KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 3, No 6, Juni 2023, Hal 786-797
DOI 10.30865/klik.v3i6.890
https://djournals.com/klik

Klasifikasi Sentimen Masyarakat di Media Sosial Twitter terhadap Calon Presiden 2024 Prabowo Subianto dengan Metode K-NN

Avaldy Rahmat Rivita*, Yusra, Muhammad Fikry

Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru , Indonesia
Email: [1,*]11950115017@students.uin-suska.ac.id, [2]yusra@uin-suska.ac.id, [3]muhammad.fikry@uin-suska.ac.id
Email Penulis Korespondensi: 11950115017@students.uin-suska.ac.id

Abstrak− Pemilihan Presiden RI Tahun 2024 adalah tahapan demokrasi untuk menentukan Presiden Negara Indonesia dan Wakil Presiden Negara Indonesia periode 2024-2029 yang dijadwalkan berlangsung pada Rabu, 14 Februari 2024. Pemilihan ini merupakan pemilihan Presiden dan Wakil Presiden langsung yang kelima di Indonesia. Beberapa partai saat ini sudah mencalonkan atau memilih calon presidennya untuk pemilihan presiden 2024. Tiga calon presiden sudah muncul, yakni Prabowo Subianto, Ganjar Pranowo, dan Anies Baswedan. Berdasarkan survei, Prabowo Subianto menjadi capres dengan elektabilitas paling tinggi dibandingkan pesaing lainnya. Tak luput dari pandangan masyarakat terhadap bakal calon presiden 2024 khususnya Prabowo Subianto banyak menimbulkan pro dan kontra. Pandangan masyarakat dapat dilihat melalui media sosial seperti Twitter. Penelitian ini memiliki tujuan untuk mengklasifikasikan sentimen masyarakat terhadap Calon Presiden (capres) Prabowo Subianto di Twitter. Jumlah data yang digunakan adalah 2100 tweet yang dikumpulkan berdasarkan kata kunci yaitu "Calon Presiden" dan "Prabowo Subianto". Penerapan metode K-Nearest Neighbor (K-NN) dengan pembobotan berupa TF-IDF dan seleksi fitur berupa Threshold akan dilakukan implementasi dengan menggunakan Google Colab. Berdasarkan hasil pengujian metode K-NN menggunakan confusion matrix pada tujuh nilai K yaitu (3,5,7,9,11,13,15) dengan perbandingan yang digunakan 70:30, 80:20, 90:10 diperoleh akurasi tertinggi pada K=5 pada rasio data latih dan data uji 80:20.

Kata Kunci: Klasifikasi Sentimen; K-NN; Twitter; Calon Presiden; Prabowo Subianto

Abstract−The 2024 Republic of Indonesia Presidential Election is a democratic stage to determine the President of the Republic of Indonesia and Vice President of the State of Indonesia for the 2024-2029 period which is scheduled to take place on Wednesday, 14 February 2024. This election is the fifth direct presidential and vice presidential election in Indonesia. Several parties have currently nominated or selected their presidential candidates for the 2024 presidential election. Three presidential candidates have emerged, namely Prabowo Subianto, Ganjar Pranowo, and Anies Baswedan. Based on a survey, Prabowo Subianto is the presidential candidate (capres) with the highest electability compared to other competitors. The society's view of the 2024 presidential candidate, especially Prabowo Subianto, has raised many pros and cons. Society's view can be seen on social media, like one of this is the Twitter. This study aims to classify public sentiment towards the Presidential Candidate (capres) Prabowo Subianto on Twitter. The amount of data used is 2100 tweets which are collected based on the keywords "Presidential Candidate" and "Prabowo Subianto". The application of the K-Nearest Neighbor (K-NN) method with weighting in the form of TF-IDF and Feature Selection in the form of Threshold will be implemented using Google Colab. Based on the results of testing the K-NN method using the confusion matrix at seven K values, namely (3,5,7,9,11,13,15) with the comparisons used 70:30, 80:20, 90:10 the highest accuracy was obtained at K = 5 at the ratio of training data and test data 80:20.

Keywords: Sentiment Classification; K-NN; Twitter; Presidential Candidates; Prabowo Subianto

1. PENDAHULUAN

Pemilihan Presiden RI Tahun 2024 adalah tahapan demokrasi untuk menentukan Presiden Negara Indonesia dan Wakil Presiden Negara Indonesia periode 2024-2029 yang dijadwalkan berlangsung pada Rabu, 14 Februari 2024. Pemilihan ini merupakan pemilihan Presiden dan Wakil Presiden langsung yang kelima di Indonesia. Presiden Joko Widodo serta Susilo Bambang Yudhoyono selaku mantan presiden RI tidak bisa ikut serta mencalonkan diri kembali sebab konstitusi dan undang-undang menetapkan batasan dua jabatan bagi seorang Presiden. Pemilihan parlemen ini akan dilakukan serentak dengan pemilihan anggota DPR RI, DPD RI, dan DPRD di seluruh Indonesia. Di sisi lain, pada hari Rabu, 27 November 2024 akan diadakan pemilihan kepala daerah yang baru [1].

Proses pemilihan Presiden dan Wakil Presiden diatur dalam Pasal 6A dan 22E Konstitusi Negara Indonesia Tahun 1945 dan Peraturan Pemilihan Umum. Pasangan Calon Presiden dan Wakil Presiden harus diusulkan oleh partai politik atau koalisi partai yang berhasil meraih paling tidak 20% kursi di parlemen atau setidaknya 25% suara nasional pada pemilihan sebelumnya. Oleh karena itu, hanya partai PDI-P yang diperbolehkan untuk mengajukan pasangan calon tanpa perlu membentuk koalisi. Pelaksanaan pemilihan Presiden dan Wakil Presiden dilakukan dalam 2 tahap, apabila tidak ada pasangan calon yang mendapatkan suara lebih dari 50% pada tahap pertama dan minimal 20% suara didistribusikan di setengah lebih wilayah provinsi di Indonesia. Sebelumnya, hanya pada Pemilihan Umum (pemilu) 2004 dilakukan pemilihan Presiden dan Wakil Presiden dengan sistem dua putaran [2].

Beberapa partai saat ini sudah mencalonkan atau memilih calon presidennya untuk pemilihan Presiden 2024. Tiga Calon Presiden sudah muncul, yakni Prabowo Subianto, Ganjar Pranowo, dan Anies Baswedan. Mengacu pada hasil Jajak Pendapat Litbang Kompas Mei 2023, kualifikasi Prabowo di kalangan Nahdliyin meningkat menjadi 25,8%, dengan Prabowo menjadi Calon Presiden paling berhak berdasarkan pemilih Nahdlatul Ulama (NU) [3]. Selain itu, Prabowo Subianto memperoleh 25,3 persen suara dalam jajak pendapat Indikator Politik Indonesia (IPI). Populi Center juga menerbitkan hasil jajak pendapat terbaru pada 29 Mei 2023, di mana Calon Presiden Prabowo mengungguli Prabowo dibandingkan dengan calon lain Ganjar dan Anies. Peneliti Populi Center Rafif Pamenang Imawan menjelaskan, data

elektabilitas Calon Presiden menunjukkan elektabilitas Prabowo Subianto semakin meningkat. Pertanyaan yang menarik adalah Prabowo Subianto menjadi tokoh paling populer yang dipilih oleh publik sebagai Calon Presiden dengan 22,8% [4].

