan teks dari *noise* atau gangguan yang tidak diperlukan.

c. *Tokenizing*

Setelah cleansing tahap selanjutnya adalah *tokenizing*, *tokenizing* merupakan proses pembagian teks menjadi bagian-bagian yang lebih kecil, yang disebut *token*. *Token* dapat berupa kata-kata atau karakter, Langkah ini memungkinkan teks untuk dipecah menjadi unit-unit yang lebih kecil, yang nantinya akan diolah dalam analisis teks atau pemodelan.

d. *Stopword Removal*

Stopword removal adalah tahap dalam pemrosesan teks yang melibatkan penghapusan kata-kata penghenti dari teks. Kata-kata penghenti merupakan kata-kata yang sangat umum dalam suatu bahasa dan cenderung tidak memberikan banyak informasi penting dalam analisis teks. Dengan menghapus kata-kata penghenti, peneliti dapat fokus pada kata-kata kunci yang lebih bermakna dalam teks untuk analisis lebih lanjut.

e. *Stemming*

Selanjutnya adalah *stemming, Stemming* merupakan langkah dalam pemrosesan teks yang melibatkan pemangkasan akhiran atau awalan kata untuk menghasilkan akar kata atau bentuk dasarnya. Fungsinya adalah untuk mengurangi variasi kata yang mungkin muncul dalam teks yang sama, sehingga kata-kata yang memiliki akar yang sama akan dianggap sama dalam analisis teks. Hal ini membantu meningkatkan konsistensi dan efektivitas dalam pemrosesan teks serta mengurangi kompleksitas dalam pengolahan data.

f. *Delete Duplicate*

Tahapan terakhir adalah menghapus data duplikat (delete duplicate) adalah langkah penting dalam proses pembersihan data untuk memastikan kualitas dan keakuratan dataset. Proses ini penting dalam analisis karena duplikasi data dapat mengganggu hasil analisis dan menyebabkan bias dalam interpretasi.

2.3.4 TF-IDF (*Term Frequency - Inverse Document Frequency*)

Selanjutnya, dilakukan pembobotan tiap kata menggunakan TF-IDF (*term frequency, inverse document frequency)*. TF-IDF Merupakan proses perhitungan atau pengekstrakan kata menjadi sebuah angka berbentuk vektor yang digunakan untuk menentukan bobot dari sebuah kata dalam sebuah dokumen atau korpus. Bobot ini berguna untuk menentukan seberapa penting kata tersebut dalam sebuah dokumen(Tri Putra *et al*., 2023). Pada proses pembobotan kata, terdapat beberapa langkah yang dilakukan yaitu mencari *term frequency, inverse document frequency* serta *term frequency – inverse document frequency*(S. Ramadhani *et al.,* 2022). Proses pembobotan kata melibatkan penggunaan *TfidfVectorizer* dari *scikit-learn* untuk mengubah teks menjadi representasi numerik, di mana representasi ini bergantung pada bobot kata-kata yang dihitung menggunakan TF-IDF. (Br Sinulingga & Sitorus, 2024). Tahapan pembobotan kata TF-IDF yang dilakukan pada penelitian ini pada (lampiran 2.10) adalah sebagai berikut :

a. Tahap pertama peneliti melakukan import library Pandas untuk membaca dan mengelola data, NumPy untuk operasi numerik, dan *TfidfVectorizer* dari *scikit-learn* untuk mengubah teks menjadi representasi TF-IDF.
b. Tahap kedua adalah pembacaan data, di mana peneliti menggunakan *Pandas* untuk membaca dataset *'dataset_uji.csv'* ke dalam *dataframe*. Selanjutnya, dari *dataframe* tersebut, peneliti mengambil kolom *'cleaned_text*' sebagai daftar dokumen yang telah melalui proses *'stemming'*, dan kolom *'Sentimen'* sebagai daftar sentimen.
c. Pada tahap ketiga, persiapan dan penerapan TF-IDF. peneliti membuat sebuah *instance TfidfVectorizer* dan mengaplikasikannya untuk mengkonversi dokumen menjadi matriks TF-IDF. Proses ini mentransformasikan teks ke dalam representasi numerik yang mencerminkan relevansi setiap kata dalam dokumen.
d. Tahap terakhir adalah print output dari metode TF-IDF. Dan langkah berikutnya adalah menampilkan array position dari dokumen dataset, dan output numerik TF-IDF untuk setiap term.

   Berikut cara penghitungan TF-IDF dapat menggunakan rumus persamaan (2.1) di bawah ini:

$$tf_{t.d} = \frac{Jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{Jumlah\ total\ kata\ dalam\ dokumen\ d} \quad (2.1)$$

Term Frequency (TF) mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Dan Terkait rumus IDF pada persamaan (2.2) yaitu:

$$idf_d = log\frac{N}{n_t} \quad (2.2)$$

Inverse Document Frequency (IDF) mengukur seberapa penting sebuah kata dengan memperhitungkan seberapa sering kata tersebut muncul dalam semua dokumen. N

melambangkan jumlah total dokumen dalam kumpulan teks. Terkait rumus TF-IDF score pada persamaan (2.3) yaitu:

$$tfidf_{t.d} = tf_{t.d} \text{ X } idf_d \quad (2.3)$$

Keterangan:

t = Kata kunci, term

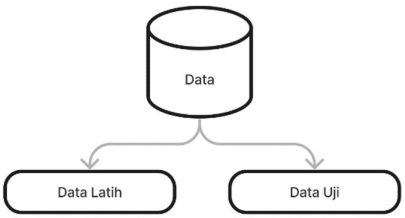
d = Dokumen

t.d = nilai TF-IDF untuk kata t dalam dokumen d

Tf = Banyaknya t (kata) yang dicari dalam dokumen

Idf = Banyak t kebalikan dari kata yang dicari

2.3.5 Split Data



Gambar 2. 3 *Split Data*

Split data adalah proses membagi dataset yang digunakan dalam penelitian menjadi dua atau lebih bagian, gambar 2.3 merupakan ilustrasi proses *split data* yang biasanya digunakan untuk menguji model atau algoritma. Dataset tersebut umumnya dibagi menjadi data latih (*training*) dan data uji (*testing*). Data latih digunakan untuk melatih algoritma, sedangkan data uji digunakan untuk mengevaluasi kinerja algoritma tersebut (Putri *et al*., 2023). Oleh karena itu pada penelitian ini data dibagi menjadi dua tahap, split data pada penelitian ini menggunakan rasio 90:10,80:20,70:30, Tahapan split data dan dilampirkan pada lampiran (2.11) sebagai berikut :

a. Langkah pertama adalah mengimpor fungsi train_test_split dari *sklearn.model* selection yang digunakan untuk membagi data.

b. Langkah kedua adalah membagi data dari term_frequency_all (fitur TF-IDF) dan kolom label df_preprocessed['Sentimen'] dari DataFrame dengan proporsi data uji sebesar 10%,20%,30% dan sisanya sebagai data latih.

c. Langkah ketiga adalah menampilkan jumlah data yang dihasilkan dari proses pembagian data tersebut.

2.3.6 *Naïve Bayes Classifier*

Naïve Bayes Classifier adalah metode klasifikasi yang didasarkan pada teorema Bayes. Metode ini menggunakan pendekatan probabilitas dan statistik yang dikembangkan oleh ilmuwan Inggris, Thomas Bayes untuk memprediksi peluang di masa depan berdasarkan data historis (Asmara *et al*., 2020). Metode ini mengasumsikan bahwa setiap fitur (kata) bersifat independen atau tidak saling bergantung satu sama lain. Naive Bayes merupakan model klasifikasi probabilistik dan statistik yang sederhana yang mengklasifikasikan data berdasarkan probabilitas tertinggi dari suatu kelas dengan mempertimbangkan nilai-nilai fitur yang diberikan(Fikri *et al*., 2020). Metode ini memiliki keunggulan karena memerlukan jumlah data latih yang relatif sedikit untuk menentukan parameter yang diperlukan dalam proses klasifikasi (Berliani & Lestari, 2024). Berdasarkan lampiran (2.11) tahapan klasifikasi sentimen menggunakan Naïve Bayes sebagai berikut :

a. Tahap pertama import library '*TfIdfVectorizer*' untuk mengubah teks menjadi representasi numerik berbasis TF-IDF. Dan '*Multinomial'* untuk inisiasi model *Naïve Bayyes.*

   Serta '*accuracy_score','classification_report'* dari '*sklearn.metrics'* untuk evaluasi performa model.

b. Tahap kedua peneliti menginisialisasi '*TfIdfVectorizer'* dengan parameter khusus, yaitu '*max_features=1000'* untuk menentukan jumlah fitur maksimum yang akan dipertahankan.

Dan '*min_df=5'* untuk mengabaikan kata-kata yang muncul kurang dari lima kali dalam dokumen, dan '*max_df=0.7'* untuk mengabaikan kata-kata yang muncul lebih dari 70% dokumen. Kemudian peneliti mengubah teks dalam data latih ('*x_train'*) dan data uji *('x_test'*) menjadi matrix TF-IDF menggunakan '*fit_transform'* dan '*transform*'.

c. Tahap ketiga peneliti menginisiasi model Naïve Bayyes dengan parameter '*alpha=0.1'* yang menentukan tingkat regularisasi *smoothing Laplace*.

d. Tahap keempat, peneliti melatih model Naïve Bayes menggunakan data latih '*x_train_tfidf'* dan label '*y_train'*.

e. Tahap kelima, melakukan prediksi pada data uji, peneliti menggunakan model yang telah di latih untuk melakukan prediksi pada data uji '*X_test_tfidf'*.

f. Tahap terakhir, peneliti melakukan evaluasi performa model dengan, menghitung akurasi menggunakan '*accuracy_score'* dan menampilkan laporan klasifikasi menggunakan '*classification_report'* yang memberikan matrix evaluasi seperti, presisi, recall dan f1-score.

