Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Analisis Sentimen X Terhadap Pemilihan Presiden Indonesia 2024 dengan Metode K-Nearest Neighbor

Tri Allan Siddiq[*], Muhammad Ikhsan

Fakultas Sains dan Teknologi, Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
Email: [1,*]triallansiddiq@gmail.com, [2]mhd.ikhsan@uinsu.ac.id
Email Penulis Korespondensi: triallansiddiq@gmail.com
Submitted: 16/08/2024; Accepted: 29/08/2024; Published: 30/08/2024

Abstrak−Pemilihan presiden (Pilpres) yang bakal berlangsung pada tahun 2024, Analisis sentimen dilakukan untuk mengetahui kecenderungan opini seseorang terhadap sebuah peristiwa atau masalah, apakah cenderung positif atau negatif. Tujuan dari penelitian ini adalah Untuk menerapkan Algoritma *K-Nearest Neighbor* pada klasifikasi opini Masyarakat terhadap pilpres 2024 Dan Menghasilkan klasifikasi dari penerapan metode algoritma *K-Nearest Neighbor* pada opini Masyarakat terhadap pilpres 2024 di media sosial X. Berdasarkan dari hasil penelitian yang telah dilakukan, maka dapat disimpulkan bahwa Analisis sentimen menggunakan metode K-Nearest Neighbor (KNN) telah terbukti efektif dalam mengidentifikasi dan memahami sentimen masyarakat terkait Pilpres Indonesia 2024. Tanggapan masyrakat terhadap fenomena tren dan opini publik terhadap calon kandidat presiden di tahun 2024 dinilai menunjukkan sikap positif, seperti yang tergambar dal**a**m analisis sentimen dengan menggunakan kamus *lexicon.* Dari sekitar 1.000 *tweet* yang telah dianalisis, sebanyak 211 di antaranya menunjukkan sentimen yang positif, 175 mengungkapkan sentimen yang negatif, sedangkan 614 lainnya mengungkapkan sentimen yang Netral. Data ini dikumpulkan mulai dari 28 November 2023 hingga 28 April 2024. Selain itu, penelitian ini juga mengidentifikasi kata-kata yang sering muncul dalam *tweet* berbahasa Indonesia**.** Pada *K-Nearest Neighbor (KNN),* hasil yang didapat akurasi *use training set* mendapatkan hasil *acuracy* sebesar 100% , *precision* sebesar 100%, *recall* sebesar 100% dan *f-measure* sebesar 100%, *10-fold cross validation* yang didapat mendapatkan hasil *acuracy* sebesar 92,5%, *precision* sebesar 100% *recall* sebesar 91% dan *f-measure* sebesar 94%, dan yang terakhir 80% *percentage split* mendapatkan hasil *acuracy* sebesar 88,55% , *precision* sebesar 100%, *recall* sebesar 87% dan *f-measure* sebesar 93,04%. Metode klasifikasi *algoritma K-Nearest Neighbor (K-NN)* menggunakan pengujian 80% *percentage split* sangat baik dalam pengujian klasifikasi memiliki *acuracy*, *precision*, *recall* dan *f-measure* lebih besar dibandingkan dengan pengujian *10-fold cross validation*.

Kata Kunci: Analisis Sentimen; Metode K-Nearest Neighbor; Pemilihan Presiden

Abstract−The presidential election which will take place in 2024, sentiment analysis is carried out to determine the tendency of a person's opinion towards an event or problem, whether it tends to be positive or negative. The purpose of this study is to apply the K-Nearest Neighbor Algorithm to the classification of public opinion on the 2024 presidential election and produce a classification of the application of the K-Nearest Neighbor algorithm method to public opinion on the 2024 presidential election on social media X. Based on the results of the research that has been done, it can be concluded that sentiment analysis using the K-Nearest Neighbor (KNN) method has proven effective in identifying and understanding public sentiment related to the 2024 Indonesian presidential election. The community's response to the phenomenon of trends and public opinion towards prospective presidential candidates in 2024 is considered to show a positive attitude, as illustrated in sentiment analysis using a lexicon dictionary. Of the approximately 1,000 tweets that have been analyzed, 211 of them show a positive sentiment, 175 express a negative sentiment, while the other 614 express a Neutral sentiment. This data was collected from November 28, 2023 to April 28, 2024. In addition, this study also identified words that frequently appear in Indonesian tweets. In K-Nearest Neighbor (KNN), the results obtained by the accuracy of the use training set get an accuracy of 100%, precision of 100%, recall of 100% and f-measure of 100%, 10-fold cross validation obtained get an accuracy of 92.5%, precision of 100% recall of 91% and f-measure of 94%, and the last 80% percentage split get an accuracy of 88.55%, precision of 100%, recall of 87% and f-measure of 93.04%. The K-Nearest Neighbor (K-NN) algorithm classification method using 80% percentage split testing is very good in classification testing has greater accuracy, precision, recall and f-measure compared to 10-fold cross validation testing.

Keywords: Sentiment Analysis; K-Nearest Neighbor Method; Presidential Election

1. PENDAHULUAN

Pemilihan umum di dalam sejarah nasional Indonesia telah dilaksanakan beberapa kali, namun pemilihan umum yang dilakukan langsung oleh masyarakat Indonesia baru pertama kali dimulai pada era reformasi setelah era orde baru runtuh yaitu tahun 2004. Pemilihan umum presiden yang akan diselenggarakan pada tahun 2024 merupakan momen yang penting untuk mewujudkan demokrasi dalam Negara Kesatuan Republik Indonesia.[1] Kandidat dan tim sukses pada pemilihan presiden 2024 ini dapat memanfaatkan media sosial untuk menyampaikan pesan kampanye, salah satu media yang aktif digunakan untuk kampanye adalah X. Pada pemilihan 2024 yang akan datang sebanyak 55% atau 112.643.972 orang merupakan pemilih pemula atau generasi milenial dan angka ini meningkat dari pemilihan presiden pada tahun 2019 yang hanya sekitar 90 juta orang, sehingga kekuatan sosial media tidak dapat dianggap remeh untuk elektabilitas ketiga pasang calon presiden dan calon wakil presiden

Pemilihan presiden (Pilpres) yang bakal berlangsung pada tahun 2024 sudah terasa mulai saat ini, seperti sosial media sebagai tempat untuk menyampaikan pandangan, sentimen, dan preferensi tokoh politik yang namanya kerap bagus dalam lembaga survei. Salah satunya, pembahasan yang sedang hangat di perbincangkan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

saat ini adalah mengenai Pilpres 2024 dari kiriman opini masyarakat di X dengan jumlah data *tweet* sangat banyak sehingga menimbulkan pandangan positif dan negatif, kemudian dari kumpulan data *tweet* mengenai pembahasan ini dapat di jadikan sebagai data untuk di olah atau di analisa sesuai dengan kebutuhan.[2]

Jejaring sosial seperti X sekarang menjadi perangkat komunikasi yang sangat populer di kalangan pengguna dunia maya dan dapat digunakan sebagai media kampanye peserta pemilihan presiden dalam menyampaikan citra positif bagi pasangan masing-masing pada calon pemilih dan pendukungnya.[1] Para kandidat yang berlaga dalam berbagai pemilihan umum di seluruh Indonesia pada tahun 2024 terlihat memanfaatkan X dan sosial media lainnya untuk membagikan slogan dan kebijakan, mengguncang popularitas saingan, dan membangun massa menjelang kampanye. Menurut laporan yang dirilis oleh *Wearesocial* pada Oktober 2023 dijelaskan bahwa pengguna internet di Indonesia adalah 212,9 juta dan pengguna sosial media aktif mencapai 167 juta. Menurut data yang dirilis juga bahwa Indonesia merupakan 5 terbesar pengguna X di dunia di bawah negara Brazil dengan angka 24 Juta pengguna.