Selain hasil pemilu di atas, opini publik tentang capres 2024, khususnya Prabowo Subianto, juga tak luput dari perhatian. Pandangan masyarakat dapat dilihat di media sosial seperti Twitter. Twitter adalah salah satu *platform* terbesar di dunia saat ini. Merujuk dari laporan We Are Social dan Hootsuite menunjukan bahwa pada Januari 2023, jumlah pengguna Twitter di seluruh dunia mencapai 556 juta. Angka tersebut naik sebesar 27,4% jika dibandingkan dengan periode yang sama pada tahun sebelumnya [5]. Sehingga di Twitter pengguna dapat dengan mudah menyampaikan dan mendapatkan informasi terbaru terkait bakal Calon Presiden 2024, dan dari pandangan masyarakat terhadap informasi yang beredar di Twitter, dapat ditemui pro dan kontra mengenai Calon Presiden Prabowo Subianto. Pro dan kontra masyarakat di Twitter kemudian menjadi data untuk dilakukan klasifikasi sentimen.

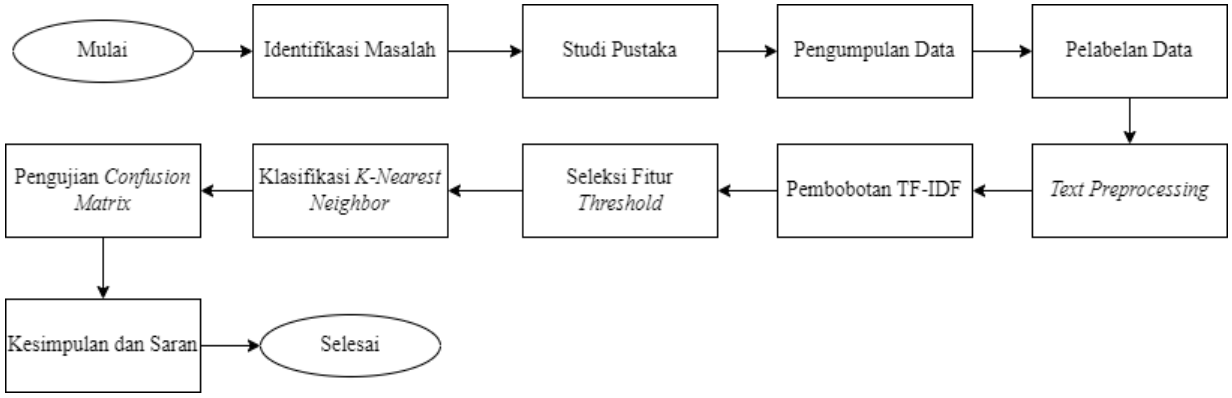
Klasifikasi sentimen digunakan untuk mengelompokkan informasi dari data yang tidak terstruktur, agar pada penelitian ini dapat teridentifikasi pandangan masyarakat khususnya di Twitter terhadap Prabowo Subianto yang akan menjadi Calon Presiden 2024. Teknik klasifikasi sentiment yang diterapkan dalam penelitian ini adalah *K-Nearest Neighbor* (K-NN). K-NN merupakan salah satu teknik sederhana untuk menyelesaikan permasalahan klasifikasi. Teknik K-NN kerap dipakai untuk melakukan klasifikasi pada data dan teks. Dalam teknik ini, dilakukan pengelompokkan objek berdasarkan data yang memiliki jarak terdekat dengan objek tersebut. Metode K-NN menggunakan klasifikasi berdasarkan tetangga terdekat sebagai nilai prediksi untuk *query instance* baru [6].

Terkait penerapan metode K-NN telah banyak dilakukan oleh penelitian sebelumnya, seperti pada penelitian [7] mengenai penerapan metode *K-Nearest Neighbor* (K-NN) terhadap sentimen pengguna Gojek. Penelitian ini memakai 1409 data dengan menerapkan metode K-NN menggunakan *confusion matrix* yang berhasil menyimpulkan bahwa tingkat ketepatan sebesar 79,43% dengan K bernilai 15. Bukti dari klasifikasi yang sukses dengan menggunakan metode K-NN ialah kemampuannya dalam mengklasifikasikan respons dari pengguna Twitter sehingga bermanfaat bagi perusahaan Gojek sebagai bahan evaluasi dan penilaian terhadap layanan Gojek. Penelitian selanjunya [8] penganalisisan sentimen pada *feedback* pengguna untuk aplikasi Bibit dan Bareksa telah dilakukan menggunakan Teknik K-NN. Penelitian ini menghasilkan hasil yang optimal dengan membagi data rasio 60:40 untuk *training* dan *testing*. Hasil *precision* dan untuk *recall* untuk aplikasi Bibit adalah 85,14%, 91,91%, dan 76,44%. Sedangkan untuk aplikasi Bareksa, hasilnya dalah 81,70%, 87,15%, dan 75,73%. Penelitian [9] mengenai pemanfaatan Teknik K-NN digunakan untuk membuat dataset token sentimen menggunakan akun Instagram Brand Elektronik. Dari hasil penelitian, dapat disimpulkan bahwa dataset memiliki tingkat akurasi sebesar 33,38% (untuk sentimen positif), 59,96% (untuk sentimen negatif), dan 56,60% (untuk sentimen netral) dengan menggunakan nilai K=1. Penelitian lainnya [10] evaluasi sentimen kepuasan konsumen dalam layanan publik dengan memanfaatkan teknik K-NN dan pendekatan *Natural Language Processing* (NLP). Hasil penelitian ini mendapat akurasi sebesar 74,00%. Serta penelitian [11] penerapan klasifikasi K-NN untuk sentimen analisis kualitas layanan akun Twitter PT PLN (Persero). Penelitian ini menggunakan 3000 data *tweet*. Data yang terkumpul akan melakukan tahap *preprocessing* dan hasilnya diimplementasikan dalam aplikasi berbasis web menggunakan bahasa pemrograman *python*. Evaluasi metode K-NN menghasilkan nilai akurasi sebesar 87,41%.

Dengan adanya variasi dalam penelitian sebelumnya, mendorong penulis untuk melaksanakan penelitian menggunakan metode K-NN dengan topik yang berbeda. Jumlah data tweet yang digunakan dalam penelitian ini sebanyak 2100 data yang diambil dari Twitter berdasarkan kata kunci "Calon Presiden" dan "Prabowo Subianto". Penelitian ini akan dilakukan beberapa tahapan untuk melakukan klasifikasi sentimen masyarakat di Twitter terhadap bakal Calon Presiden 2024 Prabowo Subianto. Tahapan dimulai dari tahap pengumpulan data, pelabelan data, *preprocessing*, klasifikasi sentimen, pembobotan TF-IDF dan seleksi fitur *threshold*, dan pengujian menggunakan *confusion matrix*.

2. METODOLOGI PENELITIAN

Pada penelitian ini terdapat metodologi penelitian yang terdiri dari beberapa tahapan. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahap Penelitian

2.1 Identifikasi Masalah

Mengidentifikasi masalah merupakan bagian penting dari penelitian seorang peneliti. Identifikasi masalah merupakan langkah awal bagi peneliti sebelum melakukan penelitian dan menuliskan hasilnya dalam publikasi ilmiah. Tanpa identifikasi masalah yang tepat dan matang, temuan penelitian dapat dengan mudah dikritik atau bertentangan dengan penelitian dan teori lain. Karena itu, peneliti harus mengidentifikasi masalah dengan benar.

2.2 Studi Pustaka

Studi pustaka adalah bagian dari artikel ilmiah yang membahas hasil-hasil penelitian sebelumnya. Penelitian kepustakaan berfungsi sebagai referensi ilmiah yang berhubungan dengan penelitian. Studi pustaka juga disebut sebagai tinjauan pustaka atau tinjauan teori. Studi pustaka ini digunakan untuk menjelaskan teori-teori penelitian sebelumnya yang berkaitan dengan topik penelitian.

