Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

# Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest

Oktaviami Manullang[1], Cahyo Prianto[2], Nisa Hanum Harani[3]

[123] Universitas Logistik dan Bisnis Internasional, Jl. Sari Asih No.54, Kota Bandung, Jawa Barat 40151, Indonesia

INFORMASI ARTIKEL

*Sejarah Artikel:*
Diterima Redaksi: 31 Juli 2023
Revisi Akhir: 08 Agustus 2023
Diterbitkan *Online*: 05 September 2023

KATA KUNCI

Analisis Sentimen
Pemilu
Random Forest
Random Forest
*Text Mining*

KORESPONDENSI

E-mail: oktaviamimanullang30@gmail.com

**A B S T R A C T**

In the context of Indonesian politics, the Presidential Election is a crucial moment. To predict the election outcome, sentiment analysis through social media can provide insights into public opinion. Twitter, as one of the popular social media platforms, becomes a rich data source during the lead-up to the Indonesian Presidential Election. In this situation, sentiment analysis can be conducted to examine the election campaign topics, considering various opinions from Twitter users, including positive, neutral, and negative sentiments. The collected tweet data from various sources undergoes preprocessing, which includes cleaning and stemming. Sentiment analysis plays a key role in understanding the public's inclinations towards the election. In this study, two methods, namely Random Forest and Lexicon Based, are used to predict sentiments towards the presidential candidates. Random Forest is employed to analyze sentiments in the textual data gathered from social media, news sites, and online forums. The research findings indicate that the Lexicon Based and Random Forest methods successfully predicted negative sentiment at 48%, positive sentiment at 96%, and neutral sentiment at 97% regarding Twitter users' opinions on the Presidential Election candidates. The overall model accuracy reached 88%. Thus, this sentiment analysis provides deeper insights into public perspectives on the Presidential Election candidates through social media, especially Twitter.

## 1. PENDAHULUAN

Saat ini, perkembangan teknologi informasi dan komunikasi berlangsung dengan pesat. Di era dinamis seperti sekarang, komunikasi telah menjadi bagian integral dari kehidupan manusia. Internet memainkan peran penting dalam meningkatkan efisiensi dan fleksibilitas teknologi komunikasi, memungkinkan manusia untuk dengan mudah mengakses informasi yang dibutuhkan. Akibatnya, lingkungan baru terbentuk di mana manusia berevolusi menjadi pencari informasi yang handal. Informasi yang tersedia juga semakin beragam, baik dalam konten maupun tautan yang menyertainya. Hal ini mendorong banyak orang untuk terus mencari informasi yang mereka perlukan secara tak henti di internet.

Pemilihan Presiden merupakan salah satu momen krusial dalam proses demokrasi suatu negara, di mana masyarakat berperan aktif dalam menentukan pemimpin yang akan memimpin negara selama periode tertentu. Dalam konteks politik modern, teknologi informasi telah memberikan dampak yang signifikan dalam mengubah cara masyarakat berinteraksi dan menyampaikan opini mereka terkait isu-isu politik, termasuk pemilihan Presiden [1]. Media sosial merupakan salah satu platform utama yang memfasilitasi interaksi dan ekspresi publik. Di era digital ini, Twitter telah menjadi salah satu media sosial paling populer yang memainkan peran penting dalam mendukung komunikasi dan berbagi informasi antara pengguna.

Twitter merupakan platform media sosial yang digunakan untuk menyampaikan berbagai opini, baik yang bersifat pribadi maupun publik. Pengguna Twitter dapat dengan bebas memposting pandangan mereka, dan postingan tersebut dapat diakses oleh orang lain [2]. Tweet adalah istilah yang digunakan untuk merujuk pada postingan pengguna di platform Twitter. Dengan banyaknya jumlah pengguna Twitter, platform ini dapat dimanfaatkan untuk memahami sentimen masyarakat terkait calon Presiden. Penentuan sentimen negatif, positif, atau netral

dari tweet dapat dilakukan secara manual, namun melihat banyaknya jumlah pengguna, jumlah opini yang dihasilkan juga meningkat secara signifikan [3]. Sehingga memerlukan investasi waktu dan usaha yang semakin besar.

Dengan begitu banyaknya informasi dan pendapat yang tersebar di media sosial, menganalisis pandangan masyarakat terhadap calon Pemilu Presiden telah menjadi semakin menantang. Oleh karena itu, dibutuhkan pendekatan analisis yang efektif dan efisien dalam mengolah data teks yang dihasilkan oleh pengguna media sosial. Di sinilah analisis sentimen hadir sebagai alat yang kuat untuk mengidentifikasi dan mengklasifikasikan sentimen masyarakat terhadap berbagai isu, termasuk pemilihan *Presiden* [4].

Penelitian ini berfokus pada penerapan analisis sentimen untuk memprediksi hasil calon *Pemilu Presiden* menggunakan dua metode, yaitu *Random Forest dan Lexion Based*. *Random Forest* adalah sebuah metode yang berbasis pada teknik ensemble learning, di mana sejumlah model pohon keputusan digabungkan untuk mencapai keputusan klasifikasi yang lebih akurat dan stabil. Sementara itu, *Lexion Based* adalah pendekatan berbasis leksikon, di mana kamus leksikon yang telah dikurasi secara manual digunakan untuk mengaitkan kata-kata dalam teks dengan sentimen tertentu [5].

Penelitian analisis sentimen sebelumnya telah dilakukan oleh Boma Bayu Baskoro, Irwan Susanto, dan Siti Khomsah. Dalam penelitian tersebut, mereka menggunakan data komentar dari pelanggan hotel di Purwokerto yang diunduh dari situs tripadvisor.co.id sebagai sumber datanya. Metode yang digunakan untuk analisis melibatkan Random Forest Classifier dan TF-IDF [6]. Dalam tahap preprocessing, dilakukan normalisasi, stemming, dan penggunaan kata-kata stopword yang tidak termasuk dalam library sastrawi. Hasil dari penelitian ini menunjukkan bahwa akurasi model mencapai 87,23%. Namun, tanpa proses stemming, akurasi model hanya mencapai 87,01% [7].

