KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 6, Juni 2024, Hal 2822-2829
DOI 10.30865/klik.v4i6.1904
https://djournals.com/klik

Analisis Sentimen Terhadap Sebuah Figur Publik di Twitter Menggunakan Metode K-Nearest Neighbor

Yenggi Putra Dinata, Yusra*, Muhammad Fikry, Febi Yanto, Eka Pandu Cynthia

Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Indonesia
Emai: [1]11950115234@students.uin-suska.ac.id, [2]muhammad.fikry@uin-suska.ac.id, [3]yusra@uin-suska.ac.id, [4]febiyanto@uin-suska.ac.id, [5]eka.pandu.cynthia@uin-suska.ac.id
Email Penulis Korespondensi: yusra@uin-suska.ac.id

Abstrak−Perkembangan media online, khususnya melalui media sosial seperti Twitter, menciptakan panggung yang luas untuk berbagai aktivitas, termasuk kampanye politik dan opini masyarakat terhadap tokoh publik. Ketika teknologi informasi berkembang pesat, opini masyarakat dapat disampaikan tanpa terbatas waktu melalui media sosial. Twitter, dengan keterbatasan karakter dan hastag "#" yang dapat digunakan oleh pengguna, dianggap lebih mudah diambil informasi tentang opini dan sentimen yang ada. Saat ini, media sosial banyak digunakan untuk berkomunikasi dan mencari teman, namun juga untuk aktivitas lainnya. Mengiklankan produk, membeli dan menjual apa pun, termasuk mengiklankan partai politik dan berkampanye untuk anggota Kongres atau calon presiden. Penelitian ini bertujuan pada analisis sentimen terhadap Puan Maharani, Ketua DPR RI, menggunakan data dari media sosial Twitter. Twitter, sebagai platform yang memungkinkan pengguna untuk mengungkapkan pendapat dalam format singkat, dijadikan sebagai sumber informasi utama dalam penelitian. Algoritma K-Nearest Neighbor untuk teknik analisis sentimen, digunakan untuk mengklasifikasikan *tweet* individu ke dalam kategori positif atau negatif mengenai pandangan Puan Maharani. Metode yang digunakan dalam penelitian ini dimulai Crawling data, pelabelan dan preprocessing data yang digunakan penelitian ini meliputi case folding, cleaning, tokenizing, negation handling, normalisasi, stopword removal, dan stemming. Untuk proses klasifikasi menggunakan metode K-nearest neighbour (KNN), feature Weighting (TF-IDF), dan feature Selection (thresholding), nilai ambang batas adalah 0,001. Data yang digunakan mencakup 9.000 *tweet* dalam bahasa Indonesia. Hasil dari pengujian yang dilakukan dalam metode K-Nearest Neighbor, yang menggunakan matriks konfusi, 6 dengan nilai K yang berbeda (3, 5, 7, 9, 11, 13), dengan mekanisme perbandingan rasio 90:10 dan 80:20 dan 70:30 mencapai perolehan akurasi yang tertinggi 90,00% dengan K = 11 dari perbandingan yang digunakan rasio 90: 10.

Kata Kunci: Puan Maharani; K-NN; Masyarakat; Twitter; Klasifikasi Sentimen

Abstract−The development of online media, particularly through social media platforms like Twitter, has created a vast stage for various activities, including political campaigns and public opinion on public figures. When information technology advances rapidly, public opinion can be conveyed without time constraints through social media. Twitter, with its character limitations and the use of hashtags by users, is considered easier to gather information about existing opinions and sentiments. Currently, social media is widely used for communication and making friends, but also for other activities. Advertising products, buying and selling anything, including advertising political parties and campaigning for members of Congress or presidential candidates. This research focuses on sentiment analysis towards Puan Maharani, the Speaker of the Indonesian House of Representatives (DPR RI), using data from the social media platform Twitter. Twitter, as a platform that allows users to express opinions in a concise format, is used as the main source of information in this research. The K-Nearest Neighbor algorithm for sentiment analysis technique is utilized to classify individual tweets into positive or negative categories regarding views on Puan Maharani. The methods used in this research include data crawling, labeling, and data preprocessing, which involve case folding, cleaning, tokenizing, negation handling, normalization, stopword removal, and stemming. For the classification process, the K-Nearest Neighbor method, feature weighting (TF-IDF), and feature selection (thresholding) are employed, with a threshold value of 0.001. The data used comprises 9,000 tweets in the Indonesian language. The results of the testing conducted in the K-Nearest Neighbor method, using confusion matrices, with 6 different values of K (3, 5, 7, 9, 11, 13), with comparison mechanisms of 90:10, 80:20, and 70:30 achieved the highest accuracy of 90.00% with K = 11 from the comparison using the 90:10 ratio.

Keywords: Puan Maharani; K-NN; Society; Twitter; Sentiment Classification

1. PENDAHULUAN

Pesatnya perkembangan media menghasilkan sejumlah besar media online, seperti media berita hingga media sosial, Twitter, Facebook, Instagram, Google, Tumblr, Linkedin, dll. Ketika teknologi informasi berkembang pesat, opini masyarakat dapat disampaikan tanpa terbatas waktu melalui media sosial. Twitter, dengan keterbatasan karakter dan hastag "#" yang dapat digunakan oleh pengguna, dianggap lebih mudah diambil informasi tentang opini dan sentimen yang ada. Saat ini, media sosial banyak digunakan untuk berkomunikasi dan mencari teman, namun juga untuk aktivitas lainnya. Mengiklankan produk, membeli dan menjual apa pun, termasuk mengiklankan partai politik dan berkampanye untuk anggota Kongres atau calon presiden. Dalam dunia politik, elektabilitas adalah subjek kontroversial yang memengaruhi individu atau bahkan pihak tertentu. Masyarakat senang memberikan tanggapan, yang seringkali diartikan sebagai komentar untuk tokoh-tokoh atau pihak tertentu, karena setiap kabar atau informasi terkini dapat diakses dengan mudah dari berbagai sumber informasi digital.

