Pengembangan Sistem Analisis Sentimen Masyarakat Terhadap Chatgpt Pada Twitter Dengan Perbandingan Metode Naive Bayes Classifier Dan K-Nearest Neighbors

Maharani Nirmala Dewi[1], Ricky Eka Putra[2]

[1,2] Teknik Informatika,Fakultas Teknik, Universitas Negeri Surabaya
[1]maharani.19032@mhs.unesa.ac.id
[2]rickyeka@unesa.ac.id

***Abstrak*— Chatgpt merupakan chatbot yang memiliki model kecerdasan buatan dan dirancang untuk menghasilkan text bahasa manusia. Chatgpt menggemparkan dunia sosial media dikarenakan dapat membantu tugas copywriter dengan banyak manfaat. Chatgpt juga dikhawatirkan membawa pengaruh negatif seperti alat pelanggaran akademik dalam ujian online. Penelitian menggunakan media Twitter untuk memperoleh dataset, *Crawling data* dari media Twitter menghasilkan 1229 data yang terdiri 629 sentimen positif 300 sentimen negatif dan 300 sentimen netral. Teknik pengujian menggunakan split data dan K-Fold Cross Validation. Hasil uji coba menunjukkan bahwa metode Naive Bayes lebih baik daripada K-Nearest Neighbors pada kedua pengujian. Pengujian dengan hasil akurasi tertinggi sebesar 82,25%, nilai presisi 81,91%, recall 82,25%, dan f1-score 81,37% untuk metode Naive Bayes dengan pengujian split data. Di sisi lain metode K-Nearest Neighbors dengan split data memiliki hasil akurasi 80,43%, nilai presisi 80,97%, recall 80,43% dan f1-score 80,33%. Metode Naive Bayes akurasinya lebih tinggi 1.82% pada pengujian split data dan lebih tinggi 4.57% untuk pengujian k-fold cross validation**

***Kata Kunci*— Chatgpt, Analisis sentimen, Naive Bayes, K-Nearest Neighbors**

I. PENDAHULUAN

Perusahaan Open AI mengembangkan Chatgpt yaitu chatbot yang memiliki model kecerdasan buatan dan dirancang untuk menghasilkan text bahasa manusia. Chatbot tersebut telah dilatih dengan data yang besar sehingga mampu memahami, menafsirkan dan menghasilkan bahasa manusia.

Chatgpt menggemparkan dunia sosial media dikarenakan dapat membantu copywriter membuat text seperti puisi, tutorial, hingga deskripsi singkat media sosial. Chatgpt mampu untuk memeriksa sebuah kode di dalam skrip dan membuat sebuah program sederhana.

Dengan banyak manfaat yang bisa dilakukan oleh Chatgpt juga dikhawatirkan membawa pengaruh negatif seperti alat pelanggaran akademik dalam ujian online. Menurut Teo Susnjak (2022) Chatgpt diyakini bisa memberikan jawaban yang menarik dan akurat untuk pertanyaan sulit yang membutuhkan analisis tingkat lanjut [1]

Dengan pengaruh Chatgpt peneliti berharap mengetahui opini masyarakat indonesia melalui analisis sentimen. Salah satu cara untuk melakukan analisis sentimen dengan menggunakan metode Naive Bayes Classifier dan K-Nearest Neighbors. Metode ini dapat digunakan untuk mengklasifikasi sentimen menjadi positif, negatif, ataupun netral berdasarkan makna dari data teks yang telah dikumpulkan. Metode tersebut dipilih karena memiliki kemampuan untuk mengklasifikasikan sentimen secara akurat dan efisien. Metode ini telah terbukti berhasil dalam berbagai penelitian analisis sentimen dan dapat mengolah data teks dalam jumlah besar dengan cepat.

Analisis sentimen masyarakat terhadap Chatgpt pada Twitter menggunakan metode Naive Bayes Classifier dan K-Nearest Neighbors dilakukan dengan cara pengolahan kata untuk melacak opini tentang Chatgpt dari tweet pengguna Twitter. pengumpulan data dilakukan dengan kata kunci Chatgpt. Selama penelitian penulis mengambil atribut username, dan isi teks.

Wongkar dan Angdresey [2], menganalisis sentimen masyarakat terhadap calon presiden Republik Indonesia 2019 di Twitter dalam penelitian *Naive Bayes* memiliki akurasi lebih tinggi daripada SVM dan KNN yaitu 80.90%, 75,58% dan 63.99%.

Prananda dan Thalib [3], dalam penelitian melakukan analisis sentimen dari aplikasi GO-JEK dan menggunakan *library Twint* berhasil mengumpulkan data 3111 tweet. Algoritma *neural network* mencapai presisi sebesar 0,52, recall sebesar 0,51 dan f1- skor 0,51. Sedangkan pada mesin *support vector* memperoleh presisi sebesar 0,54, recall sebesar 0,53 dan f1-score sebesar 0,53. *Naïve Bayes* memperoleh presisi sebesar 0,52, recall sebesar 0,52, dan f1-score sebesar 0,52. Terakhir, Decision Tree memperoleh presisi 0,55, recall 0,55, dan skor f1 0,55.

Hasri dan Alita [4], menganalisis sentimen dari dampak yang ditimbulkan akibat corona di Twitter dalam penelitian Naive Bayes memiliki akurasi lebih tinggi daripada SVM yaitu 81.07% dan 79.96%.

Penelitian oleh Prasetya, Wirawan Dwi, dan Bambang Sujatmiko[5] menggunakan proses klasifikasi dilakukan dengan cara memasukkan data ke dalam *tools Jupyter Notebook* Dan membuat rancangan proses penelitian. Dataset yang diambil oleh di Klinik Bidan Saptarum Kabupaten Jombang dengan jumlah 50 data akan diolah dengan Algoritma KNN dan *Naive Bayes.* Tahap akhir m enja dika n file dalam bentuk *Data Pickle* Agar dapat direalisasikan ke dalam sistem.Adapun hasil nilai akurasi Algoritma *KNN* dengan K=3 memiliki nilai sebesar 93%, sedangkan algoritma *Naïve Bayes* memiliki akurasi sebesar 95%.

Penelitian oleh Hasan, Fuad Nur, dkk.[6] melakukan analisis sentimen berita sepak bola dengan tokoh dunia Lionel Messi memiliki akurasi SVM (PSO) 84% sedangkan Naïve Bayes (PSO) 83%.

Penelitian oleh Agustina, Dyah Auliya, dkk.[7] melakukan analisis sentimen dibeberapa *Marketplace* seperti Tokopedia, Bukalapak, dan Shoppe dengan menggunakan Algoritma *Support Vector Machine*. Penelitian ini menunjukkan kinerja klasifikasi bahwa nilai G-mean dan AUC terbaik untuk data pengujian Bukalapak adalah 0,85 dan 0,86 pada lipatan pertama. Sedangkan nilai G-mean dan AUC terbaik untuk data pengujian Tokopedia adalah 0,76 dan 0,77 pada lipatan ketujuh dan nilai G-mean dan AUC terbaik untuk data pengujian Tokopedia adalah 0,82 dan 0,83 pada lipatan keenam.

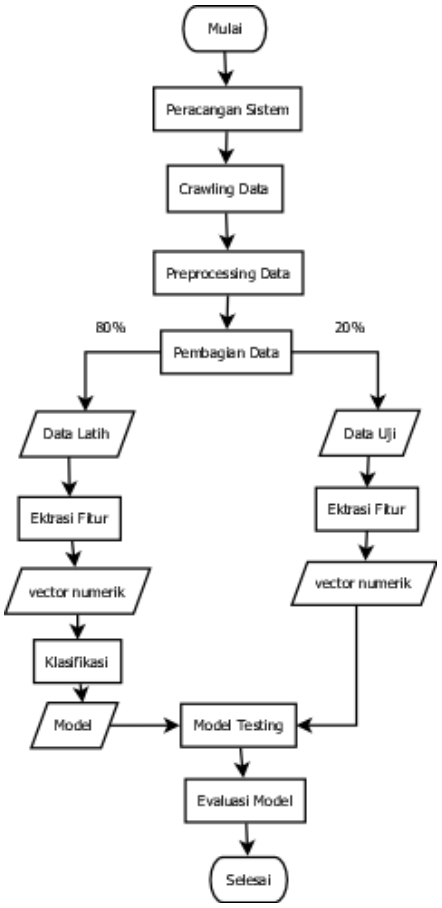
Penelitian dari Ardhiansya, Hikari. dkk.[8] melakukan analisis sentimen pendapat masyarakat terhadap PPKM DKI Jakarta dengan metode Naïve Bayes menghasilkan 87,2% netral, 4,3% positif dan 8,4% negative dan klasifikasi dengan Naïve Bayes 2 kelas mendapatkan akurasi 90% dan untuk 3 kelas mendapatkan akurasi 81%.