Berdasarkan penjelasan di atas, penelitian ini akan menginvestigasi bagaimana analisis sentimen menggunakan metode Random Forest untuk memprediksi pandangan masyarakat terhadap calon presiden. Sentimen akan dibagi menjadi tiga kelas: positif, negatif, dan netral [8]. Data untuk penelitian ini akan diambil dari komentar di Twitter dengan menggunakan Tweet Harvest, dan data teks tersebut akan melalui proses preprocessing sebelum dilakukan klasifikasi. Tujuan dari penelitian ini adalah untuk mengungkap hasil analisis sentimen pada kelas-kelas sentimen tersebut dan memberikan pemahaman yang lebih mendalam dan obyektif tentang pandangan dan sikap masyarakat terhadap calon Pemilu Presiden dengan menggunakan teknologi analisis sentimen. Hasil dari penelitian diharapkan dapat memberikan sumbangan yang berharga bagi praktisi politik, kandidat, dan partai politik dalam merumuskan strategi kampanye yang lebih efektif dan responsif terhadap aspirasi masyarakat.

## 2. TINJAUAN PUSTAKA

### 2.1 Pemilihan Umum (Pemilu)

Pemilihan umum (Pemilu) adalah sebuah proses untuk memilih seseorang yang akan mengisi jabatan politik tertentu. Jabatan tersebut memiliki beragam tingkatan, mulai dari jabatan presiden atau eksekutif, wakil rakyat atau legislatif di berbagai tingkat pemerintahan, hingga kepala desa. Pemilu merupakan salah satu cara untuk mempengaruhi rakyat secara persuasif tanpa memaksa, melalui berbagai kegiatan seperti retorika, hubungan publik, komunikasi massa, lobi, dan kegiatan lainnya [9].

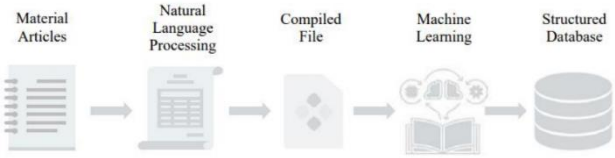
Pemilihan umum dianggap sebagai bagian penting dari proses kenegaraan. Dalam pelaksanaannya, terdapat dua manfaat yang menjadi tujuan dan sasaran langsung, yaitu pembentukan dan pemupukan kekuasaan yang sah (otoritas) serta mencapai tingkat keterwakilan politik yang tinggi (political representativeness) [10].

Ada beberapa model penyelenggaraan pemilu, di antaranya:

- Model Pemilihan Umum Legislatif
- Model Pemilihan Umum Presiden dan Wakil Presiden
- Model Pemilihan Umum Kepala Daerah

### 2.2 Text Mining

Text mining merupakan suatu teknik atau metode analisis yang digunakan untuk menggali dan mengekstraksi informasi berharga dari jumlah besar teks atau dokumen. Dalam teknik ini, algoritma dan pendekatan statistik digunakan untuk memproses, mengorganisasi, dan menyajikan data teks dalam bentuk yang dapat dimengerti dan dianalisis oleh manusia atau sistem komputer. Selain itu, text mining juga dikenal sebagai data mining teks atau penemuan pengetahuan dari database tekstual. Berdasarkan definisi dari buku "The Text Mining Handbook," text mining bisa diartikan sebagai suatu proses dimana seorang pengguna berinteraksi dengan sekumpulan dokumen menggunakan alat analisis yang merupakan komponen-komponen dari data mining [11].




Gambar 1. Proses Text Mining

Analisis sentimen adalah suatu proses dalam text mining yang menggunakan algoritma data mining untuk mengklasifikasikan data tidak terstruktur dan menghasilkan informasi mengenai sentimen secara efisien. Tujuan dari text mining adalah untuk mendapatkan informasi yang berharga dari sekumpulan dokumen. Oleh karena itu, sumber data yang digunakan dalam text mining berupa teks yang tidak terstruktur atau setidaknya semi terstruktur. Beberapa tugas khusus yang dilakukan dalam text mining antara lain adalah pengkategorisasian dan pengelompokkan teks [12]. Dalam memberikan solusi, text mining mengadopsi dan mengembangkan berbagai teknik dari bidang lain, seperti Data Mining, Information Retrieval, Statistik

dan Matematik, Machine Learning, Linguistik, Natural Language Processing, dan Visualisasi. Kegiatan riset dalam text mining meliputi ekstraksi dan penyimpanan teks, preprocessing konten teks, pengumpulan data statistik, serta indexing dan analisis sentimen.

## 2.3 Analisis Sentimen

Analisis sentimen adalah sebuah teknik atau proses yang digunakan untuk mengevaluasi dan mengidentifikasi sentimen, perasaan, atau emosi yang terkandung dalam teks, seperti dokumen, artikel, ulasan, atau pesan media sosial [5].

Tujuan dari analisis sentimen adalah untuk memahami dan mengukur reaksi emosional atau pendapat orang terhadap suatu topik, produk, layanan, atau peristiwa tertentu. Data yang telah terkumpul akan diolah melalui analisis agar dapat memperoleh pandangan yang mendalam sehingga kesimpulan dapat diambil. Hasil analisis sentimen dapat berupa sentimen positif, negatif, atau netral [13].

Analisis Sentimen dapat dibagi menjadi 3 level, yaitu [14]:

a) Level Dokumen: Pada level ini, analisis sentimen bertujuan untuk mengklasifikasikan keseluruhan dokumen ke dalam kelas positif, netral, atau negatif (Schneider, 2005).
b) Level Kalimat: Pada level ini, analisis sentimen bertujuan untuk menentukan sentimen positif atau negatif dari suatu kalimat dengan mempertimbangkan susunan kata dalam kalimat tersebut.
c) Level Aspek: Pada level ini, analisis sentimen bertujuan untuk menentukan sentimen positif, negatif, atau netral berdasarkan atribut dari suatu entitas. Dalam penelitian ini, level aspek diterapkan untuk mengkategorikan ulasan sesuai dengan aspek yang telah ditentukan.

## 2.4 Twitter

Twitter adalah platform jejaring sosial yang memungkinkan penggunanya mengirim dan membaca pesan berbasis teks hingga 140 karakter. Popularitas Twitter yang tinggi membuatnya digunakan untuk berbagai keperluan dalam berbagai aspek, seperti sarana protes, kampanye politik, pembelajaran, dan komunikasi darurat. API (Application Programming Interface) merupakan metode yang memungkinkan program komputer "berbicara" satu sama lain, memungkinkan mereka untuk saling meminta dan menyajikan informasi. Hal ini terjadi dengan memperbolehkan aplikasi perangkat lunak untuk mengakses apa yang disebut sebagai "endpoint," yaitu alamat yang terkait dengan informasi jenis tertentu yang disediakan (endpoint umumnya unik, seperti nomor telepon) [15].