Puan Maharani merupakan tokoh Partai (PDI-P) yang sekarang menjabat sebagai Ketua DPR RI dari tahun 2019 hingga 2024. Dia dilantik sebagai ketua DPR termuda ketiga, berusia 46 tahun, setelah Achmad Sjaichu dan I Gusti Gde Subamia. Sebelumnya beliau dari 2014 hingga 2019, Dia adalah menteri yang bertanggung jawab atas pembangunan manusia dan kebudayaan Indonesia. Ia merupakan perempuan termuda dan perempuan pertama yang menjabat sebagai Menteri Koordinator. Puan Maharani menjelma menjadi politisi papan atas di media sosial. Dengan mengunggah cuitan

atau tweet, orang dapat memberi tahu masyarakat tentang pendapat atau komentar tentang Puan Maharani melalui Twitter.[1]

Studi komputasional tentang pendapat, perasaan, dan perasaan yang ditulis dikenal sebagai sentiment analysis. Opinion mining dilakukan ketika sekumpulan dokumen teks berisi opini (sentimen) tentang suatu hal. Tujuan opini mining adalah komponen objek yang dikomentari dari setiap dokumen dan untuk mengetahui apakah komentar tersebut bersifat positif atau negatif. Penelitian ini menggunakan analisis sentimen untuk mengevaluasi pendapat seseorang yang ditujukan kepada Puan Maharani, yang dapat dikategorikan sebagai pendapat positif atau negatif. Pada dasarnya, perasaan yang ditujukan kepada Puan Maharani bisa dijadikan sebuah parameter opini masyarakat terhadap Puan Maharani [2].

Twitter adalah platform penting untuk komunikasi politik di Indonesia karena memungkinkan setiap orang untuk berpartisipasi dalam sebuah *tweet*, retweet, dan menanggapi apa play on words yang berkaitan dengan topik politik terkini. Ini terutama berlaku pada tahun 2013, ketika Twitter berkembang menjadi platform media sosial paling populer di Indonesia setelah penggunaan Facebook. Banyak pengguna Twitter yang aktif membahas isu politik dan berkomentar atau mengkritik politisi, termasuk Puan Maharani [3].

Untuk melakukan analisis sentimen, penelitian ini menggunakan bahasa Indonesia untuk melihat dan mengumpulkan informasi tentang opini dan pandangan masyarakat Indonesia. situs web sosial yang ditujukan kepada Puan Maharani. Dua kelompok akan terdiri dari analisis sentimen ini: positif dan negatif. Studi ini menganalisis *tweet* warganet media sosial Twitter tentang pendapat publik tentang kinerja Puan Maharani selama jabatannya sebagai Ketua DPR RI. [4].

Klasifikasi sentimen digunakan untuk menggabungkan data dari kumpulan data yang tidak terorganisir sehingga penelitian ini dapat mengidentifikasi persepsi masyarakat terhadap Puan Maharani, terutama di Twitter. Studi ini menggunakan teknik klasifikasi sentiment K-Nearest Neighbor(K-NN), metode yang salah satu penyelesain sederhana untuk menyelesaikan masalah klasifikasi. Metode ini sering digunakan untuk mengklasifikasikan teks dan data, dan menghitung nilai prediksi untuk tetangga terdekat. [5].