Secara garis besar model naïve bayes adalah (2.4) sebagai berikut:

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} \tag{2.4}$$

Dimana :

P(X/Y) = persentase X dalam Y

P(X ∩ Y) = Kata tertentu dalam kelas (tweet)

P(Y) = Jumlah kemunculan kata berlabel tertentu (Positif, Negatif)

Evaluasi pada penelitian ini menggunakan *Confusion matrix* sebagai alat untuk mengukur tingkat akurasi. *Confusion matrix* merupakan representasi visual yang *powerful* dan sangat berguna untuk memperkirakan performa model dengan menghitung jumlah prediksi yang benar dan salah, meliputi *True Positive* (TP) yaitu jumlah data positif yang di prediksi

dengan akurat, *False Positive* (FP) yaitu jumlah data negatif yang keliru diprediksi sebagai positif , *False Negative* (FN) yaitu jumlah data positif yang keliru diprediksi sebagai negatif, dan *True Negative* (TN) yaitu jumlah data negatif yang di prediksi dengan tepat sebagai negatif (Noor Hasan, 2024). Dengan menganalisis *Confusion matrix* secara cermat, peneliti memperoleh informasi berharga tentang kekuatan dan kelemahan model dalam mengklasifikasikan sentimen dengan tepat. Visualisasi Confusion Matrix ditampilkan berdasarkan lampiran (2.12).

Tabel 2. 1 *Confusion Matrix*

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

2.3.7 Evaluasi

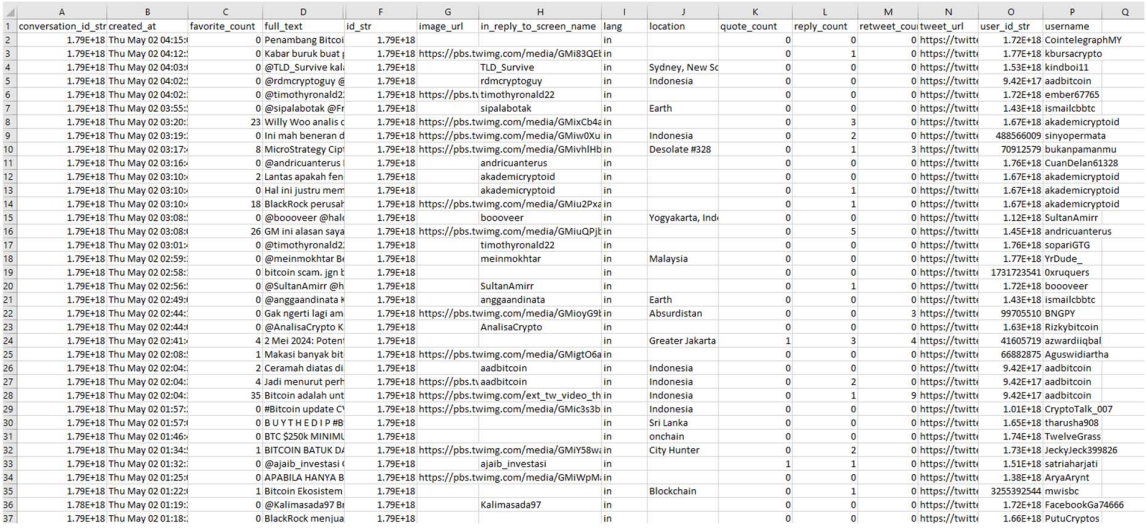
Akurasi adalah salah satu matrix evaluasi yang paling umum digunakan dalam klasifikasi, Akurasi mengukur seberapa sering model klasifikasi memberikan prediksi yang benar dibandingkan dengan keseluruhan prediksi yang dibuat. Secara umum nilai akurasi mengindikasikan rasio data *tweet* yang terdeteksi dengan benar dalam dataset pengujian. Secara sederhana, akurasi mengukur seberapa dekat prediksi sistem dengan prediksi yang dibuat oleh manusia (Azhari *et al.,* 2021). Pada penelitian ini akurasi merupakan parameter penting dalam mengevaluasi performa sistem klasifikasi, yang memberikan gambaran tentang seberapa baik sistem dapat mengklasifikasikan data. akurasi dapat dihitung dengan menggunakan persamaan (2.5) berikut:

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (2.5)$$

BAB III
HASIL ANALISIS DAN PEMBAHASAN

3.1 Dataset

Dalam penelitian ini hasil *crawling* data terdiri dari 15 kolom yang mencakup *conversation_id_str, created_at, favorite_count, full_text, id_str, image_url, in_reply_to_screen_name, lang, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, dan username*, data yang didapatkan berjumlah 558 *tweet* yang terkait dengan peristiwa bitcoin halving yang dikumpulkan dari *Twitter*. Setelah itu data *tweet* tersebut telah melalui proses preprocessing hasil akhir data yang akan berjumlah 538, data disimpan dalam format CSV. Gambar 3.1 menunjukkan hasil pengambilan data menggunakan alat pengumpul *tweet (crawling)*. Data pengambilan tersebut dapat dilihat pada gambar di bawah ini.



	A	B	C	D	F	G	H	I	J	K	L	M	N	O	P	Q
1	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_cou	tweet_url	user_id_str	username	
2	1.79E+18	Thu May 02 04:15:0	0	Penambang Bitcoi	1.79E+18			in		0	0	0	https://twitte	1.72E+18	CointelegraphMY	
3	1.79E+18	Thu May 02 04:12:2	0	Kabar buruk buat	1.79E+18	https://pbs.twimg.com/media/GMi83QEb		in		0	1	0	https://twitte	1.77E+18	kbursacrypto	
4	1.79E+18	Thu May 02 04:03:0	0	@TLD_Survive kal	1.79E+18		TLD_Survive	in	Sydney, New Sc	0	0	0	https://twitte	1.53E+18	kindboi11	
5	1.79E+18	Thu May 02 04:02:2	0	@rdmcryptoguy @	1.79E+18		rdmcryptoguy	in	Indonesia	0	0	0	https://twitte	9.42E+17	aadbitcoin	
6	1.79E+18	Thu May 02 04:02:2	0	@timothyronald2	1.79E+18	https://pbs.tv	timothyronald22	in		0	0	0	https://twitte	1.72E+18	ember67765	
7	1.79E+18	Thu May 02 03:55:2	0	@sipalabotak @Fr	1.79E+18		sipalabotak	in	Earth	0	0	0	https://twitte	1.43E+18	ismailcbbtc	
8	1.79E+18	Thu May 02 03:20:2	23	Willy Woo analis c	1.79E+18	https://pbs.twimg.com/media/GMixCb4a		in		0	3	0	https://twitte	1.67E+18	akademicryptoid	
9	1.79E+18	Thu May 02 03:19:2	0	Ini mah beneran d	1.79E+18	https://pbs.twimg.com/media/GMiw0Xu		in	Indonesia	0	2	0	https://twitte	488566009	sinyopermata	
10	1.79E+18	Thu May 02 03:17:2	8	MicroStrategy Cip	1.79E+18	https://pbs.twimg.com/media/GMivhlHb		in	Desolate #328	0	1	3	https://twitte	70912579	bukanpamanmu	
11	1.79E+18	Thu May 02 03:16:0	0	@andricuanterus	1.79E+18		andricuanterus	in		0	0	0	https://twitte	1.76E+18	CuanDelan61328	
12	1.79E+18	Thu May 02 03:10:2	2	Lantas apakah fen	1.79E+18		akademicryptoid	in		0	0	0	https://twitte	1.67E+18	akademicryptoid	
13	1.79E+18	Thu May 02 03:10:0	0	Hal ini justru mem	1.79E+18		akademicryptoid	in		0	1	0	https://twitte	1.67E+18	akademicryptoid	
14	1.79E+18	Thu May 02 03:10:0	18	BlackRock perusah	1.79E+18	https://pbs.twimg.com/media/GMiu2Pxa		in		0	1	0	https://twitte	1.67E+18	akademicryptoid	
15	1.79E+18	Thu May 02 03:08:5	0	@boooveer @halo	1.79E+18		boooveer	in	Yogyakarta, Ind	0	0	0	https://twitte	1.12E+18	SultanAmirr	
16	1.79E+18	Thu May 02 03:08:0	26	GM ini alasan saya	1.79E+18	https://pbs.twimg.com/media/GMiuQPjb		in		0	5	0	https://twitte	1.45E+18	andricuanterus	
17	1.79E+18	Thu May 02 03:01:0	0	@timothyronald2	1.79E+18		timothyronald22	in		0	0	0	https://twitte	1.76E+18	soparIGTG	
18	1.79E+18	Thu May 02 02:59:2	0	@meinmokhtar Be	1.79E+18		meinmokhtar	in	Malaysia	0	0	0	https://twitte	1.77E+18	YrDude_	
19	1.79E+18	Thu May 02 02:58:2	0	bitcoin scam. jgn t	1.79E+18			in		0	0	0	https://twitte	1731723541	0xruquers	
20	1.79E+18	Thu May 02 02:56:2	0	@SultanAmirr @h	1.79E+18		SultanAmirr	in		0	1	0	https://twitte	1.72E+18	boooveer	
21	1.79E+18	Thu May 02 02:49:0	0	@anggaandinata k	1.79E+18		anggaandinata	in	Earth	0	0	0	https://twitte	1.43E+18	ismailcbbtc	
22	1.79E+18	Thu May 02 02:44:2	0	Gak ngerti lagi am	1.79E+18	https://pbs.twimg.com/media/GMioyG9b		in	Absurdistan	0	0	3	https://twitte	99705510	BNGPY	
23	1.79E+18	Thu May 02 02:44:0	0	@AnalisaCrypto K	1.79E+18		AnalisaCrypto	in		0	0	0	https://twitte	1.63E+18	Rizkybitcoin	
24	1.79E+18	Thu May 02 02:41:2	4	2 Mei 2024: Poten	1.79E+18			in	Greater Jakarta	1	3	4	https://twitte	41605719	azwardiiqbal	
25	1.79E+18	Thu May 02 02:08:5	1	Makasi banyak bit	1.79E+18	https://pbs.twimg.com/media/GMigtO6a		in		0	0	0	https://twitte	66882875	Aguswidiartha	
26	1.79E+18	Thu May 02 02:04:2	2	Ceramah diatas di	1.79E+18		aadbitcoin	in	Indonesia	0	0	0	https://twitte	9.42E+17	aadbitcoin	
27	1.79E+18	Thu May 02 02:04:2	4	Jadi menurut perh	1.79E+18	https://pbs.tv	aadbitcoin	in	Indonesia	0	2	0	https://twitte	9.42E+17	aadbitcoin	
28	1.79E+18	Thu May 02 02:04:2	35	Bitcoin adalah unt	1.79E+18	https://pbs.twimg.com/ext_tw_video_th		in	Indonesia	0	1	9	https://twitte	9.42E+17	aadbitcoin	
29	1.79E+18	Thu May 02 01:57:2	0	#Bitcoin update C	1.79E+18	https://pbs.twimg.com/media/GMic3s3b		in	Indonesia	0	0	0	https://twitte	1.01E+18	CryptoTalk_007	
30	1.79E+18	Thu May 02 01:57:0	0	B U Y T H E D I P #B	1.79E+18			in	Sri Lanka	0	0	0	https://twitte	1.65E+18	tharusha908	
31	1.79E+18	Thu May 02 01:46:2	0	BTC $250k MINIMU	1.79E+18			in	onchain	0	0	0	https://twitte	1.74E+18	TwelveGrass	
32	1.79E+18	Thu May 02 01:34:5	1	BITCOIN BATUK DA	1.79E+18	https://pbs.twimg.com/media/GMiY58wa		in	City Hunter	0	2	0	https://twitte	1.73E+18	JeckyJeck399826	
33	1.79E+18	Thu May 02 01:32:2	0	@ajaib_investasi (	1.79E+18		ajaib_investasi	in		1	1	0	https://twitte	1.51E+18	satriaharjati	
34	1.79E+18	Thu May 02 01:25:0	0	APABILA HANYA B	1.79E+18	https://pbs.twimg.com/media/GMiWpMa		in		0	0	0	https://twitte	1.38E+18	AryaArynt	
35	1.79E+18	Thu May 02 01:22:0	1	Bitcoin Ekosistem	1.79E+18			in	Blockchain	0	1	0	https://twitte	3255392544	mwisbc	
36	1.78E+18	Thu May 02 01:19:2	0	@Kalimasada97 Be	1.79E+18		Kalimasada97	in		0	0	0	https://twitte	1.72E+18	FacebookGa74666	
37	1.79E+18	Thu May 02 01:18:2	0	BlackRock menjua	1.79E+18			in		0	0	0	https://twitte	1.66E+18	PutuCryptos	

