Jurnal TRANSFORMATIKA
Vol.22, No.2, Januari 2025, pp. 97 - 107
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Perbandingan Naïve Bayes dan K-NN dalam Analisis Sentimen Aplikasi X

Emerensye S.Y. Pandie[1*], Adriana Fanggidae[2], Ririn Lona[3]
[1]Universitas Nusa Cendana
Jl. Adisucipto Penfui, Kupang, NTT, (0380) 881580, e-mail: emerensyepandie@staf.undana.ac.id
[2]Universitas Nusa Cendana
Jl. Adisucipto Penfui, Kupang, NTT, (0380) 881580, e-mail: adrianafanggidae@staf.undana.ac.id
[3]Universitas Nusa Cendana
Jl. Adisucipto Penfui, Kupang, NTT, (0380) 881580, e-mail: lonaririn03@gmail.com

ARTICLE INFO

History of the article :
Received 24 November 2024
Received in revised form 27 November 2024
Accepted 9 January 2025
Available online 29 January 2025

Keywords:

X; Naïve bayes; K-Nearest Neighbor; Analisis sentimen.

*** Correspondece:**

Telepon:
+822-4708-9716
E-mail:
Emerensyepandie@staff.undana.ac.id

ABSTRACT

Aplikasi X, sebelumnya dikenal sebagai Twitter adalah media sosial yang memungkinkan pengguna mengirim, membalas, dan membaca pesan. Berdasarkan ulasan di *Google Play Store*, banyak pengguna mengeluhkan masalah, terutama terkait penangguhan akun setelah perubahan kepemilikan. Namun, sebagian pengguna masih merasa puas dan terbantu dengan X. Oleh karena itu, analisis sentimen dilakukan untuk mengetahui kecenderungan opini pengguna. Penelitian ini menggunakan metode *naïve bayes* dan *k-Nearest Neighbor* pada 8.723 ulasan yang kemudian diklasifikasi sebagai sentimen positif, netral, atau negatif menggunakan *K-fold cross validation*. Naïve Bayes mencapai akurasi tertinggi sebesar 88,87% pada 10-*fold*, sementara KNN dengan *k* optimal di 12-NN mencapai 90,32% pada 2-*fold*. Dalam perbandingan hasil klasifikasi dengan label pakar kedua, metode *Naïve Bayes* lebih sesuai dengan akurasi 92,56% dibandingkan KNN yang mencapai 91,73%.

1. INTRODUCTION

Di era digital yang berkembang pesat saat ini, media sosial menjadi pilihan bagi banyak individu untuk berbagi informasi, pendapat dan pengalaman tentang berbagai topik dengan mudah dan cepat. Media sosial sendiri merupakan sarana berbasis internet yang memungkinkan seseorang untuk berinteraksi, berkomunikasi serta bekerja sama dengan orang lain secara online[1]. Menurut *we are social* pengguna media sosial telah mencapai 4,95 miliar atau setara dengan 64,4% dari total 8,06 miliar orang di dunia [2]. Beberapa contoh media sosial yang populer saat ini diantaranya Instagram, X, Whatsapp, Tiktok, Youtube, dan lain-lain.

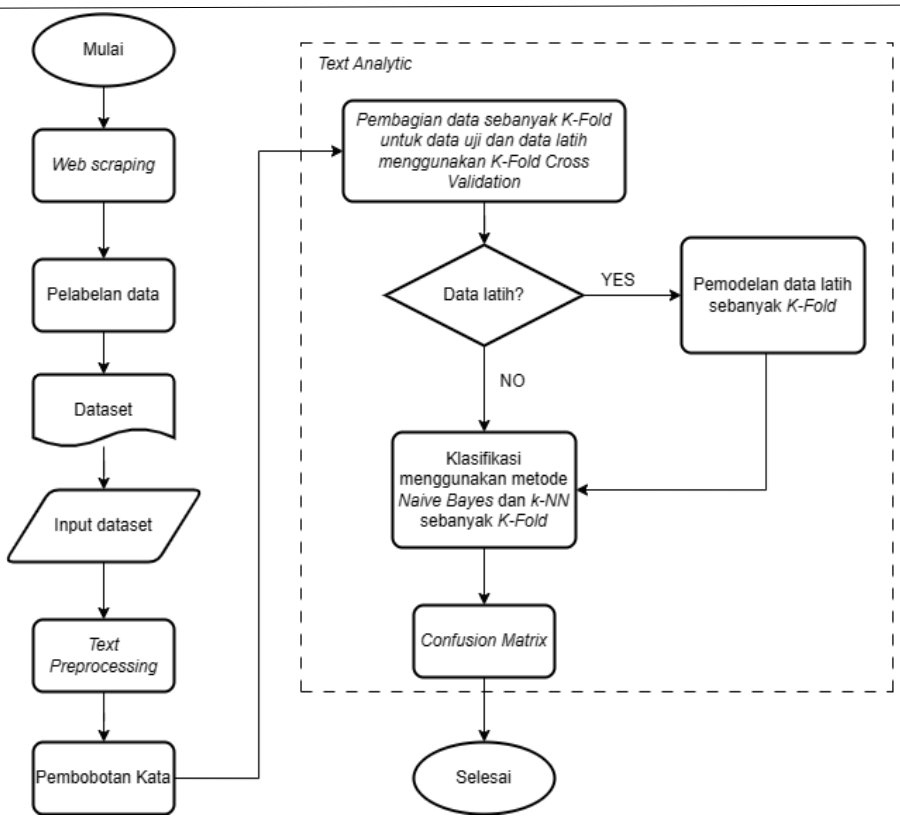
Salah satu media sosial yang banyak digunakan yaitu X atau yang sebelumnya dikenal dengan nama Twitter. X adalah media sosial yang memungkinkan pengguna untuk mengirim, membalas pesan, dan membaca postingan dari orang lain. Beranda X menggunakan algoritma kronologi waktu, sehingga informasinya bersifat terkini. Selain itu, X juga memiliki fitur trending topik yang membuat penggunanya up-to-date dengan isu yang terjadi akhir-akhir ini [3]. Menurut laporan *we are social*, pengguna X telah mencapai lebih dari 660 juta, menjadikannya sebagai media sosial ke-12 yang paling banyak digunakan [2]. Berdasarkan komentar pengguna media sosial X yang ada pada google play store, saat ini banyak keluhan yang diberikan pengguna sejak X berpindah kepemilikan, keluhan-keluhan tersebut diantaranya tentang perubahan nama dan logo, kesulitan membuat akun baru, masalah login, dan penangguhan akun. Meski demikian, masih banyak pengguna yang menyukai dan merasa terbantu dengan X. Maka dari itu perlu dilakukan analisis sentimen untuk mengetahui kecenderungan opini pengguna X agar dapat membantu calon pengguna dalam mengambil keputusan baik atau tidaknya media sosial tersebut untuk digunakan.