Twitter memungkinkan akses ke bagian dari layanannya melalui API, yang memungkinkan orang untuk membangun perangkat lunak yang terintegrasi dengan Twitter, seperti solusi untuk membantu perusahaan dalam merespons umpan balik pelanggan di Twitter. Data dari Twitter memiliki perbedaan dengan data yang dibagikan oleh kebanyakan platform sosial lain karena data tersebut mencerminkan informasi yang dipilih pengguna untuk dibagikan ke publik [16]. Platform API Twitter menyediakan akses luas ke data Twitter publik yang telah dipilih oleh pengguna untuk dibagikan ke seluruh dunia.

## 2.5 Random Forest

Random Forest merupakan pengembangan dari metode Decision Tree yang menggunakan beberapa Decision Tree. Setiap Decision Tree dilatih menggunakan sampel individu, dan setiap atribut dipilih secara acak pada setiap pohon yang dibentuk. Random Forest memiliki beberapa kelebihan, termasuk kemampuannya untuk meningkatkan akurasi hasil jika ada data yang hilang, serta ketahanan terhadap outlier dan efisiensi dalam penyimpanan data. Selain itu, Random Forest juga memiliki proses seleksi fitur yang mampu mengambil fitur terbaik sehingga dapat meningkatkan kinerja model klasifikasi [17]. Dengan adanya seleksi fitur ini, Random Forest dapat efektif bekerja pada big data dengan parameter yang kompleks.

Proses pembentukan Random Forest dapat dilakukan dengan langkah-langkah berikut [17]:

a. Mengambil sampel bootstrap sebanyak n-tree dari data.
b. Membangun pohon untuk setiap sampel bootstrap. Pada setiap simpul pohon, secara acak dipilih variabel untuk dipisahkan, dan pohon dibiarkan tumbuh hingga setiap node terminal memiliki ukuran kasus yang memadai.
c. Informasi dari masing-masing pohon dalam n-tree diagregasikan untuk memprediksi data baru, seperti melalui pemilihan mayoritas untuk klasifikasi.
d. Menghitung tingkat kesalahan out-of-bag (OOB) dengan menggunakan data yang tidak termasuk dalam sampel bootstrap.

## 2.6 Lexicon Based

Lexicon-Based (berbasis leksikon) merupakan salah satu metode dalam analisis sentimen yang menggunakan kamus leksikon yang telah dikurasi secara manual. Kamus leksikon ini berisi daftar kata-kata atau frasa yang memiliki hubungan dengan sentimen tertentu, seperti positif, negatif, atau netral [18]. Setiap kata atau frase dalam kamus diberi label sentimen yang sesuai.

Lexicon Based adalah fitur kata yang memiliki sentimen positif atau negatif berdasarkan kamus atau lexicon. Proses pelabelan data dilakukan oleh kamus Lexicon Based dengan menghitung skor sentimen. Setelah kata-kata yang mengandung sentimen positif, negatif, dan netral diidentifikasi dalam sebuah kalimat, langkah selanjutnya adalah menghitung setiap kata yang mengandung sentimen dalam kalimat tersebut, dengan menjumlahkan nilai opini. Jumlah nilai opini untuk sentimen positif bernilai 1 atau lebih, sedangkan kata-kata yang netral dalam kalimat diberi nilai = 0, dan sebaliknya, untuk kata-kata yang mengandung sentimen negatif dalam kalimat diberi nilai = -1 atau lebih [19].

## 2.7 Confusion matrix

Confusion Matrix (Matriks Konfusi) adalah tabel yang digunakan untuk mengevaluasi kinerja dari suatu sistem klasifikasi atau model prediksi [20]. Matriks ini memungkinkan kita untuk melihat seberapa baik model tersebut melakukan klasifikasi pada

set data uji dengan membandingkan hasil prediksi dengan nilai sebenarnya.

Tabel 1. Confusion Matrix

| | | Skor Terkini | |
|---|---|---|---|
| | | (Positive) | (Negative) |
| Skor Perkiraan | (Positive) | TP | FP |
| | (Negative) | FN | TN |

Keterangan pada tabel 1 terdapat empat nilai yang dikeluarkan yaitu :

a. *True Positive (TP)* adalah jumlah data positif yang diklasifikasikan dengan benar sebagai nilai positif.

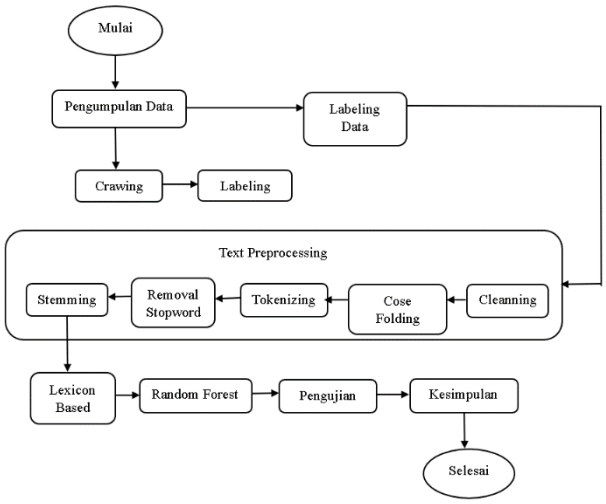
b. *False Negative* adalah jumlah data negatif yang salah diklasifikasikan sebagai nilai positif.

c. *False Positive* adalah jumlah data positif yang salah diklasifikasikan sebagai nilai negatif.

d. *True Negative* (TN) adalah jumlah data negatif yang diklasifikasikan dengan benar sebagai nilai negatif.

## 3. METODOLOGI

Penelitian dilaksanakan pada tanggal 15 Juni 2023 terhadap memprediksi pemilu calon presiden menggunakan bahasa pemrograman Python. Metode yang diimplementasikan dalam penelitian ini adalah Random Forest dan Lexicon Based, yang merupakan metode yang umum digunakan dalam penelitian terkait data mining. Penelitian ini mengikuti lima tahapan utama dalam Knowledge Discovery in Database (KDD), yaitu data selection, preprocessing, transformation, data mining, dan evaluation. Berikut adalah alur penelitian yang dilakukan.




Gambar 2. Alur Penelitian

### 3.1 Pengumpulan Data

Dataset yang akan digunakan dalam penelitian ini adalah kumpulan tweet publik dalam bahasa Indonesia atau data ekstraksi yang diperoleh dari pencarian percakapan seseorang terhadap akun resmi calon tokoh politik yang akan mencalonkan diri sebagai calon presiden Indonesia pada masa yang akan datang, scraping atau crawl data twitter yang dilakukan menggunakan tweet-harvest. Beberapa account resmi yang diambil diantaranya, @Prabowo (Prabowo Subianto), @aniesbaswedan (Anies Baswedan), @ganjarpranowo (Ganjar Pranowo). Dataset diperoleh melalui proses streaming dengan memanfaatkan API Twitter, dan selanjutnya disimpan dalam bentuk database.

