

BAB III METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian yang digunakan berupa data tweet yang berkaitan dengan Bitcoin pada media sosial X (Twitter) yang ditunjukkan pada gambar 3.1 dan 3.2.



Gambar 3. 1 Contoh Tweet Positif

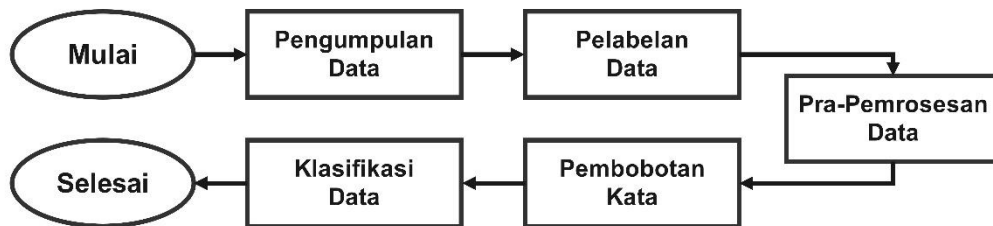


Gambar 3. 2 Contoh Tweet Negatif

Gambar di atas menampilkan contoh sentimen positif dan negatif dari pengguna Twitter terkait Bitcoin. Contoh positif (Gambar 3.1) berisi kalimat “Bitcoin keeps breaking resistance levels” yang menunjukkan optimisme terhadap kenaikan harga Bitcoin melalui frasa seperti *bullish momentum* dan *exciting times*, serta keyakinan akan pergerakan di atas \$100k dalam waktu dekat. Sebaliknya, contoh negatif (Gambar 3.2) memperlihatkan kekecewaan dan ketidakpercayaan, terlihat dari kata “tired”, “crashing”, serta frasa “too volatile and unpredictable”, menyatakan bahwa Bitcoin dianggap terlalu berisiko untuk dipercaya, sehingga diklasifikasikan sebagai sentimen negatif.

3.2 Prosedur Penelitian

Prosedur penelitian merupakan gambaran menyeluruh mengenai tahapan-tahapan penelitian yang dilakukan selama penyusunan skripsi ini. Secara visual, alur prosedur penelitian ditampilkan pada Gambar 3.3 berikut.



Gambar 3. 3 Diagram Alur Prosedur Penelitian

Tahapan penelitian ini diawali dengan pengumpulan data, kemudian dilanjutkan dengan pelabelan data untuk menentukan kategori sentimen. Data yang telah diberi label selanjutnya melalui tahap pra-pemrosesan guna membersihkan dan menormalkan teks agar siap dianalisis. Setelah itu dilakukan ekstraksi fitur menggunakan metode TF-IDF untuk mengubah data teks menjadi representasi numerik. Tahap berikutnya adalah klasifikasi data menggunakan algoritma Naive Bayes, Support Vector Machine, dan Logistic Regression. Kinerja model kemudian dievaluasi menggunakan Confusion Matrix untuk mengukur performa klasifikasi.

3.2.1 Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan metode web scraping yang ditujukan pada platform media sosial Twitter, yang kini dikenal sebagai X. Proses pengambilan data menggunakan perangkat lunak Tweet-Harvest berbasis Python untuk mengekstraksi teks secara otomatis dan tersusun rapi. Data yang dikumpulkan difokuskan pada tweet berbahasa Inggris yang mengandung kata kunci “Bitcoin” dan “BTC”, tanpa membatasi jenis akun pengirim, baik dari individu maupun lembaga. Pendekatan ini dipilih untuk menjaga keberagaman serta keterwakilan opini publik dalam analisis.