Analisis sentimen disebut juga dengan *opinion mining* (penambangan opini) yaitu proses untuk mengekstrak suatu opini atau pendapat dari dokumen untuk topik tertentu.[3] Analisis sentimen dilakukan untuk mengetahui kecenderungan opini seseorang terhadap sebuah peristiwa atau masalah, apakah cenderung positif atau negatif dan teknik yang digunakan adalah *Text Mining*. *Text Minin*g merupakan teknik penambangan data teks yang bertujuan untuk mendapatkan kembali informasi yang ada pada data teks, yang di ekstrak secara otomatis dari sumber-sumber data teks yang digunakan sebagai dataset. Pada penelitian terkait analisis sentimen, digunakan data set dari pendapat atau opini masyarakat.[4]

Peneliti melakukan literatur terhadap penelitian-penelitian terdahulu yang masih relevan terhadap permasalahan dalam penelitian ini. Adapun penelitian terdahulu yaitu : Menurut penelitian yang dilakukan Faiza R Irawan, Ahmad Jazuli dkk, 2022, dengan judul "Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearest Neighbor" di mana menghasilkan sebuah hasil mendapatkan akurasi yang tinggi dengan 79,43% dengan menggunakan 1.409 data menggunakan metode K-Nearest Neighbor dengan sentimen positif sebesar 47.59%, negative 8.66% dan netral 43.75%[5]. Pada penelitian yang dilakukan oleh Jeremy Andre Septian, Tresna Maulana Fahrudin dkk, 2019,dengan judul "Analisis Sentimen Pengguna X Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor" di mana dari 2000 data tweet terkait polemik persepakbolaan Indonesia yang diambil dari akun X PSSI dengan proporsi 790 tweet positif dan 1210 tweet negatif dengan metode K-Nearest Neighbor di dapatkan akurasi tinggi sebesar 79.99% dan eror rate 20.01%[6]. Pada penelitian yang di lakukan Susan Mayang Sari, Mhd Furqan dkk, 2022, dengan judul "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia" pada penelitian ini mendapatkan hasil penerapan dalam klasifikasi opini Masyarakat menggunakan pelabelan kelas kamus lexicon bahwa kelas sentimen positif lebih unggul berjumlah 811 di bandingkan kelas sentimen negatif berjumlah 189. Dengan kata lain Masyarakat Indonesia dalam memberi opini atau tanggapan sangat baik terhadap kebijakan new normal di era covid-19[7].Yang terakhir pada penelitian yang dilakukan oleh Haris dan Dian Eka Ratnawati dengan judul "Analisis Sentimen Berbasis Aspek terhadap Data Ulasan Menggunakan Metode *K-Nearest Neighbor* (Studi Kasus: Aplikasi Olsera POS) pada penelitian ini menghasilkan sebuah pelabelan kelas positif sebanyak 415 data dan negatif sebanyak 58 data dengan nilai metrik dari aspek *user experience* akurasi 93.9%, *precision* 93.5%, *recall* 93.9%, dan *FI-Score* 92.6%, dan dari aspek *user interface* memiliki nilai metrik akurasi 87,5%, *precision* 90.4%, *recall* 87.5%, *FI-Score* 87.2%. Hasil nilai metrik evaluasi menunjukkan bahwa model dengan algoritma KNN memiliki peforma sangat baik dalam melakukan prediksi atau klarifikasi[8].

Alasan Penulis menggunakan algoritma *K-Nearest Neighbor* yaitu algoritma tersebut merupakan algoritma klasifikasi karena mudah diimplementasikan, data yang digunakan memiliki label sehingga memudahkan dalam proses pengelompokan ke dalam kelas yang paling sesuai. Metode ini dipilih karena termasuk dalam top 10 metode data *mining* yang paling populer dan berdasarkan pada penelitian yang pernah dilakukan dengan membandingkan KNN, *Random Forest*, dan *Support Vector Machine* (SVM) diperoleh hasil bahwa metode KNN sangat direkomendasikan untuk digunakan pada sistem pengklasifikasian untuk studi kasus penelitian.[9]

Dengan latar belakang di atas serta di dukung oleh penelitian terdahulu penulis terinspirasi untuk membuat sebuah analisis sentimen terhadap pilpres diharapkan dapat memberi tambahan pengetahuan sebagai media pembelajaran tentang penerapan algoritma *k-nearets neighbor* (K-NN) dan bisa memberikan gambaran terhadap bagaimana analisis sentimen terhadap pilpres 2024 pada media sosial X untuk mengklasifikasikan opini positif dan negatif untuk berbagai kepentingan pengoptimalan informasi media sosial untuk kepentingan publik.

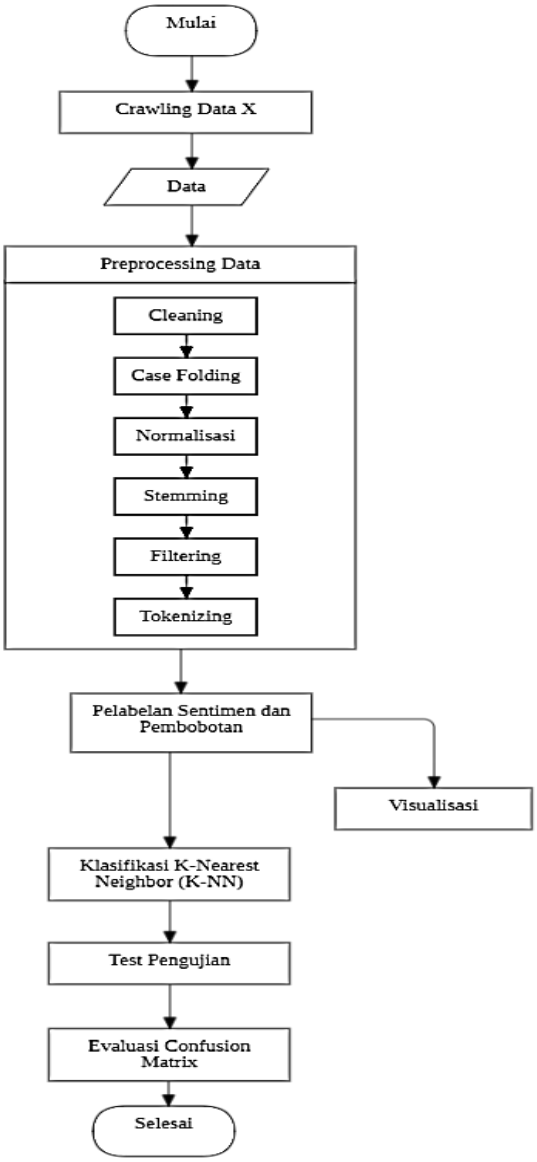
Kontribusi dari penelitian ini yaitu menganalisis tingkat akurasi dari penggunaan algoritma *K-Nearest Neighbor* terhadap sentimen Masyarakat terhadap pilpres 2024 dan diharapkan dapat memberi tambahan pengetahuan sebagai media pembelajaran tentang penerapan algoritma *K-Nearest Neighbor* (K-NN) dan bisa memberikan gambaran terhadap bagaimana analisis sentimen terhadap pilpres 2024 pada media sosial X untuk mengklasifikasikan opini positif dan negatif untuk berbagai kepentingan pengoptimalan informasi media sosial untuk kepentingan publik.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

2. METODOLOGI PENELITIAN

2.1 Kerangka Penelitian

Penelitian ini dilakukan di Universitas Islam Negeri Sumatera Utara yang berada di Jl. Lap. Golf, Kp. Tengah, Kec. Pancur Batu, Kabupaten Deli Serdang, Sumatera Utara 20353. Kerangka penelitian sebagai berikut:



Gambar 1. Kerangka Penelitian

Berdasarkan Gambar 1 Prosedur kerja yang dilakukan pada penelitian ini melalui beberapa tahapan mulai dari crawling data X sampai dengan evaluasi confusion matrix. Pada penelitian ini metode pendekatan yang digunakan adalah pendekatan metode kuantitatif yang merupakan sebuah metode penelitian yang di dalamnya mengandalkan pengukuran objektif dan analisis matematis (statistik) terhadap sampel data yang diperoleh [11] [12] [13].