Penelitian dari Putra, Ricky Eka, dkk.[9] melakukan mengklasifikasikan data dari database Messidor ke dalam kelas-kelas yang relevan dengan tingkat keparahan retinopati diabetik, yaitu kelas ringan, sedang, dan parah. Hasil dari eksperimen diperoleh dari penggunaan Homomorphic, ResNet50, dan Relief sebelum memasuki Support Vector Machine-Naïve Bayes. Homomorphic mendapatkan akurasi 85.87%, ResNet50 mencapai akurasi 86.76%, dan Relief mencapai akurasi 89.12%.

Penelitian yang dilaksanakan oleh Dyantono, Aganda Maulan Dan, dan Ricky Eka Putra[10] menggunakan komentar dari kanal YouTube yang dipandu oleh Deddy Corbuzier, membahas topik seputar ChatGPT. Pengambilan data dilakukan melalui Google Spreadsheet, sementara analisisnya dilakukan dengan RStudio dan Jupyter Notebook untuk membandingkan kinerja model berdasarkan Accuracy, Precision, Recall, F1 Score, dan ROC. Hasilnya menunjukkan bahwa model Word2vec CBOW CNN unggul dalam performa dibandingkan dengan model Sent2vec TF-IDF LR, menunjukkan perbedaan nilai yang signifikan pada metrik-metrik tersebut. Sebagai contoh, terdapat perbedaan sebesar 4,09% pada Accuracy, 6,75% pada Precision, 0,06% pada Recall, 2,81% pada F1 Score, dan 0,2% pada ROC.

II. METODE PENELITIAN

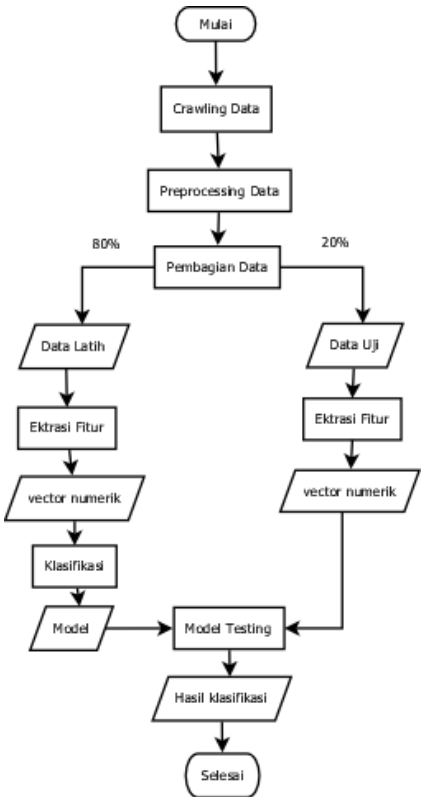
Pada bab ini membahas tentang metode penelitian yang digunakan untuk melakukan Pengembangan Sistem Analisis Sentimen Masyarakat Terhadap Chatgpt Pada Twitter Dengan Perbandingan Metode Naive Bayes Classifier Dan K-Nearest Neighbors. Metode ini dipilih karena mampu mengklasifikasikan sentimen positif, negatif, dan netral dari teks dengan akurasi yang tinggi. Gbr. 1 merupakan alur dari metode penelitian ini.



Gbr. 1 Alur metode penelitian.

A. Perancang Sistem

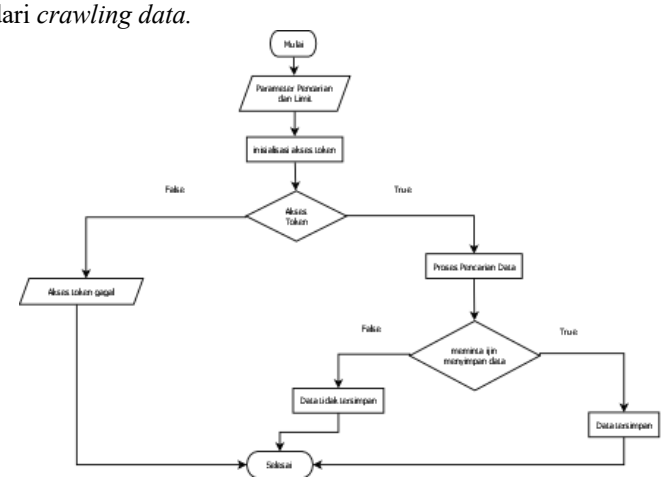
Perancang sistem bertujuan untuk merancang antarmuka grafis (GUI) dan mengaplikasikan sistem analisis sentimen. Untuk perancangan antarmuka grafis (GUI) menggunakan *library Tkinter* maupun *custom tkinter*. Pada tahap ini menjelaskan tahapan proses yang dilakukan sistem mulai dari crawling data hingga perhitungan klasifikasi dan evaluasi model. Gbr 2 menunjukkan proses pembuatan sistem.



Gbr. 2 Proses Pembuatan Sistem.

B. Crawling Data

Crawling data dilakukan dengan menggunakan *library Twint* dan *Tweepy*. *Library Twint* maupun *tweepy* merupakan library untuk mengambil data tweet. Crawling data dilakukan untuk mendapatkan sejumlah data dari media twitter dengan keyword "Chatgpt". Data yang diperoleh akan disimpan dalam bentuk file format .xlsx. Gbr 3 merupakan flowchart dari *crawling data.*



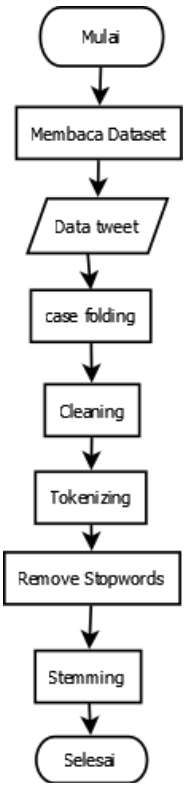
Gbr. 3 Flowchart *crawling data*

Langkah awal melibatkan penggunaan parameter pencarian dan batasan jumlah data pencarian. Setelah itu, sistem melakukan inisialisasi akses token yang terdiri dari access token, access token secret, consumer key, dan consumer secret.

Selanjutnya, sistem melakukan pemeriksaan terhadap akses token yang telah diinisialisasi untuk memastikan keabsahannya. Jika akses token dinyatakan valid, maka proses akan melanjutkan ke langkah berikutnya. Pada tahap berikutnya, sistem akan menjalankan proses pencarian data berdasarkan parameter pencarian dan batasan yang telah ditentukan. Setelah data berhasil ditemukan, sistem akan memberikan opsi kepada pengguna untuk menyimpan data tersebut. Jika pengguna memberikan izin untuk menyimpan data, maka data akan disimpan. Namun, jika pengguna tidak memberikan izin, data tidak akan disimpan. Lalu dataset tersebut dilakukan proses labeling untuk mengidentifikasi opini dari data tersebut. Pengelompokan label dilakukan dalam tiga kategori, yakni sentimen positif, sentimen negatif, dan sentimen netral.

C. Preprocessing

Preprocessing merupakan tahap awal persiapan data yang dilakukan sebelum data diolah lebih lanjut. Tujuan dari preprocessing adalah untuk menghilangkan kerusakan dalam data dan mempersiapkan data agar siap untuk diolah oleh algoritma atau model machine learning. Gbr 4 menunjukkan proses preprocessing.



Gbr. 4 Langkah-langkah proses Preprocessing.