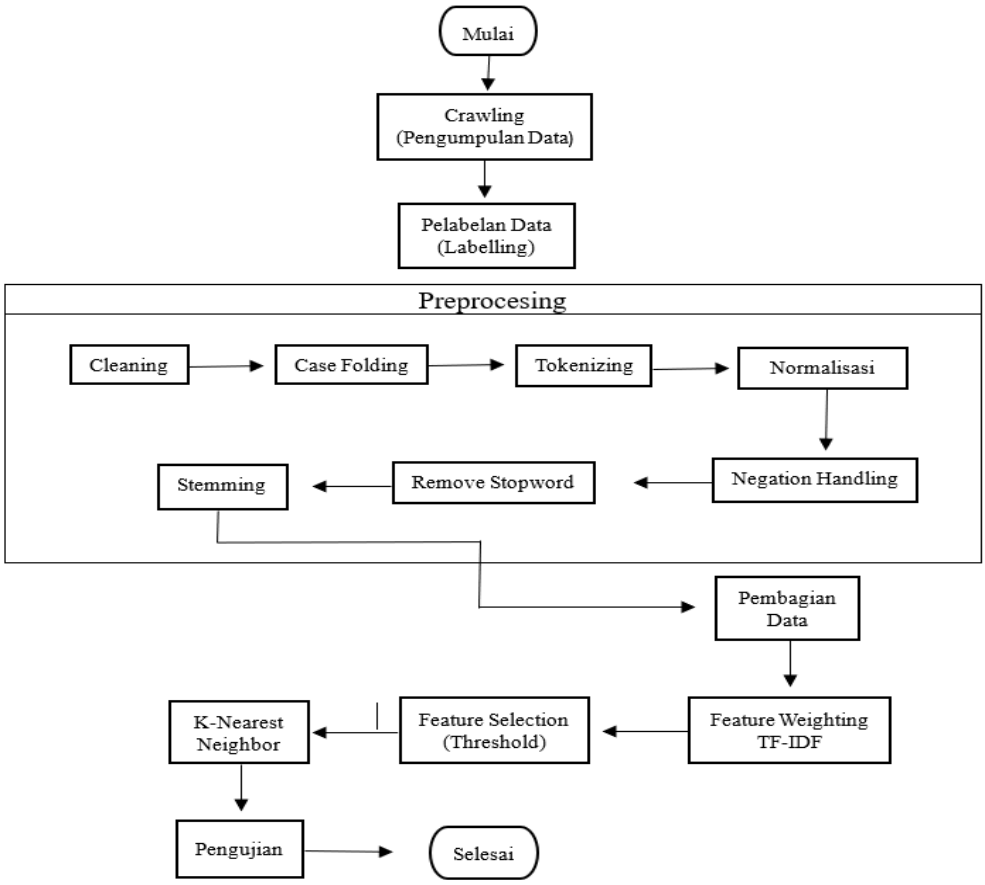
Dengan menggunakan contoh pelatihan mayoritas dari kategori Nilai K-tetangga terdekat, algoritma K-Nearest Neighbor digunakan untuk mengelompokkan objek. Nilai K adalah jumlah total objek yang ada di sekitarnya. Penelitian sebelumnya tentang metode K-Nearest Neighbor telah banyak dibahas, seperti penelitian yang dilakukan oleh Abdul Malik Zuhdi tentang penggunaan metode ini untuk menilai perasaan pengguna Twitter terhadap capres Indonesia pada tahun 2019. Dengan 1000 data, penelitian ini menemukan bahwa nilai K untuk pengujian akurasi sistem adalah 81,83%, dengan K bernilai 3 Kemampuan metode K-Nearest Neighbor untuk mengklasifikasikan respons pengguna Twitter, yang berguna sebagai sumber evaluasi dan penilaian capres di Indonesia 2019 adalah bukti penggunaan metode klasifikasi yang efektif [6]. Dalam studi kasus yang dilakukan oleh Dyah Apriliani, Analisis Pandangan Masyarakat Terhadap Vaksinasi COVID-19 dengan metode K-Nearest Neighbor(K-NN) menghasilkan akurasi 79,25%, dengan K bernilai 5. Dengan menggunakan 2241 data, hasilnya mencapai K 5 [7]. Penelitian Sayed Omas Tutus Arifta, yang membahas analisis sentiment Twitter terhadap Klasifikasi Sentimen Masyarakat di Twitter terhadap Ganjar Pranowo, menggunakan metode K-Nearest Neighbor dengan 4.000 data Twitter, menemukan bahwa akurasi klasifikasi sentimen mencapai 81,%[8]. Studi Made Surya Adi Palguna, yang serupa dengan penelitian ini, berfokus pada Analisis Sentimen Twitter Pengaruh Tokoh Politik dengan Menggunakan Metode K-Nearest Neighbor. Data yang digunakan mencakup 2.000 data yang diperoleh dari Google Collab, dengan akurasi tertinggi 84,06% dan k = 5. Penelitian ini menunjukkan bahwa teknik K-Nearest Neighbor dapat digunakan dengan sangat baik untuk menganalisis sentimen Twitter terhadap pengaruh pada tokoh politik. Hasilnya menunjukkan bahwa analisis sentimen Twitter dapat menjadi alat yang efektif untuk mengukur bagaimana tokoh politik memengaruhi masyarakat[9].

Dari Pemaparan diatas berdasarkan penelitian yang telah dilakukan oleh (Made & abdul), (Dyah & Sayed), bahwa metode ini terbukti memperoleh tingkat akurasi yang baik. Oleh karena, topik yang akan dibahas adalah cara menggunakan analisis sentimen untuk mengkategorikan komentar masyarakat tentang Puan Maharani di Indonesia. menggunakan metode K-Nearest Neighbor. Dengan 9.000 data yang telah didapat dengan sentimen negatif dan positif, tujuan dari studi ini adalah untuk mendapatkan nilai akurasi dan untuk menentukan reaksi masyarakat terhadap Puan Maharani selama jabatannya sebagai ketua DPR RI. Saat ini menjadi sangat dibicarakan di masyarakat. Karena tindakan dan keputusan Puan Maharani, sebanyak 9.000 data tweet yang dikumpulkan dari 23 April hingga 23 Mei 2023 digunakan. Studi ini akan menggunakan pendekatan K-Nearest Neighbor (K-NN) dengan bantuan alat Google Collab. Studi ini bertujuan untuk mengevaluasi ketepatan teknik K-Nearest Neighbor dalam mengkategorikan perasaan masyarakat terhadap Puan Maharani di Twitter. Selain itu, metode K-Nearest Neighbor akan diuji untuk mengetahui seberapa akurat metode tersebut dalam mengklasifikasikan sentimen masyarakat terhadap Puan Maharani.