2.2 Algoritma K-Nearest Neighbor (K-NN)

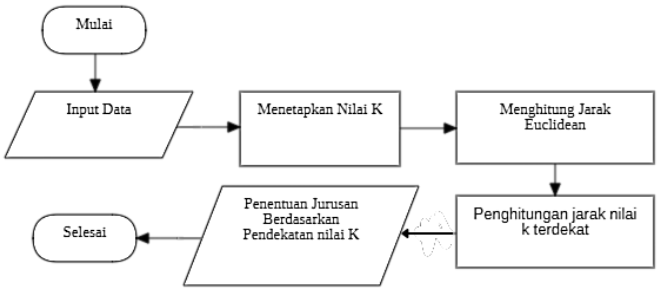
K- Nearest Neighbor (K-NN) merupakan teknik klasifikasi yang paling terkenal. Metode K-NN termasuk *supervised learning*, yaitu teknik pembelajaran yang menentukan pola hubungan antara variabel *input* dengan *output* (label) dari data *training* yang diberikan[10]. Tujuan Algoritma merupakan untuk mengklasifikasikan model baru berdasarkan atribut serta sampel dari *training* data. Metode *K-Nearest Neighbor* (K-NN) mempunya kelebihan berbentuk kestabilan dari berapa pun variasi nilai-k. Hasil yang sudah diperoleh lewat implementasi serta pengujian sistem merupakan jumlah data *training* (latih), keseimbangan jumlah jenis data *training* dan nilai k mempengaruhi ketepatan hasil dari analisis sentimen. Mencari jauhnya jarak antar titik pada kelas k akan dihitung memakai jarak *Euclidean*. Jarak *Euclidean* merupakan metode pencarian jarak antar dua titik x_1 dan x_2 akan dihitung dengan persamaan sebagai berikut ini:

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

$$d(x_1, x_2) = \sqrt{(x_{11} - x_{21})^2 + \cdots + (x_{1p} - x_{2p})^2} = \sqrt{\sum_{j=1}^{P} (x_{1j} - x_{2j})^2} \quad (1)$$

Keterangan : $d(x_1, x_1)$: jarak antara variabel $x_1$dan x_2, X : variabel, p: jumlah dimensi variabel.

2.3 Klasifikasi K-Nearest Neighbor



Gambar 2. Flowchart Algoritma K-Nearest Neighbor

Berdasarkan Gambar 2 digunakan untuk klasifikasi dengan algoritma *K-Nearest Neighbor* (*K-NN)* dengan mengambil *k*-tetangga terdekat dengan menggunakan jarak *Euclidean*. Metode ini digunakan untuk mengelompokkan objek berdasarkan contoh pelatih terdekat di ruang fitur. *K-Nearest Neighbor* dilakukan dengan mencari kelompok objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data testing.

3. HASIL DAN PEMBAHASAN

3.1 *Crawling* Data

Pengumpulan informasi dilakukan bertujuan untuk mendapatkan sebuah data yang digunakan untuk tercapainya tujuan dari riset. Informasi dari sosial media ataupun sosial media *mining* jadi salah satu metode untuk mengenali tingkatan partisipasi dari pengguna. Data diperoleh dari opini masyarakat mengenai pemilihan presiden periode 2024-2029 pada media sosial X. Data yang digunakan berbentuk *tweet* bahasa Indonesia yang diambil dari media sosial X. Proses pengambilan data menggunakan metode *crawling* pada X dengan memanfaatkan *library tweet-harvest*. Data tersebut diambil dengan dengan *keyword* "Pilpres". Berikut merupakan langkah-langkah dalam proses *crawling data* X:

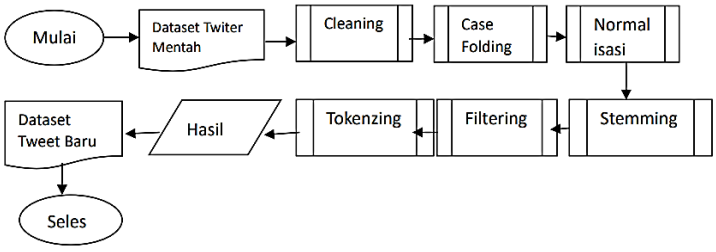
a. Proses *crawling* data dibantu dengan menggunakan *library tweet-harvest* dan melakukan input kode *AuthToken* yang sebelumnya sudah generated.

```
filename = 'pilpres.csv'
search_keyword = 'pilpres since:2023-11-28 until:2024-04-28 lang:id'
limit = 1000
!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token
{twitter_auth_token}
```

b. Data tweet digunakan dalam rentang waktu 28 November 2023 hingga 28 April 2024 dengan menggunakan kata kunci "pilpres". Data selanjutnya disimpan kedalam format *file* berbentuk csv.

3.2 *Preprocessing* Data

Preprocessing atau pra-pemrosesan adalah tahap awal dan penting dalam text mining yang dilakukan guna mengubah data teks mentah menjadi format yang sesuai dan dapat diolah lebih lanjut. *Preprocessing* bertujuan untuk meningkatkan kualitas data, menghilangkan noise dan mengubah format data agar lebih mudah diproses. Gambar 3 menampilkan alur proses *preprocessing.*



Gambar 3. *Preprocessing*

Berdasarkan Gambar 3 proses *preprocessing* dilakukan agar dihasilkan dataset tweet yang baru guna untuk memperoleh seles sesuai dengan yang di hasilkan.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

3.3 *Cleaning*

Cleaning adalah tahap awal *Preprocessing* yang bertujuan menghilangkan *noise* pada data. Tahapan ini melibatkan penghapusan karakter selain huruf, seperti simbol, tanda baca dan angka [14]. Dalam tahap ini menggunakan fungsi dari library NLTK (*Natural Language Toolkit*) pada *python*. Berikut tabel 1 merupakan hasil dari *cleaning*:

Tabel 1. Hasil Cleaning

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	faktanya emang lebih banyak yang kenal ps ketimbang anies atau bahkan ganjar dan makin ke sini nama partai emang ngak begitu relevan dengan pilpres salah satu faktornya nama jokowi
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	hari ini sidang sengketa pilpres kembali digelar agenda pemeriksan saksi dan ahli
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	sidang sengketa pilpres kembali digelar senin ini di mk kali ini kubu akan hadirkan saksi dan ahli untuk didengarkan keteranganya

3.4 *Case Folding*

Case Folding adalah proses perubahan huruf awal pada suatu kata dalam dokumen menjadi huruf kecil (*lower case)* [15]. Berikut tabel 2 merupakan hasil dari *Case Folding:*

Tabel 2. Hasil Case Folding

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	faktanya emang lebih banyak yang kenal ps ketimbang anies atau bahkan ganjar dan makin ke sini nama partai emang ngak begitu relevan dengan pilpres salah satu faktornya nama jokowi
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	hari ini sidang sengketa pilpres kembali digelar agenda pemeriksan saksi dan ahli
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	sidang sengketa pilpres kembali digelar senin ini di mk kali ini kubu akan hadirkan saksi dan ahli untuk didengarkan keteranganya

3.5 Normalisasi

Tahap berikutnya yaitu Normalisasi, yang berfungsi sebagai standarisasi kata yang memiliki makna ganda dan ambigu sehingga menghindari bias antar kata. Berikut merupakan tabel 3 hasil dari tahapan normalisasi:

Tabel 3. Hasil Normalisasi

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	faktanya emang lebih banyak yang kenal ps ketimbang anies atau bahkan ganjar dan makin ke sini nama partai emang ngak begitu relevan dengan pilpres salah satu faktornya nama jokowi
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	hari ini sidang sengketa pilpres kembali digelar agenda pemeriksan saksi dan ahli
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	sidang sengketa pilpres kembali digelar senin ini di mk kali ini kubu akan hadirkan saksi dan ahli untuk didengarkan keteranganya

