3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

PENERAPAN ALGORITMA *K-NEAREST NEIGHBORS* UNTUK MENGKLASIFIKASI SENTIEMEN MASYARAKAT TERHADAP KEBERADAAN *CHAT* GPT

Ari Ahmad Sobari[1*], Mohammad Syafrullah[2]

[1,2]Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Kota Jakarta, Indonesia
Email: [1]*ahmadsobari887@gmail.com, [2]mohammad.syafrulah@budiluhur.ac.id
(* : corresponding author)

Abstrak-Teknologi *Chat* GPT (*Generative Pre-trained Transformer*) dikembangkan oleh *OpenAI* dan menggunakan algoritma *deep learning* dalam pemrosesan bahasa alami (NLP). *Chat* GPT mampu berinteraksi dengan mesin, memahami konteks percakapan, dan menghasilkan teks bermakna seperti ucapan manusia. Namun, kehadiran *Chat* GPT juga menimbulkan kekhawatiran bahwa pekerjaan manusia dapat digantikan oleh aplikasi ini. Dalam penelitian ini, dilakukan analisis sentimen masyarakat terhadap adanya *Chat* GPT dengan menggunakan data komentar pada konten *Youtube* yang membahas teknologi tersebut. Metode yang digunakan adalah algoritma KNN (*K-Nearest Neighbors*) dalam analisis sentimen. KNN adalah algoritma *instance-based learning* yang telah banyak digunakan. Pada pengujian dilakukan dengan menggunakan confusion *matrix* dengan nilai k = 19. Hasil pengujian menunjukkan *accuracy* sebesar 62%, *precision* sebesar 60%, *recall* sebesar 98% dan f1 *score* sebesar 75%. Berdasarkan evaluasi sistem analisis sentimen, dapat disimpulkan bahwa masyarakat memiliki sentimen positif terhadap adanya *Chat* GPT yaitu bernilai positif, dengan jumlah nilai positif yaitu 82 dan nilai negatif yaitu 5, dari total data uji 87. Sistem dapat melakukan analisis sentimen dengan baik dan mampu mengenali sentimen positif dan negatif. Untuk penelitian selanjutnya, disarankan untuk menggunakan *dataset* yang lebih besar guna menghasilkan hasil yang lebih akurat. Selain itu, penggunaan metode ekstraksi fitur lainnya seperti *tf-idf*, *BoW*, *Word2Vec*, *N-Gram*, dan sebagainya. juga dapat dieksplorasi. Optimasi dalam proses *Preprocessing* juga perlu ditingkatkan untuk meningkatkan akurasi hasil analisis sentimen.

Kata Kunci: *Chat* GPT, *K-Nearest Neighbors*, *Text Mining*, Opini Masyarakat

THE IMPLEMENTATION OF THE K-NEAREST NEIGHBORS ALGORITHM FOR CLASSIFYING THE SENTIMENT OF THE COMMUNITY TOWARDS THE PRESENCE OF CHAT GPT

Abstract-*The Chat GPT (Generative Pre-trained Transformer) technology was developed by OpenAI and utilizes deep learning algorithms in natural language processing (NLP). Chat GPT is capable of interacting with machines, understanding the context of conversations, and generating meaningful text akin to human speech. However, the presence of Chat GPT also raises concerns about the potential replacement of human jobs by this application. In this research, a sentiment analysis of the public's views on Chat GPT is conducted using comment data from Youtube content discussing the technology. The method employed is the K-Nearest Neighbors (KNN) algorithm in sentiment analysis, which is an instance-based learning algorithm that has been widely used. The testing is done using a confusion matrix with a k-value of 19. The results of the testing show an accuracy of 62%, precision of 60%, recall of 98%, and an F1 score of 75%. Based on the evaluation of the sentiment analysis system, it can be concluded that the public has a positive sentiment towards Chat GPT, with a positive sentiment count of 82 and a negative sentiment count of 5, out of a total of 87 test data. The system can perform sentiment analysis well and is capable of recognizing both positive and negative sentiments. For future research, it is suggested to use a larger dataset to achieve more accurate results. Additionally, the use of other feature extraction methods such as tf-idf, BoW, Word2Vec, N-Gram, and others can also be explored. Optimization in the Preprocessing process is also needed to improve the accuracy of the sentimentanalysis results.*