### 3.2 Labeling Data

Dengan menggunakan data Twitter yang telah dikumpulkan, peneliti melakukan tahap pelabelan manual untuk menentukan label negatif atau positif. Ahli Bahasa turut membantu dalam proses pelabelan data ini [21].

### 3.3 Text Preprocessing

Text Pre-processing data adalah langkah untuk menyediakan data yang sudah diproses sesuai dengan kebutuhan yang telah ditetapkan. Pada tahap pre-processing ini, data akan diolah secara bertahap agar menjadi lebih mudah diproses oleh sistem. Preprocessing dilakukan dengan menggunakan bantuan alat RapidMiner. Selanjutnya, tahap yang dilakukan dalam preprocessing yaitu [22]:

a. *Cleaning*, merupakan proses penghapusan kata-kata yang tidak relevan seperti URL, hashtag, username, dan data duplikat.
b. *Case Folding*, adalah langkah untuk mengubah seluruh huruf pada kalimat menjadi huruf kecil.
c. *Tokenization*, merupakan tahap memisahkan kalimat menjadi kumpulan kata-kata. Setiap kalimat dipisah menjadi kata-kata individual dan menghilangkan simbol dan angka.
d. *Normalization*, adalah proses menyesuaikan kata-kata yang tidak sesuai atau tidak baku menjadi bentuk kata

baku berdasarkan Kamus Besar Bahasa Indonesia (KBBI), seperti kata-kata singkatan.

e. *Stopword*, yaitu proses untuk menghilangkan kata-kata yang tidak memiliki makna atau kata-kata sambung. Proses stopword menggunakan kamus stopword yang didapat dari hasil penelitian serta kamus stopword yang telah dibuat berdasarkan temuan kata pada dataset yang dimiliki, kemudian gabungkan menjadi satu file.
f. *Stemming*, adalah langkah untuk mengubah kata-kata berimbuhan menjadi bentuk dasarnya. Proses ini dilakukan menggunakan kamus yang dibuat berdasarkan temuan kata pada dataset dan berdasarkan KBBI.

### 3.4 Metode Lexicon Based

Metode Lexicon Based merupakan salah satu teknik analisis teks yang menggunakan kamus atau daftar kata-kata (lexicon) yang sudah dipilih dan diberi nilai sentimen atau emosi tertentu. Metode tersebut digunakan untuk mengidentifikasi dan mengevaluasi sentimen atau emosi yang terkandung dalam teks, seperti dokumen, tweet, atau ulasan [19].

Berikut adalah langkah-langkah umum dalam penerapan metode Lexicon Based:

a. Kamus Sentimen (Sentiment Lexicon)
b. Pra-pemrosesan Teks
c. Pencocokan Kata dengan Kamus Sentimen
d. Perhitungan Skor Sentimen
e. Agregasi Skor Sentimen

VADER (Valence Aware Dictionary and Sentiment Reasoner) adalah sebuah metode analisis berbasis lexicon. Metode VADER menganalisis teks berdasarkan lexicon (suatu perpustakaan kata-kata) yang menghasilkan klasifikasi sentimen berupa positif, negatif, dan netral, ditambah dengan skor total atau compound score. Kamus lexicon yang digunakan oleh VADER disebut Vader Sentiment Lexicon, yang berisi 7.500 token yang mencakup kata-kata bahasa Inggris, emoticon, akronim, dan inisial yang terkait dengan sentimen.

### 3.5 Random Forest

Salah satu metode yang digunakan untuk pengklasifikasian dan regresi adalah Random Forest. Metode Random Forest merupakan sebuah ensemble (kumpulan) metode pembelajaran yang menggunakan pohon keputusan sebagai base classifier yang dibangun dan dikombinasikan. Beberapa aspek penting dalam metode Random Forest antara lain melakukan bootstrap sampling untuk membangun pohon prediksi, masing-masing pohon keputusan memprediksi dengan prediktor acak, dan Random Forest sendiri melakukan prediksi dengan mengkombinasikan hasil dari setiap pohon keputusan melalui majority vote untuk klasifikasi dan rata-rata untuk regresi [17].

Random Forest dapat dibangun menggunakan metode bagging dengan pemilihan atribut acak. Selain itu, metode CART (Classification and Regression Tree) dapat digunakan untuk menumbuhkan pohon keputusan, dimana pohon keputusan tersebut dibiarkan tumbuh hingga mencapai ukuran maksimum dan tidak akan dipangkas, sehingga menghasilkan kumpulan pohon yang disebut forest [14].

### 3.6 Pengujian

Pada tahap Pengujian ini, peneliti akan menentukan nilai akurasi dari data Twitter sebanyak 3862 data. Melalui metode Lexicon Based dan Random Forest, keseluruhan data tersebut akan diklasifikasikan untuk mendapatkan hasil analisis sentimen. Hasil dari pengujian yang mencakup data yang diklasifikasikan dengan benar dan yang diklasifikasikan dengan salah akan direpresentasikan dalam sebuah tabel yang disebut Confusion Matrix.

## 4. HASIL DAN PEMBAHASAN

### 4.1. Pengumpulan Data

Data tweet diperoleh dari media sosial Twitter dengan menggunakan kata kunci pencarian seperti #calonpresiden, #prabowo, #anies, dan #ganjar. Data tersebut merupakan tweet berbahasa Indonesia yang diunduh dari Twitter, dengan total jumlah data sebanyak 3862 tweet. Setelah itu, data tersebut akan disimpan ke dalam database dengan format csv.

| | Tweet | Author |
|---|---|---|
| 0 | Sulsel siap dukung prabowo | @andimuhikhsan434 |
| 1 | Anies baswedan hebat | @zanimanay6442 |
| 2 | Ganjar yang menang | @jancok9564 |
| 3 | Lombok selalu dukung prabowo, walau ada Dr. TG... | @nursinsian9647 |
| 4 | Kalau saya prabowo atau anies presiden bagus, ... | @nursinsian9647 |
| ... | ... | ... |
| 3857 | Cuan cuan | @SADULURMANCING |
| 3858 | SURVEY YG TDK BISA DIPERCAYA | @abahabduh6160 |
| 3859 | Lagi musim cuan ni... Mayan klw buka usaha sur... | @fachrialbantany6210 |
| 3860 | Waktu pilgub dki menurut survei anies kalah da... | @alexrey8365 |
| 3861 | Ganjar rak payu sebagian jowo .fanatik bola Yo... | @mamasoleng886 |

Gambar 3. Dataset

Datasetnya disimpan dalam format CSV. Dataset berisi *Tweet dan Author*. Selanjutnya adalah tahap *preprocessing* adalah proses persiapan data sebelum diaplikasikan pada model atau algoritma *machine learning.* Tujuannya adalah untuk membersihkan, mengubah format, dan mempersipakan data mentah agar lebih cocok untuk analisis pemodelan. Tahap preprocessing melibatkan langkah-langkah seperti pembersihan data, tranformasi data, penghapusan duplikat, normalisasi, encoding, kategori, reduksi dimensi, dan penghapusan fitur irrelevant.