3.6 *Stemming*

Stemming adalah tahapan proses untuk mencari *root* kata (kata dasar) dari setiap kata yang harus sesuai dengan struktur morfologi bahasa Indonesia yang benar [16]. Berikut tabel 4 merupakan dari hasil *stemming:*

Tabel 4. Hasil Stemming

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang	fakta emang lebih banyak yang kenal ps

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Sebelum	Sesudah
Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	ketimbang anies atau bahkan ganjar dan makin ke sini nama partai emang ngak begitu relevan dengan pilpres salah satu faktor nama jokowi
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	hari ini sidang sengketa pilpres kembali gelar agenda pemeriksan saksi dan ahli
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	sidang sengketa pilpres kembali gelar senin ini di mk kali ini kubu akan hadir saksi dan ahli untuk dengar keteranganya

3.7 *Filtering*

Filtering untuk menghapus kata dengan jumlah karakter yang kurang dari nilai yang ditentukan. Berikut merupakan tabel 5 hasil dari tahapan *filtering:*

Tabel 5. Hasil Filtering

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	fakta emang kenal ps ketimbang anies ganjar nama partai emang ngak relevan pilpres salah faktor nama jokowi
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	sidang sengketa pilpres gelar agenda pemeriksan saksi ahli
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	sidang sengketa pilpres gelar senin mk kali kubu hadir saksi ahli dengar keteranganya

3.8 *Tokenzing*

Tokenzing adalah tahapan yang berfungsi untuk melakukan pemecahan kata dalam proses penyusunan suatu kalimat. Hal ini dilakukan agar mengetahui kemunculan dari kata tersebut. Berikut merupakan tabel 6 hasil dari tahapan *Tokenzing*:

Tabel 6. Hasil Tokenzing

Sebelum	Sesudah
Faktanya emang lebih banyak yang kenal PS ketimbang Anies atau bahkan Ganjar dan makin ke sini nama partai emang nggak begitu relevan dengan pilpres. Salah satu faktornya nama Jokowi.	[faktanya, emang, lebih, banyak, yang, kenal, ps, ketimbang, anies, atau, bahkan, ganjar, dan, makin, ke, sini, nama, partai, emang, ngak, begitu, relevan, dengan, pilpres, salah, satu, faktornya, nama, jokowi]
Hari ini Sidang Sengketa Pilpres Kembali Digelar: Agenda Pemeriksaan Saksi dan Ahli Anies-Muhaimin https://t.co/GfMibxCWXl	[hari, ini, sidang, sengketa, pilpres, kembali, digelar, agenda, pemeriksan, saksi, dan, ahli, aniesmuhaimin, htpstcogfmibxcwxl]
Sidang sengketa Pilpres kembali digelar Senin ini di MK. Kali ini kubu Anies-Muhaimin akan hadirkan saksi dan ahli untuk didengarkan keterangannya https://t.co/qMaCNiGTeq	[sidang, sengketa, pilpres, kembali, digelar, senin, ini, di, mk, kali, ini, kubu, aniesmuhaimin, akan, hadirkan, saksi, dan, ahli, untuk, didengarkan, keteranganya, htpstcoqmacnigteq]

3.9 Pelabelan Sentimen dan Pembobotan TF-IDF

Setelah tahap *preprocessing* selesai maka tahap selanjutnya yang dilakukan ialah menentukan ekstraksi fitur yang akan digunakan untuk melatih model *machine learning* pada penelitian ini, ekstraksi fitur diterapkan untuk meningkatkan akurasi dan kinerja dari model tersebut. Ekstraksi fitur yang digunakan dalam penelitian ini yaitu ekstraksi fitur TF-IDF.

3.9.1 Pelabelan Sentimen

Metode *Lexicon* juga digunakan untuk pelabelan setelah data dibersihkan pada tahap *preprocessing* teks. Penentuan tersebut dilakukan terhadap data teks berupak kalimat-kalimat yang mempunyai kata-kata dalam kamus *lexicon* yang terdiri dari kata-kata negatif dan positif. Skor kata yang teridentifikasi dalam *lexicon dictionary* (kamus *lexicon*) akan ditentukan skornya berdasarkan jumlah kata dalam setiap teks atau kalimat.

Tabel 7. Perhitungan Pelabelan Skor Sentimen

Opini	Kata Positif	Kata Negatif	Skor	Kelas Sentimen
sakit jelek anies eh pilpres suara kalah	-	Jelek	-2	Negatif

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Opini	Kata Positif	Kata Negatif	Skor	Kelas Sentimen
anies sakit darah		kalah		
	0	2		
sidang sengketa pilpres gelar agenda	ahli	-	1	Positif
pemeriksan saksi ahli	1	0		
lawan politik kerah putih mahfud md	-	-	0	Netral
aju gugat hasil pilpres mk	0	0		

Pada penelitian ini terdapat tiga sentimen data yang terdiri dari sentimen positif, negatif dan netral. Jenis sentimen positif mengandung kata-kata seperti pujian, ucapan terima kasih ataupun hal yang membanggakan lainnya. Untuk jenis sentimen negatif mengandung kata-kata seperti kekecewaan, ungkapan ketidakpuasan ataupun penghinaan dan lain sebagainya, sedangkan jenis sentimen netral mengandung kata-kata seperti opini yang memiliki statement positif dan negatif yang seimbang dan ada juga opini yang tidak memiliki statement positif maupun negatif maka dalam jenis sentimen netral semacam ungkapan tanpa sentimen, iklan dan lain sebagainya.

3.9.2 Pembobotan TF-IDF

Pada tahap ini dilakukan pembobotan *term* proses pemberian nilai pada setiap tweet yang telah melewati tahap preprocessing dan pelabelan dengan menggunakan lexicon based. Hal ini bertujuan agar pembobotan kata dengan metode *Term Frequency- Inverse Document Frequency* (TF-IDF) dapat mengukur dan mengekstrak tingkat relativitas dari kata-kata yang muncul dalam sebuah dokumen terhadap perhitungan dan hasil dari pembobotan term pada kata. Adapun implementasi dari proses pembobotan kata sebagai berikut:

a. Menghitung *TF (Term Frequency)*

Tabel 8. Proses Perhitungan Term Frequency

Kata (Term)	TF		
	D1	D2	D3
sakit	1	0	0
jelek	1	0	0
anies	2	0	0
eh	1	0	0
…	…	…	…
mk	0	0	1

Term frequency melakukan perhitungan dengan ditambahkan jumlah kata (*term*) yang muncul pada tiap dokumen. *TF* berfungi buat mencermati apakah sesuatu kata terdapat ataupun tidak pada tiap dokumen, bila kata tersebut terdapat disuatu dokumen maka diberikan nilai satu dan bertambah bila disuatu dokumen mempunyai lebih dari satu kata yang muncul

b. Menghitung *DF (Document Frequency)*

Tabel 9. Proses Perhitungan Document Frequency

Kata (Term)	TF			DF
	D1	D2	D3	
sakit	1	0	0	1
jelek	1	0	0	1
anies	2	0	0	1
eh	1	0	0	1
…	…	…	…	…
mk	0	0	1	1

Pada proses perhitungan df melakukan perhitungan jumlah dokumen yang memiliki kata (term). Dimana bila sesuatu kata terdapat disuatu dokumen maka df akanditambahkan satu hingga dokumen terakhir buat mengenali jumlah semua *df*.

c. Menghitung *IDF (Inverse Document Frequency)*

Tabel 10. Proses Perhitungan Inverse Document Frequency

Kata (Term)	TF			DF	N/DF	IDF(log N/DF)
	D1	D2	D3			
sakit	1	0	0	1	3	0,477121255
jelek	1	0	0	1	3	0,477121255
anies	2	0	0	1	3	0,477121255
eh	1	0	0	1	3	0,477121255

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Kata (Term)	TF			DF	N/DF	IDF(log N/DF)
	D1	D2	D3			
….	….	…	….	…	…	…
mk	0	0	1	1	3	0,477121255

Adapun rumus untuk menghitung *idf* sebagai berikut : $idf = log \frac{N}{df}$

Pada *inverse document freqeuncy* (idf) melakukan perhitungan dari suatu term yang terdapat dari suatu dokumen. Jumlah dokumen (N) dibagi dengan jumlah df ataupun jumlah kemunculan term disuatu dokumen

d. Menghitung bobot *(Weight)*

Tabel 11. Perhitungan Bobot (Weight)

W		
D1	D2	D3
0,47712	0	0
0,47712	0	0
…	…	…
0	0	0,47712

Untuk menghitung bobot pada tiap kata dengan rumus sebagai berikut : $W_{td} = TF_{td} * IDF_t$.

