# ANALISIS SENTIMEN MASYARAKAT PADA TWITTER TERHADAP PEMILIHAN UMUM 2024 MENGGUNAKAN ALGORITMA NAÏVE BAYES

**Salim Puad, Garno, Agung Susilo Yuda Irawan**
Program Studi Informatika, Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang, Indonesia
*1910631170043@student.unsika.ac.id*

**ABSTRAK**

Pemilihan Umum (Pemilu) merupakan mekanisme untuk menjalankan kedaulatan rakyat dengan tujuan menciptakan pemerintahan negara yang demokratis berdasarkan Pancasila dan UUD Negara RI Tahun 1945. Pemilu diselenggarakan setiap lima tahun dan melibatkan pemilihan Presiden dan Wakil Presiden, Anggota DPR, DPD, DPRD, serta kepala daerah dan wakil kepala daerah. Pemilu bertujuan untuk memilih para pemimpin yang dapat mencerminkan nilai-nilai demokrasi dan mampu mewakili aspirasi rakyat sesuai dengan perkembangan kehidupan berbangsa dan bernegara. Dalam rangka memahami pandangan masyarakat terkait Pemilihan Umum 2024, dilakukan analisis sentimen. Analisis sentimen ini bertujuan untuk memberikan gambaran tentang bagaimana masyarakat memandang Pemilihan Umum 2024 yang akan datang. Melalui penerapan algoritma Naïve Bayes dengan metodologi KDD, pendapat atau sentimen yang dianalisis menghasilkan 331 label positif, 261 label negatif, dan 825 label netral. Untuk menguji keakuratan analisis sentimen dalam menyelesaikan Pemilihan Umum 2024, digunakan metode split data dan confusion matrix untuk menguji skor. Hasilnya menunjukkan bahwa model dengan pembagian data 90:10 memiliki akurasi tertinggi. Selain menggunakan confusion matrix, juga dilakukan pengujian menggunakan grafik ROC yang menghasilkan nilai AUC tertinggi pada model dengan pembagian data 90:10, dengan nilai 0,71 yang menunjukkan bahwa model ini memiliki kualitas klasifikasi yang baik.

***Kata kunci :*** *pemilihan umum 2024, analisis sentimen, naïve bayes*

## 1. PENDAHULUAN

Pemilihan Umum (Pemilu) merupakan mekanisme yang digunakan untuk mewujudkan kedaulatan rakyat dan menghasilkan pemerintahan negara yang demokratis, sesuai dengan Pancasila dan UUD Negara RI Tahun 1945. Pemilu ini bertujuan untuk memilih Presiden dan Wakil Presiden, Anggota DPR, DPD, DPRD, serta kepala daerah dan wakil kepala daerah yang mampu mencerminkan nilai-nilai demokrasi dan mampu memperjuangkan aspirasi rakyat sesuai dengan perkembangan kehidupan berbangsa dan bernegara. Pelaksanaan Pemilu dilakukan oleh penyelenggara pemilu yang memiliki integritas, profesionalitas, dan akuntabilitas. Pemilu dijalankan dengan kualitas yang baik, sistematis, sah secara hukum, dan akuntabel, dengan melibatkan partisipasi masyarakat secara luas. Para penyelenggara pemilu, aparat pemerintah, peserta pemilu, pengawas pemilu, pemantau pemilu, pemilih, dan semua pihak terkait diharapkan untuk bersikap dan bertindak jujur sesuai dengan peraturan perundang-undangan yang berlaku. Dalam pelaksanaannya, seringkali muncul isu-isu terkait kecurangan, polarisasi politik, dan pendapat publik yang beragam terhadap calon atau partai politik tertentu. Oleh karena itu, penting bahwa pemilih dan peserta pemilu diperlakukan secara adil dan bebas dari kecurangan atau perlakuan yang tidak adil dari pihak manapun. Pemilu harus dilaksanakan dengan kualitas yang lebih baik untuk menjamin kompetisi yang sehat, partisipatif, dengan tingkat keterwakilan yang lebih tinggi, dan memiliki mekanisme pertanggungjawaban yang jelas.

Dengan demikian, pemilihan umum yang berkualitas akan memberikan jaminan terhadap pesta demokrasi yang sehat, di mana setiap pemilih memiliki peran yang sama dan dihormati, serta setiap peserta pemilu memiliki kesempatan yang adil untuk bersaing. Hal ini juga penting untuk memastikan bahwa hasil pemilihan mencerminkan kehendak dan aspirasi sebagian besar masyarakat. [1]. Kecurangan dalam pemilihan umum di Indonesia, seperti politik uang, kampanye hitam, dan penggelembungan suara, telah menjadi isu yang sering terjadi. Untuk itu, penelitian ini bertujuan untuk memberikan gambaran tentang pandangan masyarakat terkait Pemilihan Umum 2024, dengan harapan dapat memberikan kontribusi dalam memastikan keberhasilan penyelenggaraan Pemilu 2024. Pandangan masyarakat dapat berupa tanggapan positif ataupun negatif terhadap pemilihan tersebut. Dalam konteks ini, analisis sentimen melalui media sosial menjadi metode yang efektif untuk mengidentifikasi pandangan dan sentimen masyarakat terkait penyelenggaraan Pemilihan Umum 2024 [2]. Analisis sentimen adalah proses untuk mengumpulkan dan memahami pandangan individu terhadap suatu peristiwa atau topik dalam kehidupan nyata. Dalam konteks media sosial, analisis sentimen media sosial adalah teknik atau metode yang digunakan untuk memahami pendapat individu melalui platform jejaring sosial. Dengan menggunakan analisis sentimen media sosial, kita dapat mengidentifikasi dan menganalisis ekspresi emosional, opini, atau tanggapan seseorang terhadap topik tertentu yang