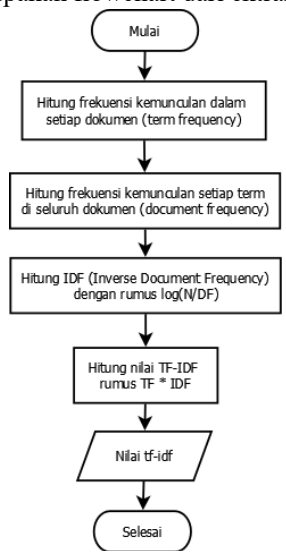
Berikut ini penjelasan dari tahap-tahap proses preprocessing:

1. *Case folding:* semua huruf dalam teks diubah menjadi huruf kecil untuk konsistensi

2. *Cleaning:* pembersihan yang mencakup penghapusan mention, link, serta simbol dan angka yang tidak relevan.
3. *Tokenizing:* teks dibagi menjadi bagian-bagian kecil yang disebut token.
4. *Remove Stopword*s: penghapusan kata-kata yang tidak memiliki makna sehingga hanya kata-kata penting yang tersisa.
5. *Stemming:* yang bertujuan untuk mengubah kata-kata dalam teks menjadi bentuk dasarnya.

D. Ektraksi Fitur

Ekstraksi fitur merupakan proses mengubah data mentah menjadi sekumpulan fitur yang dapat digunakan untuk melatih model pembelajaran mesin. Ekstraksi fitur bertujuan untuk mengidentifikasi aspek-aspek penting dari data yang dapat membantu model pembelajaran mesin membuat prediksi yang akurat. Gbr. 5 merupakan flowchart dari ektraksi fitur.



Gbr. 5 Flowchart Ektraksi Fitur

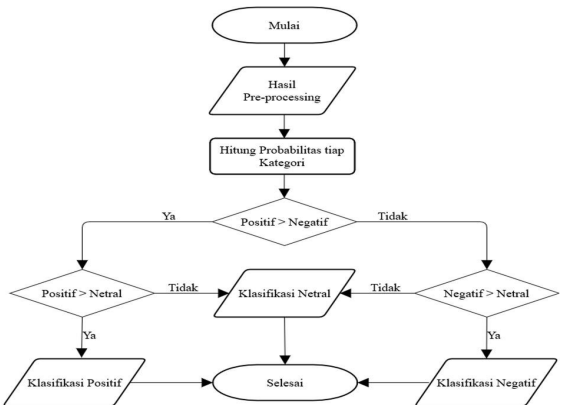
Pada tahap awal, dilakukan perhitungan Term Frequency (TF), yang mengukur frekuensi kemunculan kata-kata dalam setiap dokumen. Selanjutnya, tahap berikutnya adalah menghitung Document Frequency (DF), yang mengukur frekuensi kata-kata dalam seluruh koleksi dokumen. Setelah itu, dilakukan perhitungan Inverse Document Frequency (IDF) menggunakan rumus Log(N/DF), dimana N adalah jumlah total dokumen dalam koleksi. IDF mengukur sejauh mana kata tertentu unik dalam koleksi dokumen.

Langkah terakhir adalah menghitung nilai TF-IDF dengan mengalikan nilai TF dengan nilai IDF. Ini menghasilkan skor TF-IDF untuk setiap kata dalam setiap dokumen, yang dapat digunakan dalam berbagai aplikasi analisis teks.

Proses Ekstraksi fitur bertujuan untuk mengkonversi data tweet menjadi vektor numerik. Dalam metode ekstraksi ini, digunakan pendekatan TF-IDF (Term Frequency-Inverse Document Frequency), yang mengukur frekuensi kata dalam setiap kalimat dan memberikan nilai bobot lebih besar pada kata-kata yang sering muncul dalam data.

E. Klasifikasi Naïve Bayes Classifier

Naive Bayes Classifier adalah metode klasifikasi yang memanfaatkan dasar probabilitas untuk mengkategorikan objek ke dalam berbagai kelompok berdasarkan ciri-cirinya. Algoritma ini sering dipergunakan dalam ranah data mining dan machine learning, terutama untuk klasifikasi teks dan pengenalan pola. Gbr 6 menunjukkan cara *Naïve Bayes* memutuskan kelas.



Gbr. 6 Menentukan Kelas Naïve Bayes

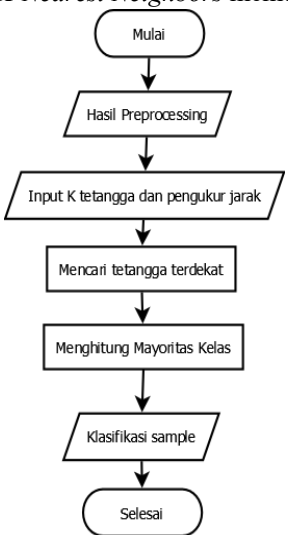
Proses klasifikasi sentimen menggunakan Naive Bayes dimulai dengan memasukkan data yang telah melewati tahap pre-processing. Pre-processing ini dapat mencakup langkah-langkah seperti pembersihan teks, tokenisasi, penghapusan kata-kata berhenti (stop words), dan normalisasi. model Naive Bayes menghitung probabilitas posterior untuk setiap kategori sentimen berdasarkan data pelatihan.

Perhitungan ini melibatkan mengalikan probabilitas prior dari setiap kategori dengan likelihood, yaitu probabilitas data teramati mengingat kategori tersebut. Model kemudian membandingkan probabilitas posterior antara kategori 'positif' dan 'negatif'. Jika 'positif' memiliki probabilitas yang lebih tinggi, model melanjutkan untuk membandingkan 'positif' dengan 'netral'. Sebaliknya, jika 'negatif' lebih tinggi, model akan membandingkannya dengan 'netral'. Kategori dengan probabilitas tertinggi dipilih sebagai hasil klasifikasi. Jika 'positif' lebih tinggi daripada 'negatif' dan juga 'netral', sampel diklasifikasikan sebagai 'positif'. Jika 'negatif' lebih tinggi daripada 'positif' dan 'netral', sampel diklasifikasikan sebagai 'negatif'. Jika tidak ada satupun yang lebih tinggi, maka sampel diklasifikasikan sebagai 'netral', menyelesaikan proses klasifikasi.

F. Klasifikasi K-Nearest Neighbors

K-Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin yang digunakan untuk masalah klasifikasi dan regresi. Algoritma ini bekerja dengan cara mencari K tetangga terdekat dari sebuah data uji (*test data*) dalam ruang fitur (*feature space*) dari data latih (*training data*), dan kemudian melakukan klasifikasi atau prediksi berdasarkan

mayoritas kelas dari K tetangga terdekat tersebut. Gbr 7 menunjukkan cara *K-Nearest Neighbors* memutuskan kelas.



Gbr. 7 Menentukan Kelas *K-Nearest Neighbors*

Pada Flowchart diatas, proses klasifikasi menggunakan k-Nearest Neighbors dimulai dengan memasukkan input file yang telah dilakukan preprocessing data. Data yang telah diproses ini kemudian siap untuk dilanjutkan ke proses klasifikasi. Berikutnya, kita melakukan input nilai k yang merepresentasikan jumlah tetangga terdekat yang akan dipertimbangkan dalam algoritma dan memilih pengukur jarak yang sesuai, seperti jarak Euclidean maupun Manhattan, yang akan digunakan untuk mengukur kedekatan antara contoh dalam ruang fitur. Setelah parameter k dan matrik jarak ditentukan algoritma. Kemudian mencari k tetangga terdekat untuk setiap sampel yang akan diklasifikasikan. Ini dilakukan dengan menghitung jarak antara sampel yang akan diklasifikasikan dan semua sampel dalam set pelatihan dan memilih k sampel dengan jarak terkecil. Dengan k tetangga terdekat yang telah diidentifikasi, algoritma selanjutnya menghitung mayoritas kelas di antara tetangga tersebut.

Kategori yang paling sering muncul di antara tetangga terdekat inilah yang akan ditetapkan sebagai kelas dari sampel yang diuji. Dengan klasifikasi kategori yang ditentukan berdasarkan suara mayoritas, sampel dianggap telah diklasifikasikan.