2.3 Pengumpulan Data

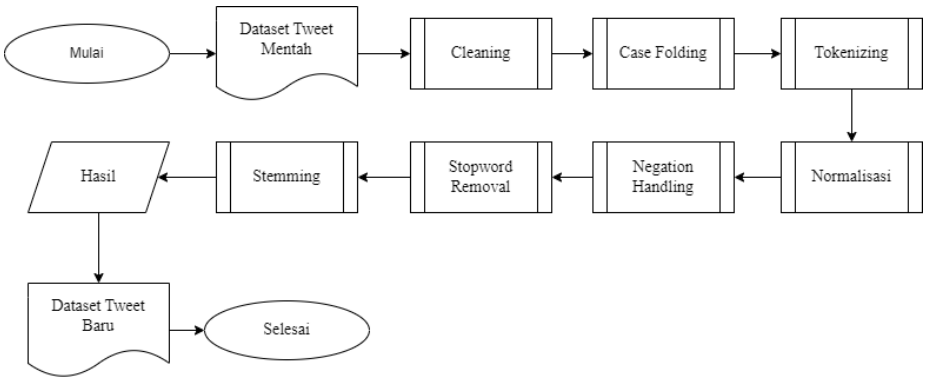
Data yang dipakai dalam penelitian ini berasal dari Twitter API. Data yang diambil adalah *tweet* masyarakat yang dicari berdasarkan kata kunci "Calon Presiden" dan "Prabowo Subianto". Periode *crawling* data yaitu dimulai 29 November 2022 – 3 Januari 2023. Penelitian ini menggunakan data sebanyak 2100 *tweet* sebagai data set sebagai bahan analis, dimana data yang digunakan cukup seimbang untuk mencapai akurasi yang optimal. Keseimbangan dan kedekatan kumpulan data, berpengaruh signifikan terhadap hasil klasifikasi, baik yang memiliki sedikit maupun banyak kelas.

2.4 Pelabelan Data

Pelabelan data dilakukan untuk mengelompokkan kata agar masuk kedalam kelompok yang tepat sesuai dengan kandungan informasinya. Dalam penelitian ini, label data merujuk pada pembagian data ke dalam dua kategori, yaitu positif dan negatif.

2.5 Tahap Text Preproccessing

Setelah tahap pengumpulan dan pelabelan data selesai, dataset tersebut masih dalam kategori data yang tidak terstruktur atau *unstructured* data. Sebelum dilakukan analisis lebih lanjut, dataset harus melalui tahap *text preprocessing* terlebih dahulu untuk menghilangkan dan menangani data yang tidak akurat (*noisy* data) agar hasil perhitungan yang dihasilkan optimal [12]. Berdasarkan dengan tahapan *text preprocessing*, antara lain meliputi *cleaning*, *case folding*, *tokenizing*, normalisasi, *negation handling*, *stopword removal*, dan *stemming*. Gambar 1 menampilkan alur *text preprocessing*.



Gambar 2. Alur *Text Preproccessing*

Alur *text preprocessing* yaitu dimulai dari dataset dari *tweet* yang masih mentah akan dilakukan proses *cleaning* terlebih dahulu, setelah atribut-atribut dibersihkan pada proses *cleaning* kemudian akan dilakukan proses *case folding*, selanjutnya teks diubah menjadi potongan kata pada proses *tokenizing*, setelah teks menjadi bentuk potongan kata selanjutnya kata tersebut akan dicek bentuk normalisasinya pada proses normalisasi, selanjutnya apabila kata yang mengandung negasi akan diubah bentuknya dalam proses *negation handling*, pada proses *stopword removal* kata yang merupakan kata hubung akan dihapuskan, selanjutnya dilakukan pengecekan kembali pada setiap kata, dimana kata akan diubah bentuknya menjadi kata dasar pada proses *stemming* sehingga akan didapatkan hasil data yang mana hasil ini merupakan dataset *tweet* baru.

2.6 Pembobotan TF-IDF

Tujuan dari pembobotan ini adalah untuk memberikan nilai pada fitur kata berdasarkan seberapa sering kata itu muncul. *Term Frequency* (TF) merupakan jumlah kemunculan atau frekuensi kata pada suatu dokumen. Sedangkan *Inverse Document Frequency* (IDF) berguna untuk menentukan relevansi kata kunci dengan istilah yang dicari. Kata kunci yang sering muncul di dalam dokumen akan memiliki dampak yang lebih rendah dalam menentukan hubungan antara kata kunci dan dokumen [13]. Perhitungan TF-IDF dihitung dalam persamaan berikut :

$$\text{Wj,i} = \frac{n_{j,i}}{\Sigma_k n_{k,i}} \; log_2 \; \frac{D}{d_j} \quad (1)$$

Keterangan :
Wj,i = Pembobotan TF-IDF untuk *term* ke j pada dokumen ke i.
$n_{j,i}$ = Banyaknya kemunculan *term* ke j pada dokumen ke i.
$\Sigma_k n_{k,i}$ = Banyaknya seluruh *term* pada dokumen ke i.
D = Jumlah dokumen yang dibangkitkan.
d_j = Jumlah dokumen yang mengandung *term* ke j.

2.7 Seleksi Fitur *Threshold*

Teknik ini dipakai guna mengurangi ataupun menghilangkan elemen yang kurang signifikan untuk meingkatkan ketepatan serta mengurangi waktu pemrosesan didalam klasifikasi sentimen. Dengan mengurangi fitur yang tidak terlalu relevan, algoritma akan lebih efisien memproses data, sehingga pemilihan seleksi fitur yang akurat dapat meningkatkan kecepatan hasil klasifikasi. Ambang batas merupakan persentase jumlah fitur yang dipilih dari seluruh fitur yang telah diurutkan [14].

2.8 *K-Nearest Neighbor*

Metode *K-Nearest Neighbor* (K-NN) adalah metode yang digunakan untuk mengelompokkan objek berdasarkan data pelatihan yang memiliki jarak terdekat dengan objek tersebut. Algoritma K-NN adalah metode pembelajaran terawasi, dan kategori dari instance baru yang di *query* ditentukan berdasarkan mayoritas kategori yang ditemukan pada algoritma K-NN. Kelas yang paling banyak muncul adalah kelas yang ditentukan dari hasil klasifikasi. Kedekatan didefinisikan dalam jarak metrik, seperti jarak *Euclidean* [15].

$$\boldsymbol{D_{x,y}} = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Keterangan :
D = Jarak kedekatan
x = Data latih
y = Data uji
n = jumlah atribut individu antara 1 s.d n
i = atribut individu antara 1 s.d n

2.9 Pengujian *Confusion Matrix*

Evalusi akan dilakukan terhadap hasil klasifikasi yang nantinya diperoleh. Nilai yang diperoleh dari evaluasi dapat digunakan untuk mengukur seberapa berhasil metode yang digunakan dalam pengujian. Salah satu cara untuk melakukan penilaian dalam analisis sentimen adalah dengan menggunakan *confusion matrix*. Definisi dari *confusion matrix* sebuah metode untuk mengevaluasi hasil yang diperoleh dari pelaksanaan algoritma klasifikasi melalui tabel [16]. Tabel 1 merupakan evaluasi dari *confusion matrix*.