dibagikan melalui media sosial [3]. Analisis sentimen adalah suatu proses pengolahan data tekstual yang bertujuan untuk memahami, mengekstraksi, dan mengolah informasi yang terkandung dalam rumor kontroversial. Melalui analisis sentimen, kita dapat mengidentifikasi sentimen atau pendapat yang terkait dengan rumor tersebut, baik itu positif, negatif, atau netral. Proses analisis sentimen ini membantu kita dalam memahami dan memperoleh informasi yang relevan dari data tekstual, sehingga dapat membantu dalam mengatasi rumor kontroversial dengan lebih baik [4]. Meskipun digunakan untuk menyelesaikan masalah klasifikasi di dunia nyata, pengklasifikasi berbasis Naive Bayes Classifier (NBC) tetap menghasilkan solusi yang dapat diimplementasikan [5]. Algoritma Naive Bayes memiliki kinerja yang luar biasa dalam melakukan klasifikasi. Untuk memastikan model Naive Bayes dapat bekerja dengan baik dalam mengklasifikasikan data yang belum pernah dilihat sebelumnya, penting untuk membagi data menjadi dua bagian, yaitu data pelatihan (training) dan data pengujian (testing) [6]. Pembagian data menjadi data training dan data testing memiliki peranan penting dalam mencegah overfitting, di mana model Naive Bayes dapat menjadi terlalu terfokus pada data pelatihan sehingga tidak mampu mengklasifikasikan data yang belum pernah dilihat sebelumnya dengan akurasi yang tinggi [7]. Dengan melakukan pembagian data tersebut, kita dapat memastikan bahwa model Naive Bayes yang telah dibentuk mampu mengklasifikasikan data yang belum pernah dilihat sebelumnya dengan tingkat akurasi yang tinggi [8]. Dalam penelitian ini, dilakukan analisis sentimen terhadap tweet di platform Twitter untuk mengklasifikasikan opini dan komentar terkait Pemilihan Umum 2024. Data yang dikumpulkan akan diolah menggunakan algoritma Naive Bayes.

## 2. TINJAUAN PUSTAKA

### 2.1. Pemilihan Umum (Pemilu)

Pemilu atau pemilihan umum adalah salah satu sarana kedaulatan atau pesta demokrasi dari rakyat untuk memilih wakilnya seperti anggota Dewan Perwakilan Rakyat (DPR), anggota Dewan Perwakilan Daerah (DPD), Presiden, dan Wakil Presiden yang dimana pelaksanaanya dapat dilangsungkan dengan cara langsung, umum, bebas, rahasia, jujur, dan adil atau yang disebut dengan Luber Jurdil dalam Negara Kesatuan Republik Indonesia yang didasarkan pada UUD Negara Republik Indonesia tahun 1945.

### 2.2. Twitter

Media sosial Twitter adalah alat jejaring sosial gratis yang memungkinkan orang untuk berbagi informasi, dalam umpan berita *real time* melalui posting komentar singkat tentang pengalaman dan pemikiran mereka Pesan publik yang dikirim dan diterima melalui Twitter atau *tweet* dibatasi tidak lebih dari 140 karakter dan dapat menyertakan tautan ke blog, halaman web, gambar, video, dan semua materi lainnya secara online. Sebagai alat komunikasi, Twitter memungkinkan pertukaran ide bebas secara nasional dan global, antara orang-orang yang tertarik pada bidang keahlian yang sama, serta memberikan kesempatan untuk terlibat dalam debat kritis [1].

### 2.3. *Data Mining*

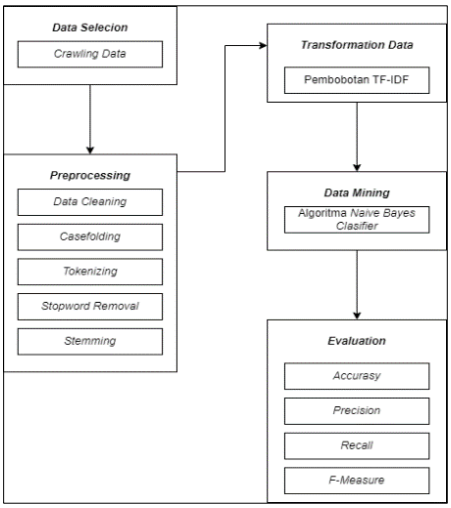
*Data Mining adalah sebuah proses dalam menemukan pola yang sangat* berguna dan juga menemukan suatu tren di dalam data set yang berukuran besar [2]. Menurut Pregibin data mining sendiri merupakan gabungan dari ilmu statistika, *Artificial Intelligence*, dan *database riset* [3]. Tugas *data mining* terbagi kedalam enam bagian, yaitu deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi [4].