2. METODOLOGI PENELITIAN

Sebagai metodologi penelitian, Dalam proses pengolahan datanya, analisis sentimen akan digunakan untuk mengevaluasi komentar sebagai positif atau negatif. Berdasarkan hasil analisis, komentar yang bersifat negatif akan diklasifikasikan dengan menggunakan algoritma K-Nearest Neighbor untuk mengklasifikasikan datanya. Dengan menghitung jarak yang ada antara dokumen uji dan dokumen latih, algoritma ini melakukan proses pengklasifikasian data. Oleh karena itu, untuk pembobotan term dari setiap dokumen, algoritma k-nearest neighbor digunakan pada sistem yang akan dibangun yang didukung dengan metode tf-idf, dan metode pemilihan fitur digunakan untuk memilih fitur terbaik. Selama proses

klasifikasi, fitur yang tidak berguna dipilih untuk dihilangkan. Proses pembuatan analisis sentimen yang didasarkan pada pendapat media sosial Twitter tentang Puan Maharan dijelaskan sebagai berikut:



Gambar 1. Metode Penelitian

Gambar 1 di atas menunjukkan proses metode K-Nearest Neighbor. Untuk mencapai hasil yang diharapkan, penelitian ini menggunakan langkah-langkah yang disusun secara sistematis. Program ini dirancang untuk dapat memaksimalkan dalam penggunaan metode KNN dengan dilakukan analisis sentimen terhadap Puan Maharani melalui Twitter. Berikut penjelasan proses metode K-Nearest Neighbor:

2.1 Crawling (pengumpulan Data)

Data penelitian dikumpulkan melalui crawling data Twitter menggunakan bahasa pemrograman Python. Dari tahapan tersebut diperoleh data sebanyak 9000 data *tweet* yang akan digunakan untuk penelitian tersebut, data *tweet yang* didapatkan menggunakan bahasa Indonesia yang dikumpulkan dari tanggal 23 April 2023 sampai tanggal 3 Mei 2023. Dalam pengambilan data hanya yang bersangkutan tentang Puan Maharani.

2.2 Pelabelan Data

Tahap berikutnya adalah pelabelan manual setelah data dikumpulkan. Pelabelan dilakukan kategori: positif dan negatif. Dengan dilakukan mempelajari makna kalimat daripada hanya memberikan label positif atau negatif. Untuk pelabelannya akan dilakukan oleh gadis Sari Elin S.pd adalah seorang guru Bahasa Indonesia. Data yang diberi label dari bulan April hingga Mei, data yang digunakan sebagai data pelatihan selama proses pelatihan model menggunakan algoritma K-Nearest Neighbor. Label diberikan secara manual oleh tenaga kerja manusia, yaitu peneliti, dengan asumsi mesin tidak dapat merasakan emosi, tetapi manusia dapat.

2.3 Processing

Setelah pelabelan, tahap selanjutnya adalah preprocessing. Data yang diambil dari proses crawling, yang berbentuk teks, dibersihkan terlebih dahulu karena bentuk teks yang tidak terstruktur[10]. Sebelum data digunakan untuk pelatihan model, mereka harus melalui tahap preprocessing setelah dipelabelan. Data dibersihkan melalui, Cleaning, Case Folding, dinormalisasi, ditokenisasi, remove stopwords. Perintah-perintah yang digunakan dalam proses preprocessing dirancang untuk membuat penerapan ke data yang digunakan lebih mudah dan lebih cepat. Tujuan dari tahap ini adalah mengolah kumpulan data sehingga algoritma yang akan digunakan dapat membacanya dan mengekstrak informasi yang diinginkan. Proses preprocessing data meliputi:

a. Cleaning

Karakter yang tidak berguna dihapus selama tahap pembersihan. Ini termasuk simbol yang tidak perlu atau mengganggu dari teks, seperti tanda baca, dan URL, emoji, dan tag HTML [11]. Tujuan dari proses pembersihan atau pembersihan adalah untuk membersihkan dan mempersiapkan data mentah agar dapat digunakan secara efisien dalam analisis atau pemodelan. Ini mencakup menemukan, menangani, dan/atau menghapus masalah data seperti nilai yang hilang, outlier, duplikat, atau kesalahan entri. Dengan membersihkan data, kita dapat memastikan bahwa data yang digunakan untuk analisis atau pembuatan model adalah akurat, konsisten, dan relevan. Dengan demikian, hasilnya menjadi lebih dapat diandalkan dan bermakna..

b. Case Folding

Menggabungkan huruf-huruf dalam teks menjadi bentuk yang konsisten dikenal sebagai case folding, biasanya dalam bentuk huruf kecil (lowercase). Dengan cara ini, dapat mengolah teks tanpa menyadari perbedaan huruf besar dan kecil [12]. Tujuan case folding untuk menyamakan bentuk huruf dalam teks, biasanya dengan mengubah semua huruf menjadi huruf kecil. Ini memungkinkan algoritma untuk menganggap kata yang sama dengan bentuk huruf yang berbeda sebagai kata yang sama karena perbedaan kecil antara huruf. Ini meningkatkan konsistensi pengolahan teks, menjadikan analisis lebih efektif dan akurat.

c. Tokenisasi

Proses tokenisasi adalah pembagian teks atau rangkaian data dibagi menjadi token yang lebih kecil. Token dapat berupa karakter, kata, atau frasa, tergantung pada tujuan aplikasinya. Tujuan utama tokenizing adalah untuk membuat teks lebih mudah dipahami atau dikelola oleh komputer, yang memungkinkan pemrosesan dan analisis lebih lanjut [13].

d. Normalisasi

Pada titik ini, kata yang dikembalikan dari bentuk tidak baku menjadi bentuk yang baku dikembalikan atau [14]. Normalisasi kalimat dilakukan untuk meningkatkan struktur atau format kalimat sehingga lebih jelas, konsisten, dan mudah dipahami. Perbaikan tata bahasa, penyesuaian gaya penulisan, atau penyempurnaan sintaksis dapat dilakukan selama proses ini untuk membuat kalimat lebih konsisten dan sesuai dengan standar linguistik yang berlaku. Normalisasi kalimat membuat komunikasi lebih efektif karena pembaca atau pendengar dapat memahami pesan yang ingin disampaikan dengan lebih baik.