Tabel 1. Rancangan Analisis Komputasi

Kelas	Prediksi	
	POSITIF	NEGATIF
POSITIF	TP	FN
NEGATIF	FP	TN

Keterangan **:**
TP (True Positif) = Dokumen yang diklasifikasikan sebagai kelas positif
TN (True Negatif) = Diklasifikasikan sebagai kelas negatif
FP (False Positif) = Dokumen negatif yang diklasifikasikan sebagai kelas positif
FN (False Negatif) **=** Diklasifikasikan sebagai kelas negatif

Pada table yang diberikan, nilai-nilai tersebut akan digunakan untuk menghitung *accuracy, precision, recall dan f1 score*. *Precision* merupakan hasil perbandingan antara jumlah prediksi benar pada data positif dengan total prediksi pada data positif. Sedangkan *recall* adalah hasil perbandingan antar jumlah prediksi benar pada data positif dengan total data positif yang diprediksi benar. *F1 score* adalah hasil rata-rata nilai *precision* dan *recall* sebagai *harmonic mean*. Berikut merupakan persamaan dari *accuracy, precision, recall,* dan *f1 score* :

$$\textbf{Accuracy} = \frac{(\boldsymbol{TP+TN})}{\boldsymbol{TP+TN+FP+FN}} \quad (3)$$

$$\textbf{Precision (P)} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP+FP})} \quad (4)$$

$$\textbf{Recall (R)} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP+FN})} \quad (5)$$

$$\textbf{F1 score} = \frac{2\ x\ P\ x\ R}{P+R} \quad (6)$$

2.10 Kesimpulan dan Saran

Kesimpulan dan saran merupakan bagian terakhir dari penelitian yang ditulis oleh peneliti, dengan isi penelitian yang telah dijelaskan pada tahapan sebelumnya. Pada bagian akhir, peneliti memaparkan secara singkat hasil penelitian yang dilakukan.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pada penelitian ini dilakukan proses pengumpulan data untuk menganalisis data sentimen Calon Presiden 2024 Prabowo Subianto. Proses pendataan terbatas pada topik dan opini publik tentang Capres Prabowo Subianto hanya di media sosial Twitter dengan menggunakan metode *crawling* Twitter, dengan memanfaatkan fungsi *Application Interface* (API) yang disediakan oleh Twitter. Data *tweet* yang diekstraksi dari Twitter berisi sebanyak 2100 data yang terdiri dari 1050 *tweet* positif dan 1050 *tweet* negatif. Gambar 2 menampilkan *load* data pada Google Colab.

	Text	Kelas
0	b'Tahun 2014 dan 2019 semua HARAM mengenakan s...	POSITIF
1	b'@kadrunmampos Tahun 2014 dan 2019 semua HARA...	POSITIF
2	b'Prabowo Subianto menanggapi santai sinyal du...	POSITIF
3	b'@geloraco Baik ganjar ato Prabowo yg jadi pr...	POSITIF
4	b'Median melakukan rilis survei elektabilitas ...	POSITIF
...	...	...
2095	b'@Miduk17 Saya gak pilih calon sebelah tapi b...	NEGATIF
2096	b'Lembaga Survei Nasional (LSN) merilis survei...	NEGATIF
2097	b'Prabowo ataupun Puan yang menjadi Presiden I...	NEGATIF
2098	b'@musniumar Gak bakalan jadi presiden mau pro...	NEGATIF
2099	b'Prabowo Subianto itu juga pernah maju sebaga...	NEGATIF

2100 rows × 2 columns

Gambar 3. Tampilan *Load* Data pada Google Colab

Gambar diatas merupakan hasil dari *load* data yang telah diimplemetasikan pada Google Colab. Proses pendataan *tweet* hanya mengambil data dalam format teks bahasa Indonesia dan tidak menyertakan gambar.

3.2 Pelabelan Data

Setelah data dikumpulkan, kemudian dilakukan pelabelan data yang dibagi menjadi dua kelas, yaitu positif dan negatif. Pelabelan dilakukan secara manual dengan dibantu tokoh ahli yaitu dosen Bahasa Indonesia Roza Afifah S.Pd selaku dosen Universitas Islam Negeri Sultan Syarif Kasim Riau. Tabel 2 menampilkan perbedaan pemisahan *tweet* positif dan *tweet* negatif.

Tabel 2. *Tweet* Positif dan *Tweet* Negatif

Tweet Positif	*Tweet* Negatif
b'@prabowo Bp CALON PRESIDEN untuk menggantikan pak Jokowi 2024 insyaallah'	b'@musniumar Gak bakalan jadi presiden mau promo kemanapun.\nAnis di siapkan jadi tumbal calon lain seperti prabowo dulu'
b'@prabowo Setelah pa jokowi selesai masa jabatannya \nAku beralih akan memilih pa prabowo sebagai calon presiden 2024'	b'@Miduk17 Saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil, prabowo presiden. \xf0\x9f\x98\x81'
b'@beritadalamedia Saya berharap pak prabowo mau jadi calon presiden lagi di tahun 2024, supaya bisa jadi mentri lagi'	b'@OposisiCerdas Fix ga pilih prabowo. Calon Presiden ku adalah yg berani menolak dan membatalkan proyek IKN'
b'Seperti pak Prabowo bismillah calon presiden 2024 \xf0\x9f\x98\x8c https://t.co/S5rePOOlhm'	b'@idextratime Liverpool ibarat Jokowi vs Prabowo, walaupun kampenye sana sini, tetep gabisa menang. '

b'@Gerindra @prabowo CALON PRESIDEN 2024 INSYAALLAH'	b'@BBCIndonesia @pantaugambut Punya malu gak @prabowo kayak gitu maksa banget pengen jadi presiden'

Berdasarkan tabel diatas, dapat disimpulkan bahwa data yang sudah di *crawling* sebelumnya, dikelompokkan menjadi dua kelompok, yaitu kelompok positif dan kelompok negatif. Pengelompokkan data tersebut dilakukan secara seimbang, yang mana 2100 data yang diambil dibagi menjadi dua, dengan pengertian 2100 data yang dimbil diantaranya 1050 data positif dan 1050 data negatif.

3.3 *Text Preprocessing*

Tahapan yang dilakukan dalam *preprocessing* yaitu:

a. *Cleaning*

Data *cleaning* atau pembersihan data yaitu proses mempersiapkan data untuk analisis menggunakan cara menghapus atau memodifikasi data yang tidak benar, tidak lengkap, tidak relevan, atau diformat dengan tidak tepat [17].

Tabel 3. Proses *Cleaning*

Data Asli	Data setelah *Cleaning*
b’@Miduk17 Saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil, Prabowo presiden. \xf0\x9f\x98\x81’	Saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil Prabowo presiden

Tabel diatas merupakan proses dari *cleaning* data, yang mana dalam prosesnya data yang masih mengandung atribut yang tidak diperlukan akan dibersihkan seperti *username*, *hashtag*, *url*, tanda baca, emoji dan *emoticon*.

b. *Case Folding*

Proses *case folding* adalah sebuah proses standarisasi data yang bertujuan untuk membuat karakter pada data menjadi seragam. Tujuan dari proses *case folding* adalah untuk mengubah semua huruf menjadi huruf kecil [18].

Tabel 4. Proses *Case Folding*

Data Asli	Data setelah *Case Folding*
Saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil prabowo presiden	saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil prabowo presiden

Tabel diatas merupakan proses dari *case folding*, yang mana dalam prosesnya semua kata diubah menjadi huruf kecil atau dikenal dengan *lowercase*.

c. *Tokenizing*

Tokenizing adalah proses membagi teks dimana teks yang panjang dipecah menjadi fragmen kecil. Bagian-bagian yang lebih kecil ini umumnya disebut sebagai token. Pengolahan akan dilanjutkan setelah kalimat-kalimat tersebut dipecah menjadi token. Proses *tokenizing* juga dapat disebut sebagai segmentasi teks atau analisis leksikal. Dengan kata lain, proses *tokenizing* adalah proses memecahan kalimat menjadi kata-kata penyusunnya (masing-masing) [18].