### 2.4. Analisis Sentimen

Analisis sentimen mencakup berbagai bidang diantaranya *Natural Language processing*, *text mining*, dan komputasi linguistik yang bertujuan untuk menganalisis suatu sentimen atau pendapat seseorang mengenai suatu topik tertentu. Sedangkan menurut (Andika et al., 2019) Analisis sentimen ialah kombinasi dari tiga metode, yaitu *natural language processing*, *information retrieval*, dan *data mining*. Tujuannya untuk mengelola dan menganalisa sentimen, emosi, dan pendapat yang disampaikan mengenai entitas tertentu dalam bentuk data teks. Tugas dasar analisis sentimen ialah mengklasifikasikan kategori teks dalam dokumen, kalimat, atau opini untuk memahami data yang memiliki makna positif, negatif dan netral [5].

## 3. METODE PENELITIAN

Dalam penelitian ini, langkah awalnya adalah mengumpulkan data dengan melakukan crawling menggunakan Python dan mengakses API Twitter. Metode penelitian yang digunakan adalah Knowledge Discovery in Database (KDD), yang terdiri dari lima tahapan utama, yaitu Seleksi Data, Pra-Pemrosesan, Transformasi, Data Mining, dan Evaluasi. Setiap tahapan memiliki perannya sendiri dalam proses penelitian, termasuk seleksi data, pra-pemrosesan data, transformasi data, penambangan data, dan evaluasi hasil yang diperoleh [9]. Pada tahapan *Preprocessing*, dilakukan penggunaan proses Text Mining untuk membersihkan data yang telah dikumpulkan. [2]. Untuk mempermudah pemahaman tentang alur penelitian, terdapat gambar 1 di bawah ini yang menjelaskan secara visual tentang proses penelitian.



Gambar 1. Metode Penlitian

### 3.1. *Data Selection*

Tahap awal dalam penelitian ini adalah seleksi data, di mana kami menggunakan teknik crawling data untuk mengumpulkan informasi yang relevan dari media sosial terkait dengan pertanyaan penelitian, terutama terkait Pemilihan Umum 2024. Kami mengumpulkan tweet yang berkaitan dengan topik tersebut dengan melakukan pencarian menggunakan Twitter API.

### 3.2. *Pre Processing*

Sebelum melakukan klasifikasi, data perlu diatur dan dibersihkan selama tahap pra-processing melibatkan langkah-langkah berikut ini:

a. *Cleaning*
   Proses pembersihan data melibatkan penghapusan emotikon atau simbol tambahan yang tidak relevan atau tidak diperlukan.
b. *Case Folding*
   Dalam langkah pertama dari prosedur ini, dilakukan upaya untuk mengubah teks menjadi huruf kecil atau merendahkan teks.
c. *Tokenizing*
   Tahap berikutnya adalah tokenisasi, di mana kalimat akan dipisahkan menjadi kata-kata secara individu. Dengan menggunakan metode ini, kalimat dapat dipecah menjadi unit-unit kata yang terpisah, memungkinkan untuk analisis lebih lanjut pada tingkat kata..
d. *Stopword Removal*
   Untuk melakukan penghapusan stopword, terdapat sebuah daftar yang berisi 32 stopword yang telah disediakan. Stopword ini akan menghapus kata apa pun dalam data yang juga terdaftar dalam daftar tersebut.
e. *Stemming*
   Prosedur yang dilakukan untuk menghilangkan imbuhan adalah dengan menghapus imbuhan setelah melakukan tokenisasi. Dalam langkah ini, imbuhan pada hasil token akan dihapus untuk mendapatkan bentuk dasar atau kata dasar dari setiap kata yang ada.

### 3.3. *Transformation Data*

Setelah dokumen telah diproses dan dimodifikasi, tahap selanjutnya adalah mengubah teks menjadi data numerik yang dapat secara akurat mewakili dokumen tersebut. Salah satu metode pembobotan yang digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF), yang memberikan nilai bagi setiap kata yang mengindikasikan seberapa sering kata kunci atau istilah lain muncul dalam dokumen tersebut [10].

### 3.4. *Data Mining*

Langkah berikutnya adalah data mining, di mana data mentah diubah menjadi informasi yang berharga. Pada tahap ini, tujuannya adalah untuk menemukan dan menganalisis pola tersembunyi dalam jumlah besar data guna mendapatkan wawasan yang berguna dan bermanfaat [11] Dalam langkah ini, digunakan pendekatan klasifikasi untuk mengkategorikan sekumpulan sentimen ke dalam kelompok positif, negatif, dan netral. Salah satu algoritma klasifikasi yang digunakan adalah Naive Bayes berdasarkan teorema Bayes. Data yang telah disiapkan akan dibagi menjadi data pelatihan dan pengujian sebagai bagian dari proses klasifikasi.