e. Negation Handling

Negation Handling berarti mengubah kata-kata seperti "tidak", "bukan", "tak", atau "tidak ada" menjadi kata baru dan kemudian mengubahnya menjadi kata baru dengan arti yang sama [15]. Penanganan negasi dilakukan untuk memastikan bahwa kata-kata negatif seperti "tidak" atau "bukan" tidak memengaruhi pemahaman yang tepat tentang makna kalimat atau frasa saat menganalisis teks. Dengan mengatasi negasi, kita dapat memastikan interpretasi yang akurat dari teks, terutama ketika melibatkan analisis sentimen.

f. Remove Stopword

Proses menghilangkan kata yang dianggap sebagai "stopword" dari teks. Stopword adalah kata biasa dan sering muncul yang tidak memberikan makna atau informasi yang signifikan selama proses pemrosesan teks. Langkah ini dilakukan untuk meningkatkan perhatian perhitungan pada kata yang lebih penting [16]. Dengan menghilangkan stopwords, pemrosesan teks menjadi lebih efisien dan akurat karena kita dapat fokus pada kata-kata kunci yang membawa makna utama dalam teks. Ini juga dapat membantu mengurangi dimensi data dalam pemodelan teks, yang dapat membuat model yang lebih sederhana dan efisien.

g. Stemming

Pada tahap ini, proses pemrosesan bahasa alami dilakukan dengan mengurangi kata ke dalam bentuk dasar atau dikenal sebagai "stems". Stemming hanya menyisakan inti atau akar kata, menghilangkan akhir atau awal kata. Dengan menggunakan stemming, variasi morfologis dalam teks dikurangi, karena kata-kata dengan akar kata yang sama dianggap identik [17]. Stemming membantu mengurangi dimensi kata dalam analisis teks, pemodelan teks yang lebih akurat dan efisien, pencarian informasi, dan tugas-tugas lainnya yang terkait dengan pemrosesan teks. Dengan menggunakannya, kita dapat menyederhanakan representasi teks tanpa kehilangan maknanya, yang memungkinkan analisis dan pengambilan informasi yang lebih efektif..

2.4 Pembagian Data

Dataset dibagi menjadi subset yang berbeda dalam proses pelatihan, validasi, dan pengujian model analisis sentimen[18]. Tujuan pembagian data adalah untuk mengukur kinerja model secara objektif, menghindari overfitting, dan memastikan generalisasi yang baik pada data yang baru. Data *tweet* dalam penelitian akan dibagi rasio 90:30, 80:20, dan 70:30.

2.5. Feature Weighting

Pada tahap TF-IDF (Term Frequency-Inverse Document Frequency), menggunakan pembobotan untuk membedakan kata (term) jarang dan sering. Metode TF-IDF (Term Frequency—Inverse Document Frequency) digunakan untuk pembobotan [19]. Proses pembobotan kata akan digunakan untuk mengubah data kata preprocessing menjadi angka. Proses ini menentukan seberapa berat setiap kata yang akan digunakan sebagai fitur; semakin banyak dokumen yang

diproses, semakin banyak fitur yang dihasilkan. Pendekatan TF-IDF digunakan untuk pembobotan kata pada setiap pembagian data, yaitu *tweet* dengan rasio 90:10, 80:20, dan 70:30 dengan data training dan data uji. Pada tahap ini, perhitungan untuk setiap kata dilakukan berdasarkan term frekuensi yang akan ditemukan pada data.

2.6 Feature Selection

Pada tahap ini proses pemilihan subset teks tweet yang paling relevan dan informatif untuk digunakan dalam model analisis sentimen. Seleksi fitur berpengaruh langsung terhadap hasil klasifikasi [20]. Model-model tersebut mempresentasikan kata frase dalam sebuah tweet yang dapat memberikan informasi sentimental penting di dalamnya. Pada Feature Selection dilakukan pada setiap pembagian data yang sudah melewati tahap pembobotan Feature Weighting tweet dengan rasio 70:30, 80:20 dan 90:10 dengan data uji dan data latih. Tujuan analisis sentimen ini adalah untuk menemukan dan memahami perasaan atau opini dalam teks tweet yang telah melalui proses preprocessing.

2.7 K-Nearest Neighbor

Pada tahap ini, setelah data melalui tahap preprocessing, data telah siap untuk diolah dengan metode K-Nearest Neighbor. Metode ini adalah salah satu algoritma klasifikasi yang diajarkan dengan pengawasan; mayoritas kelas tetangga terdekat menentukan hasil klasifikasi.[21]. Metode K-Nearest Neighbor mengklasifikasikan data yang dimasukkan ke dalam kelas yang telah ditetapkan sebelumnya dengan tingkat kemiripan atau jarak terdekatnya dengan data latih atau dataset yang ada. Nilai "k" jarak terdekat mengklasifikasikan data yang dimasukkan ke dalam kelas yang telah ditetapkan sebelumnya [22].

3. HASIL DAN PEMBAHASAN

Hasil penelitian adalah analisis sentimen melalui klasifikasi ulasan Twitter yang terkait dengan Puan Maharani menggunakan metode algoritma K-Nearest Neighbor. Penelitian ini bertujuan untuk mendapatkan hasil klasifikasi data dengan menggunakan algoritma K-Nearest Neighbor (K-NN).untuk dapat hasil klasifikasi data yang paling akurat berdasarkan data sentimen Puan Maharani di media sosial Twitter.