G. Evaluasi Model

Pada tahap evaluasi model dilakukan untuk mengukur seberapa akurat model dalam memprediksi sentimen. Evaluasi model penting guna memastikan bahwa model yang dihasilkan dapat digunakan dengan baik untuk menganalisis sentimen. Evaluasi model dilakukan pengujian dengan K-Fold Cross Validation dan split data.

K-Fold Cross Validation adalah salah satu teknik validasi model yang umum digunakan dalam machine learning untuk mengukur kinerja model dan meminimalkan overfitting pada data training. Metode ini melibatkan pembagian data training menjadi k kelompok yang sama besar, atau biasa disebut "fold". Setiap fold kemudian bergiliran sebagai data validasi sementara k-1 fold yang lainnya digunakan sebagai data training. Proses ini diulang sebanyak k kali, sehingga setiap fold digunakan sebagai data validasi tepat satu kali[11].

III. HASIL DAN PEMBAHASAN

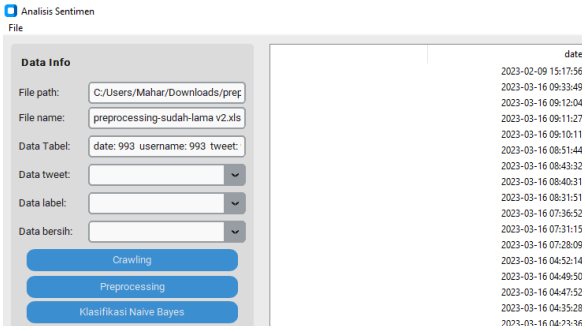
Hasil penelitian serta penerapan pengembangan sistem untuk menganalisis sentimen masyarakat terhadap ChatGPT di *platform* Twitter. Analisis akan mencakup perbandingan penggunaan metode klasifikasi *Naive Bayes Classifier* dan *K-Nearest Neighbors.*

A. Hasil Perancangan Sistem

Perancangan sistem dengan menggunakan library *Tkinter* dan *customtkinter*. *Tkinter* maupun *customtkinter* merupakan *library* yang digunakan untuk membuat antarmuka grafis (GUI). Berikut ini tampilan antarmuka yang dibutuhkan:

1. Halaman Beranda

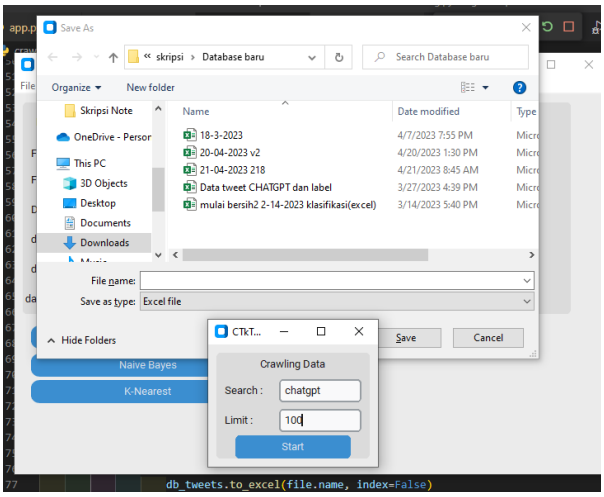
Halaman beranda akan menampilkan beberapa tombol dan informasi terkait data yang akan diolah. Di halaman tersebut akan terdapat tombol preprocessing, tombol klasifikasi *Naive Bayes*, dan tombol klasifikasi *K-Nearest Neighbors*. Selain itu, informasi tentang data akan ditampilkan seperti nama file, letak file, besarnya data, dan tabel data. Gbr. 8 menunjukkan gambar halaman beranda.



Gbr. 8 Halaman Beranda

2. Halaman *Crawling Data*

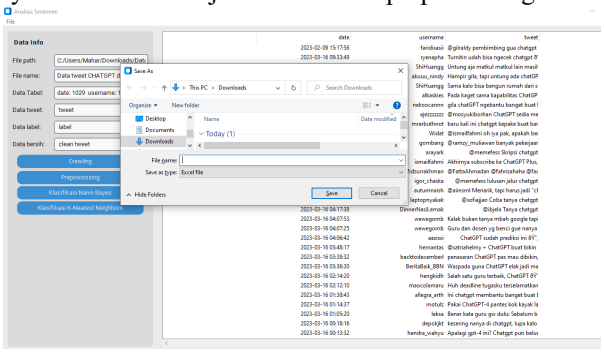
Halaman ini dirancang untuk pengguna dalam melakukan pencarian data dengan memasukkan input tertentu dan menentukan batasan jumlah data yang ingin dicari. Selanjutkan akan dilakukan validasi token API key valid jika tidak valid maka selesai dan jika valid akan dilakukan pencarian. Setelah pencarian selesai maka ada validasi untuk menyimpan data tersebut. Dengan demikian, pengguna dapat memperoleh data yang relevan dan sesuai dengan kebutuhan mereka. Gbr. 9 merupakan gambar halaman *crawling* data.



Gbr. 9 Halaman *Crawling Data*

3. Halaman Preprocessing

Pada tahap *preprocessing*, dilakukan langkah-langkah untuk menghindari gangguan dari proses lainnya seperti menonaktifkan tombol yang tidak diperlukan. Hal ini dilakukan karena tahap *preprocessing* membutuhkan waktu yang cukup banyak, sekitar 30-40 menit untuk memproses 1000 data, sehingga untuk memastikan proses berjalan dengan lancar, langkah-langkah tersebut perlu dilakukan. Setelah tahap preprocessing selesai dilakukan, data akan disimpan dalam bentuk file atau format yang lebih mudah oleh proses selanjutnya. Gbr. 10 menunjukkan halaman preprocessing.

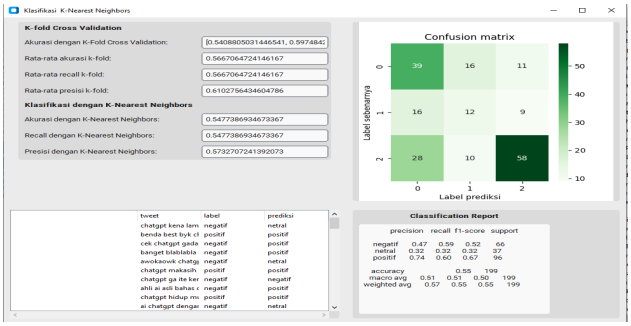


Gbr. 10 Halaman *Preprocessing*

4. Halaman Klasifikasi

Halaman ini dirancang untuk menampilkan hasil evaluasi klasifikasi *Naive Bayes* ataupun *K-Nearest Neighbors.* Pada halaman ini, akan ditampilkan nilai akurasi, recall, dan presisi dari *k-fold cross validation* serta nilai akurasi, recall, dan presisi dari klasifikasi. Selain itu, halaman ini juga akan menampilkan *confusion matrix* dan *classification report* sebagai gambaran visual dan detail tentang kinerja model klasifikasi.

Selain itu, halaman ini akan menampilkan tabel prediksi dari data uji yang telah diolah. Tabel ini akan berisi prediksi dari model klasifikasi terhadap data uji dan dapat membantu pengguna untuk mengevaluasi kinerja model pada data yang belum pernah dilihat sebelumnya. Gbr. 11 menunjukkan halaman untuk klasifikasi *Naive Bayes* dan *K-Nearest Neighbors.*



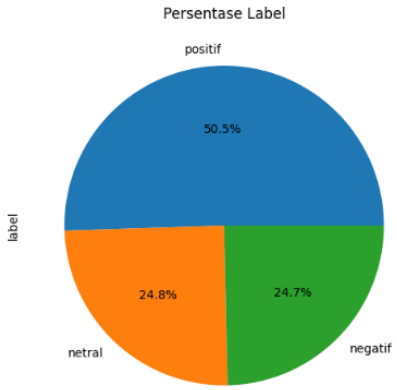
Gbr. 11 Pop up metode klasifikasi

B. *Crawling Data*

Crawling data dilakukan dengan menggunakan *library Twint* dan *Tweepy*. Library Twint maupun tweepy merupakan library untuk mengambil data tweet dari Twitter. Perbedaannya kedua *library* yaitu *library Twint* tidak perlu memiliki token dari Twitter sedangkan *library Tweepy* untuk mengakses API Twitter resmi perlu mengajukan permohonan untuk akses pengembang dan mendapatkan *API key* dari Twitter.