Gambar 3. 1 *Hasil Crawling*

Untuk mempermudah analisis data pada tahap selanjutnya, hanya atribut *full_text* yang digunakan. Berikut adalah Tabel 3.1 yang menampilkan atribut dataset yang digunakan dalam penelitian ini.

Tabel 3. 1 Dataset

No	***Full_text***

1	Penambang Bitcoin @RiotPlatforms melaporkan suku rekod dengan pendapatan bersih sebanyak $211.8 juta tetapi tidak mencapai anggaran pendapatan analis. https://t.co/VW3jSxLusT
2	Kabar buruk buat para pecinta kripto! Harga Bitcoin merosot karena prospek suku bunga yang tetap tinggi. Koreksi bulanan terdalam sejak kolapsnya FTX membuat aset digital turun sekitar 16 persen di bulan April 2024. https://t.co/eTrWAZl9JH
3	@TLD_Survive kalau saya dikasih 100juta dalam kondisi ini uang nganggur ya. saya all in beli bitcoin saat sentiment lagi fear mungkin beberapa bulan kedepan saat harga sedang dibottomnya tunggu 2025 atau 2027 and earn money from that
4	@rdmcryptoguy @crypto_radiz @yanzero_ Berarti sekilas aja smart contract bug risk ama liquidity risk aja palingan.. Risk pasti ga 0% tp mitigasi resiko mestinya bisa.. Udh pas lah bro @yanzero_ lah yg lbh paham masalah defi ini dbanding aye..
5	@timothyronald22 GN akibat ikutin borong bitcoin konsisten https://t.co/j4pAyYrFZr
…..	…..
557	Sebagian orang menanti dan merayakan Bitcoin Halving seperti turunnya Nabi Isa. *random rant udah lama nyimpen tweet ini
558	Kalau tahun ini bakalan keluar ntar bulan Januari klw habis itu uang disetor ke rdo tunggu bitcoin hancur luluh lantah dan bearish lalu beli bitcoin tak lupa sebagian uangnya modal untuk buka usaha dan buat pengembangan kompetensi diri sendiri

3.2 *Labeling*

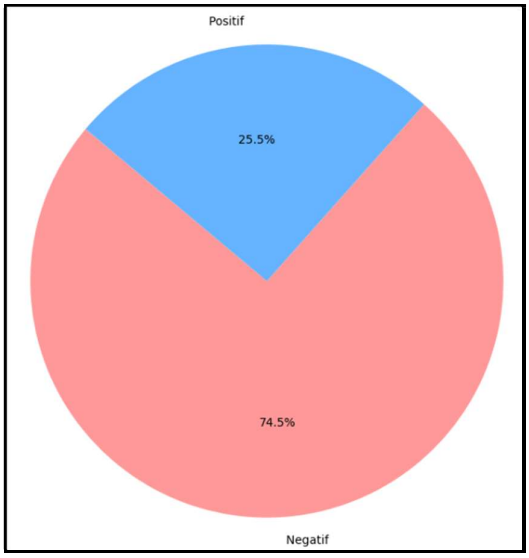
Dataset kemudian akan melalui tahap pemberian label, pelabelan sebanyak 538 data dilakukan secara manual oleh *expert* di bidang bahasa dengan mengkategorikan tweet menjadi dua kategori yaitu positif dan negatif berdasarkan sentimen. Tabel 3.2 merupakan hasil text yang telah diberikan label sebagai berikut :

Tabel 3. 2 *Labeling Dataset*

No	***Full_text***	***Sentimen***
1	Penambang Bitcoin @RiotPlatforms melaporkan suku rekod dengan pendapatan bersih sebanyak $211.8 juta tetapi tidak mencapai anggaran pendapatan analis. https://t.co/VW3jSxLusT	Negatif
2	Kabar buruk buat para pecinta kripto! Harga Bitcoin merosot karena prospek suku bunga yang tetap tinggi. Koreksi bulanan terdalam sejak kolapsnya FTX membuat aset digital turun sekitar 16 persen di bulan April 2024. https://t.co/eTrWAZl9JH	Negatif
3	@TLD_Survive kalau saya dikasih 100juta dalam kondisi ini uang nganggur ya. saya all in beli bitcoin saat sentiment lagi fear mungkin beberapa bulan kedepan saat harga sedang dibottomnya tunggu 2025 atau 2027 and earn money from that	Positif

No	*Full_text*	*Sentimen*
4	@timothyronald22 GN akibat ikutin borong bitcoin konsisten https://t.co/j4pAyYrFZr	Positif
5	bitcoin scam. jgn beli. ponzi	Negatif
…..	…..	…..
557	Sebagian orang menanti dan merayakan Bitcoin Halving seperti turunnya Nabi Isa. *random rant udah lama nyimpen tweet ini	Negatif
558	Kalau tahun ini bakalan keluar ntar bulan Januari klw habis itu uang disetor ke rdo tunggu bitcoin hancur luluh lantah dan bearish lalu beli bitcoin tak lupa sebagian uangnya modal untuk buka usaha dan buat pengembangan kompetensi diri sendiri	Negatif

Dari total dataset yang telah diberikan label, terdapat 25.5% *tweet* yang memiliki sentimen positif, sedangkan 74.5% *tweet* lainnya memiliki sentimen negatif. Visualisasi distribusi data sentimen dapat dilihat pada gambar 3.2.



Gambar 3. 2 *Visualisasi Persentase Sentimen*

Setelah dataset terlabeli, maka proses selanjutnya adalah preprocessing, dimana proses preprocessing tersebut terbagi menjadi beberapa tahap, yaitu case folding, cleansing, tokenzing, stopword removal, stemming dan delete duplicates.

3.3 *Pre-processing*

Proses pembersihan atau *preprocessing* bertujuan untuk membersihkan data. Pada tahapan ini, data hasil *crawling* akan diolah dengan memilih data yang relevan untuk digunakan dan menghapus data yang tidak berguna, sehingga data menjadi lebih terstruktur dan siap untuk diproses lebih lanjut Tahap preprocessing terdiri dari *case folding*, *cleansing*, *tokenizing*, *stopword removal*, *stemming* dan *delete duplicates* untuk menghapus dataset yang muncul dua kali atau lebih. Berikut hasil teks setelah *preprocessing* sebagai berikut:

a. *Case folding*

berdasarkan pada tabel 3.3, semua huruf kapital dalam data tweet telah diubah menjadi huruf kecil. Perubahan ini dilakukan untuk mempermudah proses pembacaan oleh mesin terhadap korpus dan mengurangi waktu yang dibutuhkan.

Tabel 3. 3 *Hasil case folding*

No	***Case Folding***
1	penambang bitcoin @riotplatforms melaporkan suku rekod dengan pendapatan bersih sebanyak $211.8 juta tetapi tidak mencapai anggaran pendapatan analis. https://t.co/vw3jsxlust
2	kabar buruk buat para pecinta kripto! harga bitcoin merosot karena prospek suku bunga yang tetap tinggi. koreksi bulanan terdalam sejak kolapsnya ftx membuat aset digital turun sekitar 16 persen di bulan april 2024. https://t.co/etrwazl9jh
3	@tld_survive kalau saya dikasih 100juta dalam kondisi ini uang nganggur ya. saya all in beli bitcoin saat sentiment lagi fear mungkin beberapa bulan kedepan saat harga sedang dibottomnya tunggu 2025 atau 2027 and earn money from that
4	@rdmcryptoguy @crypto_radiz @yanzero_ berarti sekilas aja smart contract bug risk ama liquidity risk aja palingan.. risk pasti ga 0% tp mitigasi resiko mestinya bisa.. udh pas lah bro @yanzero_ lah yg lbh paham masalah defi ini dbanding aye..
5	@timothyronald22 gn akibat ikutin borong bitcoin konsisten https://t.co/j4payyrfzr
…..	……
557	sebagian orang menanti dan merayakan bitcoin halving seperti turunnya nabi isa. *random rant udah lama nyimpen tweet ini
558	kalau tahun ini bakalan keluar ntar bulan januari klw habis itu uang disetor ke rdo tunggu bitcoin hancur luluh lantah dan bearish lalu beli bitcoin tak lupa sebagian uangnya modal untuk buka usaha dan buat pengembangan kompetensi diri sendiri

b. *Cleansing*

Cleansing telah dilakukan pada Tabel 3.4, Cleansing merupakan penghapusan karakter yang tidak diperlukan dari teks. Langkah ini dilakukan untuk membersihkan teks dari karakter-karakter yang tidak relevan dan dapat mengganggu proses analisis.