3.1 Pengumpulan data

Bahasa pemrograman Python digunakan untuk mengcrawling data di Twitter. Data ini dikumpulkan dari 23 April 2023 hingga 3 Mei 2023. Pengambilan data hanya yang bersangkutan tentang Puan Maharani. Setelah data dikumpulkan dari 23 April hingga 3 Mei 2023, setiap tweet yang memiliki sentimen disortir untuk dimasukkan ke dalam dataset yang digunakan untuk penelitian analisis sentimen ini, yang terdiri dari 9.000 tweet.

3.2 Pelabelan Manual

Tahap berikutnya adalah pelabelan manual setelah data dikumpulkan. Komentar dikategorikan ke dalam kategori positif dan negatif untuk melakukan pelabelan. Pelabelan ini dilakukan dengan memahami makna kalimat daripada memberikan label positif atau negatif setiap kata. Pelabelan akan dilakukan oleh Gadis Sari Elin S.pd, seseorang yang mengajar bahasa Indonesia. Hasil pelabelan dari validator bahasa Indonesia menunjukkan bahwa 7.800 data diberi label positif dan 1.200 data diberi label negatif.

3.3 Preprocessing

Pada tahapan ini, tabel 1 menunjukkan hasil dari tahapan preprocessing data ini: Proses preprocessing diperlukan untuk mendapatkan data bersih karena tweet yang diambil dari Twitter merupakan data mentah. Preprocessing teks bertujuan untuk meningkatkan kualitas data karena merupakan tahap awal dalam pemrosesan data sampai data sesuai dengan kebutuhan analisis dan siap untuk tahap berikutnya. Hal-hal berikut telah dilakukan:

Tabel 1. Hasil Preprocessing

Tahapan	Sebelum	Sesudah
Cleaning	“Tan_Mar Kalo gitu barisan oposisi sangat tidak cerdas berarti Harusnya pihak oposisi yg mencalon presidenkan puan Maharani dari pada mencalonkan Anies”	Kalo gitu barisan oposisi sangat tidak cerdas berarti Harusnya pihak oposisi yg mencalon presidenkan puan Maharani dari pada mencalonkan Anies
Case Folding	Kalo gitu barisan oposisi sangat tidak cerdas berarti Harusnya pihak oposisi yg mencalon presidenkan puan Maharani dari pada mencalonkan Anies	**kalo** gitu barisan oposisi sangat tidak cerdas berarti **harusnya** pihak oposisi yg mencalon presidenkan puan **maharani** dari pada mencalonkan **anies**
Tokenizing	kalo gitu barisan oposisi sangat tidak cerdas berarti harusnya pihak oposisi yg mencalon	“kalo” , ”gitu” , “barisan” , “oposisi” , “sangat” , “tidak” , “cerdas” , “berarti” , “harusnya” , “pihak” , “oposisi” , “yg” , “mencalon” , “presidenkan” ,

	presidenkan puan maharani dari pada mencalonkan anies"	"puan" , "maharani" , "dari" , "pada" , "mencalonkan" , "anies"
Normalisasi	"kalo" , "gitu" , "barisan" , "oposisi" , "sangat" , "tidak" , "cerdas" , "berarti" , "harusnya" , "pihak" , "oposisi" , "yg" , "mencalon" , "presidenkan" , "puan" , "maharani" , "dari" , "pada" , "mencalonkan" , "anies"	"**kalau**" , "**begitu**" , "barisan" , "oposisi" , "sangat" , "tidak" , "cerdas" , "berarti" , "**seharusnya**" , "pihak" , "oposisi" , "**yang**" , "mencalon" , "presidenkan" , "puan" , "maharani" , "dari" , "pada" , "mencalonkan" , "anies"
Negation Handling	kalau begitu barisan oposisi sangat tidak cerdas berarti	kalau begitu barisan oposisi sangat **bodoh** berarti
Remove Stopword	"kalau" , "begitu" , "barisan" , "oposisi" , "sangat" , "bodoh" , "berarti" , "seharusnya" , "pihak" , "oposisi" , "yang" , "mencalon" , "presidenkan" , "puan" , "maharani" , "dari" , "pada" , "mencalonkan" , "anies"	"barisan" , "oposisi" , "sangat" , "bodoh" , "seharusnya" , "pihak" "oposisi" , "mencalon" , "presidenkan" , "puan" , "maharani"
Stemming	barisan oposisi sangat bodoh seharusnya presidenkan mencalonkan	**baris** oposisi sangat bodoh **harus** **presiden** **calon**

Pada tahap ini, data yang diambil dari proses crawling, yang berbentuk teks, harus diproses terlebih dahulu. Hal ini disebabkan oleh fakta bahwa teks yang tidak terstruktur dan dan hal-hal yang berisik harus dihilangkan terlebih dahulu. Preprocessing dilakukan untuk memastikan bahwa data yang digunakan untuk proses pencarian pengetahuan berkualitas lebih baik, sehingga pengetahuan yang dihasilkan juga lebih rendah.

3.4 Feature Weighting

Selanjutnya, pendekatan TF-IDF digunakan untuk membatasi kata pada setiap pembagian data, yaitu data tweet, dengan menggunakana rasio 90:10, 80:20, dan 70:30 untuk data uji dan latih. Feature Weighting, Setiap kata atau term dari hasil preprocessing akan dibobot menggunakan TF-IDF setelah tahap preprocessing teks selesai.