### 3.5. *Evaluation*

Pada langkah terakhir, model yang telah dibentuk akan dievaluasi untuk memastikan konsistensi dengan hipotesis sebelumnya. Untuk mengukur akurasi algoritma, kami menggunakan confusion matrix. Confusion matrix digunakan untuk menggambarkan prediksi model dan seberapa baik metode tersebut diterapkan. Dengan menggunakan confusion matrix, kita dapat menganalisis kinerja model dalam empat dimensi dan mendapatkan pemahaman yang lebih jelas tentang performa yang dicapai:

a. *Akurasi*: Untuk mengukur sejauh mana nilai yang diharapkan dan nilai aktual memiliki kemiripan satu sama lain.
b. *Presisi* : Untuk mengevaluasi keakuratan atau presisi suatu model berdasarkan prediksi positifnya. Presisi adalah metrik yang berguna saat model memiliki tingkat Positif Palsu yang tinggi.
c. *Recall* : Untuk menilai berapa banyak nilai Positif Aktual yang telah dikenali sebagai True Positif oleh model melalui pelabelan. Recall adalah metrik yang digunakan saat terdapat tingkat False Negatif yang tinggi. Recall menjadi metrik penting dalam menentukan model yang terbaik.
d. *F-Measure*: Perbandingan antara presisi dan recall dapat diukur menggunakan F1-score, yang merupakan perolehan hasil rata-rata tertimbang dari kedua metrik tersebut. F1-score memberikan gambaran keseluruhan tentang keseimbangan antara presisi dan recall dalam penilaian kinerja model.

### *F. Receiver Operating Characteristic (ROC)*

Kurva ROC adalah grafik yang menggambarkan hubungan antara spesifisitas (nilai benar negatif) pada sumbu x dan sensitivitas (nilai benar positif) pada sumbu y. Kurva ROC sering digunakan karena dapat digunakan untuk memvalidasi hasil prediksi dan memilih klasifikasi berdasarkan kinerja algoritma. Di dalam kurva ROC, terdapat nilai Area Under Curve (AUC) yang mewakili bagian dari area persegi dengan nilai antara 0 hingga 1. Nilai AUC dapat dikelompokkan ke dalam beberapa kategori berikut ini [7] :

*a.* 0.90 – 1.00 = *Excellent Classification*
*b.* 0.80 – 0.90 = *Good Classification*
*c.* 0.70 – 0.80 = *Fair Classification*
*d.* 0.60 – 0.70 = *Poor Classification*
*e.* 0.50 – 0.60 = *Failur Classification*

## 4. HASIL DAN PEMBAHASAN

Dalam penelitian ini, dilakukan analisis sentimen terhadap data tweet terkait pemilihan umum 2024. Data tweet tersebut kemudian diklasifikasikan ke dalam tiga kelas, yaitu positif, negatif, dan netral, dengan menggunakan algoritma klasifikasi Naive Bayes Classifier. Evaluasi sistem dilakukan dengan menggunakan Confusion Matrix untuk mendapatkan nilai akurasi, presisi, recall, dan f-measure dari model tersebut.

### 4.1. *Data Selection*

Dataset awal yang diperoleh melalui teknik Crawling data adalah koleksi tweet dari media sosial Twitter dengan kata kunci "Pemilihan Umum 2024" dalam bahasa Indonesia. Proses pengambilan data ini dilakukan melalui akses Twitter API. Contoh sampel dari dataset awal yang dihasilkan dapat dilihat pada Gambar 2.

| | username | tweetcreatedts | text |
|---|---|---|---|
| 0 | MIrvanDarwin1 | 2023-05-24 23:58:27+00:00 | https://t.co/mH6Eu88UXj \n\nIlah ancama demo... |
| 1 | krokotpink | 2023-05-24 23:58:16+00:00 | Serba-Serbi MMC\n\nPemilu 2024 diikuti oleh ba... |
| 2 | DanangLester | 2023-05-24 23:58:07+00:00 | Serba-Serbi MMC\n\nPemilu 2024 diikuti oleh ba... |
| 3 | DanangCampbell | 2023-05-24 23:57:48+00:00 | Serba-Serbi MMC\n\nPemilu 2024 diikuti oleh ba... |
| 4 | Indogo7 | 2023-05-24 23:57:44+00:00 | BREAKING NEWS\n\nDin Syamsuddin dan Jusuf Kall... |
| ... | ... | ... | ... |
| 5995 | DesiRahayuPutr1 | 2023-05-24 10:59:48+00:00 | Tahapan Pemilu 2024 akan berlangsung sesuai ja... |
| 5996 | DesiRahayuPutr1 | 2023-05-24 10:59:47+00:00 | Tahapan Pemilu 2024 akan berlangsung sesuai ja... |
| 5997 | DesiRahayuPutr1 | 2023-05-24 10:59:47+00:00 | Tahapan Pemilu 2024 akan berlangsung sesuai ja... |
| 5998 | DesiRahayuPutr1 | 2023-05-24 10:59:46+00:00 | Tahapan Pemilu 2024 akan berlangsung sesuai ja... |
| 5999 | DesiRahayuPutr1 | 2023-05-24 10:59:45+00:00 | Tahapan Pemilu 2024 akan berlangsung sesuai ja... |

6000 rows × 3 columns

Gambar 2. Dataset hasil *crawling*

Data awal yang dihasilkan melalui proses Crawling memiliki total 6000 data. Proses Crawling dilakukan dalam rentang waktu mulai dari tanggal 9 Februari hingga 25 Mei 2023. Dataset ini terdiri dari beberapa atribut, termasuk username, tweetcreatedts (tanggal dan waktu tweet dibuat), dan text (isi teks tweet). Informasi lebih rinci tentang setiap atribut tersebut dapat ditemukan di Tabel 1.