3.2.2 Pra-Pemrosesan Teks

Pra-pemrosesan teks merupakan tahap krusial dalam penelitian ini, karena data mentah dari Twitter sering kali berisi informasi yang tidak relevan, noise, dan variasi penulisan. Tahapan ini bertujuan untuk membersihkan, menstandarisasi, dan menyederhanakan teks sehingga dapat diubah menjadi representasi numerik yang efektif untuk algoritma Machine Learning. Proses pra-pemrosesan memastikan model belajar dari informasi yang relevan, konsisten, dan bebas bias, sehingga meningkatkan akurasi klasifikasi sentimen. Proses pra-pemrosesan teks pada penelitian ini digambarkan pada Gambar 3.4 Diagram Alur Pra-Pemrosesan Teks, yang menunjukkan urutan tahapan dari data mentah hingga teks siap dianalisis.

a. *Case Folding*

Pada tahapan *case folding*, semua karakter diubah menjadi huruf kecil, ukuran setiap karakter diperiksa dari awal hingga akhir, dan jika ditemukan karakter yang menggunakan huruf besar, karakter diubah akan menjadi huruf kecil. Berikut alur dari proses *case folding* ditunjukkan pada Tabel 3.1.

Tabel 3. 1 Proses *Case Folding*

<i>Case Folding</i>
Input : Kalimat
Proses :
1. Baca kalimat
2. Merubah kalimat kapital pada kalimat menjadi huruf nonkapital
Output : kalimat menjadi nonkapital

b. *Cleansing*

Tahapan *cleansing* adalah fondasi penting dalam text preprocessing. Dengan membersihkan teks dari elemen-elemen yang tidak relevan seperti angka, simbol, URL, emoji, dan stopwords, data teks menjadi

lebih terstruktur, konsisten, dan siap diproses lebih lanjut untuk analisis sentimen, clustering, atau pembobotan kata (*TF-IDF*).

Tabel 3. 2 Proses *Cleansing*

Proses <i>Cleansing</i>
Input : Kalimat mentah (mengandung angka, simbol, URL, emoji, spasi berlebih, dll.)
Proses :
<ol style="list-style-type: none"> 1. Baca teks mentah 2. Hapus Angka Pada teks 3. Hapus tanda baca dan teks husus 4. Hapus URL/tautan 5. Hapus emoji/emotikon 6. Hapus spasi berlebih
Output : Kalimat bersih (hanya berisi teks utama yang siap diproses lebih lanjut)

c. *Tokenizing*

Tokenizing pada tahapan ini merupakan tahapan dalam memecahkan string atau input terhadap suatu teks berdasarkan tiap kata yang Menyusun nya dan menghilangkan *mention* yang ada pada dokumen *tweet*. Beserta menghilangkan tanda titik (.), koma (,), tanda seru (!), tanda tanya (?).

Tabel 3. 3 Proses *Tokenizing*

Proses <i>Tokenizing</i>
Input : kalimat
Proses :
<ol style="list-style-type: none"> 1. Baca Kalimat 2. Menghilangkan Tanda Baca
Output : Kalimat menjadi tidak ada tanda baca dan karakter lain nya

d. *Stopword Removal*

Kata-kata yang sering muncul secara umum dan tidak relevan dengan teks dihilangkan. Tahap ini disebut juga dengan stop word removal, yaitu menghilangkan kata-kata yang tidak berarti dan tidak berdampak pada analisis sentiment.

Tabel 3. 4 Proses Stopword Removal

Proses <i>Filtering</i>
Input : kalimat
Proses :
1. Baca Kalimat
2. Menghilangkan kalimat yang tida penting
Output : kealimat yang tidak penting menjadi menghilang

3.2.3 Labeling Data

Proses labeling data adalah tahap memberikan kategori atau kelas tertentu pada setiap data agar lebih terstruktur dan mudah dianalisis. Dalam analisis sentimen misalnya, teks yang bernuansa positif akan diberi label “positif”, sedangkan yang bernuansa negatif akan diberi label “negatif”. Labeling dapat dilakukan secara manual oleh manusia, semi-otomatis dengan bantuan aturan atau kamus, maupun otomatis menggunakan model machine learning. Tujuannya adalah menyediakan data berlabel sebagai acuan (*ground truth*) sehingga proses analisis, klasifikasi, maupun evaluasi dapat dilakukan secara lebih akurat dan konsisten.