Pada awalnya akan memasukkan parameter kata yang akan dilakukan pencarian dan batas jumlah data pencarian. Setelah itu, dilakukan inisialisasi aksen token berupa API key dari Twitter. Selanjutnya, API key digunakan untuk membuktikan identitas jika API key dinyatakan benar, maka proses akan melanjutkan ke langkah berikutnya.

Pada tahap berikutnya, sistem akan menjalankan proses pencarian data berdasarkan parameter pencarian dan batas jumlah yang telah ditentukan. Hasil pencarian akan menyimpan data username dan tweet. Data akan disimpan dalam bentuk file format .xlsx dari dataset telah disimpan ditemukan 1229 data yang digunakan dalam penelitian ini. Setelah itu data tersebut akan diberi label seperti positif, negatif, dan netral. Gbr 12 menunjukkan persentase masing-masing label.



Gbr. 12 Persentase label

C. *Hasil Preprocessing*

Pada dataset yang telah diberikan label akan dilakukan preprocessing untuk mempersiapkan data agar siap dilakukan

klasifikasi metode. Contoh hasil dari preprocessing terdapat pada Tabel I.

TABEL I
PROSES PREPROCESSING

Tweet asli	
Sama kalo bisa bangun rumah dari sekarang, anak sipil angkatanku (aku dan temanku) ngerjain uts pake chatgpt https://t.co/zDgFhNvYvU	
Tahap Preprocessing	Hasil Preprocessing
Case Folding	sama kalo bisa bangun rumah dari sekarang, anak sipil angkatanku (aku dan temanku) ngerjain uts pake chatgpt https://t.co/zdgfhnvyvu
Cleaning	sama kalo bisa bangun rumah dari sekarang anak sipil angkatanku aku dan temanku ngerjain uts pake chatgpt
Tokenizing	['sama', 'kalo', 'bisa', 'bangun', 'rumah', 'dari', 'sekarang', 'anak', 'sipil', 'angkatanku', 'aku', 'dan', 'temanku', 'ngerjain', 'uts', 'pake', 'chatgpt']
Remove Stopwords	['kalo', 'bangun', 'rumah', 'anak', 'sipil', 'angkatanku', 'temanku', 'ngerjain', 'uts', 'pake', 'chatgpt']
Stemming	['kalo', 'bangun', 'rumah', 'anak', 'sipil', 'angkat', 'teman', 'ngerjain', 'uts', 'pake', 'chatgpt']
Hasil Preprocessing	
kalo bangun rumah anak sipil angkat teman ngerjain uts pake chatgpt	

D. *Klasifikasi Naïve Bayes*

Hasil klasifikasi dengan metode *Naive Bayes* menggunakan *library ComplementNB* maupun *MultinomialNB*. Hasil performa akurasi model klasifikasi *Naive Bayes* pada data uji dengan menggunakan rasio *data split* 80:20 menunjukkan nilai akurasi sebesar 81,7%. Hasil Klasifikasi *Naive Bayes* menggunakan *multinomialNB()* dan *ComplementNB()* pada Tabel II.

TABEL II
KLASIFIKASI NAÏVE BAYES

	Multinomial	Complement
Hasil Akurasi	64,2%	81,7%
Hasil Presisi	72,9%	81,9%
Hasil Recall	64,2%	81,7%
Hasil F1-score	68,2%	81,7%

Model *ComplementNB* memiliki kinerja yang lebih baik daripada *MultinomialNB* dengan dataset yang ada. Dengan akurasi data uji sekitar 81,7%, presisi 81,9%, recall 81,7%, dan F1-score 81,7%. Performa yang tinggi dari *ComplementNB* bisa disebabkan oleh kemampuan yang lebih baik dalam menangani data.

E. *Klasifikasi K-Nearest Neighbors*

Hasil klasifikasi *K-Nearest Neighbors* dengan jarak *Euclidean* dan *Manhattan.* Hasil performa akurasi model dengan menggunakan rasio *data split* 80:20 menunjukkan bahwa jarak *Euclidean* lebih baik daripada jarak *Manhattan* dengan dataset yang ada. Hasil dari akurasi untuk masing-masing jarak terdapat pada Tabel III.

TABEL III
KLASIFIKASI K-NEAREST NEIGHBORS

K	*Euclidean*	*Manhattan*
2	72,7%	37,8%
3	76,1%	58,2%
4	75,7%	53,2%
5	80,4%	52,3%
6	78,2%	54%
7	78,7%	40%
8	80%	40%
9	75%	35,3%
10	73,6%	33,1%

F. *Hasil Pengujian Split Data*

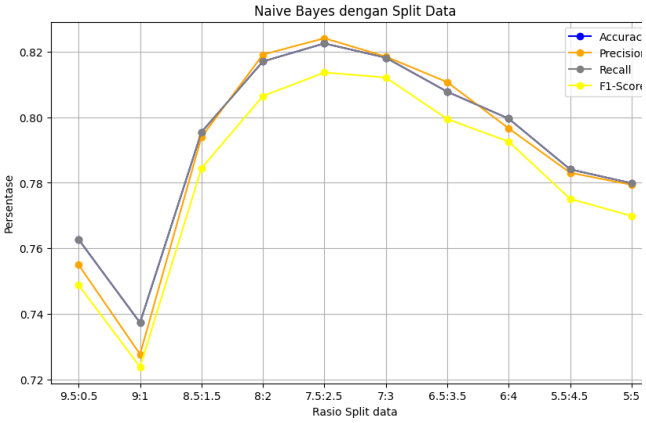
Pengujian split data akan dilakukan pada metode *Naive Bayes Classifier* dan *K-Nearest Neighbors.* Pada pengujian *split data* dengan *Naïve Bayes* menggunakan model *ComplementNB.* Hasil pengujian dengan metode *Naive Bayes* pada Tabel IV.

. TABEL IV
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *SPLIT DATA*

Rasio	Akurasi	Presisi	Recall	F1-Score
95:5	76.27%	75.50%	76.27%	74.88%
90:10	73.73%	72.77%	73.73%	72.38%
85:15	79.55%	79.39%	79.55%	78.44%
80:20	81.70%	81.91%	81.70%	80.65%
75:25	82.25%	82.41%	82.25%	81.37%
70:30	81.82%	81.85%	81.82%	81.21%

65:35	80.78%	81.07%	80.78%	79.94%
60:40	79.96%	79.66%	79.96%	79.26%
55:45	78.41%	78.31%	78.41%	77.51%
50:50	77.99%	77.94%	77.99%	76.99%

Berdasarkan hasil pengujian *Naïve Bayes* menggunakan *split data* memiliki akurasi tertinggi dengan rasio 75:25 yaitu sebesar 82.25%, presisi sebesar 82,41%, recall senilai 82,25% dan f1-score 81,37%. Gbr. 13 merupakan grafik hasil *Naïve Bayes* dengan pengujian *split data.*



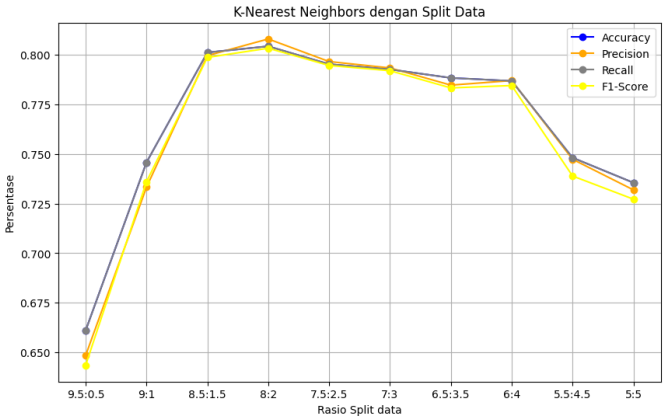
Gbr. 13 Grafik *Naïve Bayes* dengan *split data*

Pada pengujian *split data* dengan *K-Nearest Neighbors* jarak Euclidean. Hasil pengujian *split data dengan K-Nearest Neighbors* pada Tabel V.