Tabel 1. Atribut pada Dataset

| No | Atribut | Penjelasan |
|---|---|---|
| 1 | Username | Username dari akun Twitter yang membuat cuitan tersebut. |
| 2 | Text | Merupakan cuitan dari pengguna Twitter |
| 3 | Tweetcreatedts | Waktu dan tanggal dari pembuatan cuitan tersebut. |

Data yang telah di-crawling masih dalam bentuk yang tidak terstruktur dan mengandung banyak noise, seperti tanda baca, angka, simbol, dan kata-kata tidak baku yang tidak diperlukan dalam proses klasifikasi. Oleh karena itu, dataset ini akan melalui tahap Pre-Processing selanjutnya untuk diolah lebih lanjut.

### 4.2. Hasil *Pre processing*

Berikut ini adalah contoh hasil pengambilan sampel teks yang akan diproses melalui tahap preprocessing. Dimulai dari data asli yang diperoleh dari proses crawling data, kemudian dilanjutkan dengan proses case folding, tokenizing, penghapusan stopword, dan stemming. Informasi lebih lanjut dapat ditemukan dalam Tabel 2.

Tabel 2. Hasil Sample Text

| *Preprocessing* | *Content* |
|---|---|
| Data Mentah | Ngeri!! Polri Temukan Skandal Indikasi Dana Politik Dari Jaringan Narkotika Untuk Peserta Pemilu 2024 https://t.co/XwYLkWIaDA |
| Cleaning | Ngeri Polri Temukan Skandal Indikasi Dana Politik Dari Jaringan Narkotika Untuk Peserta Pemilu 2024 |
| Case Folding | ngeri polri temukan skandal indikasi dana politik dari jaringan narkotika untuk peserta pemilu 2024 |
| Tokenizing | ‘ngeri’, ‘polri’, ‘temukan’, ‘skandal’, ‘indikasi’, ‘dana’,’politik’, ‘dari’, ‘jaringan’, ‘narkotika’ ‘untuk’, ‘peserta’,’pemilu’ |
| Stopword Removal | ‘ngeri’, ‘polri’, ‘temu’, ‘skandal’, ‘indikasi’, ‘dana’, ‘politik’, ‘jaring’,’narkotika’,’serta’,’pemilu’ |
| Stemming | ‘ngeri’, ‘polri’, ‘skandal’, ‘indikasi’, ‘dana’, ‘politik’, ‘jaring’, ‘serta’, ‘pemilu’ |

### 4.3. *Transformation Data*

Setelah melalui tahap preprocessing, langkah berikutnya adalah tahap transformasi data. Sebelum melakukan transformasi data, dilakukan proses pembobotan sentimen untuk memberi label pada data. Proses ini melibatkan penghitungan skor sentimen dengan mencari jumlah kata yang bersentimen positif dan kata yang bersentimen negatif dalam dataset. Skor sentimen diperoleh dari jumlah polaritas sentimen positif dan negatif yang terdapat dalam dataset. Dalam penelitian ini, digunakan kamus Lexicon dan Negative Word yang telah digunakan dalam penelitian sebelumnya [12] dalam rangka memberikan bobot yang lebih signifikan pada sentimen dari setiap tweet

dan mengevaluasi sentimen tersebut. Langkah selanjutnya adalah perhitungan skor sentimen dengan menghitung polaritas setiap kalimat yang terdapat dalam data tweet tersebut. Hasil dari pembobotan sentimen menggunakan kamus lexicon dan negative words dapat dilihat pada Gambar 3. Dengan menggunakan metode ini, komputer dapat memberikan penilaian yang lebih baik terhadap sentimen yang terdapat dalam data tweet.

| | Tweet | label | weight |
|---|---|---|---|
| 0 | ancama demokrasi sungguh sodara | positive | 0 + 1 + 0 + 0 = 1 |
| 1 | serba serbi mmc milu ikut deret kepala daerah ... | positive | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0... |
| 2 | breaking news din syamsuddin jusuf kalla senti... | neutral | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0... |
| 3 | benak me tanda milu anggota legislatif pamer p... | neutral | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0... |
| 4 | ngeri polri temu skandal indikasi dana politik... | neutral | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0 |
| ... | ... | ... | ... |
| 1412 | said tanggap survei litbang kompas tempat ganj... | positive | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0... |
| 1413 | hasil tang oligarki kpu bawaslu backing jahat ... | negative | 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + (1 * -1 ) + 0 + ... |
| 1414 | batas regenerasi hakim mk keren politik kuasa ... | positive | 0 + 0 + 0 + 0 + 1 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0... |
| 1415 | marm tunda milu rampok uang rakyat | neutral | 0 + 0 + 0 + 0 + 0 + 0 + 0 = 0 |
| 1416 | baidhawie timur gagal sblum milu | neutral | 0 + 0 + 0 + 0 + 0 + 0 = 0 |

1417 rows × 3 columns

✓ 0 d selesai pada 23.59

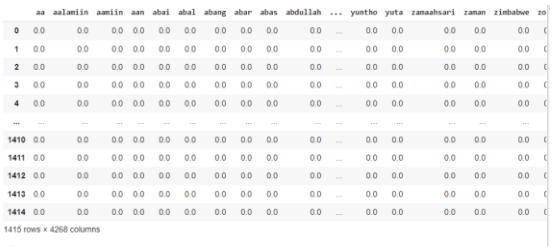
Gambar 3. Hasil Pembobotan Sentimen

Selanjutnya sentimen dapat ditemukan pada Tabel 3.