Tabel 5. Proses *Tokenizing*

Data Asli	Data setelah *Tokenizing*
saya gak pilih calon sebelah tapi blm setuju juga dgn white hair cocok jadi wakil prabowo presiden	[‘saya’, ‘gak’, ‘pilih’, ‘calon’, ‘sebelah’, ‘tapi’, ‘blm’, ‘setuju’, ‘juga’, ‘dgn’, ‘white’, ‘hair’, ‘cocok’, ‘jadi’, ‘wakil’, ‘prabowo’, ‘presiden’]

Tabel diatas merupakan proses dari *tokenizing*, yang mana dalam prosesnya teks dipisahkan menjadi potongan-potongan kata atau disebut dengan *term*.

d. Normalisasi

Normalisasi data merupakan unsur pokok dalam data *mining* yang bertujuan memastikan konsistensi *record* pada dataset tetap terjaga. Dalam proses normalisasi diperlukan perubahan data atau konversi data awal ke dalam format yang memugkinkan untuk diproses secara efisien. Tujuan utama dari normalisasi data yakni menghilangkan redundansi data (pengulangan) dan menstandarisasi informasi untuk alur kerja data yang lebih baik [18].

Tabel 6. Proses Normalisasi

Data Asli	Data setelah Normalisasi
saya	saya
gak	**tidak**
pilih	pilih
calon	calon
sebelah	sebelah

tapi	tapi
blm	**belum**
setuju	setuju
juga	juga
dgn	**dengan**
white	white
hair	hair
cocok	cocok
jadi	jadi
wakil	wakil
prabowo	prabowo
presiden	presiden

Tabel diatas merupakan proses dari normalisasi, yang mana dalam prosesnya setiap kata atau *term* di cek apakah kata tersebut sudah dalam bentuk normalnya atau belum. Apabila kata belum dalam bentuk normalnya makan akan dilakukan normalisasi.

e. *Negation Handling*

Negation handling merupakan sebuah proses dimana menghapus kata negatif atau kata yang mengandung makna negatif, lalu kemudian mengubah kata tersebut menjadi kata yang baku [19].

Tabel 7. Proses *Negation Handling*

Data Asli	Data setelah *Negation Handling*
[‘saya’, ‘**tidak’, ‘pilih’, ‘**calon’, ‘sebelah’, ‘tapi’, ‘**belum’, ‘setuju’, ‘**juga’, ‘dengan’, ‘white’, ‘hair’, ‘cocok’, ‘jadi’, ‘wakil’, ‘prabowo’, ‘presiden’]	[‘saya’, ‘**membantah’, ‘**calon’, ‘sebelah’, ‘tapi’, ‘**menolak’, ‘**juga’, ‘dengan’, ‘white’, ‘hair’, ‘cocok’, ‘jadi’, ‘wakil’, ‘prabowo’, ‘presiden’]

Tabel diatas merupakan proses dari *negation handling*, yang mana dalam prosesnya dilakukan pencarian kata yang mengandung negasi dan diubah menjadi makna yang serupa tanpa mengubah sentimennya.

f. *Stopword Removal*

Stopword removal merupakan proses dimana kata-kata penghubung dihapuskan dalam sebuah kalimat, seperti kata “yang”, “dan”, “dengan” dan lainnya [20].

Tabel 8. Proses *Stopword Removal*

Data Asli	Data setelah *Stopword Removal*
[‘saya’, ‘membantah’, ‘calon’, ‘sebelah’, **‘tapi’**, ‘menolak’, **‘juga’**, **‘dengan’**, ‘white’, ‘hair’, ‘cocok’, **‘jadi’**, ‘wakil’, ‘prabowo’, ‘presiden’]	[‘saya’, ‘membantah’, ‘calon’, ‘sebelah’, ‘menolak’, ‘white’, ‘hair’, ‘cocok’, ‘wakil’, ‘prabowo’, ‘presiden’]

Tabel diatas merupakan proses dari *stopword removal*, yang mana dalam prosesnya dilakukan pencarian untuk kata yang mengandung kata hubung sehingga kata tersebut di hapus dari kumpulan *term* tersebut.

g. *Stemming*

Stemming merupakan sebuah proses menghapus kata imbuhan atau mengubah kata menjadi bentuk kata dasarnya sesuai dengan Kamus Besar Bahasa Indonesia [20].

Tabel 9. Proses *Stemming*

Data Asli	Data setelah *Stemming*
[‘saya’, **‘membantah’**, ‘calon’, ‘sebelah’, ‘**menolak**’, ‘white’, ‘hair’, ‘cocok’, ‘wakil’, ‘prabowo’, ‘presiden’]	[‘saya’, **‘bantah’**, ‘calon’, ‘sebelah’, **‘tolak’**, ‘white’, ‘hair’, ‘cocok’, ‘wakil’, ‘prabowo’, ‘presiden’]

Tabel diatas merupakan proses dari *stemming*, yang mana dalam prosesnya semua kata yang bentuknya bukan kata dasar akan diubah menjadi bentuk kata dasar.

3.4 Pembobotan TF-IDF

Kata-kata yang sudah melalui proses perataan kata kemudian dinilai nilai bobotnya dengan menggunakan TF-IDF. Bobot bertujuan untuk memberikan nilai terhadap kejadian berulang kata yang sama. *Term Frequency* (TF) adalah gagasan pengukuran dengan menentukan seberapa sering (frekuensi) sebuah istilah muncul dalam satu dokumen. Tiap dokumen memiliki ukuran yang berbeda, mungkin saja suatu kata muncul lebih sering di dokumen yang lebih besar dibandingkan dengan dokumen yang lebih kecil. Oleh karena itu, frekuensi kata sering kali dibagi dengan jumlah kata dalam dokumen (yaitu total kata yang terdapat dalam dokumen tersebut). Sementara itu, *Document Frequency* (DF) merujuk pada jumlah dokumen dimana suatu istilah muncul. Semakin jarang kemunculannya, maka semakin rendah pula nilai bobotnya. Saat melakukan perhitungan frekuensi kata, setiap kata dianggap memiliki tingkat penting yang sama. Setelah melalui proses

text preprocessing, langkah selanjutnya adalah melakukan seleksi fitur dan klasifikasi dengan menggunakan metode *K-Nearest Neighbor*. Gambar 3 menampilkan hasil pembobotan TF-IDF pada Google Colab.

	00	000t	002	02	07	08	10	100	110	115	...	yusril	ywh	zalim	zhone	zhonk	zon	zonauang	zonky	zulkifli	Kelas
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF

5 rows × 3533 columns

Gambar 4. TF-IDF pada Google Colab

Gambar diatas merupakan hasil tampilan proses pembobotan menggunakan TF-IDF yang telah diimplementasikan pada Google Colab. Dapat dilihat pada gambar hasil bobot untuk setiap kata atau *term*, yang mana hasil pembobotan akan digunakan untuk melakukan klasifikasi.

3.5 Seleksi Fitur *Threshold*

Seleksi fitur, yang juga dikenal sebagai seleksi variabel, seleksi atribut, atau pemilihan subset fitur, adalah proses memilih fitur yang relevan dalam data *learning* untuk suatu masalah target. Uji seleksi fitur bertujuan untuk mengevaluasi pengaruh jumlah *term* yang dipilih pada proses klasifikasi. Pada seleksi fitur *threshold* yang digunakan yaitu 0,001. Gambar 4 menampilkan hasil seleksi fitur dengan menggunakan *threshold* pada Google Colab :

	08	20102	2024	2029	agama	ahmad	akrab	aku	alonely	ancam	...	unggul	ungkap	usung	wacana	wakil	warga	widodo	yea	yess	Kelas
0	0.0	0.0	0.030228	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF
1	0.0	0.0	0.030228	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF
2	0.0	0.0	0.089887	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.260091	0.0	0.0	POSITIF
3	0.0	0.0	0.062738	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	POSITIF
4	0.0	0.0	0.089277	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.192855	0.0	0.000000	0.0	0.0	POSITIF
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2095	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.207126	0.0	0.000000	0.0	0.0	NEGATIF
2096	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF
2097	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF
2098	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	NEGATIF
2099	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.163733	0.0	0.000000	0.0	0.0	NEGATIF

2100 rows × 175 columns

Gambar 5. Seleksi Fitur pada Google Colab

Gambar diatas merupakan hasil tampilan proses seleksi fitur menggunakan *threshold* yang telah diimplementasikan pada Google Colab. Dapat dilihat pada gambar hasil seleksi fitur yang dipilih berdasarkan *threshold* yang digunakan yaitu 0,001.