Keywords*: Chat GPT, K-Nearest Neighbors, Text Mining, Public Opinion*

1. PENDAHULUAN

Chat GPT akhir-akhir ini ramai sekali diperbincangkan oleh masyarakat, terutama pada kalangan programmer, adanya *chat* GPT ini para pekerja terancam digantikan oleh aplikasi tersebut. Menurut Dekan FMIPA UGM Kuwat Triyono, cepat atau lambat semua pekerjaan manusia bisa digantikan oleh AI, sampai saat ini AI telah banyak digunakan, terutama dalam pekerjaan yang berulang [1]. Sosial Media merupakan salah satu anak dari dunia maya yang kini menjadi sebuah tren yang begitu kuat pengaruhnya terhadap perkembangan pemikiran manusia [2]. Salah satu media sosial yang sering digunakan adalah *Youtube*, dengan menggunakan data komentar pada konten *Youtube* yang membahas teknologi *chat* GPT atau video yang mendemonstrasikan kemampuan *chat* GPT untuk menghasilkan teks, analisis sentimen tentang adanya *chat* GPT ini dapat dilakukan

untuk mengetahui bagaimana sentimen masyarakat. Dalam konteks analisis sentimen, Algoritma KNN (*K-Nearest Neighbors*) merupakan salah satu algoritma yang sudah banyak digunakan. KNN (*K-Nearest Neighbors*) ini termasuk ke dalam kategori *instance-based learning.* Metode KNN (*K-Nearest Neighbors*) adalah teknik *lazy learning* [3]. Dalam analisis sentimen, metode pembanding digunakan untuk membandingkan pendapat atau sentimen yang berkaitan dengan suatu hal atau masalah tertentu. Metode ini melibatkan perbandingan pendapat atau sentimen seseorang dengan nilai tertentu, seperti positif, negatif, atau tinggi rendah. Seperti yang dilakukan oleh [4] penelitian ini dilakukan untuk menganalisis sentimen *Stay Home*. Algoritma *Naive Bayes*, *support vector machine* dan *K-nearest neighbors* digunakan dalam penelitian ini. Pada proses pelabelan *Vader* dilakukan proses untuk mengolah data sehingga setiap tweet mendapat tag sesuai nilainya. Pada penelitian ini ditambahkan *operator SMOTE Upsampling Toolbox* pada setiap komputer hasil akurasi algoritma, sebuah proses yang digunakan agar data tweet menjadi lebih seimbang antara tweet positif dan negatif. Selain itu penelitian yang dilakukan M Syafruddin [5] Tujuan survei ini adalah menganalisis opini masyarakat terhadap Covid-19. Algoritma *Naive Bayes* dan *K-Nearest Neighbor* (KNN) digunakan dalam penelitian ini. Khusus digunakan untuk kesulitan TIF-IDF dan GINI INDEX. Ini adalah algoritma yang biasa digunakan untuk menghitung bobot setiap kata dan terdiri dari model klasifikasi *Naive Bayes* dan KNN atau model klasifikasi data data mining menggunakan *RapidMiner*. Selain itu, pengujian data dilakukan menggunakan *RapidMiner* untuk menunjukkan secara jelas nilai akurasi, presisi, recall, dan AUC dari setiap model klasifikasi. Model klasifikasi KNN kemudian dibandingkan dengan model klasifikasi *Naive Bayes*. Model-model ini mengklasifikasikan data tweet COVID-19 dengan paling akurat. Oleh karena itu penulis analisis sentimen untuk mengklasifikasi, pendapat masyarakat atau opini publik keberadaan *Chat* GPT dengan menggunakan algoritma *K-Nearest Neighbors*. pada saluran *YouTube* Web Programming UNPAS dan Dea Afrizal. Serta pengujian menggunakan *Confusion Matrix* dan ekstraksi fitur *CountVectorizer*. Analisis sentimen dengan mengumpulkan komentar video dan menggunakan algoritma KNN (*K-Nearest Neighbors*) untuk menentukan seberapa positif atau negatif sentimen masyarakat terhadap adanya *chat* GPT. Analisis opini *chat* GPT dapat memberikan informasi yang berguna untuk pengembangan teknologi dan regulasi yang lebih baik di masa yang akan mendatang. Selain itu, analisis sentimen juga dapat membantu untuk memahami pendapat orang tentang teknologi ini dan memberikan pedoman penyebaran informasi yang efektif melalui teknologi *chat* GPT.