. TABEL V
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *SPLIT DATA*

Rasio	Akurasi	Presisi	Recall	F1-Score
95:5	66.10%	64.84%	66.10%	64.35%
90:10	74.58%	73.33%	74.58%	73.60%
85:15	80.11%	79.95%	80.11%	79.86%
80:20	80.43%	80.79%	80.43%	80.33%
75:25	79.52%	79.65%	79.52%	79.46%
70:30	79.26%	79.33%	79.26%	79.19%
65:35	78.83%	78.47%	78.83%	78.33%
60:40	78.68%	78.70%	78.68%	78.45%
55:45	74.81%	74.73%	74.81%	73.89%
50:50	73.55%	73.19%	73.55%	72.72%

Berdasarkan hasil pengujian K-Nearest Neighbors menggunakan split data memiliki akurasi tertinggi dengan rasio 80:20 yaitu sebesar 80,43%, presisi sebesar 80,79%, recall senilai 80,42% dan f1-score 80.33%. Gbr. 14 merupakan grafik hasil *K-Nearest Neighbors* dengan pengujian *split data*



Gbr. 14 Grafik *K-Nearest Neighbors* dengan *split data*

G. *Hasil Pengujian K-Fold Cross Validation*

Dalam penggunaan *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* maupun *K-Nearest Neighbors,* langkah ini dilakukan untuk mencegah *overfitting* pada data latih tertentu dan menguji kinerja model pada data yang berbeda. Proses *K-Fold Cross Validation* melibatkan pembagian data latih menjadi 11 subset dengan ukuran yang sama.

Model kemudian diuji sebanyak sebelas kali, di mana pada setiap iterasi, satu subset digunakan sebagai data uji, sementara sepuluh subset lainnya digunakan sebagai data latih.

Berikut ini proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan beberapa rasio dataset:

1. Rasio 95:5

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 95:5 menghasilkan rata-rata akurasi 82,74% dan 79,98%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 89,10% dan 84,16%. Tabel VI dan Tabel VII merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL VI
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81,37%	81.56%	81.37%	81.22%
2	87,25%	87.11%	87.25%	87.15%
3	87,12%	86.95%	87.13%	86.90%
4	80,19%	79.75%	80.20%	79.46%
5	76,23%	77.12%	76.24%	75.77%
6	81,18%	81.51%	81.19%	80.61%
7	78,21%	77.79%	78.22%	77.88%
8	80,19%	80.64%	80.20%	79.74%

9	82,17%	82.25%	82.18%	82.12%
10	89,10%	89.04%	89.11%	89.02%
11	87,12%	87.57%	87.13%	86.91%

TABEL VII
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81.37%	80.88%	81.37%	80.90%
2	81.37%	82.49%	81.37%	81.50%
3	77.23%	77.27%	77.23%	76.98%
4	79.21%	78.94%	79.21%	78.49%
5	72.28%	72.67%	72.28%	71.24%
6	74.26%	75.29%	74.26%	73.62%
7	76.24%	76.58%	76.24%	76.34%
8	77.23%	78.20%	77.23%	76.84%
9	75.25%	75.33%	75.25%	74.68%
10	84.16%	85.05%	84.16%	83.97%
11	79.21%	79.00%	79.21%	78.53%

2. Rasio 90:10

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 90:10 menghasilkan rata-rata akurasi 83,5% dan 78,18%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 92,63% dan 82,29%. Tabel VIII dan Tabel IX merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL VIII
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81,37%	81.56%	81.37%	81.22%
2	87,25%	87.11%	87.25%	87.15%
3	87,12%	86.95%	87.13%	86.90%
4	80,19%	79.75%	80.20%	79.46%
5	76,23%	77.12%	76.24%	75.77%
6	81,18%	81.51%	81.19%	80.61%
7	78,21%	77.79%	78.22%	77.88%
8	80,19%	80.64%	80.20%	79.74%
9	82,17%	82.25%	82.18%	82.12%
10	89,10%	89.04%	89.11%	89.02%
11	87,12%	87.57%	87.13%	86.91%

TABEL IX
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	79.17%	79.91%	79.17%	79.23%
2	82.29%	83.09%	82.29%	82.55%
3	81.25%	81.68%	81.25%	80.87%
4	79.17%	79.66%	79.17%	78.53%
5	75.00%	74.97%	75.00%	74.51%
6	72.92%	74.11%	72.92%	72.34%
7	78.12%	78.20%	78.12%	77.91%
8	76.04%	76.94%	76.04%	74.98%
9	75.00%	75.37%	75.00%	75.02%
10	82.11%	83.23%	82.11%	81.69%
11	78.95%	78.62%	78.95%	78.29%

3. Rasio 85:15

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 85:15 menghasilkan rata-rata akurasi 83,54% dan 77,81%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 90% dan 84,44%. Tabel X dan Tabel XI merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL X
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	86.81%	86.43%	86.81%	86.55%
2	86.81%	86.75%	86.81%	86.70%
3	76.92%	76.49%	76.92%	76.27%
4	80.22%	80.58%	80.22%	79.78%
5	79.12%	79.47%	79.12%	77.86%
6	81.32%	81.35%	81.32%	81.32%
7	81.11%	82.13%	81.11%	80.11%
8	83.33%	83.57%	83.33%	83.20%
9	83.33%	83.68%	83.33%	83.46%
10	90.00%	90.02%	90.00%	89.90%
11	90.00%	90.27%	90.00%	90.04%

TABEL XI
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	83.52%	83.90%	83.52%	83.62%
2	79.12%	79.70%	79.12%	78.71%
3	76.92%	77.34%	76.92%	75.79%
4	74.73%	76.57%	74.73%	73.74%
5	74.73%	74.74%	74.73%	73.73%
6	78.02%	78.79%	78.02%	77.55%
7	66.67%	65.89%	66.67%	65.29%
8	76.67%	76.42%	76.67%	76.38%
9	78.89%	79.03%	78.89%	78.91%
10	82.22%	82.57%	82.22%	81.68%
11	84.44%	85.06%	84.44%	84.11%

4. Rasio 80:20

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 80:20 menghasilkan rata-rata akurasi 83,45% dan 78,97%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 89,41% dan 88,23% Tabel XII dan Tabel XIII merupakan

hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XII
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	86,04%	85,82%	86,04%	85,85%
2	83,72%	83,87%	83,72%	83,20%
3	77,64%	78,58%	77,64%	77,09%
4	82,35%	82,50%	82,35%	81,90%
5	74,11%	74,77%	74,11%	73,55%
6	83,52%	83,42%	83,52%	83,42%
7	83,52%	85,23%	83,52%	82,92%
8	85,88%	85,67%	85,88%	85,72%
9	84,70%	86,32%	84,70%	84,93%
10	89,41%	89,41%	89,41%	89,39%
11	87,05%	87,99%	87,05%	87,11%

TABEL XII
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81,39%	81,71%	81,39%	80,88%
2	80,23%	80,69%	80,23%	79,46%
3	77,64%	79,93%	77,64%	76,58%
4	70,58%	69,81%	70,58%	69,98%
5	75,29%	76,13%	75,29%	75,01%
6	83,52%	83,30%	83,52%	83,09%
7	70,58%	71,38%	70,58%	69,49%
8	76,47%	77,04%	76,47%	76,20%
9	83,52%	84,09%	83,52%	83,70%
10	81,17%	81,17%	81,17%	80,76%
11	88,23%	88,86%	88,23%	87,86%

5. Rasio 75:25

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 75:25 menghasilkan rata-rata akurasi 82,71% dan 77,14%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 92,5% dan 87,34%. Tabel XIV dan Tabel XV merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XIV
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	83.75%	83.83%	83.75%	83.76%
2	77.50%	76.47%	77.50%	76.42%
3	81.25%	82.03%	81.25%	81.17%
4	81.25%	81.76%	81.25%	80.78%
5	76.25%	76.21%	76.25%	76.01%
6	83.75%	83.77%	83.75%	83.62%
7	78.75%	79.90%	78.75%	78.61%
8	85.00%	87.73%	85.00%	85.25%
9	82.50%	83.89%	82.50%	82.58%
10	92.50%	92.66%	92.50%	92.45%
11	87.34%	87.50%	87.34%	87.06%