3.6 Klasifikasi *K-Nearest Neighbor*

Uji coba pengaruh nilai K dilaksanakan guna menentukan nilai K terbaik dalam menjalankan proses klasifikasi *K-Nearest Neighbor*. Pada pengujian ini menggunakan nilai K yang bervariatif yaitu K = 3,5,7,9,11,13, dan 15. Tiap nilai K yang dipakai, akan jadi variabel dalam pengujian, sehingga bisa diketahui pengaruhnya terhadap ketepatan. Hasil klasifikasi positif dan negatif pada *tweet* dikatakan "Benar" jika sistem berhasil mengenali dan mengelompokkan komentar ke dalam kelas yang sesuai dengan kelas data latih. Gambar 5 menampilkan hasil klasifikasi K-NN.

Hasil Klasifikasi

	08	20102	2024	2029	agama	ahmad	akrab	aku	alonely	ancam	...	ungkap	usung	wacana	wakil	warga	widodo	yea	yess	Kelas Target	Hasil Klasifikasi
0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF
1	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.529496	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF
2	0.000000	0.000000	0.068377	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF	POSITIF
3	0.000000	0.000000	0.123453	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF	NEGATIF
4	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
415	0.000000	0.000000	0.086769	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	POSITIF	POSITIF
416	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF
417	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	POSITIF
418	0.602420	0.261853	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF
419	0.462578	0.201068	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NEGATIF	NEGATIF

420 rows × 176 columns

Gambar 6. Klasifikasi K-NN pada Google Colab

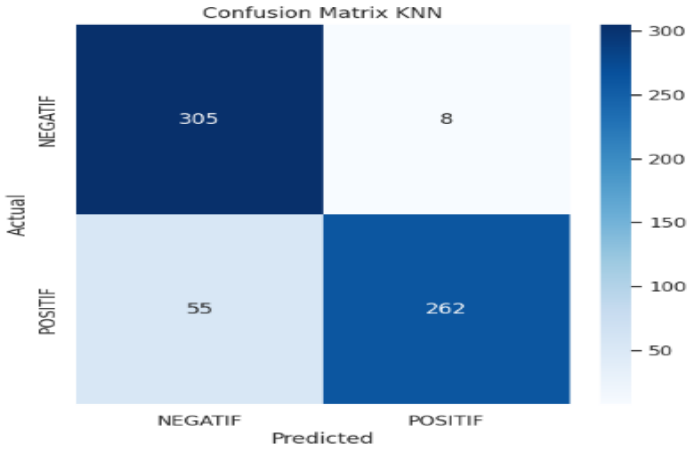
Gambar diatas merupakan hasil tampilan proses klasifikasi dengan menggunakan metode K-NN yang telah diimplementasikan pada Google Colab. Dapat dilihat pada gambar yang mana terdapat kolom Kelas Target yaitu kelas yang dilabel sebelum dilakukan klasifikasi dan kolom Hasil Klasifikasi yaitu lebel yang didapat setelah dilakukan proses klasifikasi.

3.7 Pengujian *Confusion Matrix*

Setelah melakukan proses pembobotan kata, seleksi fitur, dan pengklasifikasian dengan metode K-NN, maka dilakukan pengujian confusion matrix dengan tujuan mencari nilai *accuracy, precision, recall,* dan *f1 score*. *Confusion matrix* berguna untuk menentukan pengujian mana yang optimal untuk membangun model klasifikasi menggunakan metode *K-Nearest Neighbor* (K-NN). Pada penelitian pengujian menggunakan *confusion matrix* untuk rasio yang digunakan yaitu 70:30, 80:20, dan 90:10.

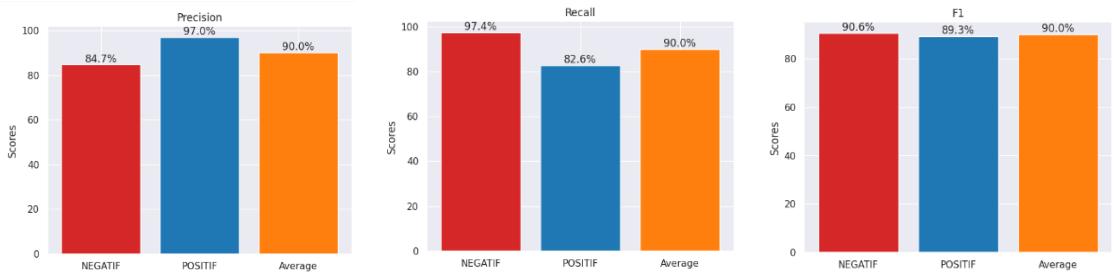
a. Perbandingan 70:30

Pada pengujian yang dilakukan dengan perbandingan data uji dan data latih yaitu 70:30 dengan K yang digunakan yaitu K = 3,5,7,9,11,13, dan 15 didapatkan akurasi tertinggi pada K=5 yaitu 90,0%. Gambar 6 menampilkan pengujian *confusion matrix* yang telah diimplementasikan pada Google Colab.



Gambar 7. *Confusion Matrix* 70:30

Setelah mendapatkan akurasi, dilanjutkan dengan melakukan pengujian untuk *precision, recall* dan *f1 score*. Gambar 7 menampilkan hasil pengujian *precision, recall* dan *f1 score* pada nilai K=5 dengan rasio data latih dan data uji yaitu 70:30.

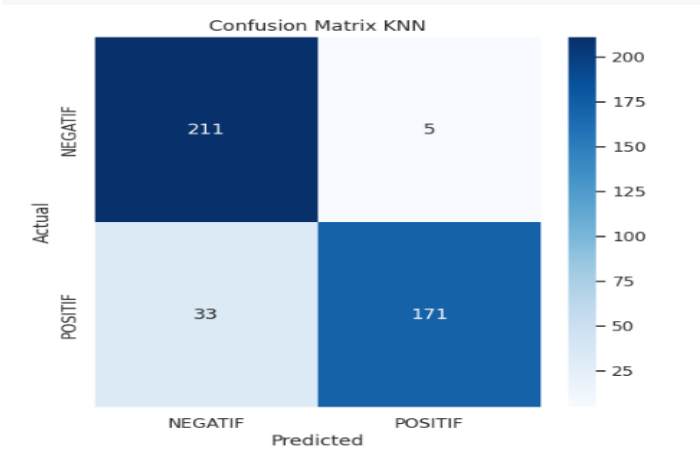


Gambar 8. Hasil *Precision*, *Recall*, dan *F1 score* 70:30

Dari Gambar didapat hasil *precision* untuk *tweet* negatif yaitu 84,7%, sedangkan untuk *tweet* positif yaitu 97,0%, sehingga dihitung rata-rata *precision* yaitu 90,0%. *Recall* yang didapatkan untuk *tweet* negatif yaitu 97,4%, sedangkan untuk *tweet* positif yaitu 82,6%, sehingga dihitung rata-rata *recall* yaitu 90,0%. *F1 score* yang didapatkan untuk *tweet* negatif 90,6%, sedangkan untuk *tweet* positif 89,3%, sehingga dihitung rata-rata *f1 score* yaitu 90,0%.