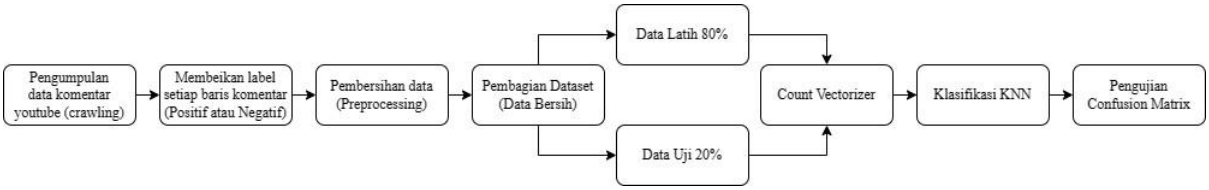
2. METODE PENELITIAN

2.1 Data Penelitian

Data yang digunakan dalam penelitian ini referensi dari data komentar pada 2 konten *youtube* yang membahas tentang *chat* GPT. Data yang digunakan adalah dari *Channel Youtube Web Programming* UNPAS dan Dea Afrizal dengan.Data yang digunakan yaitu sebanyak 500, namun data setelah melalui tahapan *Preprocessing* menjadi 474 data, selanjutnya data akan dibagi 2 dengan rasio 80:20, dengan pembagian 80% data latih dan 20% data uji, dengan hasil 394 data latih dan 87 data uji.

2.2 Penerapan Metode

Adapun beberapa tahapan dari penelitian ini untuk merancang web atau program analisis sentimen. Pada tahapan ini mewakili setiap rancangan dan proses penelitian, berikut penerapan metode dapat dilihat pada gambar 1.



Gambar 1. Penerapan Metode

Pada Gambar 1 merupakan penerapan metode, pada tahapan pertama dilakukan crawling atau mengumpulkan data dari komentar konten *youtube* yang digunakan sebagai *dataset*, kemudian *dataset* disimpan ke dalam *Database* pada table *dataset*,. Selanjutnya *dataset* di *export* kedalam bentuk *excel* dan dilakukan proses *Label* manual dengan 2 *Label* yaitu positif dan negatif. Selanjutnya tahapan *Preprocessing* yaitu tahap pembersihan data meliputi Cleansing, Case Folding, Tokenizing, Slangword, menghapus *Stopword* dan *Stemming* yang akan menghasilkan data komentar bersih. Kemudian data komentar bersih ini dibagi menjadi 2 yaitu dengan perbandingan data uji sebesar 20% dan data latih sebesar 80%, kemudian data latih diproses pada tahapan ekstraksi fitur *CountVector*, tujuannya untuk menghitung jumlah kemunculan setiap kata pada dokumen. Selanjutnya hasil dari *CountVector* diproses pada tahapan metode *K-Nearest Neighbors* yaitu dengan menghitung jarak yang paling dekat dengan objek, rumus yang digunakan pada algoritma *K-Nearest Neighbors* yaitu rumus *Euclidean Distance*, dan akan memperoleh hasil model klasifikasi.

2.3 Pengumpulan Data

Pengumpulan data yang digunakan pada tahapan ini adalah data komentar dari konten *youtube* dengan memanfaatkan *API key* yang didapatkan dari *API google console*. *API key* dimplementasikan pada program yang akan disimpan pada *Database*. Data yang digunakan adalah dari *Channel Youtube Web Programming* UNPAS dengan *video id* (gDkrQp-zQCU) dan Dea Afrizal dengan *video id* (JyOvLyasDlE).

2.4 Manual *Labeling*

Pada tahapan pelabelan data komentar ini peneliti menentukan nilai dari data komentar yang bernilai 0 (nol) untuk *Label* negatif dan 1 (satu) untuk *Label* positif, pada tahapan *manual Labeling* pada *dataset* 2 orang ahli bahasa Indonesia dan 1 orang ahli bahasa Inggris memberi label. Dalam proses ini, responden diminta untuk menentukan apakah data komentar YouTube positif atau negatif. Saat ketiga responden diminta memberi label, penulis menggunakan sentiment paling banyak untuk menentukan hasil pelabelan.