Tabel 2. Preprocessing

| No | Sebelum *preprocessing* | Sesudah *preprocessing* | Candidat |
|---|---|---|---|
| 1. | Sulsel siap dukung prabowo | sulsel siap dukung prabowo | Prabowo |
| 2. | Dari tiga calon Anis terbaik cerdas dan berprestasi dunia luar mengakuinya | dari tiga calon Anis terbaik cerdas dan berprestasi dunia luar mengakuinya | Anis |
| 3. | GANJAR SEMAKIN DI DEPAN presiden 2024..... | ganjar semakin di depan presiden 2024 | Ganjar |

### 4.1.1 Cleaning

Pada proses ini, pesan teks pada dataset dijalani pembersihan untuk menghilangkan teks yang tidak diinginkan, seperti URL atau tautan, username atau mentions, dan hashtag.

Tabel 3. Cleaning

| Sebelum |
|---|
| @hagoapk6118 emang beda banget.sekarang PENDUKUNG pak Prabowo bener2 mencerminkan kedamaian dan kerukuna sesama bangsa Indonesia biarpun beda pilihan |
| **Sesudah** |
| emang beda banget.sekarang PENDUKUNG pak Prabowo bener2 mencerminkan kedamaian dan kerukuna sesama bangsa Indonesia biarpun beda pilihan |

### 4.1.2 Case Folding

Pada tahap ini, dataset yang telah dibersihkan sebelumnya akan dilakukan langkah untuk mengubah semua huruf menjadi huruf kecil (lower case).

Tabel 4. Case Folding

| Sebelum |
|---|
| Kalau saya prabowo atau anies presiden bagus, tapi ganjar no ngomong gaya megawati, mega jadi presiden bukan pihan rakyat, karena pak gusdur, sby end jokowi karena rakyat. |
| **Sesudah** |
| kalau saya prabowo atau anies presiden bagus, tapi ganjar no ngomong gaya megawati, mega jadi presiden bukan pihan rakyat, karena pak gusdur, sby end jokowi karena rakyat. |

### 4.1.3 Normalization

Selanjutnya, tahap normalization dilakukan untuk menyesuaikan kata-kata yang tidak sesuai berdasarkan Kamus Besar Bahasa Indonesia (KBBI), seperti kata-kata singkatan. Proses ini menggunakan operator dari RapidMiner dan kamus, serta mengelola dataset dengan melakukan normalisasi kata-kata singkatan dan kata-kata gaul.

Tabel 5. Normalization

| Sebelum |
|---|
| Saya akan pilih prabowo krna misi parbowo ingin membangkitkan mrnjadikan pertanian salah satu kekuatan senjata indonesia bersaing di ekonomi dunia bila prabowo terpilih semoga tepat janji. |
| **Sesudah** |
| Saya akan pilih prabowo karena misi parbowo ingin membangkitkan menjadikan pertanian salah satu kekuatan senjata indonesia bersaing di ekonomi dunia bila prabowo terpilih semoga tepat janji. |

### 4.1.4 Tokenize

Kemudian, dilanjutkan dengan tahap tokenize, di mana kalimat akan dipisah menjadi kata-kata individual dan simbol serta angka akan dihilangkan.

| normalized_text | tokenized_text |
|---|---|
| sulsel siap dukung prabowo | [sulsel, siap, dukung, prabowo] |
| anies baswedan hebat | [anies, baswedan, hebat] |
| ganjar yang menang | [ganjar, yang, menang] |
| lombok selalu dukung prabowo walau ada dari tg... | [lombok, selalu, dukung, prabowo, walau, ada, ... |
| kalau saya prabowo atau anies presiden bagus t... | [kalau, saya, prabowo, atau, anies, presiden, ... |
| ... | ... |
| cuan cuan | [cuan, cuan] |
| survey yang tidak bisa dipercaya | [survey, yang, tidak, bisa, dipercaya] |
| lagi musim cuan ni mayan klw buka usaha survey ya | [lagi, musim, cuan, ni, mayan, klw, buka, usah... |
| waktu pilgub dki menurut survei anies kalah da... | [waktu, pilgub, dki, menurut, survei, anies, k... |
| ganjar rak payu sebagian jowo fanatik bola yo ... | [ganjar, rak, payu, sebagian, jowo, fanatik, b... |

Gambar 4. Tokenize

### 4.1.5 Stopword

Untuk pada tahap stopword, dataset akan mengalami proses pemfilteran kata yang tidak memiliki makna, seperti kata sambung. Langkah ini menggunakan operator stopword yang tersedia di RapidMiner dan kamus. Kamus yang digunakan merupakan gabungan dari kamus stopword hasil penelitian dan kamus yang dibuat oleh peneliti berdasarkan temuan kata tanpa makna dari dataset, dengan format berupa file ".txt".

Tabel 6. Contoh Stopword

| Sebelum |
|---|
| Lombok selalu dukung prabowo, walau ada Dr. TGB. Muhammad Zainul Majdi dipihak ganjar, lombok tetap dukung prabowo. |
| **Sesudah** |
| Lombok selalu dukung prabowo, ada Dr. TGB. Muhammad Zainul Majdi dipihak ganjar, lombok tetap dukung prabowo. |

## 4.2. Labeling

Labeling data yang tepat dan akurat sangat penting untuk menciptakan model pembelajaran mesin yang efektif dan

memiliki performa yang baik dalam memprediksi atau mengklasifikasikan data baru. Setelah proses preprocessing selesai, langkah selanjutnya adalah melakukan pelabelan dengan menggunakan metode Lexicon based dan Random Forest dengan memanfaatkan library vader sentiment. Untuk mendapatkan pelabelan yang sesuai yaitu dengan cara menghitung akurasi dari *lexicon based* dan melakukan pelabelan sentimen pada setiap teks, yang mana dengan cara melihat kata mana yang positif, negatif, dan netral pada data yang sudah ada.