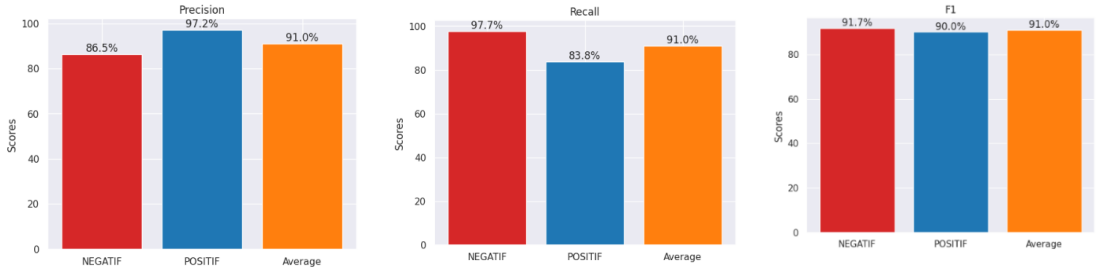
b. Perbandingan 80:20

Pada pengujian yang dilakukan dengan perbandingan data uji dan data latih yaitu 80:20 dengan K yang digunakan yaitu K = 3,5,7,9,11,13, dan 15 didapatkan akurasi tertinggi pada K=5 yaitu 91,0%. Gambar 8 menampilkan pengujian *confusion matrix* yang telah diimplementasikan pada Google Colab.



Gambar 9. *Confusion Matrix* 80:20

Setelah mendapatkan akurasi, dilanjutkan dengan melakukan pengujian untuk *precision, recall* dan *f1 score*. Gambar 9 menampilkan hasil pengujian *precision, recall* dan *f1 score* pada nilai K=5 dengan rasio data latih dan data uji yaitu 80:20.

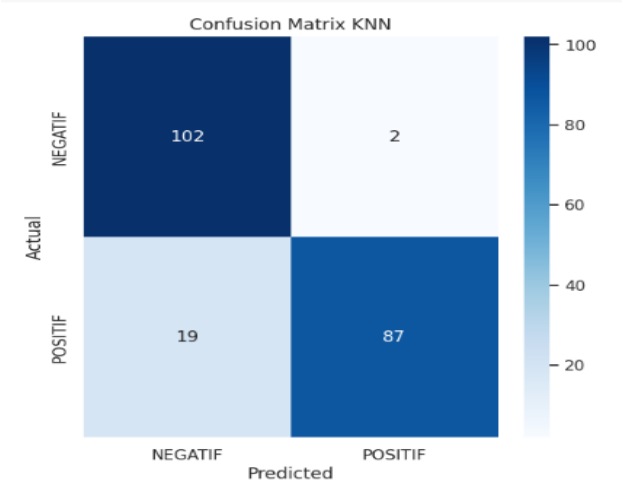


Gambar 10. Hasil *Precision*, *Recall*, dan *F1 score* 80:20

Dari Gambar didapat hasil *precision* untuk *tweet* negatif yaitu 86,5%, sedangkan untuk *tweet* positif yaitu 97,2%, sehingga dihitung rata-rata *precision* yaitu 91,0%. *Recall* yang didapatkan untuk *tweet* negatif yaitu 97,7%, sedangkan untuk *tweet* positif yaitu 83,8%, sehingga dihitung rata-rata *recall* yaitu 91,0%. *F1 score* yang didapatkan untuk *tweet* negatif 91,7%, sedangkan untuk *tweet* positif 90,0%, sehingga dihitung rata-rata *f1 score* yaitu 91,0%.

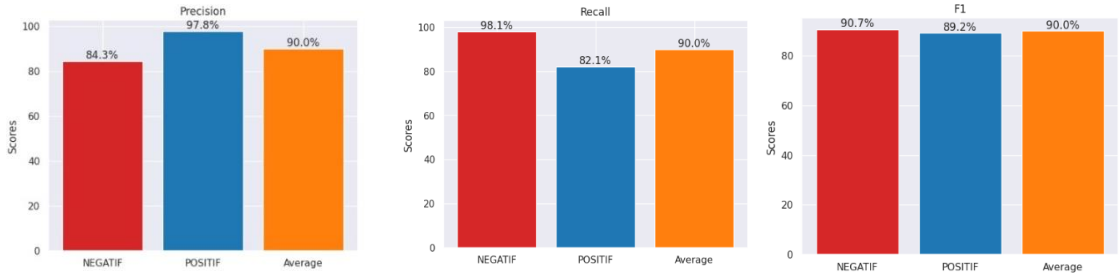
c. Perbandingan 90:10

Pada pengujian yang dilakukan dengan perbandingan data uji dan data latih yaitu 90:10 dengan K yang digunakan yaitu K = 3,5,7,9,11,13, dan 15 didapatkan akurasi tertinggi pada K=5 yaitu 90,2%. Gambar 10 menampilkan pengujian *confusion matrix* yang telah diimplementasikan pada Google Colab.



Gambar 11. *Confusion Matrix* 90:10

Setelah mendapatkan akurasi, dilanjutkan dengan melakukan pengujian untuk *precision, recall* dan *f1 score*. Gambar 11 menampilkan hasil pengujian *precision, recall* dan *f1 score* pada nilai K=5 dengan rasio data latih dan data uji yaitu 90:10.

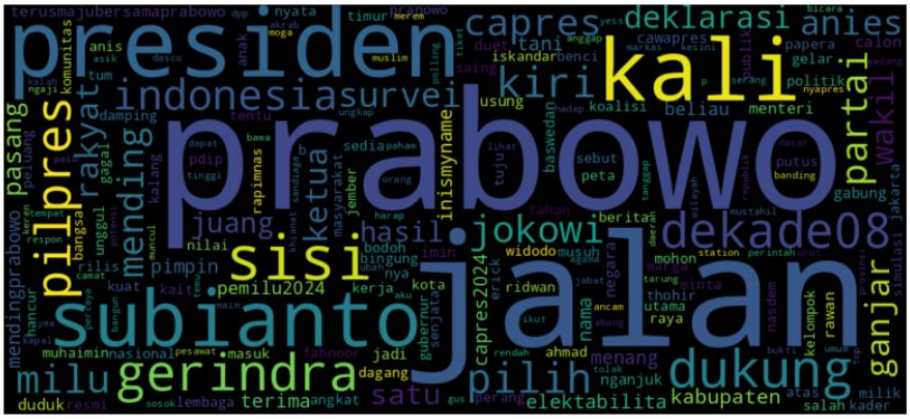


Gambar 12. Hasil *Precision*, *Recall*, dan *F1 score* 90:10

Dari Gambar didapat hasil *precision* untuk *tweet* negatif yaitu 84,3%, sedangkan untuk *tweet* positif yaitu 97,8%, sehingga dihitung rata-rata *precision* yaitu 90,0%. *Recall* yang didapatkan untuk *tweet* negatif yaitu 98,1%, sedangkan untuk *tweet* positif yaitu 82,1%, sehingga dihitung rata-rata *recall* yaitu 90,0%. *F1 score* yang didapatkan untuk *tweet* negatif 90,7%, sedangkan untuk *tweet* positif 89,2%, sehingga dihitung rata-rata *f1 score* yaitu 90,0%.

d. *Word Cloud*

Setelah melalui seluruh proses, data kemudian diproses untuk menampilkan *word cloud* yang akan menghasilkan visualisasi kata-kata yang paling banyak muncul, sehingga dapat diketahui seberapa sering kata-kata tersebut digunakan oleh masyarakat dalam mengungkapkan pandangan mereka tentang Calon Presiden 2024, Prabowo Subianto, di media sosial Twitter. Semakin besar ukuran kata tersebut, semakin sering pula kata tersebut muncul dalam kata. Gambar 12 menampilkan hasil *word cloud* yang telah diimplementasi pada Google Colab.