Pada perhitungan bobot (*Weight*) dilakukan perhitungan dokumen ke-d terhadap kata (*term*). Sedangkan *tf* merupakan jumlah kemunculan *term* (t) dalam dokumen (d). *IDF* diperoleh dari hasil perhitungan proses sebelumnya. Hasil dari *term frequency* dikalikan dengan hasil IDF akan memperoleh setiap bobot dari masing-masing kata.

3.10 Klasifikasi *K-Nearest Neighbor*

Setelah melakukan pembobotan dari masing-masing *tweet* dan *vector* dari masing-masing *tweet*, kemudian akan masuk ketahap dengan menghitung akurasi menggunakan metode algoritma K-Nearest Neighbor (K-NN).

a. *Confussiom Matriks*
Pada penelitian ini akan mengevaluasi hasil dan mengukur kinerja suatu metode klasifikasi dengan menggunakan *confusion matrix*. Setelah melakukan perhitungan *K-Nearest Neighbor* maka akan dilakukan pengujian akurasi menggunakan *confusion matrix* untuk mengetahui keakuratan hasil klasifikasi dan untuk mengetahui seberapa besar keberhasilan sistem dalam melakukan klasifikasi. Dalam *confusion matrix* akan dihitung *acuracy, precision, recall dan f-measure.*

Tabel 12. Contoh Perhitungan Confusion Matrix

Prediksi	Nilai	
	Positif	Negatif
Positif	2	0
Negatif	0	1

Bersadarkan table 12 mengenai perhitungan confusion matrixs dengan prediksi positif dan negative maka perhitungannya sebagai berikut:

$\text{Acuracy} = \frac{2+1}{2+0+0+1} \text{ x } 100\% = 100\%$

$\text{Precision} = \frac{2}{2+0} \text{ x } 100\% = 100\%$

$\text{Recall} = \frac{2}{2+0} \text{ x } 100\% = 100\%$

$\text{F} - \text{measure} = 2\ x\ \frac{100 \text{ x } 100}{100+100} = 100\%$

Hasil acuracy, precision, recall dan f-measure yang telah di dapatkan yaitu 100%.

3.11 Hasil

Pada penelitian ini, perhitungan pelabelan sentimen telah mendapatkan hasil perbandingan jumlah data kelas sentimen seperti terlihat pada Tabel 13 :

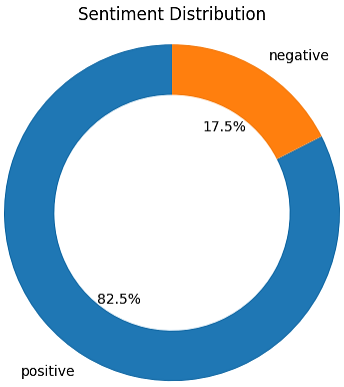
Tabel 13. Perbandingan Jumlah Data Kelas Sentimen

Kelas Sentimen Sementara	Jumlah
Positif	211
Negatif	175
Netral	614

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Kelas Sentimen Sementara	Jumlah
Total	1000

Pada Tabel 13 terdapat kelas sentimen yaitu positif, negatif dan juga netral. Selanjutnya akan melakukan proses reduksi pada kelas sentimen netral dimasukan kedalam sentimen positif yang dibuat secara manual. Untuk hasil pelabelan kelas sentimen terhadap Pilpres Indonesia 2024 dapat dilihat sebagai berikut :



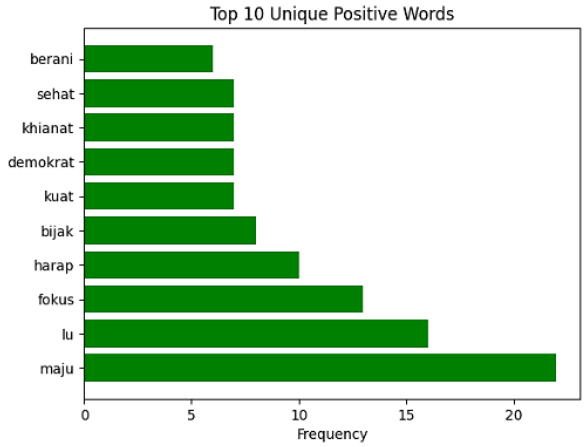
Gambar 4. Hasil Kelas Sentimen Terhadap Pilpres Indonesia 2024

Berdasarkan Gambar 4 diatas terlihat jumlah pada data kelas sentimen terhadap new normal yang terdiri dari 1000 data opini. Dengan persentase pada kelas sentimen positif mendapatkan 82.5% dan 17.5% pada kelas sentimen negatif. Dalam hal ini opini masyarakat terhadap Pilpres Indonesia 2024 mendapatkan kelas sentimen positif lebih unggul dibandingkan kelas sentimen negatif. Dengan kata lain masyarakat Indonesia dalam memberi opini atau tanggapan sangat baik terhadap Pilpres Indonesia 2024.

3.12 Visualisasi

Setelah melakukan pelabelan menggunakan kamus *lexicon dataset*, selanjutnta divisualisasikan ke dalam bentuk *wordcloud* dari kelas positif maupun negatif. Dengan Tujuan visualisasi ini untuk memudahkan pemahaman dan interpretasi hasil analisis sentimen secara visual[17]

1. Opini Positif
   Adapun hasil visualisasi untuk keseluruhan data *tweet* opini positif adalah seperti berikut:

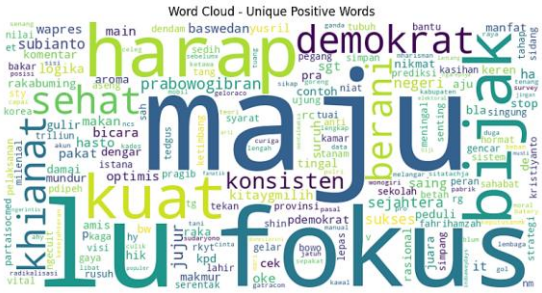


Gambar 5. Frekuensi kata Positif yang sering muncul

Pada Gambar 5 menunjukkan 10 kata positif yang paling sering muncul dalam teks yang telah diolah, dengan fokus pada kata-kata yang unik dan tidak muncul dalam kategori negatif. Dalam bar chart ini, kata "maju" muncul sebagai kata positif yang paling sering digunakan, dengan frekuensi mencapai lebih dari 20 kali. Ini mengindikasikan bahwa kata "maju" memiliki konotasi yang kuat dalam konteks positif, mungkin terkait dengan perkembangan, kemajuan, atau peningkatan. Kata "lu" juga memiliki frekuensi tinggi, diikuti oleh "fokus" dan "harap". Ini mungkin menggambarkan adanya harapan, kepercayaan diri, dan perhatian yang kuat terhadap suatu tujuan atau visi tertentu. Kata seperti "bijak", "kuat", dan "sehat" juga menunjukkan nilai-nilai positif yang dihargai dalam diskusi atau narasi yang dianalisis. Sementara itu, kata-kata seperti "demokrat", "khianat", dan "berani" meskipun frekuensinya lebih rendah, tetap menjadi bagian dari spektrum positif dalam analisis ini, menunjukkan apresiasi terhadap tindakan berani, integritas, dan prinsip-prinsip demokrasi.