2.5 *Preprocessing*

Preprocessing adalah tahapan awal pada *text mining* dan *data mining*, tujuannya yaitu membersihkan data yang tidak diperlukan, agar pada saat proses pengolahan data menjadi lebih mudah, pada proses ini data disesuaikan dengan isi data sesuai hasil dari tahapan *crawling* data komentar. Adapun tahapan dari *Preprocessing* yaitu *Case Folding* adalah perubahan keseluruhan data huruf kapital (*uppercase*) dalam dokumen menjadi huruf kecil (*lowercase*). *Cleansing* merupakan tahapan untuk membersihkan komponen yang tidak digunakan dalam data seperti tanda baca dan simbol [6]. Cara kerjanya mengurangi *noise* pada *dataset*. Contoh karakter yang dihapus seperti tanda baca seperti koma (,), titik (.) atau tanda baca yang lainnya [7]. *Tokenizing* merupakan proses pemecahan data, kalimat awal akan dipecah menjadi kalimat-kalimat atau pemutusan urutan *string* menjadi beberapa bagian, seperti kata-kata berdasarkan tiap kata penyusunnya. *Slangword* merupakan tahapan untuk mengidentifikasi dan mengganti kata-kata *slang* atau bahasa gaul dengan kata-kata yang lebih umum atau formal. Tujuan dari tahapan *Slangword* adalah untuk meningkatkan pemahaman dan analisis teks yang dilakukan selanjutnya. *Stopword Removal* adalah proses pembersihan data dokumen dari kata yang tidak penting, misalnya kata imbuhan yang tidak memiliki kata seperti ‘kami’, ‘aku’, ‘kalau’ dan lain-lain. Berikut contoh penerapan pada proses *Stopword Removal*. *Stemming* merupakan proses untuk menghilangkan imbuhan sehingga merubah kata-kata dalam kalimat menjadi kata dasar.

2.6 Pembagian Data

Setelah data melalui tahapan *manual Labeling* dan tahapan *preprocessing*, selanjutnya data yang sudah bersih akan dibagi menjadi 2 yaitu dengan perbandingan data uji sebesar 20% dan data latih sebesar 80%, pembagian data ini dilakukan secara acak. Tujuannya yaitu untuk membedakan data yang akan diuji pada proses pemodelan klasifikasi KNN dan pengujian *confusion matrix*.

2.7 *Count Vectorizer*

Count Vectorizer mengubah *Bug Of Word* menjadi vektor. Pisahkan kata-kata dalam dokumen menjadi kata-kata individual yang menyusunnya dan hitung seberapa sering setiap kata muncul di setiap dokumen. Setiap dokumen diwakili oleh vektor yang ukurannya sama dengan jumlah kata dalam kamus, dan entri dalam vektor untuk dokumen tertentu mewakili jumlah kata dalam dokumen tersebut [8]. Berikut adalah contoh perhitungan *countvectorizer* pada 2 contoh data.

Dokumen 1

ai cerdas buat anda ilmu tahu manusia program ai tahu ai kembang pengetahuanya dapat ai robot mampu ai kuasa daya manusia

Dokumen 2

ai guna tangan dimanfatkan ai bahaya masyarakat doktrin film tema ai lawan manusia teknologi ai tolak aku masyarakat sehinga tangan nya maksimal bahaya terkadang bagus untung ama

Tabel 1. Hasil Count *Vectorizer*

Term	Count Dokument 1	Count Dokument 2
ai	3	5
cerdas	1	0
buat	1	0
anda	1	0
ilmu	1	0
tahu	1	0
manusia	1	2
program	1	0
kembang	1	0
pengetahuanya	1	0
dapat	1	0
robot	1	0
mampu	1	0
kuasa	1	0
daya	1	0
guna	0	1
tangan	0	2
dimanfaatkan	0	1
bahaya	0	2
masyarakat	0	1
doktrin	0	1
film	0	1
tema	0	1
lawan	0	1
teknologi	0	1
tolak	0	1
aku	0	1
sehingga	0	1
maksimal	0	1
terkadang	0	1
bagus	0	1
untung	0	1
aman	0	1

2.8 *K-Nearet Neighbors*

Pada tahap K-Nearest Neighbors ini data latih dan data uji yang telah melalui ekstraksi fitur menggunakan *Count Vectorizer*. dihitung jarak *Euclidean* antara *vector* data uji dengan setiap *vector* data latih. Kemudian, jarak-jarak tersebut diurutkan secara berurutan ascending. Selanjutnya, hanya k tetangga terdekat yang diambil dari jarak-jarak tersebut. Setelah itu, *Label* dari k tetangga terdekat diambil dan disimpan dalam *array Labels*. Kemudian dihitung *Label* yang paling sering muncul dalam tetangga terdekat (prediksi kelas) dan akan menghasilkan prediksi klasifikasi.