Tabel 3. Jumlah Pada Setiap Label

| Label/Kelas | Jumlah |
|---|---|
| Positif | 331 |
| Negatif | 261 |
| Netral | 825 |
| **Total** | **1417** |

pembobotan sentimen menggunakan kamus lexicon dan negative words, selanjutnya adalah melakukan pembobotan dengan menerapkan metode TF IDF.



| | aa | aalamiin | aamiin | aan | abai | abal | abang | abar | abas | abdullah | ... | yuntho | yuta | zamaahsari | zaman | zimbabwe | zo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1410 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 1411 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 1412 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 1413 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |
| 1414 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 |

1415 rows × 4268 columns

Gambar 4. Hasil Pembobotan TF-IDF

Gambar 4 menampilkan daftar kata/term yang ada dalam korpus, diurutkan secara alfabetis, dan dihitung kemungkinan kemunculan kata tersebut dalam dokumen.

### 4.4. *Data Mining*

Tahap ini melibatkan klasifikasi data menggunakan model Multinomial Naive Bayes, yang merupakan salah satu algoritma Naive Bayes yang diimplementasikan menggunakan library sklearn pada Python. Implementasi algoritma Multinomial Naive Bayes dapat ditemukan dalam Gambar 4 berikut.

```
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB().fit(X_train_df, y_train)
prediction_mi = model.predict(X_test_df)
prediction_proba_mi = model.predict(X_test_df)
```

Gambar 5. Library Multinomial Bayes

Dalam penelitian ini, data diproses menggunakan algoritma Naive Bayes dengan skenario 90:10, 80:20, 70:30, dan 60:40. Naive Bayes classifier digunakan untuk mengklasifikasikan sentimen terkait Pemilihan Umum 2024, dan Confusion Matrix digunakan untuk menghasilkan nilai akurasi prediksi mesin terhadap data yang telah diproses. Berikut adalah pembagian data menjadi data training dan data testing.

Tabel 4. Tabel Skenario Pembagian Split Data

| Presentasi Data | | Jumlah Data | |
|---|---|---|---|
| **Training** | **Testing** | **Training** | **Testing** |
| 90% | 10% | 1275 | 142 |
| 80% | 20% | 1134 | 283 |
| 70% | 30% | 992 | 425 |
| 60% | 40% | 850 | 567 |
| **Total** | | **1417** | |

### 4.5. *Evaluation*

Setelah selesai pembuatan model, model yang telah dibuat dari masing-masing skenario akan diuji. Hasil uji model ini akan memberikan nilai akurasi, presisi, recall, dan f-measure (f1-score). Berikut ini adalah hasil perhitungan evaluasi dari keempat model skenario yang telah diuji.

a. Skenario 1 (90%:10%)
b. Skenario 2 (80%:20%)
c. Skenario 3 (70%:30%)
d. Skenario 4 (60%:40%)

```
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00        20
     neutral       0.65      1.00      0.79        89
    positive       1.00      0.15      0.26        33

    accuracy                           0.66       142
   macro avg       0.55      0.38      0.35       142
weighted avg       0.64      0.66      0.56       142

Multinomial NB Accuracy :  0.6619718309859155
Multinomial NB Precision :  0.551470588235294
Multinomial NB Recall :  0.38383838383838387
Multinomial NB F1-Score :  0.3514230019493178
```

Gambar 6. Hasil Skenario 90:10

Berdasarkan Gambar 6, hasil klasifikasi menggunakan Naïve Bayes pada perbandingan 90:10 menunjukkan nilai akurasi sebesar 66%, presisi sebesar 55%, recall sebesar 38%, dan f-measure sebesar 35%.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.00 | 0.00 | 0.00 | 51 |
| neutral | 0.63 | 0.99 | 0.77 | 173 |
| positive | 0.88 | 0.12 | 0.21 | 59 |
| accuracy | | | 0.63 | 283 |
| macro avg | 0.50 | 0.37 | 0.33 | 283 |
| weighted avg | 0.57 | 0.63 | 0.51 | 283 |

```
Multinomial NB Accuracy :  0.6325088339222615
Multinomial NB Precision :  0.500912408759124
Multinomial NB Recall :  0.37095457365860024
Multinomial NB F1-Score :  0.32617672265072845
```