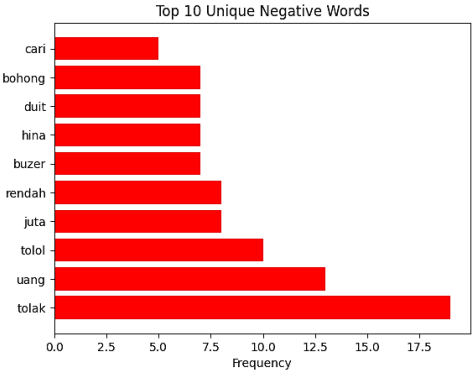
Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Secara keseluruhan, visualisasi ini menggambarkan bagaimana berbagai konsep positif tersebar dan diapresiasi dalam data yang dianalisis, dengan beberapa kata kunci yang sangat menonjol dalam konotasi positifnya seperti berikut:



Gambar 6. Wordcloud Positif

2. Opini Negatif
Adapun hasil visualisasi untuk keseluruhan data tweet opini positif adalah seperti berikut:



Gambar 6. Frekuensi kata negatif yang sering muncul

Pada gambar 6 menunjukkan 10 kata negatif yang paling sering muncul dalam teks yang telah diolah, dengan fokus pada kata-kata yang unik dan tidak muncul dalam kategori negatif. Dalam bar chart ini, kata "tolak" muncul sebagai kata negatif yang paling sering digunakan, dengan frekuensi mencapai lebih dari 17 kali, mencerminkan adanya resistensi atau penolakan yang signifikan terhadap suatu ide atau tindakan. Ini menunjukkan adanya ketidaksetujuan yang meluas dalam konteks pembahasan. Kata "uang" juga muncul sebagai kata negatif, yang dalam konteks ini mungkin mengindikasikan kekhawatiran terhadap pengaruh uang, seperti dalam kasus korupsi atau ketidakadilan ekonomi. Selanjutnya, kata "tolol" memperkuat sentimen negatif dengan menyiratkan penghinaan atau ejekan, mengungkapkan ketidakpuasan yang mendalam terhadap sesuatu yang dianggap bodoh atau tidak kompeten.

Kata "juta", meskipun sering terkait dengan angka besar, dalam konteks ini membawa konotasi negatif yang mungkin merujuk pada kekhawatiran tentang pemborosan atau penyalahgunaan sumber daya dalam skala besar. Terakhir, kata "rendah" digunakan untuk menggambarkan sesuatu yang dianggap berada di bawah standar, baik dalam hal kualitas, moralitas, atau status, menegaskan pandangan yang negatif dan merendahkan terhadap subjek yang dibahas. Secara keseluruhan, kata-kata ini mencerminkan sikap kritis dan ketidakpuasan yang mendalam terhadap isu-isu yang diangkat dalam teks tersebut. Secara keseluruhan, visualisasi ini menggambarkan bagaimana berbagai konsep negatif tersebar dan diapresiasi dalam data yang dianalisis, dengan beberapa kata kunci yang sangat menonjol dalam konotasi negatifnya seperti berikut:

Gambar 7. Wordcloud Negatif

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

3.13 Pengujian

Pada penelitian ini klasifikasi data akan menggunakan Aplikasi *Weka* (*Waikato Environment Analysis Version 3.8.4*) aplikasi yang menyediakan berbagai *tools* yang berguna dalam ranah *data mining*[18]. Sebelum melakukan proses pengujian, dataset yang sudah mempunyai label kelas sentimen yang sudah disimpan dalam bantuk *file.csv* akan diubah ke dalam bentuk *file.arff* dengan menggunakan tools pada Aplikasi *Weka*, hal ini berguna untuk memudahkan untuk proses dalam melakukan analisis. Dalam klasifikasi perlu dilakukan proses pembobotan kata (*term*), pembobotan dalam penelitian ini menggunakan *TF-IDF* (*Term Frequency-Inverse Document Frequency*).

Pada klasifikasi algoritma *K-Nearest Neighbor (K-NN)* akan dilakukan dengan nilai k=1 pada pengujian tes yaitu *use training set, 10 fold cross validation* dan 80% *percentage split*, sehingga penelitian ini membandingkan mana pengujian yang memiliki akurasi yang cukup baik. Menu pengujian yang digunakan adalah *default* yang terdapat pada *tools* di Aplikasi *Weka*. Berikut hasil dari keseluruhan evaluasi dari beberapa pengujian menggunakan *Algoritma K-Nearest Neighbor (K-NN) :*

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     positive
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     negative
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 825   0 |   a = positive
   0 175 |   b = negative
```

Gambar 8. Hasil Klasifikasi K-NN dengan Use Training Set

Dari gambar 8 menunjukan bahwa hasil sistem klasifikasi *K-NN* pada pengujian *use training set* dalam mengelompokan klasifikasi positif dan negatif yang berhasil terindentifikasi. Jumlah kata positif yang benar yaitu 825, sedangkan negatif berjumlah 175. Pada pengujian *use training set* merupakan pengujian atau pengetesan dengan menggunakan data *training* itu sendiri.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.429    0.917      1.000    0.957      0.724    0.911     0.972     positive
               0.571    0.000    1.000      0.571    0.727      0.724    0.911     0.740     negative
Weighted Avg.  0.925    0.354    0.931      0.925    0.916      0.724    0.911     0.931

=== Confusion Matrix ===

   a   b   <-- classified as
 825   0 |   a = positive
  75 100 |   b = negative
```

Gambar 9. Hasil Klasifikasi K-NN dengan 10-Fold Cross-Validation

Pada pengujian 10-*fold cross validation* melakukan pengetesan dengan menggunakan pilihan banyaknya *fold*. Dari Gambar 9 menunjukan bahwa hasil sistem klasifikasi K-NN pada pengujian 10-*fold cross validation* dalam melakukan klasifikasi positif dan negatif. Jumlah kata positif yang benar yaitu sebanyak 825 sedangkan untuk kata negatif berjumlah 100.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.561    0.874      1.000    0.933      0.619    0.864     0.939     positive
               0.439    0.000    1.000      0.439    0.610      0.619    0.864     0.634     negative
Weighted Avg.  0.885    0.446    0.900      0.885    0.866      0.619    0.864     0.876