Tabel 3. 4 *Hasil Cleansing*

No	*Cleansing*
1	penambang bitcoin melaporkan suku rekod dengan pendapatan bersih sebanyak juta tetapi tidak mencapai anggaran pendapatan analis
2	kabar buruk buat para pecinta kripto harga bitcoin merosot karena prospek suku bunga yang tetap tinggi koreksi bulanan terdalam sejak kolapsnya ftx membuat aset digital turun sekitar persen di bulan april
3	kalau saya dikasih juta dalam kondisi ini uang nganggur ya saya all in beli bitcoin saat sentiment lagi fear mungkin beberapa bulan kedepan saat harga sedang dibottomnya tunggu atau and earn money from that
4	berarti sekilas aja smart contract bug risk ama liquidity risk aja palingan risk pasti ga tp mitigasi resiko mestinya bisa udh pas lah bro lah yg lbh paham masalah defi ini dbanding aye
5	gn akibat ikutin borong bitcoin konsisten
…..	…..
557	sebagian orang menanti dan merayakan bitcoin halving seperti turunnya nabi isa random rant udah lama nyimpen tweet ini
558	kalau tahun ini bakalan keluar ntar bulan januari klw habis itu uang disetor ke rdo tunggu bitcoin hancur luluh lantah dan bearish lalu beli bitcoin tak lupa sebagian uangnya modal untuk buka usaha dan buat pengembangan kompetensi diri sendiri

c. *Tokenizing*

Tokenizing telah dilakukan seperti pada Tabel 3.5 Data tweet yang awalnya berbentuk kalimat telah dipecah menjadi kata-kata individu. Proses *tokenizing* dilakukan untuk memudahkan tahap transformasi sehingga pemrosesan dilakukan berdasarkan kata demi kata, bukan berdasarkan kalimat.

Tabel 3. 5 *Hasil Tokenizing*

No	*Tokenizing*
1	['penambang', 'bitcoin', 'melaporkan', 'suku', 'rekod', 'dengan', 'pendapatan', 'bersih', 'sebanyak', 'juta', 'tetapi', 'tidak', 'mencapai', 'anggaran', 'pendapatan', 'analis']

No	*Tokenizing*
2	['kabar', 'buruk', 'buat', 'para', 'pecinta', 'kripto', 'harga', 'bitcoin', 'merosot', 'karena', 'prospek', 'suku', 'bunga', 'yang', 'tetap', 'tinggi', 'koreksi', 'bulanan', 'terdalam', 'sejak', 'kolapsnya', 'ftx', 'membuat', 'aset', 'digital', 'turun', 'sekitar', 'persen', 'di', 'bulan', 'april']
3	['kalau', 'saya', 'dikasih', 'juta', 'dalam', 'kondisi', 'ini', 'uang', 'nganggur', 'ya', 'saya', 'all', 'in', 'beli', 'bitcoin', 'saat', 'sentiment', 'lagi', 'fear', 'mungkin', 'beberapa', 'bulan', 'kedepan', 'saat', 'harga', 'sedang', 'dibottomnya', 'tunggu', 'atau', 'and', 'earn', 'money', 'from', 'that']
4	['berarti', 'sekilas', 'aja', 'smart', 'contract', 'bug', 'risk', 'ama', 'liquidity', 'risk', 'aja', 'palingan', 'risk', 'pasti', 'ga', 'tp', 'mitigasi', 'resiko', 'mestinya', 'bisa', 'udh', 'pas', 'lah', 'bro', 'lah', 'yg', 'lbh', 'paham', 'masalah', 'defi', 'ini', 'dbanding', 'aye']
5	['gn', 'akibat', 'ikutin', 'borong', 'bitcoin', 'konsisten']
…..	…..
557	['sebagian', 'orang', 'menanti', 'dan', 'merayakan', 'bitcoin', 'halving', 'seperti', 'turunnya', 'nabi', 'isa', 'random', 'rant', 'udah', 'lama', 'nyimpen', 'tweet', 'ini']
558	['kalau', 'tahun', 'ini', 'bakalan', 'keluar', 'ntar', 'bulan', 'januari', 'klw', 'habis', 'itu', 'uang', 'disetor', 'ke', 'rdo', 'tunggu', 'bitcoin', 'hancur', 'luluh', 'lantah', 'dan', 'bearish', 'lalu', 'beli', 'bitcoin', 'tak', 'lupa', 'sebagian', 'uangnya', 'modal', 'untuk', 'buka', 'usaha', 'dan', 'buat', 'pengembangan', 'kompetensi', 'diri', 'sendiri']

d. ***Stopword Removal***

Kata-kata sambung seperti "yang", "dan", "di", "dari" telah dihilangkan melalui proses stopword, terlihat pada Tabel 3.6. Kata-kata tersebut tidak muncul lagi. Proses stopword bertujuan untuk menyaring kata-kata yang tidak relevan dalam klasifikasi, seperti kata penghubung "yang", "dan", "di", dan lainnya. Proses ini dilakukan dengan bantuan library Sastrawi yang tersedia dalam bahasa pemrograman Python.

Tabel 3. 6 *Hasil Stopword Removal*

No	*Stopword_removal*
1	['penambang', 'bitcoin', 'melaporkan', 'suku', 'rekod', 'pendapatan', 'bersih', 'juta', 'mencapai', 'anggaran', 'pendapatan', 'analis']
2	['kabar', 'buruk', 'pecinta', 'kripto', 'harga', 'bitcoin', 'merosot', 'prospek', 'suku', 'bunga', 'koreksi', 'bulanan', 'terdalam', 'kolapsnya', 'ftx', 'aset', 'digital', 'turun', 'persen', 'april']
3	['dikasih', 'juta', 'kondisi', 'uang', 'nganggur', 'ya', 'all', 'in', 'beli', 'bitcoin', 'sentiment', 'fear', 'kedepan', 'harga', 'dibottomnya', 'tunggu', 'and', 'earn', 'money', 'from', 'that']
4	['sekilas', 'aja', 'smart', 'contract', 'bug', 'risk', 'ama', 'liquidity', 'risk', 'aja', 'palingan', 'risk', 'ga', 'tp', 'mitigasi', 'resiko', 'mestinya', 'udh', 'pas', 'bro', 'yg', 'lbh', 'paham', 'defi', 'dbanding', 'aye']
5	['gn', 'akibat', 'ikutin', 'borong', 'bitcoin', 'konsisten']

No	*Stopword_removal*
…..	…..
557	['orang', 'merayakan', 'bitcoin', 'halving', 'turunnya', 'nabi', 'isa', 'random', 'rant', 'udah', 'nyimpen', 'tweet']
558	['ntar', 'januari', 'klw', 'habis', 'uang', 'disetor', 'rdo', 'tunggu', 'bitcoin', 'hancur', 'luluh', 'lantah', 'bearish', 'beli', 'bitcoin', 'lupa', 'uangnya', 'modal', 'buka', 'usaha', 'pengembangan', 'kompetensi']

e. *Stemming*

Tahap selanjutnya dalam *preprocessing* adalah *stemming*. Proses ini melibatkan pengubahan kata-kata menjadi bentuk dasarnya dengan menghilangkan imbuhan awalan, sisipan, akhiran. Tujuannya adalah untuk menyeragamkan bentuk kata sehingga variasi kata yang memiliki makna serupa dapat dikenali sebagai entitas yang sama. Berikut adalah tabel 3.7 hasil text setelah *stemming*:

Tabel 3. 7 *Hasil Stemming*

No	*Stemming*
1	tambang bitcoin lapor suku rekod dapat bersih juta capai anggar dapat analis
2	kabar buruk cinta kripto harga bitcoin merosot prospek suku bunga koreksi bulan dalam kolaps ftx aset digital turun persen april
3	kasih juta kondisi uang nganggur ya all in beli bitcoin sentiment fear depan harga dibottomnya tunggu and earn money from that
4	kilas aja smart contract bug risk ama liquidity risk aja paling risk ga tp mitigasi resiko mesti udh pas bro yg lbh paham defi dbanding aye
5	gn akibat ikutin borong bitcoin konsisten
…..	…..
557	orang raya bitcoin halving turun nabi isa random rant udah nyimpen tweet
558	ntar januari klw habis uang setor rdo tunggu bitcoin hancur luluh lantah bearish beli bitcoin lupa uang modal buka usaha kembang kompetensi

f. *Delete Duplicates*

Dataset yang sebelumnya berjumlah 558, setelah melewati proses delete duplicate, mengalami pengurangan sebanyak 20 baris. Penghapusan ini dilakukan karena baris-baris tersebut muncul dua kali atau lebih dalam dataset, yang dapat menyebabkan bias dalam analisis data dan penurunan performa model. Setelah proses penghapusan duplikasi, jumlah akhir

dataset menjadi 538 baris, memastikan bahwa setiap entri dalam dataset adalah unik dan representatif. Dengan demikian, dataset yang telah dibersihkan dari duplikasi ini akan memberikan dasar yang lebih kuat dan akurat untuk penelitian lebih lanjut.

DATA DUPLIKAT YANG TELAH DI HAPUS:

	full_text
62	@DappRadar Setelah halving Bitcoin tetap menja...
203	@andricuanterus Bang kan kata lu bitcoin harus...
209	@andricuanterus Bang kan kata lu bitcoin harus...
220	Halving Bitcoin Ancam Bisnis Penambang Kecil d...
228	@anggaandinata Demand untuk bitcoin semakin me...
234	Cek suara dulu Yang kemarin: - bilang kalau tu...
249	Analis Prediksi Nasib Bitcoin Usai Halving Kee...
262	Prediksi Prospek Pasar Crytocurrency Indonesia...
301	Selasa 30 April 2024 10 hari setelah halving #...
344	Halving Selesai Harga Bitcoin Akan Tetap Naik ...
367	Trader pasca halving menjadi seperti: #Bitcoin...
406	Terjemahan Persediaan Bitcoin sedang menuju go...
448	@DappRadar Setelah halving Bitcoin tetap menja...
454	Halving Bitcoin Ancam Bisnis Penambang Kecil d...
456	@anggaandinata Demand untuk bitcoin semakin me...
457	Cek suara dulu Yang kemarin: - bilang kalau tu...
458	Analis Prediksi Nasib Bitcoin Usai Halving Kee...
460	Prediksi Prospek Pasar Crytocurrency Indonesia...
463	Selasa 30 April 2024 10 hari setelah halving #...
465	Halving Selesai Harga Bitcoin Akan Tetap Naik ...