K-Nearest Neighbors merupakan algoritma klasifikasi data yang didasarkan pada data *training* dengan jarak objek yang paling dekat. Perhitungan algoritma KNN menggunakan rumus *Euclidean Distance* [9]. Adapun Teorema *K-Nearest Neighbors* akan dijelaskan pada persamaan 1 :

$$d(x,y) = \sqrt{\left(\sum_{i=1}^{n} (xi - yi)^2\right)} \quad (1)$$

Keterangan:

d : Jarak
x : Data uji
y : Data Latih
i : Variabel data
n : Dimensi data

2.9 *Confusion Matrix*

Confusion Matrix adalah teknik untuk mengevaluasi klasifikasi menilai suatu pernyataan sebagai benar atau salah. Matriks prediksi dibandingkan dengan kelas asli yang berisi data aktual dan nilai skor prediksi [10]. *Confusion Matrix* bertujuan untuk menghasilkan nilai seperti *accuracy*, *precision*, *recall* dan *f1 score*.

a. Accuracy

Accuracy adalah total dokumen yang diklasifikasikan dengan benar, baik benar positif maupun benar negatif [5]. Berikut rumus untuk menghitung *accuracy* akan dijelaskan pada persamaan 2 :

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} * 100 \quad (2)$$

b. Precision

Precision adalah seberapa banyak hasil pemrosesan relevan dengan informasi yang dicari. Dengan kata lain, *precision* adalah klasifikasi *True Positive*, dan semua data diproyeksikan sebagai kelas positif [5]. Berikut adalah rumus untuk menghitung *precision* akan dijelaskan pada persamaan 3 :

$$Precision = \frac{TP}{TP+FP} * 100 \quad (3)$$

c. Recall

Recall adalah jumlah dokumen relevan yang dikumpulkan oleh sistem. Dengan kata lain, *recall* adalah jumlah dokumen positif benar dari seluruh dokumen positif benar (termasuk *False Negatif*) [5]. Berikut adalah rumus untuk menghitung *precision* akan dijelaskan pada persamaan 4 :

$$Recall = \frac{TP}{TP+FN} * 100 \quad (4)$$

d. F1-Score

F-measure atau *F1-Score* merupakan harmonic mean dari *precision* dan *recall* [11]. Berikut rumus untuk menghitung *precision* akan dijelaskan pada persamaan 5 :

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} * 100 \quad (5)$$

3. HASIL DAN PEMBAHASAN

500 data digunakan berdasarkan data yang lolos tahap *preprocessing*, yaitu. tidak kurang dari 474 data *preprocessing*. Selain itu, data bersih dibagi menjadi dua bagian yaitu 20% data uji dan 80% data latih untuk pembanding, sehingga menghasilkan 394 baris data latih dan 87 baris data uji.

3.1 Tahap Pengumpulan Data

Pengumpulan data yang digunakan pada tahapan ini adalah data komentar dari konten *youtube* dengan memanfaatkan *API key* yang didapatkan dari *API google console*. *API key* dimplementasikan pada program dengan mencantumkan *video_id youtube* dengan jumlah 600 data komentar yang akan disimpan pada *Database*, pada tahapan ini jumlah data yang didapatkan yaitu sebesar 500 data komentar

3.2 Tahap *Manual Labeling*

Hasil yang didapatkan dari *manual Labeling* ini didapatkan dengan keseluruhan komentar yaitu 500 baris

data, dengan jumlah komentar bersentimen Negatif yaitu 169 dan 331 komentar bersentimen Positif, selanjutnya *dataset* akan melalui tahapan *Preprocessing*.

3.3 Tahap *Preprocessing*

Preprocessing adalah tahapan awal pada *text mining* dan *data mining*, tujuannya yaitu membersihkan data yang tidak diperlukan, agar pada saat proses pengolahan data menjadi lebih mudah, pada proses ini data disesuaikan dengan isi data sesuai hasil dari tahapan *crawling* data komentar.

Tabel 2. Sebelum dan Sesudah Tahap *Preprocessing*

Sebelum	Sesudah
DRONE + AI + ROBOT SENJATA YG MENAKUTKAN????????????	drone ai robot senjata takut

3.4 Tahap *Count Vectorizer*

Pada tahapan ini langkah pertama adalah pengumpulan korpus yang merupakan seluruh kata unik dari semua komentar yang telah dibersihkan. Langkah ini dilakukan dengan memisahkan kata-kata dalam setiap komentar, lalu menambahkan kata-kata yang belum ada dalam korpus. Kemudian, dilakukan perhitungan jumlah kemunculan setiap kata dalam komentar dan pembuatan vektor berdasarkan hasil perhitungan tersebut. Kata-kata pada komentar dipecah menjadi *array* kata-kata, dan setiap kata pada korpus diperiksa apakah ada dalam komentar. Jika ada, jumlah kemunculan kata tersebut dihitung dan ditambahkan ke dalam vektor. Berikut adalah contoh sampel data latih dan data uji pada tahapan ekstraksi fitur *Count Vectorizer* dapat dilihat pada tabel 3, 4 dan hasil *count vectorizer* pada tabel 5 menggunakan persamaan 1.