Gambar 13. Hasil *Word Cloud*

Gambar diatas merupakan hasil tampilan word cloud untuk keseluruhan data yang telah diimplementasikan pada Google Colab. Dapat dilihat pada gambar hasil word cloud untuk keseluruhan data didapatkan kata yang paling banyak muncul yaitu kata Prabowo, jalan, presiden, kali, dan subianto. Pengujian *word cloud* selanjutnya dilakukan juga untuk data *tweet* positif dan data *tweet* negatif. Gambar 13 menampilkan *word cloud tweet* positif dan *tweet* negatif.

Gambar 14. (a) Hasil *Word Cloud* Positif (b) Hasil *Word Cloud* Negatif

Pada Gambar 13 (a) merupakan *word cloud* dari keseluruhan tanggapan positif masyarakat Indonesia mengenai Calon Presiden Prabowo Subianto yang diperoleh pada bulan November 2022 sampai dengan Januari 2023 melalui media sosial Twitter. Pada hasil *word cloud* positif didapatkan kesimpulan kata yang paling banyak muncul yaitu jalan, parabowo, presiden, kali, dan dukung. Gambar (b) merupakan *word cloud* dari tanggapan negatif masyarakat Indonesia mengenai Calon Presiden Prabowo Subianto yang diperoleh pada bulan November 2022 sampai dengan Januari 2023 melalui media sosial Twitter. Pada hasil *word cloud* negatif didapatkan kesimpulan kata yang paling banyak muncul yaitu prabowo, kali, jalan, sisi, dan subianto.

4. KESIMPULAN

Berdasarkan hasil pembahasan dari penelitian ini, dapat disimpulkan bahwa metode *K-Nearest Neighbor* (K-NN) bisa dimanfaatkan untuk mengklasifikasikan sentimen dengan topik Calon Presiden 2024, Prabowo Subianto. Hal ini dilakukan melalui beberapa tahapan, seperti pengumpulan data, *text preprocessing*, pembobotan dengan TF-IDF, pemilihan fitur dengan *threshold*, pengklasifikasian dengan metode K-NN, serta pengujian dengan *confusion matrix*. Dengan dilakukan pengujian menggunakan *confusion matrix* dengan nilai K yaitu K=3,5,7,9,11,13, dan 15, serta perbandingan yang digunakan yaitu 70:30, 80:20, dan 90:10, diperoleh pengujian yang paling terbaik dengan nilai akurasi tertinggi dari proses klasifikasi yaitu 91,0% pada K=5 dengan *precision* 91,0% , *recall* 91,0%, dan *f1 score* 91,0% pada rasio data latih dan data uji 80:20. Dari kesimpulan yang telah diambil, pembaca dapat memperluas penelitian selanjutnya yang dapat dijadikan pertimbangan, seperti melaksanakan penelitian lebih mendalam untuk analisis sentimen dengan menggunakan teknik klasifikasi guna meningkatkan akurasi hasil, terutama dalam tahap *preprocessing*. Kumpulan data yang dipakai bisa ditingkatkan dengan menggunakan kumpulan data lain seperti kumpulan data dalam bentuk tulisan yang bukan menggunakan huruf latin. Dibutuhkan uji coba dengan memakai metode seleksi fitur yang berbeda selain *threshold* untuk memastikan apakah metode seleksi fitur lainnya memiliki akurasi yang lebih baik atau tidak. Selain itu, bisa dilakukan implementasi menggunakan *tools* lainnya.

REFERENCES

[1] W. A. Wibawana, "Sejarah Pemilu di Indonesia dari Awal Sampai Sekarang," *detiknews*, 2023. https://news.detik.com/pemilu/d-6526532/sejarah-pemilu-di-indonesia-dari-awal-sampai-sekarang#:~:text=Pilpres 2024 adalah pemilihan umum yang akan menjadi,secara serentak pada tanggal 14 Februari 2024 mendatang. (accessed Jun. 09, 2023).

[2] F. C. Farisa, "Sejarah Dimulainya Pemilu Presiden dan Wakil Presiden Secara Langsung," *KOMPAS.com*, 2022. https://nasional.kompas.com/read/2022/05/31/12112831/sejarah-dimulainya-pemilu-presiden-dan-wakil-presiden-secara-langsung (accessed Jun. 09, 2023).

[3] F. C. Farisa, "Pilpres 2024 Diprediksi Diikuti 3 Capres: Ganjar, Prabowo, dan Anies," *KOMPAS.com*, 2023.

[4] F. C. Farisa, "Survei Indikator: Elektabilitas Prabowo Bersaing Ketat dengan Ganjar, Anies Urutan Ketiga," *KOMPAS.com*, 2023.

[5] M. Aulidya, "Manfaat, Fitur, dan Fungsi Twitter: Mengenal Lebih Dekat Media Sosial yang Populer Saat ini," *kompasiana*, 2023. https://www.kompasiana.com/mishelaulidya7261/643bb5c9a7e0fa33b71c51e3/manfaat-fitur-dan-fungsi-twitter-mengenal-lebih-dekat-media-sosial-yang-populer-saat-ini (accessed Jun. 09, 2023).

[6] R. T. Prasetio, "Seleksi Fitur dan Optimasi Parameter K-NN Berbasis Algoritma Genetika pada Dataset Medis," *J. Responsif Ris. Sains dan Inform.*, vol. 2, no. 2, pp. 213–221, 2020, doi: 10.51977/jti.v2i2.319.

[7] F. Rizqi Irawan, "Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearset Neighbors," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 1, pp. 62–68, 2022, doi: 10.33387/jiko.v5i1.4267.

[8] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.

[9] K. A. Nugraha, "Pembentukan Dataset Token Sentimen Berdasarkan Akun Instagram," *J. Buana Inform. Vol.*, vol. 12, no. April, pp. 68–77, 2021.

[10] E. H. Muktafin and P. Kusrini, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 1, pp. 146–154, 2021, doi: 10.12928/TELKOMNIKA.V19I1.17417.

[11] R. Damarta, A. Hidayat, and A. S. Abdullah, "The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012002.

[12] N. G. Yudiarta, M. Sudarma, and W. G. Ariastina, "Pengelompokan Berita Pada Unstructured Textual Data," vol. 17, no. 3, pp. 339–344, 2018.

[13] F. Rozi, F. Sukmana, and M. N. Adani, "Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF)," vol. 6, no. 3, pp. 1–5, 2022.

[14] V. Bolón-canedo and A. Alonso-betanzos, "Ensembles for feature selection : A review and future trends," *ELSEVIER*, vol. 52, no. May 2018, pp. 1–12, 2019, doi: 10.1016/j.inffus.2018.11.008.

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

[16] A. Yoga Pratama, Y. Umaidah, and Voutama, "Analisis Sentimen Media Sosial Twitter dengan Algoritma K-Nearest Neighbor dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 897–910, 2021.

[17] T. D. Arista, M. Fikry, and L. Oktavia, "Klasifikasi Sentimen Masyarakat di Twitter terhadap Kenaikan Harga BBM dengan Metode K-NN," *JUKI*, vol. 5, pp. 140–150, 2023.

[18] A. F. Rahman, "Klasifikasi Tweet di Twitter dengan Menggunakan Metode K-Nearest Neighbor," *J. Sist. Inf. dan Teknol.*, vol. 4, pp. 64–69, 2022, doi: 10.37034/jsisfotek.v4i2.125.

[19] M. T. Diwandanu, "Analisis Sentimen terhadap Twit Maxim pada Twitter Menggunakan R Programming dan K-Nearest Neighbors," *J. Ilm. Inform. Komput.*, vol. 28, pp. 1–16, 2023.

[20] R. Kosasih and A. Alberto, "Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor," *InfoTekJar J. Nas. Inform. dan Teknol. Jar.*, vol. 6, no. 1, pp. 134–139, 2021.