=== Confusion Matrix ===

   a   b   <-- classified as
 159   0 |   a = positive
  23  18 |   b = negative
```

Gambar 10. Hasil Klasifikasi K-NN dengan 80% Percentage Split

Dari pengujian 80% *percentage split* melakukan pengetesan dengan k% dari data, dimana k merupakan proporsi dari *dataset* yang digunakan untuk data *training*. Persentase pada pengujian 80% untuk data latih (*training*) dan 20% untuk data uji (*testing*). Dari Gambar 10 menunjukan bahwa hasil sistem klasifikasi *K-NN* pada pengujian 80% *percentage split* dalam melakukan klasifikasi positif sebanyak 159 dan untuk label negatif sebanyak 18.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Tabel 14. Perbandingan Keseluruhan Evaluasi Dari Beberapa Pengujian Test Menggunakan Algoritma K-Nearest Neighbor (K-NN)

Data Set	Algoritma	Mode Test	Correctly Classified Instances		Incorrectly Classified Instances		Mean Absolute Error
			(n)	%	(n)	%	
Pilpres.csv	K-Nearest Neighbor (K-NN) Lazy - IBK	Use Training Set	1000	100%	0	0	0,0006
		10 Fold Cross Validati on	925	92.5%	75	7.5%	0,1205
		80% Percentage Split	177	88,5%	23	11,5%	0,1607

Perbandingan dari keseluruhan evaluasi dari beberapa pengujian menggunakan algoritma K-Nearest Neighbor (K-NN) dapat dilihat pada Tabel 14. Informasi yang terdapat terdiri dari pengujian yang digunakan untuk dataset yang terdiri dari pengujian use training set, 10 fold cross validation dan 80% percentage split. Informasi ukuran akurasi juga bisa didapatkan dari Tabel 14 pada kolom *correctly classified instance* dan *incorrectly classified instance*. *Mean absolute error* juga merupakan kolom yang menyediakan informasi rata-rata error yang adapada beberapa jenis algoritma ketika membangun model klasifikasi untuk algoritma *K-Nearest Neighbor* (K-NN) dalam Tabel 14.

Pada pengujian *use training set* yang ada pada Tabel 14 pada penelitian ini nilai correctly classified instance yang didapatkan 100% yang terdiri dari 1000 instance yang terklasifikasi benar dan memiliki nilai *mean absolute error* sebesar 0,0006 dikarenakan pengujian dilakukan menggunakan data training itu sendiri. Pada pengujian 10-*fold cross-validation* mendapatkan hasil *correctly classified instance* yaitu sebesar 92,5% yang terdiri dari 925 *instance* yang terklasifikasi benar dan *incorrectly classified instance* sebesar 7,5% data yang terklasifikasi salah sebesar 75 instance dari total semua dataset berjumlah 1000 instance, mencapai nilai *mean absolute error* sebesar 0,1205.

Pada 80% percentage split menggunakan pembagian nilai jumlah data latih dan data ujiialah 80% untuk data latih dan 20% untuk data uji. Mendapatkan hasil correctly classified instance yaitu sebesar 88,5% yang terdiri dari 177 *instance* yang terklasifikasi benar dan *incorrectly classified instance* sebesar 11.5% yang dimana klasifikasi data salah hanya sebesar 23 *instance*, mencapai nilai mean absolute error sebesar 0.1607.

Tabel 15. Waktu Yang Dibutuhkan Untuk Membangun Model

Algoritma	Pengujian	Pilpres Indonesia 2024 (Detik)
K-Nearest Neighbor (K-NN) Lazy - IBK	Use Training Set	0,07
	10 Fold Cross Validation	0,01
	80% Percentage Split	0,01

Tabel 15 menampilkan informasi waktu yang diperlukan untuk membangun model pada pengujian menggunakan algoritma *K-Nearest Neighbor (K-NN).* Satuan waktu yang digunakan yaitu satuan detik, yang dimana pengujian yang digunakan yaitu *use trainingg set*, *10-fold cross-validation* dan 80% *percentage split*. Pada Tabel 15 dapat dilihat pada pengujian *use training set* untuk dataset Pilpres Indonesia 2024 menghasilkan catatan waktu sebesar 0,07 detik untuk membuat model klasifikasi. Pada pengujian *test 10 fold cross validation* menghasilkan catatan waktu yang sangat cepat dari lainnya sebesar 0,01 detik dan 80% *percentage split* untuk membangun model klasifikasi untuk *dataset* Pilpres Indonesia 2024 memiliki catatan waktu yaitu sebesar 0,01 detik.

3.14 Evaluasi

Untuk mengukur ketepatan model dalam mengklasifikasikan sentimen, digunakan nilai akurasi dari sebuah *confusion matrix*. *Confusion matrix* adalah *tool* yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah *matrix* dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai aktual dan prediksi pada klasifikasi [19]. Hasil dri klasifikasi yang nantinya didapatkan akan dievaluasi dimana nimai yang didapat dari evaluasi dapat digunakan untuk mengukur tingkat keberhasilan dari suatu metode yang dilakukan dalam pengujian. Salah satu metode untuk melakukan evaluasi dalam analisis sentimen [20]. Berikut *confusion matrix* untuk pembagian data training dan data testing 80:20 :