Tabel 3. Sampel Data Latih

Komentar	***Label***
ofline bot nya	0
ai program ai langsung ai ulang kali	1
keren sayang salah chat gpt nyoba cari program arduino eror library library disebutin	1
cari ganti bahasa ah pakai bahasa indonesia pakai bahasa ingris	1
besok toto gelap hongkong angka coba gpt	0

Tabel 4. Sampel Data Uji

Komentar
takut salah retas data data rahasia pribadi
video rusak sinkron suara video

Tabel 5. Hasil *Count Vectorizer*

	Data Latih							**Data Uji**	
No	**Term**	**y1**	**y2**	**y3**	**y4**	**y5**		**x1**	**x2**
1	ofline	1	0	0	0	0		0	0
2	bot	1	0	0	0	0		0	0
3	nya	1	0	0	0	0		0	0
4	ai	0	3	0	0	0		0	0
5	program	0	1	1	0	0		0	0
6	langsung	0	1	0	0	0		0	0
7	ulang	0	1	0	0	0		0	0
8	kali	0	1	0	0	0		0	0
9	keren	0	0	1	0	0		0	0
10	sayang	0	0	1	0	0		0	0
11	salah	0	0	1	0	0		1	0
12	chat	0	0	1	0	0		0	0
13	gpt	0	0	1	0	1		0	0
14	nyoba	0	0	1	0	0		0	0
15	cari	0	0	1	1	0		0	0

16	arduino	0	0	1	0	0		0	0
17	eror	0	0	1	0	0		0	0
18	library	0	0	2	0	0		0	0
19	disebutin	0	0	1	0	0		0	0
20	ganti	0	0	0	1	0		0	0
21	bahasa	0	0	0	3	0		0	0
22	ah	0	0	0	1	0		0	0
23	pakai	0	0	0	2	0		0	0
24	indonesia	0	0	0	1	0		0	0
25	ingris	0	0	0	1	0		0	0
26	besok	0	0	0	0	1		0	0
27	toto	0	0	0	0	1		0	0
28	gelap	0	0	0	0	1		0	0
29	hongkong	0	0	0	0	1		0	0
30	angka	0	0	0	0	1		0	0
31	coba	0	0	0	0	1		0	0
32	takut	0	0	0	0	0		1	0
33	retas	0	0	0	0	0		1	0
34	data	0	0	0	0	0		2	0
35	rahasia	0	0	0	0	0		1	0
36	pribadi	0	0	0	0	0		1	0
37	video	0	0	0	0	0		0	2
38	rusak	0	0	0	0	0		0	1
39	sinkron	0	0	0	0	0		0	1
40	suara	0	0	0	0	0		0	0

3.5 Tahap *K-Nearest Neighbors*

Tahapan selanjutnya setelah tahapan implementasi metode *Count Vectorizer* yaitu tahapan implementasi metode *K-Nearest Neighbors*. Pada tahapan ini hasil *Count Vectorizer* data uji dan data latih akan dihitung jarak antara *vector* data latih dan vektor data uji. Berdasarkan pada sub bab (2.8), proses perhitungan jarak melibatkan *Euclidean Distance* dengan menggunakan persamaan (1). Berikut adalah contoh penerapan *Euclidean Distance* untuk menghitung jarak pada vektor data uji (x1) dengan *vector* data latih (y1):

$$\mathrm{d(uji1,latih1)} = \sqrt{\begin{array}{l}(0-1)^2+(0-1)^2+(0-1)^2+(0-0)^2+(0-0)^2+\\(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+\\(1-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+\\(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+\\(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+\\(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+\\(0-0)^2+(1-0)^2+(1-0)^2+(2-0)^2+(1-0)^2+\\(1-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2\end{array}}$$

$$\mathrm{d(uji1,latih1)} = \sqrt{12}$$

$$\mathrm{d(uji1,latih1)} = \sqrt{3{,}464101615}$$

Berdasarkan hasil perhitungan *Euclidean Distance*, maka hasil perhitungan setiap jarak akan diurutkan berdasarkan data terkecil ke terbesar, perhitungan *Euclidean Distance* setelah diurutkan jarak antara vektor uji dengan vektor latih dapat dilihat pada tabel 6.