Gambar 3. 3 *Delete Duplicates*

3.4 *Wordcloud*

Seberapa sering suatu kata muncul dalam data akan menentukan ukuran kata tersebut dalam *wordcloud*. Semakin sering kata itu muncul dalam data, semakin besar pula ukurannya dalam *wordcloud*. Wordcloud yang ditampilkan memberikan visualisasi frekuensi kata dalam teks terkait peristiwa Bitcoin halving. Kata-kata seperti "turun" dan "naik" di tampilkan dengan ukuran besar, menunjukkan bahwa diskusi sering berfokus pada fluktuasi harga. Selain itu, kata-kata seperti "nih", "kripto", "harga", dan "peristiwa" sering muncul, menunjukkan topik utama yang dibahas adalah harga kripto dan peristiwa terkait. Kata-kata seperti "analisis" dan "investor" menunjukkan adanya diskusi teknis dan mendalam mengenai fenomena ini. Visualisasi *wordcloud* ditampilkan pada gambar 3.4.

Gambar 3. 4 *Visualisasi Wordcloud*

3.5 Pembobotan Kata TF-IDF

Setelah melalui tahap *preprocessing*, kata-kata perlu diubah menjadi angka agar dapat diproses oleh mesin. Hal ini memerlukan penggunaan pembobotan TF-IDF (*term frequency – inverse document frequency*), dimana nilai TF-IDF yang telah diperoleh akan ditampilkan dalam hasil perhitungan. Dan hasil TF-IDF akan tercantum dalam pada tabel 3.8, menunjukkan array position dan nilai TF-IDF untuk setiap term atau kata dalam kalimat tersebut yang digunakan untuk memberikan bobot pada kata-kata. Proses ini melibatkan perhitungan nilai *Term Frequency* (TF), *Inverse Document Frequency* (IDF), dan *Document Frequency - Inverse Document Frequency* (TF- IDF) secara berurutan.

Tabel 3. 8 *Hasil TF-IDF*

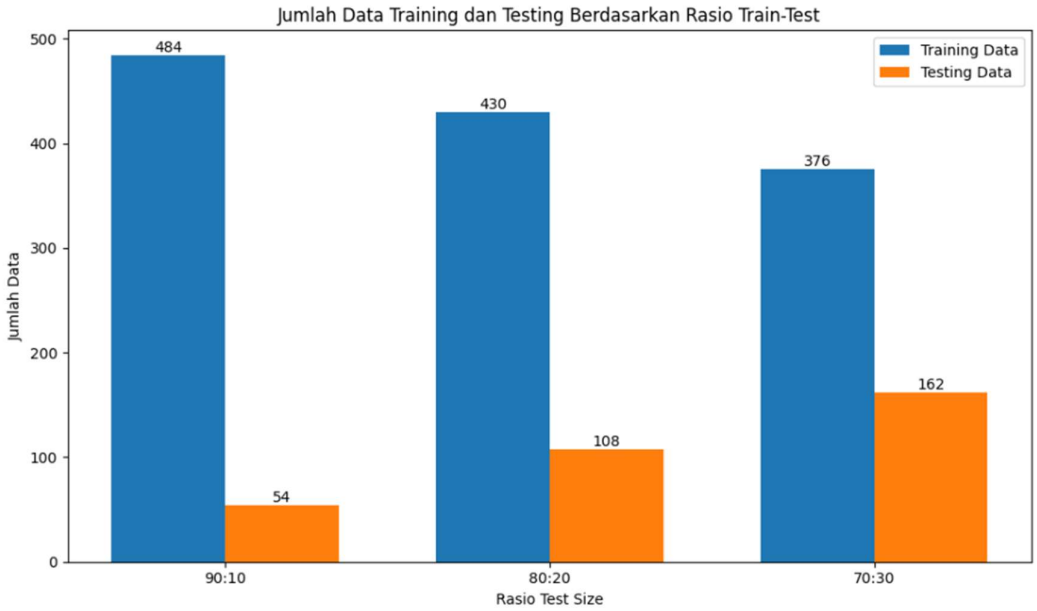
Array Position	*TF*	*IDF*	*TF-IDF*	*Term*
342	0.050000	6.633002	0.291165	Buruk
1088	0.050000	6.633002	0.291165	Kolaps
397	0.050000	6.227537	0.273367	Cinta
685	0.050000	6.227537	0.273367	Ftx
987	0.050000	6.227537	0.273367	kabar
464	0.050000	5.939855	0.260738	Dalam
1328	0.050000	5.939855	0.260738	Merosot
1686	0.050000	5.534390	0.242940	Prospek
1605	0.050000	5.380239	0.236173	Persen

332	0.050000	5.380239	0.236173	Bulan
2013	0.050000	5.023564	0.220516	Suku
337	0.050000	4.928254	0.216333	Bunga
1109	0.050000	4.687092	0.205747	Koreksi
521	0.050000	4.618099	0.202718	Digital
128	0.050000	3.892162	0.170852	Aser
116	0.050000	3.688563	0.161915	April
2185	0.050000	3.374906	0.148146	Turun
1117	0.050000	3.318816	0.145684	Kripto
803	0.050000	3.092043	0.135730	Harga
268	0.050000	1.415353	0.062129	bitcoin

Hasil perhitungan TF-IDF ini menunjukkan bobot pentingnya kata-kata dalam dokumen dataset. Array position menunjukan posisi term didalam index dokumen, term seperti "buruk","kolaps" memiliki nilai IDF tertinggi sebesar 6.633002, menunjukkan bahwa kata tersebut sangat jarang muncul dalam dokumen. Nilai TF-IDF yang tinggi, seperti pada kata "buruk" dan "kolaps" dengan nilai 0.291165, menunjukkan bahwa meskipun frekuensi kemunculannya rendah, kata tersebut sangat penting didalam dokumen. Term-term ini memiliki signifikansi tinggi dalam klasifikasi sentimen pada dokumen.

3.6 Split Data

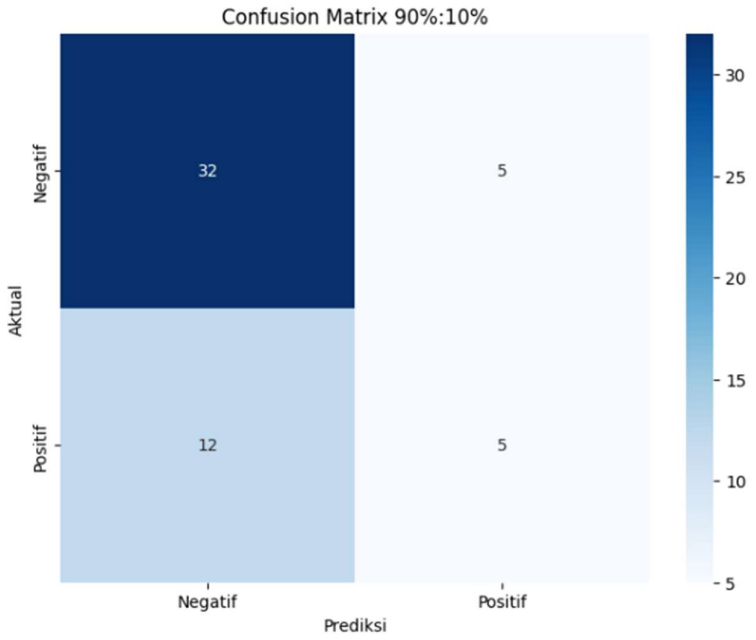
Pada penelitian ini, 538 data akan dibagi dengan beberapa rasio, yaitu 90:10,80:20, 70:20 Dari total data tersebut digunakan untuk Pembagian data dengan rasio 90:10 data training berjumlah 484, dan data testing berjumlah 54. Untuk 80:20 jumlah data training 430 dan data testing berjumlah 108 data. Terakhir untuk 70:30 jumlah data training 376 data dan 162 digunakan sebagai data testing. Visualisasi split data ditampilkan pada gambar 3.6.



Gambar 3. 5 *Split Dataset*

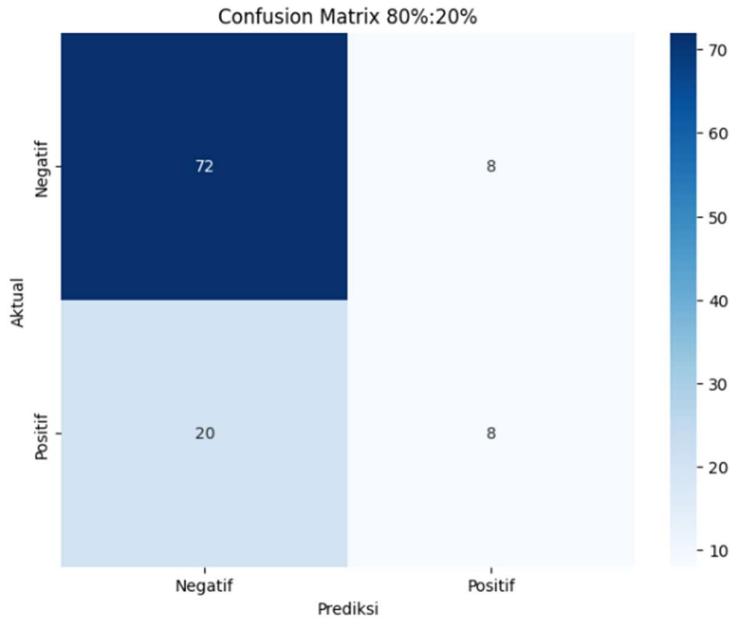
3.7 Klasifikasi

Model Naïve Bayes yang digunakan dalam penelitian ini telah diuji dengan berbagai rasio pembagian data, yaitu 90:10, 80:20, dan 70:30, untuk menentukan rasio yang paling optimal dalam klasifikasi sentimen. Berikut hasil klasifikasi ditampilkan melalui Confusion Matrix.