TABEL XV
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81.25%	80.74%	81.25%	80.66%
2	76.25%	75.93%	76.25%	75.66%
3	68.75%	69.65%	68.75%	68.03%
4	72.50%	73.43%	72.50%	72.74%
5	77.50%	80.03%	77.50%	76.89%
6	76.25%	76.38%	76.25%	76.10%
7	73.75%	73.21%	73.75%	72.76%
8	73.75%	74.63%	73.75%	73.13%
9	75.00%	75.37%	75.00%	75.12%
10	86.25%	86.28%	86.25%	86.05%
11	87.34%	87.89%	87.34%	86.84%

6. Rasio 70:30

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 70:30 menghasilkan rata-rata akurasi 81,24% dan 75,75%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 90,54% dan 82,43%. Tabel XVI dan Tabel XVII merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XVI
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	77.33%	77.54%	77.33%	76.44%
2	74.67%	75.25%	74.67%	74.74%
3	82.67%	82.10%	82.67%	81.96%
4	74.67%	76.28%	74.67%	74.33%
5	78.67%	78.98%	78.67%	78.77%
6	78.67%	79.74%	78.67%	78.25%
7	79.73%	80.99%	79.73%	79.55%
8	83.78%	84.71%	83.78%	84.01%
9	86.49%	87.75%	86.49%	86.59%
10	90.54%	90.93%	90.54%	90.44%
11	86.49%	86.24%	86.49%	86.34%

TABEL XVII
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	81.33%	82.76%	81.33%	80.46%
2	70.67%	71.12%	70.67%	69.75%
3	70.67%	71.49%	70.67%	70.98%
4	72.00%	72.53%	72.00%	71.80%
5	78.67%	79.05%	78.67%	78.17%

6	68.00%	66.59%	68.00%	66.32%
7	81.08%	81.51%	81.08%	80.53%
8	67.57%	67.81%	67.57%	67.47%
9	79.73%	79.45%	79.73%	79.53%
10	82.43%	82.59%	82.43%	81.86%
11	81.08%	80.89%	81.08%	80.96%

7. Rasio 65:35

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 65:35 menghasilkan rata-rata akurasi 79,77% dan 75,57%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 86,96% dan 82,61% Tabel XVIII dan Tabel XIX merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XVIII
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	77.14%	76.79%	77.14%	76.72%
2	75.71%	76.43%	75.71%	75.68%
3	86.96%	87.75%	86.96%	86.31%
4	68.12%	68.20%	68.12%	68.13%
5	76.81%	76.81%	76.81%	76.71%
6	75.36%	74.78%	75.36%	74.08%
7	86.96%	88.51%	86.96%	87.38%
8	75.36%	77.50%	75.36%	75.68%
9	84.06%	85.10%	84.06%	84.35%
10	84.06%	85.09%	84.06%	83.75%
11	86.96%	89.63%	86.96%	87.68%

TABEL XIX
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	77.14%	77.91%	77.14%	76.21%
2	65.71%	65.74%	65.71%	64.78%
3	78.26%	80.22%	78.26%	77.52%
4	71.01%	73.72%	71.01%	69.89%
5	82.61%	82.51%	82.61%	82.37%
6	63.77%	66.13%	63.77%	61.81%
7	79.71%	82.53%	79.71%	78.64%
8	73.91%	76.89%	73.91%	73.45%
9	81.16%	80.85%	81.16%	80.94%
10	76.81%	79.39%	76.81%	76.10%
11	81.16%	80.51%	81.16%	80.41%

8. Rasio 60:40

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 60:40 menghasilkan rata-rata akurasi 79,38% dan 74,55%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 90,62% dan 85,94%. Tabel XX dan Tabel XXI merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XX
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	79.69%	81.14%	79.69%	79.64%
2	75.00%	74.94%	75.00%	74.95%
3	75.00%	74.76%	75.00%	74.85%
4	75.00%	76.22%	75.00%	75.42%
5	81.25%	81.12%	81.25%	80.85%
6	76.56%	78.04%	76.56%	76.18%
7	79.69%	79.44%	79.69%	79.30%
8	73.44%	76.25%	73.44%	73.87%
9	90.62%	90.76%	90.62%	90.66%
10	82.81%	83.19%	82.81%	82.48%
11	84.13%	86.37%	84.13%	84.14%

TABEL XXI
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	67.19%	70.99%	67.19%	66.47%
2	64.06%	66.00%	64.06%	64.78%
3	71.88%	72.91%	71.88%	71.76%
4	85.94%	86.33%	85.94%	85.86%
5	71.88%	71.26%	71.88%	71.12%
6	70.31%	73.67%	70.31%	68.92%
7	75.00%	78.23%	75.00%	72.83%
8	81.25%	82.00%	81.25%	81.37%
9	79.69%	78.79%	79.69%	79.13%
10	71.88%	72.50%	71.88%	71.74%
11	80.95%	80.52%	80.95%	80.47%

9. Rasio 55:45

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 55:45 menghasilkan rata-rata akurasi 78,27%. dan 72,85%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 89,66%. dan 82,76%. Tabel XXII dan Tabel XXIII merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XXII
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	77.97%	79.43%	77.97%	78.23%
2	76.27%	75.73%	76.27%	75.63%
3	66.10%	67.42%	66.10%	66.58%
4	84.75%	84.72%	84.75%	84.68%

5	74.58%	74.53%	74.58%	73.17%
6	79.66%	82.16%	79.66%	80.30%
7	68.97%	70.51%	68.97%	69.24%
8	77.59%	80.56%	77.59%	78.24%
9	89.66%	89.82%	89.66%	89.68%
10	82.76%	83.06%	82.76%	82.79%
11	82.76%	86.96%	82.76%	83.50%

TABEL XXIII
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	66.10%	65.41%	66.10%	65.19%
2	64.41%	70.57%	64.41%	64.10%
3	74.58%	76.22%	74.58%	73.12%
4	74.58%	74.09%	74.58%	73.17%
5	62.71%	64.44%	62.71%	60.47%
6	81.36%	84.50%	81.36%	79.57%
7	74.14%	77.84%	74.14%	73.46%
8	70.69%	71.42%	70.69%	70.88%
9	82.76%	83.27%	82.76%	82.04%
10	70.69%	73.53%	70.69%	70.10%
11	79.31%	80.70%	79.31%	77.38%

10. Rasio 50:50

Pada proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* dan *K-Nearest Neighbors* dengan rasio 50:50 menghasilkan rata-rata akurasi 77,66% dan 71,67%. Sedangkan untuk akurasi tertinggi *K-Fold Cross Validation Naive Bayes* dan *K-Nearest Neigbors* yaitu 88,68% dan 79,63%. Tabel XXIV dan Tabel XXV merupakan hasil pengujian *Naïve Bayes* dan *K-Nearest Neighbors* dengan K-Fold Cross Validation.

TABEL XXIV
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	74.07%	74.10%	74.07%	72.92%
2	70.37%	70.70%	70.37%	70.50%
3	81.48%	82.34%	81.48%	81.63%
4	77.36%	79.42%	77.36%	77.82%
5	66.04%	73.06%	66.04%	65.95%
6	83.02%	83.83%	83.02%	83.26%
7	71.70%	73.66%	71.70%	71.99%
8	83.02%	85.46%	83.02%	83.07%
9	88.68%	88.51%	88.68%	88.46%
10	77.36%	77.67%	77.36%	77.32%
11	81.13%	86.07%	81.13%	82.65%

TABEL XXV
HASIL PENGUJIAN *K-NEAREST NEIGHBORS* DENGAN *K-FOLD CROSS VALIDATION*

Nilai K	Akurasi	Presisi	Recall	F1-Score
1	68.52%	74.78%	68.52%	68.73%
2	68.52%	68.37%	68.52%	67.25%
3	79.63%	81.57%	79.63%	78.64%
4	58.49%	59.23%	58.49%	57.03%
5	73.58%	76.92%	73.58%	72.20%
6	73.58%	73.85%	73.58%	72.02%
7	73.58%	74.73%	73.58%	73.32%
8	71.70%	71.52%	71.70%	71.57%
9	73.58%	73.15%	73.58%	72.60%
10	67.92%	70.65%	67.92%	67.74%
11	79.25%	81.13%	79.25%	76.71%

Dengan berbagai rasio dengan *split data*, proses *K-Fold Cross Validation* pada klasifikasi *Naive Bayes* memiliki nilai terbaik pada rasio 80:20. Dengan rasio 80:20 *Naive Bayes* memiliki rata-rata akurasi *K-Fold Cross Validation* dan akurasi terbaik. Pada Tabel XXVI menunjukkan pemaduan dari rata-rata akurasi *K-Fold Cross Validation* dan *split data*.