Tabel 6. Hasil Jarak *Euclidean Distance*

	Data Uji 1	Label
Data Latih 1	3,464101615	Negatif

Data Latih 5	4	Negatif
Data Latih 2	4,69041576	Positif
Data Latih 3	4,69041576	Positif
Data Latih 4	5,196152423	Positif

Berdasarkan tabel 6 akan dihitung *Label* yang paling sering muncul dalam tetangga terdekat (prediksi kelas) dan akan menghasilkan prediksi klasifikasi. Dapat disimpulkan bahwa jika ditentukan dengan nilai k = 3 maka hasil prediksi klasifikasi adalah negatif, namun jika ditentukan dengan nilai k = 5 maka hasil prediksi klasifikasi adalah positif.

3.6 Pengujian

Pengujian *Confusion Matrix* dihitung berdasarkan prediksi hasil dari pemodelan data latih dan data uji. Dengan menggunakan data latih dengan total keseluruhan 394 baris data dengan aktual *Label* 274 untuk positif dan aktual *Label* 120 *Label* negatif, sedangkan data uji dengan total keseluruhan 87 baris data dengan aktual *Label* 51 untuk positif dan aktual *Label* 36 *Label* negatif. Berikut contoh tabel 7 hasil pemodelan klasifikasi KNN.

Tabel 7. Hasil Pemodelan Klasifikasi KNN

No	Komentar	Aktual Label	Predict Label
1	Bro coba minta no togel	Negatif	Negatif
2	ini yang akan membuat mahasiswa s1 lulusan 2022 ga kepake karena perusahaan hrd akan berpikir mereka membuat skripsi dengan bantuan chat gpt	Negatif	Positif
3	Bg chat gpt gk bisa bahasa indo kah di android soalnya gw pakek ko gk bisa indo yal	Positif	Positif
4	Negeri paman biden emang luar biasa mslh tehnologi.. cuma kaum yg suka sok pintar yg g percaya giliran terbukti mereka membisu	Negatif	Negatif
-	----------	----------	----------
87	shi...	Negatif	Positif

Pada tabel 7 pada kolom aktual *Label* diambil dari field id_Label dan *predict Label* diambil dari hasil klasifikasi KNN. untuk keseluruhan data uji ada 87 baris data. Lalu akan direpresentasikan pada *Confusion Matrix*. Pada pengujian pengujian *Confusion Matrix* mempresentasikan dengan nilai k adalah 19 yang dapat dilihat pada tabel 8 berikut.

Tabel 8. Pengujian *Confusion Matrix*

		Nilai Aktual	
		Positif	Negatif
Nilai Prediksi	Positif	TP 50	FP 32
	Negatif	TN 1	TN 4

Dari perolehan tabel 8 tabel pengujian *Confusion Matrix* k = 19 maka akan dikalkulasikan dengan nilai *accuracy*, *precision*, *recall* dan *f1 score*. Tabel kalkulasi nilai pengujian k = 19 dapat dilihat pada tabel 9.

Tabel 9. Kalkulasi Nilai Pengujian k = 19

Pengujian		
Accuracy	$\frac{50+4}{50+1+32+4} * 100$	0.62 (62%)

Precision	$\frac{50}{50+32} * 100$	0.60 (60%)
Recall	$\frac{50}{50+1} * 100$	0.98 (98%)
F1 Score	$2 * \frac{60.98 * 98.04}{60.98 + 98.04} * 100$	0.75 (75%)

Hasil pengujian pada tabel 9 dilakukan secara berkala, dengan menguji nilai k secara berbeda. Hasil pengujian diketahui secara keseluruhan dapat dilihat pada tabel 10.

Tabel 10. Hasil Pengujian

Nilai k	Accuracy	Precision	Recall	F1 Score
3	57%	60%	82%	69%
5	55%	58%	82%	68%
7	55%	58%	84%	68%
9	52%	56%	84%	67%
11	51%	56%	82%	66%
13	58%	60%	84%	70%
15	60%	61%	90%	73%
17	59%	60%	92%	72%
19	**62%**	**60%**	**98%**	**75%**

Berdasarkan tabel 10 diatas dapat disimpulkan bahwa nilai *accuracy* tertinggi didapatkan 62% dengan nilai k = 19 dan nilai terendah didapatkan 51% dengan nilai k = 11. Selanjutnya nilai *precision* tertinggi didapatkan 61% dengan nilai k = 15 dan nilai terendah didapatkan 56% dengan nilai k = 9 dan 11. Selanjutnya nilai *recall* tertinggi didapatkan 98% dengan nilai k = 19 dan nilai terendah didapatkan 82% dengan nilai k = 3 dan dengan nilai k = 5. Dan yang terakhir nilai *f1 score* tertinggi didapatkan 75% dengan nilai k = 19 dan nilai terendah didapatkan 66% dengan nilai k = 11.