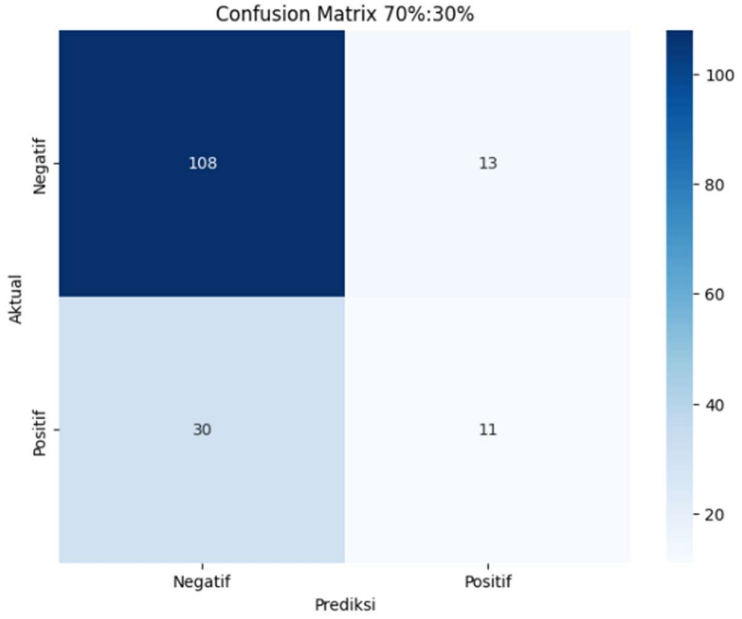
Gambar 3. 6 *Confusion Matrix Rasio 90:10*

Pada gambar 3.7, pengujian rasio 90:10, model Naive Bayes dapat mengidentifikasi 32 true negatives dan 5 true positives, namun juga terdapat 12 false negatives dan 5 false positives. Hasil tersebut menunjukkan bahwa meskipun model cenderung efektif dalam mengklasifikasikan kasus negatif, namun kemampuannya dalam mengenali kasus positif model belum maksimal.



Gambar 3. 7 *Confusion Matrix 80:20*

Pada gambar 3.8, rasio pengujian 80:20, terdapat peningkatan jumlah true positives menjadi 8, namun false negatives juga meningkat menjadi 20. Hal ini mengindikasikan bahwa dengan memperluas data uji, model mampu meningkatkan kemampuannya dalam mengklasifikasikan kasus positif dengan lebih baik daripada sebelumnya.

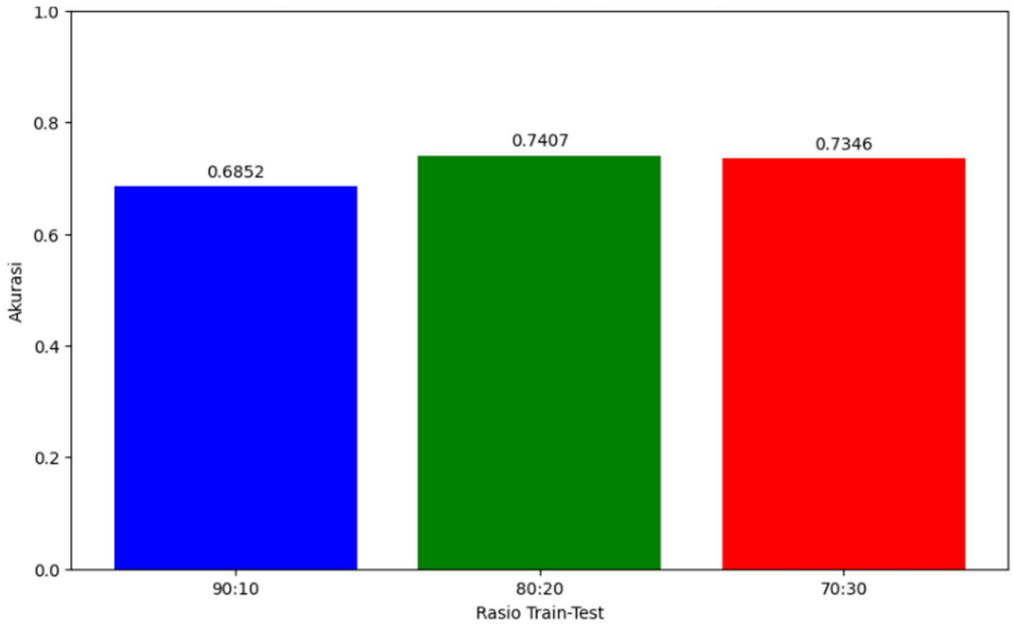


Gambar 3. 8 *Confusion Matrix Rasio 70:30*

Pada gambar 3.8, pengujian dengan rasio 70:30, model berhasil meningkatkan true positives menjadi 11 dengan false positives rendah sebanyak 13, namun terdapat peningkatan false negatives menjadi 30. Meskipun demikian, penambahan data uji menunjukkan potensi untuk meningkatkan kemampuan model dalam mengenali kasus positif dengan lebih baik.

3.8 Evaluasi

Setelah dilakukan pengujian dengan 3 model rasio yaitu 90:10,80:20,70:30 didapatkan akurasi untuk masing masing rasio sebesar 68.5% untuk pengujian menggunakan 90:10, dan 74% untuk pengujian dengan rasio 80:20, dan 73.4% untuk pengujian dengan rasio 70:30, dengan berbagai model rasio pengujian tersebut, menunjukkan bahwa model Naive Bayes yang dilatih dengan teknik teks *preprocessing* yang benar serta pembobotan TF-IDF memberikan performa yang cukup baik. Gambar 3.10 menunjukan perbandingan akurasi model dalam melakukan klasifikasi dengan rasio yang telah di tentukan:



Gambar 3. 9 *Perbandingan Accuracy Naïve Bayyes*

Hasil evaluasi menunjukkan bahwa model *Naive Bayes* dengan rasio pengujian 70:30 mencapai akurasi sebesar 0.7346. Ini berarti model berhasil mengklasifikasikan sekitar 73.46% data dengan benar. Pada rasio pengujian 80:20, model menunjukkan peningkatan performa dengan akurasi sebesar 0.7407, mengindikasikan bahwa sekitar 74.07% data diklasifikasikan dengan benar. Sementara pada rasio pengujian 90:10, akurasi model menurun menjadi 0.6852, berarti model mampu mengklasifikasikan sekitar 68.52% data dengan benar. Perbedaan akurasi ini menunjukkan bahwa performa model sedikit bervariasi tergantung pada rasio pembagian data latih dan data uji. Pada rasio 70:30 dan 80:20, model menunjukkan performa yang lebih baik dibandingkan dengan rasio 90:10. Hal tersebut dapat disebabkan oleh berbagai faktor, seperti jumlah data latih yang lebih besar pada rasio 70:30 dan 80:20, yang memungkinkan model untuk mempelajari lebih banyak pola dari data sehingga kemampuan model dalam mengklasifikasikan meningkat. Dan untuk rasio 90:10, jumlah data latih yang lebih sedikit mungkin tidak cukup untuk menangkap variasi yang ada dalam data, sehingga menyebabkan penurunan akurasi. Secara keseluruhan, hasil ini menunjukkan pemilihan rasio yang tepat dalam pembagian data latih dan data uji untuk mendapatkan performa model yang optimal.

Pemilihan rasio yang baik dapat membantu model mempelajari pola dengan lebih baik dan menghasilkan prediksi yang lebih akurat.

Berikut penghitungan model akurasi rasio 90:10 pada persamaan rumus (4.3).

$$Akurasi = \frac{TP + TN}{Total} + \frac{5 + 32}{32 + 5 + 12 + 5} = \frac{37}{54} = 0.685 \quad (4.3)$$

Berikut penghitungan model akurasi rasio 80:20 pada persamaan rumus (4.4).

$$Akurasi = \frac{TP + TN}{Total} + \frac{8 + 72}{72 + 8 + 20 + 8} = \frac{80}{108} = 0.740 \quad (4.4)$$

Berikut penghitungan model akurasi rasio 70:30 pada persamaan rumus (4.5).

$$Akurasi = \frac{TP + TN}{Total} + \frac{11 + 108}{108 + 13 + 30 + 11} = \frac{119}{162} = 0.734 \quad (4.5)$$

BAB 4
PENUTUP

4.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penggunaan algoritma *Naïve Bayes Classifier* dengan pembobotan fitur menggunakan TF-IDF (*Term Frequency – Inverse Document Frequency*) dapat mengklasifikasikan sentimen publik terhadap peristiwa Bitcoin Halving secara efektif. Dengan menggunakan data dari 538 tweet yang diperoleh melalui proses *crawling* di media sosial *Twitter*, dan setelah dilakukan berbagai tahap preprocessing serta pembagian data untuk pelatihan dan pengujian model, penelitian ini menunjukkan hasil yang signifikan. Tiga rasio data latih dan data uji yang berbeda 90:10, 80:20, dan 80:20 digunakan dalam pengujian model. Hasilnya menunjukkan bahwa model *Naïve Bayes* dengan rasio 80:20 memberikan akurasi terbaik yaitu 74% dari beberapa model rasio lainnya. Ini menunjukkan bahwa rasio data latih dan uji yang cukup beragam memberikan model untuk membaca dan mempelajari dataset dengan baik sehingga performa yang dihasilkan akurasi yang cukup optimal. Dan pada penelitian ini berhasil mendapatkan akurasi serta menunjukkan bahwa metode *text mining*, khususnya algoritma *Naïve Bayes* dengan TF-IDF, dapat diterapkan dengan baik dalam analisis sentimen publik terhadap topik yang sedang trending di media sosial. Hasil ini dapat digunakan sebagai dasar untuk penelitian lebih lanjut dalam bidang analisis sentimen serta pengembangan model yang lebih kompleks untuk meningkatkan akurasi dan performa klasifikasi.

4.2 Saran

Berdasarkan hasil penelitian, beberapa rekomendasi dapat dipertimbangkan untuk meningkatkan kinerja model klasifikasi teks peristiwa bitcoin halving di *Twitter*. Pertama, Mengatasi ketidakseimbangan kelas melalui teknik *oversampling* dan undersampling mungkin bermanfaat. Kedua, memperbesar dataset sehingga model klasifikasi dapat belajar lebih banyak lagi dalam memahami suatu pola untuk melakukan klasifikasi serta. Ketiga, Penggunaan

metode ekstraksi fitur alternatif seperti *word embeddings* bisa membantu menangkap makna kata dengan lebih baik. Keempat, Membandingkan performa dengan algoritma klasifikasi lain seperti *Gradient Boosting* atau *Random Forest* dapat memberikan wawasan baru. Kelima, Peningkatan proses preprocessing, terutama untuk bahasa informal, juga bisa menjadi area perbaikan. Memperbesar dataset dan menerapkan *cross-validation* mungkin akan meningkatkan generalisasi dan akurasi estimasi performa model. Implementasi dari saran-saran ini diharapkan dapat meningkatkan akurasi dan keseimbangan model secara signifikan.

DAFTAR PUSTAKA

Asmara, R., Ardiansyah, M. F., & Anshori, M. (2020). *Analisa Sentiment Masyarakat terhadap Pemilu 2019 berdasarkan Opini di Twitter menggunakan Metode Naive Bayes Classifier*. 193–204.

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Jurnal Media Informatika Budidarma*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937

Barus, S. G. (2022). *KLASIFIKASI SENTIMEN DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN K-NEAREST NEIGHBOR PADA ULASAN*. 162–173.