TABEL XXVI
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Rasio	Rata- Rata K-fold				Akurasi Naive Bayes
	Akurasi	presisi	Recall	f1-Score	
95:5	82.75%	82.84%	82.75%	82.43%	76.27%
90:10	83.5%	83.53%	83.5%	83.24%	73.73%
85:15	82.54%	83.7%	83.54%	83.2%	79.55%
80:20	83.45%	83.97%	83.45%	83.2%	81.7%
75:25	82.71%	83.25%	82.71%	82.52%	82.25%
70:30	81.24%	81.86%	81.24%	81.04%	81.82%
65:35	79.77%	80.6%	79.77%	79.68%	80.78%
60:40	79.38%	80.2%	79.38%	79.3%	79.96%
55:45	78.28%	79.53%	78.28%	78.37%	78.41%
50:50	77.66%	79.53%	77.66%	77.78%	77.99%

Hasil *K-Fold Cross Validation* dengan metode *Naïve Bayes* memiliki nilai tertinggi mengggunakan rasio 80:20. Dengan rata-rata akurasi sebesar 83.45%, presisi senilai 83,97%, recall 83,45% dan f1-score 83,2%. Sedangkan akurasi pada data uji memiliki 81.7%.

Dengan berbagai rasio, proses *K-Fold Cross Validation* pada klasifikasi *K-Nearest Neighbors* memiliki nilai terbaik pada rasio 80:20. Dengan rasio 80:20 *K-Nearest Neighbors* memiliki rata-rata akurasi *K-Fold Cross Validation* dan akurasi terbaik. Pada Tabel XXVII menunjukkan rata-rata akurasi *K-Fold Cross Validation* dan *split data.*

TABEL XXVII
HASIL PENGUJIAN *NAÏVE BAYES* DENGAN *K-FOLD CROSS VALIDATION*

Rasio	Rata- Rata K-fold				Akurasi K-Nearest Neighbors
	Akurasi	presisi	recall	f1-Score	
95:5	77.98%	78.34%	77.98%	77.55%	66.1%
90:10	78.18%	78.71%	78.18%	77.81%	74.58%
85:15	77.81%	78.18%	77.81%	77.23%	80.11%
80:20	78.97%	79.47%	78.97%	78.46%	80.43%
75:25	77.14%	77.59%	77.14%	76.72%	79.52%
70:30	75.75%	75.98%	75.75%	75.26%	79.26%
65:35	75.57%	76.94%	75.57%	74.74%	78.83%
60:40	74.55%	75.75%	74.55%	74.04%	78.68%
55:45	72.85%	74.73%	72.85%	71.77%	74.81%
50:50	71.67%	73.26%	71.67%	70.71%	73.55%

Hasil *K-Fold Cross Validation* dengan metode *K-Nearest Neighbors* memiliki nilai tertinggi mengggunakan rasio 80:20. Dengan rata-rata akurasi sebesar 78.97%, presisi senilai 79.47%, recall 78,97% dan f1-score 78,46%. Sedangkan akurasi pada data uji memiliki 80.43%.

IV. *KESIMPULAN*

Pada penelitian ini penulis dapat membangun aplikasi desktop untuk analisis sentimen Seperti yang sudah dirumuskan pada latar belakang, berikut ini beberapa kesimpulan antara lain:

1. Pada bab keempat penelitian ini, telah berhasil mengimplementasikan kedua metode yaitu *Naive Bayes Classifier* dan *K-Nearest Neighbors* untuk melakukan klasifikasi sentimen pada teks yang terkait dengan Chatgpt.
2. Analisis sentimen masyarakat terhadap Chatgpt pada platform Twitter dilakukan dengan membandingkan dua metode klasifikasi, yaitu *Naive Bayes Classifier* dan *K-Nearest Neighbors* (KNN). Penggunaan dataset terdiri dari total 1229 data tweet yang terdiri dari 629 label positif, 300 label negatif, dan 300 label netral. Hasil eksperimen menunjukkan bahwa saat nilai k tetangga = 5 pada KNN, metrik evaluasi yang dihasilkan adalah akurasi sebesar 80,4%, presisi sebesar 80,7%, dan recall sebesar 80,4%. Di sisi lain *Naïve Bayes* menghasilkan nilai akurasi sebesar 81,7%, presisi sebesar 81,9%, dan recall sebesar 81,7%. Dalam konteks analisis sentimen terhadap Chatgpt di Twitter, *Naive Bayes Classifier* mampu mencapai tingkat akurasi yang lebih tinggi dibandingkan dengan *K-Nearest Neighbors* pada eksperimen ini.

V. SARAN

Pada penelitian ini, berdasarkan hasil penelitian didapatkan beberapa saran untuk pengembangan penelitian ini agar lebih baik untuk kedepannya. Berikut saran yang diberikan:

1. Pembuatan aplikasi desktop yang memiliki desain user interface yang lebih menarik dan mudah dalam penggunaannya.
2. Disarankan untuk mengumpulkan lebih banyak data yang berkaitan dengan analisis sentimen. Hal ini akan membantu dalam meningkatkan variasi data dan menjaga keseimbangan antara jumlah data pada setiap kelasnya.

REFERENSI

[1] Susnjak, Teo. "ChatGPT: The end of online exam integrity?." *arXiv preprint arXiv:2212.09292* (2022)..

[2] Wongkar, Meylan, and Apriandy Angdresey. "Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter." *2019 Fourth International Conference on Informatics and Computing (ICIC)*. IEEE, 2019.

[3] Prananda, Alifia Revan, and Irfandy Thalib. "Sentiment analysis for customer review: Case study of GO-JEK expansion." *Journal of Information Systems Engineering and Business Intelligence* 6.1 (2020): 1.

[4] Hasri, Cholid Fadilah, and Debby Alita. "Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter." Jurnal Informatika dan Rekayasa Perangkat Lunak 3.2 (2022): 145-160.

[5] Prasetya, Wirawan Dwi, and Bambang Sujatmiko. "Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes." Journal of Informatics and Computer Science (JINACS) 3.04 (2022): 515-525.

[6] Hasan, Fuad Nur, and Mochamad Wahyudi. "Analisis sentimen artikel berita tokoh sepak bola dunia menggunakan algoritma support vector machine dan naive bayes berbasis particle swarm optimization." *Akrab Juara: Jurnal Ilmu-ilmu Sosial* 3.4 (2018): 42-55.

[7] Agustina, Dyah Auliya, Sri Subanti, and Etik Zukhronah. "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine." Indonesian Journal of Applied Statistics 3.2 (2021): 109-122.

[8] Ardhiansya, Hikari. "Analisis Sentimen Pendapat Masyarakat Terhadap PPKM DKI Jakarta Dengan Metode Naïve Bayes." *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)* 10.2 (2023): 60-70.

[9] Putra, Ricky Eka, Handayani Tjandrasa, and Nanik Suciati. "Severity Classification of Non-Proliferative Diabetic Retinopathy Using Convolutional Support Vector Machine." International Journal of Intelligent Engineering & Systems 13.4 (2020).

[10] Dyantono, Aganda Maulan Dan, and Ricky Eka Putra. "Perbandingan Sent2vec TF-IDF Logistic Regression dan Word2vec CNN pada hasil Sentiment Analysis Youtube Comment." *Journal of Informatics and Computer Science (JINACS)* 5.01 (2023): 63-72.

[11] Farras, Muhammad, Viny Christanti Mawardi, and Tri Sutrisno. "Aplikasi Analisis Sentimen Komentar Pengguna Genshin Impact Di Play Store." *Jurnal Ilmu Komputer dan Sistem Informasi* 11.2 (2023).