| final_text | label_words |
|---|---|
| sulsel siap dukung prabowo | positif |
| anies baswedan hebat | positif |
| ganjar menang | positif |
| lombok selalu dukung prabowo ada tgb muhammad zainul majdi dipihak ganjar lombok tetap dukung prabowo | positif |
| kalau prabowo anies presiden bagus ganjar no ngomong gaya megawati mega jadi presiden bukan pihan rakyat pak gusdur sby end jokowi rakyat | positif |
| prabowo | positif |
| prabowo jelas | positif |
| prabowocalonpemimpinsemakinberkelassemogahbapakri | netral |
| belum tau siapa paling unggul daerah jelas paling unggul ganjar pranowo lah | negatif |
| kalo survey anies baswedan paling dibawah karna pernah disurvey kalau survey prabowo ganjar selalu diatas karna kalian kuasai media | positif |
| jangann indonesia dipegang anies | positif |
| sah prabowo sumatra win tinggal daerah jawa pak semoga clear hari h | positif |
| pabowo menang satos | positif |
| nomer saya idamkan | netral |
| pastikan dulu jawa barat menang telak amin | positif |

Gambar 4. Labeling

### 4.3 Topic Modelling

Topic modelling sangat bermanfaat dalam mengenali pola, tren, dan wawasan yang terkandung dalam dataset teks yang besar dan rumit. Dengan mengelompokkan dokumen ke dalam topik-topik yang saling terkait, topic modelling memungkinkan analis untuk melakukan eksplorasi dan pemahaman terhadap data teks yang kompleks dengan lebih efisien dan efektif [23].

Di samping analisis sentimen, penelitian ini juga melibatkan analisis mengenai topic modelling. Selanjutnya, untuk melihat topik-topik apa saja yang menjadi bahasan pada pencarian, digunakan metode wordcloud. Topic Modelling merupakan suatu teknik dalam analisis teks yang bertujuan untuk mengidentifikasi dan menemukan pola topik atau tema yang tersembunyi dalam kumpulan besar dokumen atau teks. Tujuan utama dari topic modelling yaitu untuk mengelompokkan dokumen yang serupa berdasarkan konten mereka dan mengidentifikasi kata-kata kunci atau topik utama yang muncul dalam setiap kelompok tersebut.


Gambar 5. Wordcloud Positif LB


Gambar 6. Wordcloud Negatif LB


Gambar 7. Wordcloud Netral LB

Dari hasil gambar diatas menggunakan Lexicon Based yang terdapat gambar pertama Wordcloud Positif yang mana ini menunjukan kata positif dari data, lalu gambar kedua Wordcloud Negatif ini menunjukan kata negatif apa saja yang ada pada data, dan gambar terakhir Wordcloud Netral menunjukan kata netral pada data yang sudah ada.


Gambar 8. Wordcloud Positif RF


Gambar 9. Wordcloud Negatif RF


Gambar 10. Wordcloud Netral RF

Dari hasil gambar diatas menggunakan Lexicon Based yang terdapat gambar pertama Wordcloud Positif yang mana ini menunjukan kata positif dari data, lalu gambar kedua Wordcloud Negatif ini menunjukan kata negatif apa saja yang ada pada data, dan gambar terakhir Wordcloud Netral menunjukan kata netral pada data yang sudah ada.

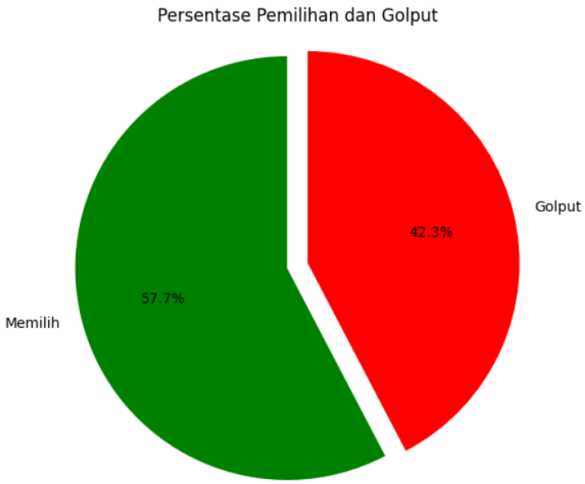
### 4.3. Analysis Distribusi Data Sesuai Dengan Kandidat Calon Presiden

Pada tahap ini kita melakukan proses menganalisis dan memvisualisasikan data yang berkaitan dengan dukungan atau preferensi pemilih terhadap calon presiden tertentu. Tujuan dari analisis ini adalah untuk memahami sebaran atau pola dari dukungan pemilih terhadap masing-masing kandidat calon presiden. Analisis distribusi data sesuai dengan kandidat calon presiden merupakan salah satu aspek penting dalam pemilihan umum atau proses pemilihan kepala negara. Hasil analisis ini dapat membantu pemilih, analis politik, dan calon presiden dalam memahami pola dukungan pemilih dan merancang strategi kampanye yang lebih efektif.

Tabel 7. Analisis Sentimen

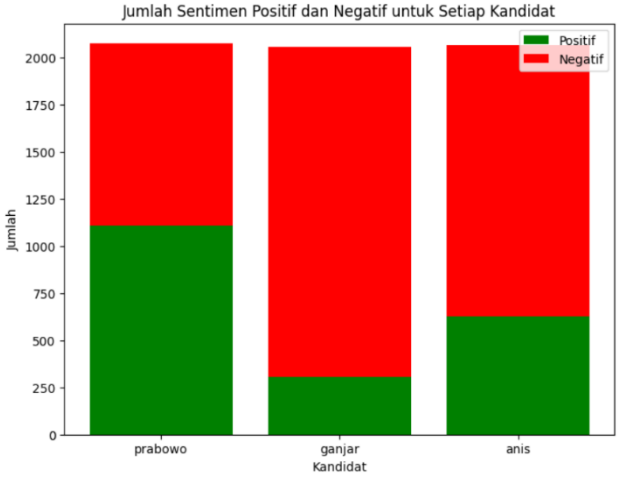
| No | Kandidat | Positif | Negatif |
|---|---|---|---|
| 1. | Prabowo | 1108 | 968 |
| 2. | Anis | 304 | 1754 |
| 3. | Ganjar | 628 | 1439 |

Pada gambar dibawah ini menunjukan bahwa dalam persentase yang dilakukan peneliti memiliki 57,7% yang memilih dan 42,3% yang golput.