Gambar 7. Hasil Skenario 80:20

Dilihat dari Gambar 7, hasil klasifikasi menggunakan Naïve Bayes pada perbandingan 80:20 menghasilkan nilai akurasi sebesar 63%, presisi sebesar 50%, recall sebesar 37%, dan f-measure sebesar 32%.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.00 | 0.00 | 0.00 | 78 |
| neutral | 0.61 | 0.99 | 0.76 | 254 |
| positive | 0.86 | 0.13 | 0.22 | 93 |
| accuracy | | | 0.62 | 425 |
| macro avg | 0.49 | 0.37 | 0.33 | 425 |
| weighted avg | 0.55 | 0.62 | 0.50 | 425 |

```
Multinomial NB Accuracy :  0.6211764705882353
Multinomial NB Precision :  0.49009384775808135
Multinomial NB Recall :  0.37371941410549486
Multinomial NB F1-Score :  0.32739793408755535
```

Gambar 8. Hasil skenario 70:30

Melihat Gambar 8, hasil klasifikasi menggunakan Naïve Bayes pada perbandingan 70:30 menunjukkan nilai akurasi sebesar 62%, presisi sebesar 49%, recall sebesar 37%, dan f-measure sebesar 33%.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.00 | 0.00 | 0.00 | 100 |
| neutral | 0.61 | 0.99 | 0.76 | 340 |
| positive | 0.83 | 0.08 | 0.14 | 126 |
| accuracy | | | 0.61 | 566 |
| macro avg | 0.48 | 0.36 | 0.30 | 566 |
| weighted avg | 0.55 | 0.61 | 0.49 | 566 |

```
Multinomial NB Accuracy :  0.6148409893992933
Multinomial NB Precision :  0.4811472121941436
Multinomial NB Recall :  0.3578275754746343
Multinomial NB F1-Score :  0.3003598871705087
```

Gambar 9. Hasil Skenario 60:40

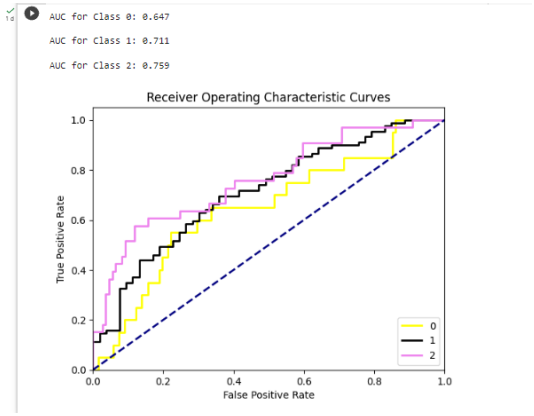
Dari Gambar 9, dapat dilihat bahwa hasil klasifikasi menggunakan Naïve Bayes pada perbandingan 60:40 menunjukkan nilai akurasi sebesar 61%, presisi sebesar 48%, recall sebesar 36%, dan f-measure sebesar 30%.

Nilai AUC untuk semua model Multinomial Naive Bayes dapat ditemukan dalam Tabel 5.

Tabel 5. Hasil Nilai AUC Setiap Skenario

| Skenario | AUC | Kualitas |
|---|---|---|
| 90:10 | 0.71 | *Good Classification* |
| 80:20 | 0.69 | *Poor Classification* |
| 70:30 | 0.69 | *Poor Classification* |
| 60:40 | 0.66 | *Poor Classification* |

Dari Tabel 5, hampir semua model skenario menunjukkan nilai AUC dengan kualitas poor classification. Model skenario 60:40 memiliki nilai AUC terendah sebesar 0.66, sedangkan model skenario 90:10 memiliki nilai AUC tertinggi sebesar 0.71 dengan kualitas Good classification. Berikut ini adalah grafik ROC dari model skenario 90:10.




Gambar 10. Grafik ROC Hasil Skenario 90:10

Dalam Gambar 10, terdapat nilai AUC untuk masing-masing kelas pada model tersebut. Untuk kelas positif, nilai AUC adalah 0.647 dengan garis warna kuning. Untuk kelas netral, nilai AUC adalah 0.711 dengan garis warna hitam. Sedangkan untuk kelas negatif, nilai AUC adalah 0.759 dengan garis warna merah muda. Berdasarkan nilai-nilai ini, ROC AUC score pada model tersebut adalah:

$$\text{ROC AUC } Score = \frac{0.647+0.711+0.759}{3} = 0.71$$

Hasil klasifikasi analisis sentimen terhadap ulasan tentang pemilihan umum 2024 dapat divisualisasikan menggunakan word cloud untuk memberikan gambaran atau informasi umum tentang data sentimen terkait pemilihan umum 2024. Berikut ini adalah pembahasan mengenai visualisasi kata-kata dari masing-masing kelas sentimen.


Gambar 11. Wordcloud keyword ‘pemilihan umum 2024’

Dalam Gambar 11, ditampilkan Wordcloud yang menunjukkan kata-kata yang sering muncul dalam dataset. Kata-kata yang paling sering muncul adalah "pemilu", "presiden", "indonesia", dan "partai". Ukuran kata dalam Wordcloud menunjukkan frekuensi penggunaannya, di mana kata-kata dengan ukuran yang lebih besar memiliki frekuensi yang lebih tinggi. Hal ini menunjukkan bahwa kata-kata tersebut sering digunakan oleh masyarakat dalam mengirimkan cuitan di Twitter. Dapat disimpulkan bahwa dalam sentimen ini, hasil netral mendominasi dalam dataset. Sentimen netral mengindikasikan ketiadaan sentimen yang kuat, dengan tidak adanya kecenderungan yang signifikan baik positif maupun negatif dalam cuitan yang dikirimkan oleh masyarakat terkait Pemilihan Umum 2024. Hal ini menunjukkan bahwa mayoritas cuitan yang dianalisis tidak menunjukkan ekspresi yang kuat dalam mendukung atau menentang Pemilihan Umum 2024.