Berliani, S., & Lestari, S. (2024). Analisis Sentimen Masyarakat Terhadap Isu Pecat Sri Mulyani Pada Twitter Menggunakan Metode Naive Bayes Dan Support Vector Machine. *Jurnal Sains Dan Teknologi*, *5*(3), 951–960. https://doi.org/10.55338/saintek.v5i3.2746

Br Sinulingga, J. E., & Sitorus, H. C. K. (2024). Analisis Sentimen Opini Masyarakat terhadap Film Horor Indonesia Menggunakan Metode SVM dan TF-IDF. *Jurnal Manajemen Informatika (JAMIKA)*, *14*(1), 42–53. https://doi.org/10.34010/jamika.v14i1.11946

Fauzianto, R. A., Informatika, P. S., Yogyakarta, U. M., Network, N., Sentimen, A., Logistik, R., Forest, R., Network, N., Forest, R., Network, N., Analysis, S., Regression, L., Forest, R., Network, N., Bayes, N., & Admiration, S. (2023). *Analisis Sentimen Opini Masyarakat Terhadap Tech Winter Pada Twitter*. *4*(9), 1577–1585. https://doi.org/10.46799/jsa.v3i9.909

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Comparison of Naïve Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis. *Smatika Jurnal*, *10*(02), 71–76.

Firasari, E., Khasanah, N., Khultsum, U., Kholifah, D. N., Komarudin, R., & Widyastuty, W. (2020). Comparation of K-Nearest Neighboor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012077

Fitriyah, Z., & Kartikasari, M. D. (2023). Text Classification of Twitter Opinion Related To Permendikbud 30/2021 Using Bidirectional Lstm. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(2), 1113–1122. https://doi.org/10.30598/barekengvol17iss2pp1113-1122

Humam, C., & Laksito, A. D. (2023). Implementasi Aplikasi Sentimen Pada Data Twitter Jelang Pemilu 2024. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(2), 105–112. https://doi.org/10.30591/jpit.v8i2.5051

Imelda, I., & Arief Ramdhan Kurnianto. (2023). Naïve Bayes and TF-IDF for Sentiment

Analysis of the Covid-19 Booster Vaccine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *7*(1), 1–6. https://doi.org/10.29207/resti.v7i1.4467

Julianto, I. T., Kurniadi, D., Nashrulloh, M. R., & Mulyani, A. (2022). Twitter Social Media Sentiment Analysis Against Bitcoin Cryptocurrency Trends Using Rapidminer. *Jurnal Teknik Informatika (Jutif)*, *3*(5), 1183–1187. https://doi.org/10.20884/1.jutif.2022.3.5.289

Meynkhard, A. (2019). Fair market value of bitcoin: Halving effect. *Investment Management and Financial Innovations*, *16*(4), 72–85. https://doi.org/10.21511/imfi.16(4).2019.07

Noor Hasan, F. (2024). Analisis Sentimen Pengguna Aplikasi CapCut Pada Ulasan di Play Store Menggunakan Metode Naïve Bayes. *Media Online)*, *4*(4), 2272–2280. https://doi.org/10.30865/klik.v4i4.1555

Putri, A., Hardiana, C. S., Novfuja, E., Siregar, F. T. P., Rahmaddeni, R., Fatma, Y., & Wahyuni, R. (2023). Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 20–26. https://doi.org/10.57152/malcom.v3i1.610

Ramadhani, M. H. Z. K. (2022). The Impact of Bitcoin Halving Day on Stock Market in Indonesia. *Journal of International Conference Proceedings*, *5*(3), 127–137. https://doi.org/10.32535/jicp.v5i3.1800

Ramadhani, S., Azzahra, D., & Z, T. (2022). Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(1), 24–33. https://doi.org/10.31849/digitalzone.v13i1.9292

Ramos, D., Zanko, G., & -Bogotá, M. (2007). *BTC Halving: A Review of its Consequences in the Environment of Cryptocurrency Trading*.

Septiarini, T. W., Taufik, M. R., Afif, M., & Rukminastiti Masyrifah, A. (2020). A comparative study for Bitcoin cryptocurrency forecasting in period 2017-2019. *Journal of Physics: Conference Series*, *1511*(1). https://doi.org/10.1088/1742-6596/1511/1/012056

Srividya, K., & Mary Sowjanya, A. (2019). Aspect based sentiment analysis using POS tagging and TFIDF. *International Journal of Engineering and Advanced Technology*, *8*(6), 1960–1963. https://doi.org/10.35940/ijeat.F7935.088619

Tri Putra, K., Amin Hariyadi, M., & Crysdian, C. (2023). Perbandingan Feature Extraction Tf-Idf Dan Bow Untuk Analisis Sentimen Berbasis Svm. *Jurnal Cahaya MAndalika*, 1449.

Yuniarossy, B. A., Hindrayani, K. M., & Terza, A. (2024). *ANALISIS SENTIMEN TERHADAP ISU FEMINISME DI TWITTER MENGGUNAKAN MODEL CONVOLUTIONAL NEURAL NETWORK*. *5*(1), 477–491.

Zhafira, D. F., Rahayudi, B., & Indriati, I. (2021). Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube. *Jurnal Sistem Informasi, Teknologi Informasi, Dan Edukasi Sistem Informasi*, *2*(1), 55–63. https://doi.org/10.25126/justsi.v2i1.24

RIWAYAT HIDUP PENULIS

Andi Nur Halim atau biasa di panggil Halim, Lahir pada tanggal 14 Juli 2002 dari pasangan bapak Abrani dan Ibu Fitriah, penulis merupakan anak kedua dari 3 bersaudara, berkebangsaan Indonesia dan beragama Islam. Penulis berasal dari desa Jantur Kecamatan Muara Muntai. Beralamat di Jalan Pulau Keramat Desa Jantur kecamatan Muara Muntai Kabupaten Kutai Katanegara. Penulis Menempuh pendidikan dari SD 007 Kecamatan Muara Muntai (2008-20014), selanjutnya menempuh pendidikan sekolah menengah pertama di SMP Negeri 2 Kecamatan Muara Muntai (2014-2017), kemudian untuk pendidikan sekolah menengah atas di SMAN 1 Fillial 1 Desa Jantur (2017-2020), untuk pendidikan perguruan tinggi penulis tempuh di Universitas Muhammadiyah Kalimantan Timur (UMKT), dari 2020 sampai sekarang.

LAMPIRAN

Lampiran 1. 1 CV Expert Labelling

Irfan Abdul Hakim / +62 821-3535-7602/ hirfan825@gmail.com

SUMMARY

A young professional who is interested in social projects, corporate social responsibilities, social research, learning development, and education. Highly articulative and creative with strong interpersonal communication. Experienced in education, observing, Human resource and interviewing.

EDUCATION

Bachelor of Sociology (S. Sos) | University of Gajah Mada | **GPA:** 3.40 (out of 4.00)

Master of Arts (M. A.) | University of Gadjah Mada| GPA: 3.50 (out of 4.00)

WORK EXPERIENCE

OFFICE STAFF

Panitia Pengawas Pemilu (Panwaslu) Kecamatan Semboro | 2022

- Report election violations
- Documenting administration report about election violations

TEACHER

MTs Ali Maksum Yogyakarta | 2021

SMA Nurul Muslim Batealit Jepara| 2023

- Collecting and analyzing data, providing statistical reports, and interview
- Microteaching and observation
- Designed effective teaching tools for learning development

RESEARCH ASSISTANT

Pusat Studi Pancasila Universitas Pembangunan Nasional Yogyakarta | 2019

Collecting and analyzing data, providing statistical reports, and interview

PROJECT EXPERIENCE

MODERATOR

National Seminar Entrepreneur of PMII Hasyim Asy'arie UNY| 2020

LOGISTIC

Acceptance of new students MA Ali Maksum Krapyak Yogyakarta | 2020

COORDINATOR ENUMERATOR

Customer Satisfaction Survey in PDAM Yogyakarta | 2020

FINANCE MANAGER

Children's Party | 2020

LOGISTIC

Farewell Party of Sosiology | 2020

Irfan Abdul Hakim / +62 821-3535-7602/ hirfan825@gmail.com

PROFESSIONAL SKILL

RESEARCH SKILL

- Designed effective teaching tools
- Microteaching
- observation
- interview
- statistical analysis

PERSONAL TRAIT

- Creative
- Fast Learner
- Team Player
- Highly Motivated
- Adaptive

LANGUAGE

- Indonesian | **Native**
- English | **Professional Working Proficiency**

AWARD

- Finalist essay's competition of Jala PRT| **2022**

RESEARCH& PUBLICATION

- Strategi Dakwah Komunitas Arus Informasi Santri Nusantara| **2020**

ORGANIZATION

- Gerakan Mahasiswa Satu Bangsa (GEMASABA) Kab. Sleman | Vice Chairman | **2021 – 2022,**
- Dormitory Administrator of MTs Ali Maksum Pondok Pesantren Krapyak Yogyakarta | Chief | **2015-2022**
- Ikatan Alumni MA Ali Maksum Yogyakarta | Chief | **2017**
- Pergerakan Mahasiswa Islam Indonesia (PMII) Gadjah Mada | Manager of Caderitation |

Lampiran 2. 1 Code Crawling Twitter

```
#@title Twitter Auth Token

twitter_auth_token = '**********' # change this auth
token
```

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key
| sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg
```

```
!NODE_MAJOR=20 && echo "deb [signed-
by=/etc/apt/keyrings/nodesource.gpg]
https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo
tee /etc/apt/sources.list.d/nodesource.list

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

```
# Crawl Data

filename = 'dataset.csv'
search_keyword = 'Bitcoin Halving lang:id'
limit = 600

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --
tab "LATEST" -l {limit} --token {twitter_auth_token}
```

```
import pandas as pd

# Specify the path to your CSV file
file_path = f"tweets-data/{filename}"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path, delimiter=",")

# Display the DataFrame
display(df)
```

```
# Cek jumlah data yang didapatkan

num_tweets = len(df)
print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")
```

Lampiran 2. 2 Import Library & Pip Install

```
!pip install Sastrawi
import pandas as pd
import numpy as np
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from sklearn.feature_extraction.text import TfidfVectorizer
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
```

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score, classification_report


# Download data NLTK
nltk.download('punkt')
nltk.download('stopwords')
```

Lampiran 2. 3 Read Dataset & Count Sentiment

```
df = pd.read_csv('dataset_uji.csv')