Gambar 8. Visualisasi Pie Chart

Kemudian peneliti melakukan Visualisasi Data Sentimen dengan menggunakan Bar plot untuk mengetahui siapa kandidat calon presiden yang dipilih oleh masyarakat atau netizen pada pemilu 2024 mendatang.



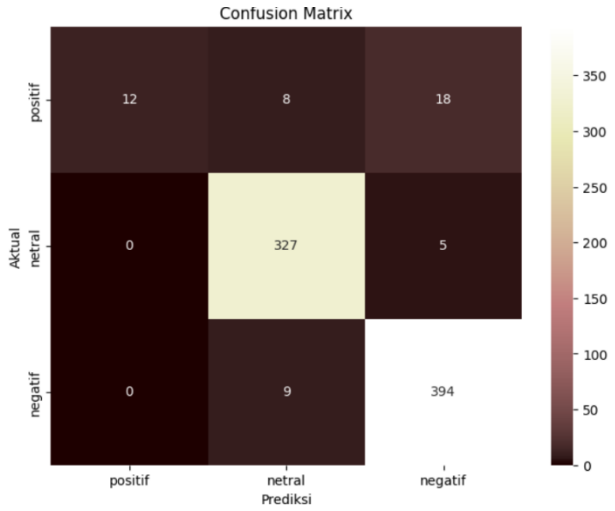

Gambar 10. Analisis Sentimen

Dari hasil analisis sentiment tersebut terdapat data sebanyak 57.6% masayarak memilih kandidat nya dan 42.4% masyarakat tidak mendukung kandidat satu pun, dengan demikian dukungan terbanyak yang terdapat pada data tersebut adalah untuk prabowo kemudian anies baswedan, dan posisi terkahir adalah ganjar.

### 4.4. Pengujian Menggunakan Confusion Matrix

Pada tahap pemodelan klasifikasi, digunakan 80% data untuk training dan 20% data untuk testing dengan metode random forest. Hasil klasifikasi menggunakan metode random forest akan digunakan untuk mengukur performa atau kinerja model dengan menggunakan confusion matrix, yang merupakan hasil dari transformasi data review menjadi data numerik.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negatif | 1.00 | 0.32 | 0.48 | 38 |
| netral | 0.95 | 0.98 | 0.97 | 332 |
| positif | 0.94 | 0.98 | 0.96 | 403 |
| accuracy | | | 0.95 | 773 |
| macro avg | 0.97 | 0.76 | 0.80 | 773 |
| weighted avg | 0.95 | 0.95 | 0.94 | 773 |

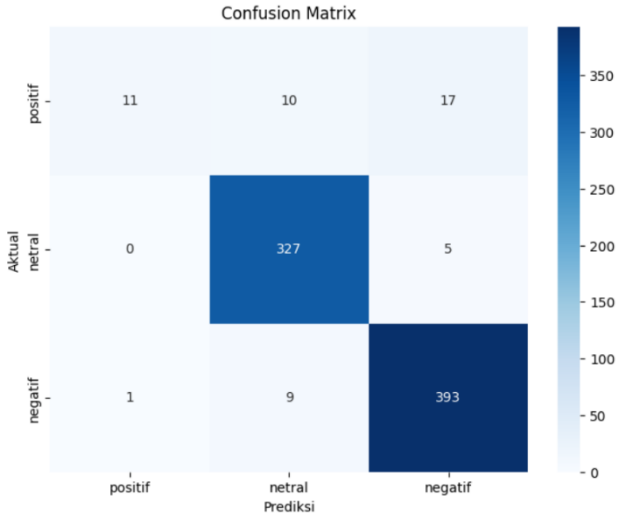



Gambar 11. Confusion Matrix Lexicon Based

Terdiri dari 3 kelas: "negatif", "positif", dan "netral". Diagonal utama dari tabel tersebut menunjukkan jumlah data yang diprediksi dengan benar untuk setiap kelas. Misalnya, terdapat 12 data yang sebenarnya "positif dan diprediksi dengan benar sebagai "positif", 327 data yang sebenarnya "netral" dan diprediksi dengan benar sebagai "netral", dan 394 data yang sebenarnya "negatif" dan diprediksi dengan benar sebagai "negatif".

Accuracy Random Forest : 0.9456662354463131
Confusion Matrix:
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negatif | 0.92 | 0.29 | 0.44 | 38 |
| netral | 0.95 | 0.98 | 0.96 | 332 |
| positif | 0.95 | 0.98 | 0.96 | 403 |
| accuracy | | | 0.95 | 773 |
| macro avg | 0.94 | 0.75 | 0.79 | 773 |
| weighted avg | 0.94 | 0.95 | 0.94 | 773 |




Gambar 12. Confusion Matrix Random Forest

Dalam kasus ini, terdapat beberapa kesalahan prediksi yang terlihat dari nilai di luar diagonal utama. Misalnya, terdapat 10 data yang sebenarnya "negatif", namun salah diprediksi sebagai "positif". Terdapat pula 17 data yang sebenarnya "positif", namun salah diprediksi sebagai "negatif". Demikian pula, ada 5 data yang sebenarnya "netral", namun salah diprediksi sebagai "negatif", dan seterusnya.

## 5. KESIMPULAN

Dari hasil yang diperoleh menggunakan metode Lexicon Based dan Random Forest, berikut adalah kesimpulan dan saran yang dapat diambil:

1. Metode Random Forest menunjukkan tingkat akurasi yang tinggi, dengan nilai akurasi mencapai 94%.
2. Dalam penelitian ini, dua metode yang digunakan untuk analisis sentimen adalah Random Forest dan Lexicon Based. Random Forest adalah salah satu algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi, sementara Lexicon Based adalah teknik yang menggunakan kamus atau daftar kata-kata dengan nilai sentimen untuk mengidentifikasi emosi dalam teks.
3. Sistem hasil dari analisis sentiment memprediksi kandidat calon presiden sudah tepat dan sesuai dengan sentiment.
4. Berdasarkan penelitian yang dilakukan pada dataset kumpulan opini pengguna Twitter tentang calon presiden, dapat disimpulkan dari hasil perbandingan penerapan metode Lexicon Based dan Random Forest pada analisis sentimen bahwa Lexicon Based menghasilkan tingkat sentimen negatif sebesar 48%, sentimen positif sebesar 96%, dan sentimen netral sebesar 97%, dengan akurasi total model mencapai 88%.
   Sedangkan menggunakan *Random Forest* pada analisis sentiment dapat disimpulkan bahwa model memiliki performa yang sangat baik dalam mengenali data dengan sentimen "negatif" (Precision, Recall, dan F1-Score untuk kelas "negatif" mendekati 1.00).