3.5 Feature Selection

Pada Feature Selection dilakukan pada setiap pembagian data yang sudah melewati tahap pembobotan Feature Weighting *tweet* dengan rasio 90:10, 80:20 dan 70:30 dengan data latih dan data uji, kemudian ambang batas akan digunakan untuk memilih fitur terbaik. Seleksi fitur dilakukan dengan tujuan menghapus fitur yang tidak berguna dalam proses klasifikas. Ambang batas untuk penelitian ini adalah 0,001.

3.6 Klasifikasi K-Nearest neighbor

Pengujian dilakukan menggunakan confusion matriks untuk melakukan pengujian pada data latih dan data uji. Pada titik ini, metode K-Nearest Neighbor (KNN) digunakan untuk menghitung jarak ke tetangga terdekat. Nilai K bervariasi, yaitu bilangan ganjil antara 3 dan 13. Pengujian pengaruh nilai K menggunakan nilai K yang bervariasi untuk menentukan nilai K ideal untuk melakukan proses klasifikasi K-Nearest Neighbor terhadap pada hasil akurasi di sistem. Nilai K digunakan sebagai parameter pengujian, menunjukkan dampak mereka terhadap akurasi sistem. Mekanisme perbandingan 90:10, 80:20, dan 70:30 untuk menemukan nilai akurasi terbaik. Dengan menggunakan dua pengujian yaitu, *unbalance* dan *balance*.

3.6.1 Pengujian Data *Unbalance*

Pengujian menggunakan 6 nilai K yang bervariasi, (3,5,7,9,11,13), dengan mekanisme komparasi 90:10, 80:20, dan 70:30 untuk 9.000 data. 1.200 data kelas negatif dan 7.800 data kelas positif digunakan. Tabel berikut menunjukkan hasil didapatkan dari pengujian.

Tabel 2. Hasil Pengujian data *Unbalance*

Nilai K	Accuracy		
	70:30	**80:20**	**90:10**
3	86.67	87.22	86.78
5	88.22	89.00	87.67

7	88.44	88.72	88.56
9	88.52	89.33	89.44
11	88.81	89.28	90.00
13	88.78	89.33	89.78

Tabel 2 merupakan dari hasil pengujian *Unbalance* setiap hasil dari pengujian memiliki nilai akurasi yang berbeda, seperti yang ditunjukkan dalam Tabel 2. Nilai K 11 memiliki akurasi paling tinggi sebesar 90,00%, 90.00, skor precision 90.0%. skor recall 90.0% dan skor f1 score 90.0%, dengan perbandingan digunakan 90:10, dan Nilai K 3 memiliki akurasi paling rendah sebesar 86,67%, dengan perbandingan 70:30.

3.6.2 Pengujian Data *Balance*

2.000 data digunakan untuk pengujian ini, terdiri dari 1.000 data kelas negatif dan 1.000 data kelas positif, seperti yang tercantum dalam tabel berikut;

Tabel 3. Hasil Pengujian Data *Balance*

Nilai K	Accuracy		
	70:30	**80:20**	**90:10**
3	74.14%	75.65%	75,00%
5	75.00%	76.08%	78.02%
7	74.71%	75.00%	78.45%
9	76.29%	75.22%	76.29%
11	74.71%	75.43%	78.42%

Tabel 3 merupakan dari hasil pengujian *Balance* menghasilkan hasil akurasi yang tertinggi rasio 90:10 pada nilai K=7 dengan skor accuracy 78.45%, skor precision 78.4%, skor recall 78.4%, dan skor f1 score 78.4%. Dalam setiap pengujian nilai K memiliki variasi dalam hasil akurasi antara setiap nilai K. Hasil dari setiap nilai K yang berbeda dapat disebabkan oleh sifat unik dari dataset, trade-off antara underfitting dan overfitting, randomness dalam algoritma, perbedaan dalam ukuran dan kualitas data uji, serta metrik evaluasi yang digunakan. Faktor-faktor ini dapat menyebabkan fluktuasi dalam performa model di setiap nilai k.

4. KESIMPULAN

Berdasarkan analisis yang dilakukan, dapat dikatakan bahwa algoritma K-Nearest Neighbor memiliki kemampuan yang lebih baik untuk mengklasifikasikan sentimen masyarakat tentang Puan Maharani. Pengujian *Unbalance*, yang dilakukan dalam dua pengujian, memiliki tingkat akurasi terbaik dalam proses klasifikasi di pengujian *Unbalance*, yaitu 90.00% pada K=11 pada K=11 dengan analisis data pelatihan dan pengujian 90:10. Pengujian *Balance*, yang diklasifikasikan menggunakan dengan metode K-Nearest Neighbor, memiliki nilai akurasi proses klasifikasi, yaitu 78,45% pada K=7 dengan perbandingan data latif dan data uji 90:10. Pada saat percobaan *Unbalance* data yang pengujian *Unbalance*, hasil klasifikasi K-Nearest Neighbor lebih banyak memilih pada kelas mayoritas positif, sehingga dalam klasifikasi menyebabkan kesalahan tweet negatif. Tujuan penelitian adalah untuk mendapatkan nilai akurasi melalui dua tes.

REFERENCES

[1] I. Febriansyah, M. Fikry, and Yusra, "Analisis Sentiment di Twitter terhadap Anies Baswedan sebagai Bakal Calon Presiden 2024 Menggunakan Metode K-Nearest Neighbor," G-Tech: Jurnal Teknologi Terapan, vol. 7, no. 3, pp. 1061–1070, Jul. 2023, doi: 10.33379/gtech.v7i4.2723.

[2] Z. Ulfah Siregar, R. Ruli, A. Siregar, and R. Arianto, "KLASIFIKASI SENTIMENT ANALYSIS PADA KOMENTAR PESERTA DIKLAT MENGGUNAKAN METODE K-NEAREST NEIGHBOR," vol. 8, no. 1, 2019.