Tabel 3. 5 Proses Labeling Data

Proses Labeling Data
Input : Data teks mentah (ulasan, komentar, atau kalimat)
Proses :
1. Tentukan kategori label yang akan digunakan (misalnya: positif, negatif)
2. Baca setiap teks pada dataset
3. Analisis isi teks sesuai pedoman labeling

4. Berikan label pada teks berdasarkan makna (positif/negatif)

Output : Data teks dengan label yang sesuai (misalnya: "Produk ini bagus" → positif)

3.2.4 Pembobotan Kata

Langkah selanjutnya adalah menggunakan metode TF-IDF untuk menghitung bobot setiap kata atau kata berdasarkan seberapa sering kata tersebut muncul dalam dokumen. TF-IDF adalah statistik numerik yang menampilkan kata kunci dengan kata-kata tertentu. Selain itu, TF-IDF dapat menemukan kata-kata yang sering muncul dalam dokumen.

Tabel 3. 6 Proses Pembobotan Kata

Proses TF-IDF
Input : k: jumlah kluster, inisiasi jumlah claster awal k =2 nilai <i>centroid</i> awal Menggunakan metode mean, data set n objek.
Proses : Membobotkan kata
Output : Memunculkan Bobot Nilai pada kata

3.2.5 Proses Klasifikasi Model

Pada fase ini ada tiga algoritma klasifikasi yang diterapkan, yaitu Algoritma *Naïve Bayes (NB)*, *Support Vector Machine (SVM)*, dan *Logistic Regression (LR)* untuk melaksanakan klasifikasi berdasarkan nilai bobot yang telah diperoleh sebelumnya. Ketiga algoritma ini digunakan untuk menganalisis kinerja model dalam mengkategorikan data pembelajaran. Berikut adalah langkah-langkah dari masing-masing algoritma :

1. *Naïve Bayes (NB)*

Proses klasifikasi dilakukan dengan menghitung probabilitas posterior setiap kelas menggunakan pendekatan Teorema Bayes. Tahapan yang dilakukan meliputi :

- a. Menentukan probabilitas awal (*prior probability*) dari setiap kelas berdasarkan distribusi data pelatihan.

- b. Menghitung probabilitas kemunculan setiap fitur terhadap kelas tertentu (*likelihood*).
- c. Mengalikan nilai probabilitas *prior* dengan *likelihood* untuk memperoleh probabilitas *posterior* masing-masing kelas.
- d. Menentukan hasil klasifikasi berdasarkan nilai probabilitas posterior terbesar.

2. Support Vector Machine

Proses klasifikasi dilakukan dengan menentukan batas pemisah optimal yang memaksimalkan margin antar kelas. Tahapan yang dilakukan sebagai berikut :

- a. Memetakan data ke dalam ruang fitur berdimensi tertentu.
- b. Mengidentifikasi Support Vector sebagai titik data yang berada paling dekat dengan batas pemisah.
- c. Menghitung margin optimal untuk memaksimalkan jarak antara dua kelas yang berbeda.
- d. Menentukan *hyperplane* sebagai batas klasifikasi dengan menggunakan fungsi *kernel linear*.

3. Logistic Regression

Proses klasifikasi dilakukan dengan memodelkan hubungan antara variabel independen dan probabilitas kelas menggunakan fungsi logistik. Tahapan yang dilakukan diantara lain :

- a. Menghitung kombinasi linear antara fitur dan parameter bobot.
- b. Menerapkan fungsi Sigmoid untuk mengubah nilai linear menjadi probabilitas dalam rentang 0 hingga 1.
- c. Menentukan nilai ambang batas (*threshold*) untuk memetakan probabilitas ke dalam kelas tertentu.
- d. Menghasilkan label klasifikasi berdasarkan hasil probabilitas yang diperoleh.