```
=== Confusion Matrix ===

   a   b   <-- classified as
 825   0 |   a = positive
   0 175 |   b = negative
```

Gambar 11. *Confusion Matrix*

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

Berdasarkan gambar 11, dapat dilihat bahwa model mengklasifikasikan sentimen dengan benar sebanyak 211 sentimen untuk kelas positif dan 175 sentimen untuk kelas negatif. Dan juga model memprediksi 614 sentimen netral.

Tabel 16. Confusion Matrix

Prediksi	Use Training Set		10-Fold Cross Validation		80% Percentage Split	
	Positive	Negative	Positive	Negative	Positive	Negative
Positive	825	0	825	0	159	0
Negative	0	175	75	100	23	18

Dari data pada Tabel 16 confusion matrix maka use training set dapat dihitung acuracy, precision, recall dan f-measure sebagai berikut :

$$\text{Acuracy} = \frac{825+175}{825+175} \text{ x } 100\% = 100\%$$

$$\text{Precision} = \frac{825}{825+0} \text{ x } 100\% = 100\%$$

$$\text{Recall} = \frac{825}{825+0} \text{ x } 100\% = 100\%$$

$$\text{F} - \text{measure} = 2\,x\, \frac{100 \text{ x } 100}{100+100} = 100\%$$

Dari data pada Tabel 16 *confusion matrix* maka *10-fold cross-validation* dapat dihitung *acuracy, precision, recall* dan *f-measure* sebagai berikut :

$$\text{Acuracy} = \frac{825+100}{825+75+100} \text{ x } 100\% = 92{,}5\%$$

$$\text{Precision} = \frac{825}{825+0} \text{ x } 100\% = 100\%$$

$$\text{Recall} = \frac{825}{825+75} \text{ x } 100\% = 91\%$$

$$\text{F} - \text{measure} = 2\,x\, \frac{100 \text{ x } 91}{100+91} = 94\%$$

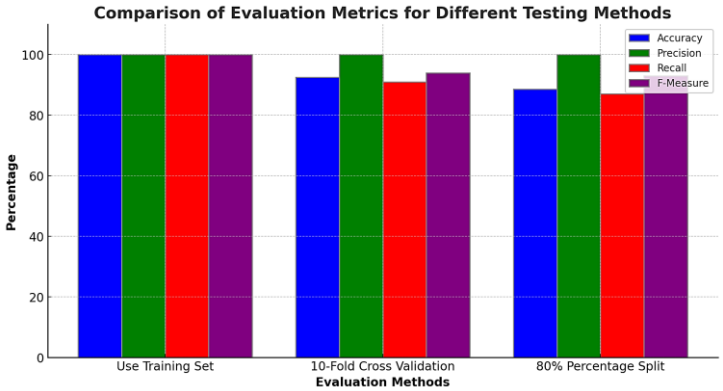
Dari data pada Tabel 16 confusion matrix maka 80% percentage split dapat dihitung acuracy, precision, recall dan f-measure sebagai berikut :

$$\text{Acuracy} = \frac{159+18}{159+18+23} \text{ x } 100\% = 88{,}5\%$$

$$\text{Precision} = \frac{159}{159+0} \text{ x } 100\% = 100\%$$

$$\text{Recall} = \frac{159}{159+23} \text{ x } 100\% = 87\%$$

$$\text{F} - \text{measure} = 2\,x\, \frac{100 \text{ x } 87}{100+87} = 93{,}04\%$$



Gambar 12. Grafik Evaluasi

Dari hasil diatas yang didapat akurasi *use training set* mendapatkan hasil *acuracy* sebesar 100% , *precision* sebesar 100%, *recall* sebesar 100% dan *f-measure* sebesar 100%, *10-fold cross validation* yang didapat mendapatkan hasil *acuracy* sebesar 92,5%, *precision* sebesar 100% *recall* sebesar 91% dan *f-measure* sebesar 94%, dan yang terakhir 80% *percentage split* mendapatkan hasil *acuracy* sebesar 88,55% , *precision* sebesar

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

100%, *recall* sebesar 87% dan *f-measure* sebesar 93,04%. Metode klasifikasi *algoritma K-Nearest Neighbor (K-NN)* menggunakan pengujian 80% *percentage split* sangat baik dalam pengujian klasifikasi memiliki *acuracy, precision, recall* dan *f-measure* lebih besar dibandingkan dengan pengujian *10-fold cross validation*.

4. KESIMPULAN

Berdasarkan dari hasil penelitian yang telah dilakukan, maka dapat disimpulkan bahwa Analisis sentimen menggunakan metode K-Nearest Neighbor (KNN) telah terbukti efektif dalam mengidentifikasi dan memahami sentimen masyarakat terkait Pilpres Indonesia 2024. Tanggapan masyrakat terhadap fenomena tren dan opini publik terhadap calon kandidat presiden di tahun 2024 dinilai menunjukkan sikap positif, seperti yang tergambar dalam analisis sentimen dengan menggunakan kamus *lexicon.* Dari sekitar 1.000 *tweet* yang telah dianalisis, sebanyak 211 di antaranya menunjukkan sentimen yang positif, 175 mengungkapkan sentimen yang negatif, sedangkan 614 lainnya mengungkapkan sentimen yang Netral. Data ini dikumpulkan mulai dari 28 November 2023 hingga 28 April 2024. Selain itu, penelitian ini juga mengidentifikasi kata-kata yang sering muncul dalam *tweet* berbahasa Indonesia. Pada *K-Nearest Neighbor (KNN),* hasil yang didapat akurasi *use training set* mendapatkan hasil *acuracy* sebesar 100% , *precision* sebesar 100%, *recall* sebesar 100% dan *f-measure* sebesar 100%, *10-fold cross validation* yang didapat mendapatkan hasil *acuracy* sebesar 92,5%, *precision* sebesar 100% *recall* sebesar 91% dan *f-measure* sebesar 94%, dan yang terakhir 80% *percentage split* mendapatkan hasil *acuracy* sebesar 88,55% , *precision* sebesar 100%, *recall* sebesar 87% dan *f-measure* sebesar 93,04%. Metode klasifikasi *algoritma K-Nearest Neighbor (K-NN)* menggunakan pengujian 80% *percentage split* sangat baik dalam pengujian klasifikasi memiliki *acuracy, precision, recall* dan *f-measure* lebih besar dibandingkan dengan pengujian *10-fold cross validation***.**

REFERENCES

[1] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Eksplora Informatika*, vol. 9, no. 1, pp. 1–10, Sep. 2019, doi: 10.30864/eksplora.v9i1.237.

[2] T. Rosyida, H. P. Putro, and H. Wahyono, "Analisis Sentimen Terhadap Pilpres 2024 Berdasarkan Opini dari Twitter Menggunakan Naive Bayes dan SVM," *Teknokris*, vol. 26, no. 1, pp. 23–32, 2023

[3] A. Malik Zuhdi, E. Utami, and S. Raharjo, "Analisis Sentiment Twitter Terhadap Capres Indonesia 2019 Dengan Metode K-NN," *Jurnal INFORMA Politeknik Indonusa Surakarta*, vol. 5, no. 2, pp. 1–7, 2019.

[4] F. A. Wenando, R. Hayami, and A. J. Anggrawan, "Analisis Sentimen Pada Pemerintahan Terpilih Pada Pilpres 2019 Di Twitter Menggunakan Algoritma NaiveBayes," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 7, no. 1, pp. 99–104, Dec. 2020, doi: 10.33330/jurteksi.v7i1.851.

[5] F. R. Irawan, A. Jazuli, and T. Khotimah, "ANALISIS SENTIMEN TERHADAP PENGGUNA GOJEK MENGGUNAKAN METODE K-NEARSET NEIGHBORS," *Jurnal Informatika dan Komputer) Akreditasi KEMENRISTEKDIKTI*, vol. 5, no. 1, 2022, doi: 10.33387/jiko.

[6] Reddy, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor", 2020

[7] M. Furqan, and S. Mayang Sari "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia", 2022

[8] D. Eka Ratnawati, "Analisis Sentimen berbasis Aspek terhadap Data Ulasan menggunakan Metode K-Nearest Neighbor (Studi Kasus: Aplikasi Olsera POS)," 2023. [Online]. Available: http://j-ptiik.ub.ac.id

[9] J. A. Samudra, S. Anraeni, and D. Herman, "Penerapan Metode K-Nearest Neighbor untuk Memprediksi Tingkat Kelulusan Mahasiswa Berbasis Web pada Fakultas Ilmu Komputer UMI," vol. 1, no. 4, pp. 230–237, 2020.

[10] D. Agustina and F. Rahmah, "Analisis Sentimen pada Sosial Media Twitter terhadap MRT Jakarta Menggunakan Machine Learning," *Insearch*, vol. 2, no. 1, pp. 1–6, 2022.

[11] R. Mustaqilah, O. Widyaningtyas, and T. Wantoro, "Efektivitas Penggunaan Twitter Sebagai Sarana Peningkatan Berpikir Kritis Mahasiswa Ilmu Komunikasi," *MUKASI: Jurnal Ilmu Komunikasi*, vol. 2, no. 1, pp. 18–28, 2023, doi: 10.54259/mukasi.v2i1.1346.

[12] M. J. Aridiyanto and P. Penagsang, "Analisis Faktor-Faktor Yang Mempengaruhi Kinerja Koperasi (Studi Kasus : Koperasi di Surabaya Utara)," *Jurnal Ekonomi dan Bisnis*, vol. 7, no. 1, pp. 27–40, 2022.

[13] A. Halim, Y. Yusra, M. Fikry, M. Irsyad, and E. Budianita, "Klasifikasi Sentimen Masyarakat Di Twitter Terhadap Prabowo Subianto Sebagai Bakal Calon Presiden 2024 Menggunakan M-KNN," *Journal of Information System Research (JOSH)*, vol. 5, no. 1, pp. 202–212, 2023, doi: 10.47065/josh.v5i1.4054.

[14] A. F. Sabily, P. P. Adikara, and M. A. Fauzi, "Analisis Sentimen Pemilihan Presiden 2019 pada Twitter menggunakan Metode Maximum Entropy," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 5, pp. 4204–4209, 2019.

[15] Z. Mahendra and A. Ridok, "Analisis Sentimen Opini Masyarakat Terhadap Fenomena TikTokShop di Indonesia Menggunakan Metode K-Nearest Neighbor berbasis N-gram dengan Seleksi Fitur Information Gain," vol. 1, no. 1, pp. 1–10, 2022.

[16] I. Indriati and A. Ridok, "Sentiment Analysis for Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn)," *Journal of Enviromental Engineering and Sustainable Technology*, vol. 3, no. 1, pp. 23–32, 2021, doi: 10.21776/ub.jeest.2016.003.01.4.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1064-1078
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5802

[17] S. Abdullah, "Penulis Pertama: Visualisasi Data Analisa Sentimen … 261 Visualisasi Data Analisa Sentimen RUU Omnibus law Kesehatan Menggunakan KNN dengan Software RapidMiner," vol. 8, no. 3, 2023.

[18] I. Mulia and M. Muanas, "Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 dan Software Weka," *JAS-PT (Jurnal Analisis Sistem Pendidikan Tinggi Indonesia)*, vol. 5, no. 1, pp. 57–64, Jun. 2021, doi: 10.36339/jaspt.v5i1.417.

[19] I. Nawangsih, I. Melani, S. Fauziah, and A. I. Artikel, "Pelita Teknologi Prediksi Pengangkatan Karyawan Dengan Metode Algoritma C5.0 (Studi Kasus Pt. Mataram Cakra Buana Agung," *Jurnal Pelita Teknologi*, vol. 16, no. 2, pp. 24–33, 2021.

[20] G. F. Grandis, Y. Arumsari, and Indriati, "Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 8, pp. 3507–3514, 2021.