4. KESIMPULAN

Berdasarkan hasil evaluasi sistem analisis sentimen opini masyarakat, dapat disimpulkan bahwa, hasil prediksi klasifikasi dapat disimpulkan adanya *chat* GPT yaitu bernilai positif, dengan jumlah nilai positif yaitu 82 dan nilai negatif yaitu 5, dari total data uji 87, sistem dapat bekerja dengan selayaknya, dapat menganalisa sentimen bernilai positif dan negatif, dengan nilai pengujian tertinggi menggunakan metode *confusion matrix* k = 19 yaitu dengan nilai *accuracy* 62%, *precision* 60%, *recall* 98% dan *f1 score* 75%. Adapun saran untuk penelitian kedepannya yaitu menggunakan *dataset* lebih banyak agar hasilnya lebih akurat, menggunakan ekstraksi fitur yang lainnya, misalnya *tf-idf*, *BoW*, *Word2Vec*, *N-Gram* atau lain sebagainya, mengoptimalkan pada proses Preprocessing agar hasilnya lebih akurat.

DAFTAR PUSTAKA

[1] R. Ramadhan and K. S Kurniato, “Cepat Atau Lambat, Pekerjaan Manusia akan Diganti oleh AI,” Dec. 17, 2022. https://kumparan.com/kumparantech/cepat-atau-lambat-pekerjaan-manusia-akan-diganti-oleh-ai-1zSBrRLX8WO/full (accessed May 18, 2023).

[2] N. Ainiyah, “REMAJA MILLENIAL DAN MEDIA SOSIAL: MEDIA SOSIAL SEBAGAI MEDIA INFORMASI PENDIDIKAN BAGI REMAJA MILLENIAL,” 2018.

[3] D. Cahyanti, A. Rahmayani, and S. Ainy Husniar, “Indonesian Journal of Data and Science Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara,” vol. 1, no. 2, pp. 39–43, 2020.

[4] I. Hakim, A. Nugroho, S. Hadi Sukmana, W. Gata, and S. Nusa Mandiri, “Sentimen Analisis Stay Home menggunakan metode klasifikasi Naive Bayes, Support Vector Machine, dan k-Nearest Neighbor,” vol. 22, no. 2, 2020, doi: 10.31294/p.v21i2.

[5] M. Syarifuddin, “ANALISIS SENTIMEN OPINI PUBLIK TERHADAP EFEK PSBB PADA TWITTER DENGAN ALGORITMA DECISION TREE,KNN, DAN NAÏVE BAYES,” *INTI Nusa Mandiri*, vol. 15, no. 1, pp. 87–94, Aug. 2020, doi: 10.33480/inti.v15i1.1433.

[6] A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia, “Automated Bahasa Indonesia essay evaluation with latent semantic analysis,” in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1742-6596/1235/1/012100.

[7] M. Syarifuddinn, "ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN," *INTI Nusa Mandiri*, vol. 15, no. 1, pp. 23–28, Aug. 2020, doi: 10.33480/inti.v15i1.1347.

[8] S. Khomsah and Agus Sasmito Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 648–654, Aug. 2020, doi: 10.29207/resti.v4i4.2035.

[9] T. Wiratama Putra and A. Triayudi, "Analisis Sentimen Pembelajaran Daring menggunakan Metode Naïve Bayes, KNN, dan Decision Tree," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 6, no. 1, p. 2022, 2022, doi: 10.35870/jti.

[10] M. Riefky and A. R. Anandyani, "KLASIFIKASI PERSEPSI PENGGUNA TWITTER TERHADAP TUNTUTAN KERINGANAN PEMBAYARAN UANG KULIAH TUNGGAL (UKT) PADA MASA PANDEMI COVID-19 MENGGUNAKAN K-NEAREST NEIGHBOR," *Seminar Nasional Official Statistics*, vol. 1, pp. 247–257, 2020, doi: 10.34123/semnasoffstat.v2020i1.443.

[11] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst Appl*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.