## 5. KESIMPULAN

Memuat Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan beberapa hal, diantaranya: Dalam pengolahan opini atau sentimen menggunakan algoritma Naïve Bayes dengan menggunakan metodologi KDD, data dari media sosial Twitter telah diklasifikasikan ke dalam tiga kelas, yaitu Positif, Netral, dan Negatif. Hasil klasifikasi menunjukkan bahwa terdapat 331 label dengan sentimen positif, 261 label dengan sentimen negatif, dan 825 label dengan sentimen netral. Dari hasil tersebut, dapat dilihat bahwa kelas sentimen Netral memiliki jumlah data yang paling banyak. Hal ini menunjukkan bahwa sentimen terhadap Pemilihan Umum 2024 cenderung masuk ke dalam kelas Netral. Tingkat akurasi analisis sentimen pada Pemilihan Umum 2024 menggunakan algoritma Naïve Bayes diuji dengan menggunakan metode Split data, dengan pembagian data menjadi empat model, 90:10, 80:20, 70:30, dan 60:40. Hasil pengujian menggunakan confusion matrix untuk skor menunjukkan bahwa model dengan pembagian 90:10 memiliki nilai akurasi tertinggi. Selain itu, model 90:10 juga memiliki nilai presisi tertinggi. Selain menggunakan confusion matrix, model juga diuji menggunakan grafik ROC. Hasilnya menunjukkan bahwa model dengan pembagian 90:10 memiliki nilai AUC tertinggi, yaitu 0.71, yang menunjukkan kualitas model ini dalam klasifikasi yang baik (good classification).

## DAFTAR PUSTAKA


[1] Adiba, F. I., Islam, T., & Kaiser, M. S. (2020). Effect of Corpora on Classification of Fake News using Naive Bayes Classifier. *Int J Auto AI Mach Learn*, *1*(1), 80. https://www.researchgate.net/publication/352551511

[2] Afrizal, S., Irmanda, H. N., Falih, N., & Isnainiyah, I. N. (2020). Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap. *Informatik : Jurnal Ilmu Komputer*, *15*(3), 157. https://doi.org/10.52958/iftk.v15i3.1454

[3] Ali Fauzi, M., & Candra Brata, K. (2018). *Sistem Temu Kembali Informasi Pasal-Pasal KUHP (Kitab Undang-Undang Hukum Pidana) Berbasis Android Menggunakan Metode Synonym Recognition dan Cosine Similarity Mental Model for foreign Language M-Learning View project Twitter Sentiment Analysis View project*. *February*. https://www.researchgate.net/publication/322963820

[4] Andika, L. A., Azizah, P. A. N., & Respatiwulan, R. (2019). Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier. *Indonesian Journal of Applied Statistics*, *2*(1), 34. https://doi.org/10.13057/ijas.v2i1.29998

[5] Aulia, G. N., & Patriya, E. (2019). Implementasi Lexicon Based Dan Naive Bayes Pada Analisis Sentimen Pengguna Twitter Topik Pemilihan Presiden 2019. *Jurnal Ilmiah Informatika Komputer*, *24*(2), 140–153. https://doi.org/10.35760/ik.2019.v24i2.2369.

[6] Daniel T. Larose, C. D. L. (n.d.). *Discovering Knowledge in Data: An Introduction to Data Mining*.

[7] Maclean, F., Jones, D., Carin-Levy, G., & Hunter, H. (2013). Understanding twitter. *British Journal of Occupational Therapy*, *76*(6), 295–298. https://doi.org/10.4276/030802213X13706169933021

[8] Merawati, D., & Rino. (2019). Penerapan data mining penentu minat Dan bakat siswa Smk dengan metode C4 . 5. *Jurnal Algor*, *1*(1), 28–37.

[9] Ratiasasadara, P. W., Sudarno, S., & Tarno, T. (2023). Analisis Sentimen Penerapan Ppkm Pada Twitter Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi-Square. *Jurnal Gaussian*, *11*(4), 580–590. https://doi.org/10.14710/j.gauss.11.4.580-590

[10] Rozi, I. F., Ardiansyah, R., & Rebeka, N. (2019). Penerapan Normalisasi Kata Tidak Baku Menggunakan Levenshtein Distance pada Analisa Sentimen Layanan PT. KAI di Twitter. *Seminar Informatika Aplikatif*, 106–112. http://jurnalti.polinema.ac.id/index.php/SIAP/article/view/563

[11] Stewart, A. (2016). *Python Programming: Python Programming for Intermediates*.

[12] Sumantri, R. B. B., & Utami, E. (2020). Penentuan Status Tahapan Keluarga Sejahtera Kecamatan Sidareja Menggunakan Teknik Data Mining. *Respati*, *15*(3), 71. https://doi.org/10.35842/jtir.v15i3.375.