5. Lexicon based Cenderung tidak akurat untuk teks yang mengandung banyak kalimat majemuk atau kalimat yang ambigu secara sentimen. Sedangkan Random Forest dapat memberikan hasil yang lebih stabil dan generalisasi yang lebih baik, tetapi masih memiliki ruang untuk peningkatan dalam mengklasifikasikan sampel pada kelas "negatif".

## DAFTAR PUSTAKA


[1] U. Rauta, "Menggagas Pemilihan Presiden yang Demokratis dan Aspiratif," *J. Konstitusi*, vol. 11, no. 3, p. 600, 2016, doi: 10.31078/jk11310.

[2] F. F. Sultan and Silviana Purwanti, "Pembentukan Opini Publik Pada Akun Twitter Pribadi Novel Baswedan," *eJournal lmu Komunikasi, 2022, 10 155-164 ISNN 2502-597X ISNN 2502-5961 ejournal.ip.fisip-unmul.org © Copyr. 2022*, vol. 10, no. 4, pp. 155–164, 2022.

[3] R. Vindua and A. U. Zailani, "Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python," *JURIKOM (Jurnal Ris. Komputer)*, vol. 10, no. 2, p. 479, Apr. 2023, doi: 10.30865/jurikom.v10i2.5945.

[4] C. Juditha, "Political Marketing dan Media Sosial," *J. Stud. Komun. dan Media*, vol. 19, no. 2, pp. 225–242, 2015.

[5] P. A. Permatasari, L. Linawati, and L. Jasa, "Survei Tentang Analisis Sentimen Pada Media Sosial," *Maj. Ilm. Teknol. Elektro*, vol. 20, no. 2, p. 177, 2021, doi: 10.24843/mite.2021.v20i02.p01.

[6] B. Bayu Baskoro *et al.*, "Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)," *Inista*, vol. 3, no. 2, pp. 021–029, 2021, doi: 10.20895/INISTA.V3.

[7] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *... Teknol. Inf. dan ...*, vol. 6, no. 9, pp. 4305–4313, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[8] C. P. Yanti, N. W. E. Agustini, and ..., "Perbandingan Metode K-NN Dan Metode Random Forest Untuk Analisis Sentimen pada Tweet Isu Minyak Goreng di Indonesia," *J. Media ...*, vol. 7, no. April, pp. 756–765, 2023, doi: 10.30865/mib.v7i2.5900.

[9] A. B. Azed, "Sistem Pemilihan Umum di Indonesia," *J. Huk. dan Pembang.*, vol. 17, no. 2, pp. 170–180, 1987.

[10] A. E. Subiyanto, "Pemilihan Umum Serentak yang Berintegritas sebagai Pembaruan Demokrasi Indonesia," *J. Konstitusi*, vol. 17, no. 2, p. 355, 2020, doi: 10.31078/jk1726.

[11] A. Firdaus and W. I. Firdaus, "Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan)," *J. JUPITER*, vol. 13, no. 1, p. 66, 2021.

[12] F. Nurhuda, S. W. Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," vol. 2, no. 2, 2013.

[13] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 2, p. 183, 2020, doi: 10.26418/justin.v8i2.36776.

[14] M. H. Chyntia, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 4, pp. 1702–1708, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[15] L. I. Ambronsius and A. Rahman, "Menganalisis Pengaruh Platform Twitter Sebagai Alat Komunikasi Kpopers dalam Berspesialisasi Penyebaran Informasi Analyzing the Influence of the Twitter Platform as a Kpopers Communication Tool in Specializing in Information Dissemination," *Pinisi J. Art, Humanit. Soc. Stud.*, vol. 2, no. 6, pp. 234–240, 2022.

[16] M. Hidayatullah *et al.*, "Sentiment Analysis of Police Performance On Twitter Users Using Naïve Bayes Method," *RISTEC Res. Inf. Syst. Technol.*, vol. 2, no. 2, pp. 113–125, 2021, doi: 10.31980/ristec.v2i2.1945.

[17] R. M. Awangga and N. H. Khonsa', "Analisis Performa Algoritma Random Forest dan Naive Bayes Multinomial pada Dataset Ulasan Obat dan Ulasan Film," *InComTech J. Telekomun. dan Komput.*, vol. 12, no. 1, p. 60, 2022, doi: 10.22441/incomtech.v12i1.14770.

[18] D. Winarso, Yanda Noor Yudha, and Syahril, "Analisis Sentimen Masyarakat Pada Twiter Terhadap Isu Covid-19 Menggunakan Metode Lexicon Based," *J. Fasilkom*, vol. 11, no. 2, pp. 97–103, 2021, doi: 10.37859/jf.v11i2.2772.

[19] P. A. Sumitro, Rasiban, D. I. Mulyana, and W. Saputro, "Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based," *J-ICOM - J. Inform. dan Teknol. Komput.*, vol. 2, no. 2, pp. 50–56, 2021, doi: 10.33059/j-icom.v2i2.4009.

[20] E. V. Tjahjadi and B. Santoso, "Klasifikasi Malware Menggunakan Teknik Machine Learning," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 60–70, 2023.

[21] N. Sari, P. Juana, E. Haerani, F. Syafria, and E. Budianita, "Analisis Sentimen Tanggapan Masyarakat Terhadap Calon Presiden 2024 Ridwan Kamil Menggunakan Metode Naive Bayes Classifier," vol. 4, pp. 570–576, 2024, doi: 10.30865/json.v4i4.6168.

[22] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16–22, 2021, doi: 10.30813/jbase.v4i2.3000.

[23] A. Nurlayli and M. A. Nasichuddin, "Topik Modeling Penelitian Dosen Jptei Uny Pada Google Scholar Menggunakan Latent Dirichlet Allocation," *Elinvo (Electronics, Informatics, Vocat. Educ.*, vol. 4, no. 2, pp. 154–161, 2019, doi: 10.21831/elinvo.v4i2.28254.


## BIODATA PENULIS

**Oktaviami Manullang**
Mahasiswi Pendidikan D4 Teknik Informatika Universitas Logistik dan Bisnis Internasional Bandung.

**Cahyo Prianto, S.Pd., M.T .,CDSP, SFPC.**

Merupakan Kasubag. Akademik dan Akreditasi Program Studi Sarjana Terapan Teknik Informatika Sekaligus Dosen Universitas Logistik dan Bisnis Internasional Bandung.

**Nisa Hanum Harani, S.Kom., M.T., CDSP, SFPC**
Dosen Program Studi Sarjana Terapan Teknik Informatika, Universitas Logistik dan Bisnis Internasional Bandung.