[3] F. Zahria Emeraldien, R. Jefri Sunarsono, R. Alit, J. Raya Rungkut Madya, G. Anyar, and J. Timur, "TWITTER SEBAGAI PLATFORM KOMUNIKASI POLITIK DI INDONESIA." 2019

[4] D. Aby Vonega, A. Fadila, and D. Ely Kurniawan, "Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024," 2022. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[5] R. T. Prasetio, "SELEKSI FITUR DAN OPTIMASI PARAMETER k-NN BERBASIS ALGORITMA GENETIKA PADA DATASET MEDIS," JURNAL RESPONSIF, vol. 2, no. 2, pp. 213–221, 2020, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti

[6] A. Malik Zuhdi, E. Utami, and S. Raharjo, "ANALISIS SENTIMENT TWITTER TERHADAP CAPRES INDONESIA 2019 DENGAN METODE K-NN," 2019. doi: Vol 5 No 2 (2019): Juni.

[7] D. Apriliani, A. Susanto, M. Fikri Hidayattullah, and G. Wiro Sasmito, "Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors," vol. 8, no. 1, 2023.

[8] S. Omas Tutus Arifta and M. Fikry, "Klasifikasi Sentimen Masyarakat di Twitter terhadap Ganjar Pranowo dengan Metode K-Nearest Neighbor," JSAI : Journal Scientific and Applied Informatics, vol. 06, no. 02, 2023, doi: 10.36085.

[9] "Analisis Sentimen Twitter Pengaruh Tokoh Politik dengan Menggunakan Metode K-Nearest Neighbor." doi: Vol 2 No 2 (2024): JNATIA Vol. 2, No. 2, Februari 2024.

[10] M. Furqan, S. Mayang Sari, and P. Ilmu Komputer Fakultas Sains dan Teknologi, "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia Sentiment Analysis using K-Nearest Neighbor towards the New Normal During the Covid-19 Period in Indonesia," 2022. doi: 10.33633/tc.v21i1.5446.

[11] A. Yoga Pratama et al., "Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," 2021. doi: Vol 5, No 2 (2021).

[12] R. M. Candra and A. Nanda Rozana, "Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor," IT Journal Research and Development, vol. 5, no. 1, pp. 45–52, Jul. 2020, doi: 10.25299/itjrd.2020.vol5(1).4962.

[13] A. Noviyanti, Y. Umaidah, R. Mayasari, U. Singaperbangsa, and K. Abstract, "Analisis Sentimen Pada Pembelajaran Daring Menggunakan Metode K-Nearest Neightbour (Studi Kasus: SMA Negeri 3 Cikampek)," Jurnal Ilmiah Wahana Pendidikan, 2022, doi: 10.5281/zenodo.6943200.

[14] M. Sholeh, D. Andayati, R. Yuliana Rachmawati, P. Studi Informatika, and F. Teknologi Informasi dan Bisnis, "DATA MINING MODEL KLASIFIKASI MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR DENGAN NORMALISASI UNTUK PREDIKSI PENYAKIT DIABETES." 2022, doi: 12(02):77-87.

[15] D. Ferarizki, M. Fikry, F. Yanto, and F. Insani, "Klasifikasi Sentimen Masyarakat di Twitter Terhadap Ancaman Resesi Ekonomi 2023 dengan Metode K-Nearest Neighbor," vol. 4, no. 2, 2023, doi: 10.30865/klik.v4i2.1315.

[16] A. Deviyanto, M. R. Didik Wahyudi, and T. Informatika UIN Sunan Kalijaga Yogyakarta Jl Marsda Adi Sucipto No, "PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR," Jurnal Informatika Sunan Kalijaga, vol. 3, no. 1, pp. 1–13, 2018

[17] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION, vol. 1, no. 1, pp. 43–49, 2019, doi: https://doi.org/10.52985/insyst.v1i1.36.

[18] C. Heltroyce, G. Feoh, and I. Made Dwi Ardiada, "SENTIMENT ANALYSIS ON THE INCREASE OF FUEL OIL PRICES USING THE K-NEAREST NEIGHBOR ALGORITHM ANALISIS SENTIMEN TERHADAP KENAIKAN HARGA BAHAN BAKAR MINYAK MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR," 2024. doi: Vol. 3, No.1 April 2024.

[19] A. Asro'i and H. Februariyanti, "Analisis Sentimen Pengguna Twitter terhadap Perpanjangan PPKM Menggunakan Metode K-Nearest Neighbor," Jurnal Khatulistiwa Informatika, vol. 10, no. 1, pp. 17–24, 2022.

[20] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 6, no. 1, p. 49, Jan. 2022, doi: 10.30865/mib.v6i1.3312.

[21] N. Basidt, E. Supriyadi, and A. Susilo, "Perbandingan Algoritma Klasifikasi dalam Analisis Sentimen Opini Masyarakat tentang Kenaikan Harga BBM" 2023

[22] R. Noviantho, A. Siswo, R. Ansori, and R. R. Septiawan, "ANALISIS SENTIMEN PADA KOMENTAR VIDEO ULASAN MAKANAN DARI SALURAN YOUTUBE BERBAHASA INDONESIA MENGGUNAKAN K-NEAREST NEIGHBOR SENTIMENT ANALYSIS ON VIDEO COMMENTS ABOUT FOOD REVIEW FROM INDONESIAN YOUTUBE CHANNELS USING K-NEAREST NEIGHBOR." 2021