# jumlah sentimen
sentimen_counts = df['Sentimen'].value_counts()
print(sentimen_counts)

for sentimen, count in sentimen_counts.items():
    print(f"Jumlah Sentimen {sentimen}: {count}")
```

Lampiran 2. 4 Preprocessing

```
# Load dataset
file_path = 'dataset_uji.csv'
data = pd.read_csv(file_path)

# Mengganti NaN dengan string kosong
data['full_text'] = data['full_text'].fillna('')

# Inisialisasi stemmer untuk Bahasa Indonesia
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Fungsi untuk case folding
def case_folding(text):
    return text.lower()

# Fungsi untuk cleansing
def cleansing(text):
    text = re.sub(r'http\S+', '', text)
    text = re.sub(r'@\w+|#\w+', '', text)
    text = re.sub(r'[^a-z\s]', '', text)
    return text

# Fungsi untuk tokenizing
def tokenizing(text):
    return word_tokenize(text)
```

```
# Fungsi untuk stopword removal
def stopword_removal(tokens):
    stop_words = set(stopwords.words('indonesian'))
    return [word for word in tokens if word not in stop_words]

# Fungsi untuk stemming
def stemming(tokens):
    return ' '.join([stemmer.stem(word) for word in tokens])

# Preprocessing teks
data['case_folding'] = data['full_text'].apply(case_folding)
data['cleansing'] = data['case_folding'].apply(cleansing)
data['tokenizing'] = data['cleansing'].apply(tokenizing)
data['stopword_removal'] = data['tokenizing'].apply(stopword_removal)
data['stemming'] = data['stopword_removal'].apply(stemming)

# Tampilkan hasil preprocessing
data[['full_text', 'case_folding', 'cleansing', 'tokenizing',
'stopword_removal', 'stemming', 'Sentimen']].head()
```

Lampiran 2. 5 Code Untuk Menyimpan Hasil Teks Preprocessing

```
output_path = 'preprocessed_text.csv'
columns_to_save = ['full_text', 'case_folding', 'cleansing',
'tokenizing', 'stopword_removal', 'stemming', 'Sentimen']
data[columns_to_save].to_csv(output_path, index=False)
print(f"Data hasil preprocessing telah disimpan di: {output_path}")

# Menampilkan hasil preprocessing
data[['full_text', 'case_folding', 'cleansing', 'tokenizing',
'stopword_removal', 'stemming', 'Sentimen']].head()
```

Lampiran 2. 6 Code Delete Duplicate

```
df_preprocessed = pd.read_csv('preprocessed_text.csv')

df_preprocessed['stemming'] = df_preprocessed['stemming'].fillna(' ')
df_preprocessed.dropna(subset=['stemming'], inplace=True)
df_preprocessed.drop_duplicates(subset=['stemming'], inplace=True)

# Menyimpan data yang telah dibersihkan ke dalam file baru
df_preprocessed.to_csv('cleaned_text_no_duplicates.csv', index=False)

# Menampilkan data yang telah dihapus duplikatnya
print("Data setelah duplikat dihapus:")
```

```
print(df_preprocessed)
```

Lampiran 2. 7 Code Visualisasi Persentase Sentimen

```
df = pd.read_csv('cleaned_text_no_duplicates.csv')
sentimen_counts = df['Sentimen'].value_counts()

# pie chart
plt.figure(figsize=(8, 8))
plt.pie(sentimen_counts, labels=sentimen_counts.index,
autopct='%1.1f%%', startangle=140, colors=['#ff9999','#66b3ff'])
plt.axis('equal')
plt.show()
```

Lampiran 2. 8 Code Wordcloud Sebelum Teks Preprocessing

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud
df = pd.read_csv('dataset.csv')

text = " ".join(review for review in df['full_text'])
wordcloud = WordCloud(width=800, height=400, background_color='white',
stopwords={'bitcoin', 'halving'}).generate(text)

# Menampilkan WordCloud
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

Lampiran 2. 9 Code Wordcloud Setelah Teks Preprocessing

```
df = pd.read_csv('cleaned_text_no_duplicates.csv')

bitcoin_halving_texts = '
'.join(df[df['stemming'].str.contains('bitcoin halving',
na=False)]['stemming'].dropna())

# Define stopwords
stopwords = set(STOPWORDS)
stopwords.update(['bitcoin', 'halving', 'bitcoin halving'])
wordcloud = WordCloud(width=800, height=400, background_color='white',
stopwords=stopwords).generate(bitcoin_halving_texts)

# Display the WordCloud
```

```
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

Lampiran 2. 10 TF-IDF (Term Frequency – Inverse Document Frequency)

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer

df = pd.read_csv('cleaned_text_duplicates.csv')
tfidf_vectorizer = TfidfVectorizer()

tfidf_matrix = tfidf_vectorizer.fit_transform(df['stemming'])

sample_index = 1
sample_text = df['stemming'].iloc[sample_index]
tfidf_sample_matrix = tfidf_vectorizer.transform([sample_text])

terms = tfidf_vectorizer.get_feature_names_out()
tfidf_values = tfidf_sample_matrix.toarray()[0]

idf_values = tfidf_vectorizer.idf_
tf_values = (tfidf_sample_matrix > 0).astype(int).toarray()[0] /
len(sample_text.split())
tfidf_df = pd.DataFrame({
    'TF': tf_values,
    'IDF': idf_values,
    'TF-IDF': tfidf_values,
    'Term': terms
})

tfidf_df = tfidf_df[tfidf_df['TF-IDF'] > 0]
tfidf_df = tfidf_df.sort_values(by='TF-IDF', ascending=False)

print(f"Show TFIDF sample ke-{sample_index}")
print(tfidf_df)
print(f"\nTFIDF Table for Sample Index {sample_index}:\n")
print(f"{'Array Position':<20}{'TF':<10}{'IDF':<10}{'TF-
IDF':<10}{'Term':<20}")
print("="*60)
for i, row in tfidf_df.iterrows():
    print(f"{i:<20}{row['TF']:<10.6f}{row['IDF']:<10.6f}{row['TF-
IDF']:<10.6f}{row['Term']:<20}")
```

Lampiran 2. 11 Naive Bayes Classification

```
def evaluate_naive_bayes(df, test_size):
    X = df['stemming']
    y = df['Sentimen']
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=test_size, random_state=42)

    vectorizer = TfidfVectorizer(max_features=1000, min_df=5,
max_df=0.7)
    X_train_tfidf = vectorizer.fit_transform(X_train)
    X_test_tfidf = vectorizer.transform(X_test)

    naive_bayes = MultinomialNB(alpha=0.1)
    naive_bayes.fit(X_train_tfidf, y_train)

    y_pred = naive_bayes.predict(X_test_tfidf)

    accuracy = accuracy_score(y_test, y_pred)
    print(f'Naive Bayyes Accuracy {1-test_size:.0%}:{test_size:.0%}:
{accuracy:.4f}')
    print(classification_report(y_test, y_pred, zero_division=1))

ratios = [0.1, 0.2, 0.3]
accuracies = []

for ratio in ratios:
    accuracy = evaluate_naive_bayes(df, test_size=ratio)
    accuracies.append(accuracy)
```

Lampiran 2. 12 Code Confusion Matrix

```
 # Hitung confusion matrix
    cm = confusion_matrix(y_test, y_pred)

    # Plot confusion matrix
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
xticklabels=['Negatif', 'Positif'], yticklabels=['Negatif',
'Positif'])
    plt.xlabel('Prediksi')
    plt.ylabel('Aktual')
    plt.title(f'Confusion Matrix {1-test_size:.0%}:{test_size:.0%}')
    plt.show()
```

Lampiran 3. 1 Kartu Kendali Bimbingan

KARTU KENDALI BIMBINGAN LAPORAN KARYA ILMIAH

Nama	:	Andi Nur Halim
NIM	:	2011102441038
Nama Dosen Pembimbing	:	Rudiman, S.Kom., M.Sc
Judul Penelitian	:	Analisis Sentimen Opini Publik Terhadap Peristiwa Bitcoin Halving Pada Data Teks Twitter Menggunakan Metode Naïve Bayes Dan Pembobotan Fitur TF-IDF

No	Tanggal	Uraian Pembimbingan	Paraf Dosen
1	7/2/2024	Persetujuan bimbingan dengan dosen	
2	14/2/2024	Mencari topik permasalahan yang akan digunakan sebagai objek penelitian.	
3	22/2/2024	Evaluasi objek penelitian	
4	29/2/2024	Menentukan Judul penelitian dan Latar belakang	
5	9/3/2024	melakukan penulisan Latar belakang masalah sebuah Judul dan arahan dosen	
6	13/03/2024	Revisi penulisan Latar belakang	
7	18/3/2024	Memberikan arahan dalam penulisan canvas pengajuan Judul.	
8	27/3/2024	Memperbaiki revisi dan saran di bab 1-2	
9	5/4/2024	Revisi penulisan bab 2	
10	29/4/2024	memulai penulisan bab 3 dan membuat code TF-IDF sesuai arahan	
11	26/5/2024	dosen pembimbing memberikan masukan di bab 3	
12	17/07/2024	Memberikan revisi menyesuai Jurnal. dan Skripsi	
13	30/07/224	Melakukan revisi Jurnal yang akan di submit oleh dosen pembimbing	

Dosen Pembimbing

Rudiman, S.Kom., M.Sc
NIDN. 1105068202

Mengetahui
Ketua Program Studi

Arbansyah, S.Kom., M.Ti
NIDN. 1119019203

SKRIPSI ANDI NUR HALIM

by Teknik Informatika UMKT

Submission date: 25-Jul-2024 09:24AM (UTC+0800)
Submission ID: 2422037094
File name: SKRIPSI_ANDI_NUR_HALIM.docx (1.03M)
Word count: 6837
Character count: 44291

SKRIPSI ANDI NUR HALIM

ORIGINALITY REPORT

19%	16%	11%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ojs.cahayamandalika.com Internet Source	1%
2	Submitted to Universitas Negeri Semarang - iTh Student Paper	1%
3	ejournal.itn.ac.id Internet Source	1%
4	jurnalsyntaxadmiration.com Internet Source	1%
5	www.ejurnal.stmik-budidarma.ac.id Internet Source	1%
6	dspace.umkt.ac.id Internet Source	1%
7	Almi yulistia Alwanda, Ema Utami, Ainul Yaqin. "Analisis Klasifikasi Konsentrasi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor", Infotek: Jurnal Informatika dan Teknologi, 2024 Publication	1%