Analisis sentimen merupakan proses untuk mengolah data teks secara otomatis untuk mendapatkan informasi mengenai sentimen yang terkandung dalam opini-opini yang diberikan pengguna[4]. Analisis sentimen dapat dilakukan dengan menggunakan metode *naïve bayes* dan *k-Nearest Neighbor*. Metode *naïve bayes* digunakan untuk menghitung probabilitas kata dan probabilitas kelas untuk klasifikasi data dengan cara mengalikan probabilitas kata dengan probabilitas kelas, untuk kemudian mengklasifikasikan data uji berdasarkan probabilitas tertinggi[5]. Di sisi lain, metode *k-Nearest Neighbor* bekerja berdasarkan data latih yang memiliki jarak paling dekat dengan objek data uji [6]. Pada penelitian yang menggunakan 1000 data latih yang terdiri dari 500 data sentimen negatif dan 500 data sentimen positif, kemudian digunakan 1500 untuk data uji [7]. Penelitian tersebut melakukan pengujian menggunakan metode *naïve bayes* menghasilkan tingkat akurasi sebesar 97,13% dan waktu komputasi yang dihasilkan dalam proses klasifikasi selama 0,4 detik. Sedangkan, penelitian yang menggunakan 2000 data dengan label kelas anjuran, larangan dan informasi yang mana masing-masing kelas memiliki jumlah data 376, 204, dan 1977, kemudian dibagi menjadi data latih dan data uji menggunakan 5-*fold cross validation* yang mana masing-masing *fold* terdiri dari 1600 data latih dan 400 data uji [8]. Setelah proses pengujian menggunakan metode *k-Nearest Neighbor*, menghasilkan nilai akurasi sebesar 90,23% dan waktu komputasi sebesar 37 menit 06 detik. Kedua metode ini efektif dalam melakukan klasifikasi dibuktikan dengan tingkat keakuratan dalam penelitian sebelumnya, namun metode-metode ini juga memiliki kelebihan dan kelemahan yang didasarkan pada kecepatan komputasi, dimana metode *naïve bayes* dikenal memiliki kelebihan waktu komputasi yang lebih cepat [9]. dibandingkan dengan *k-Nearest Neighbor* yang cenderung membutuhkan waktu komputasi yang lebih lama[10].

Oleh karena itu, penelitian ini akan membandingkan metode *naïve bayes* dan *k-Nearest Neighbor* dalam mencari nilai akurasi tertinggi dari kedua metode tersebut serta mengetahui kecenderungan opini pengguna media sosial X.

RESEARCH METHODS

Langkah-langkah yang dilalui dalam penelitian ini sangat penting untuk memastikan penelitian ini dapat berlangsung dengan lancar. Metodologi penelitian yang dilakukan dalam penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Metodologi Penelitian

A. *Web Scraping*

Web scraping merupakan proses pengumpulan dokumen semi terstruktur secara otomatis dari internet [11]. Data dalam penelitian ini dikumpulkan dari situs *google play store* yaitu data ulasan pengguna aplikasi X.

B. Pelabelan Data

Pelabelan data merupakan tahap dimana data diberi label sesuai dengan kelasnya. Proses pelabelan data dilakukan oleh ahli bahasa dengan tujuan menghindari pendapat subjektif pada proses pelabelan. Adapun dalam penelitian ini terdapat dua pakar yang melabeli data, pakar pertama akan melabeli data asli dan pakar kedua akan melabeli data yang telah melalui tahap *text preprocessing*.

C. *Text Preprocessing*

Text preprocessing bertujuan untuk mempersiapkan data sebelum dilakukan pengolahan Data mentah yang diperoleh biasanya mengandung banyak karakter yang tidak dibutuhkan dalam proses pengolahan data misalnya tanda baca, angka, kata berimbuhan, kata-kata tidak baku dan lain sebagainya. Adapun tahapan yang perlu dilakukan dalam *text preprocessing* diantaranya:

1. *Cleaning* merupakan tahap untuk menghapus karakter-karakter yang tidak diperlukan dalam proses analisis. Misalnya tanda baca, angka, simbol dan lain-lain.
2. *Case folding* adalah tahap untuk mengkonversi seluruh data ke dalam bentuk yang sama. Biasanya *case folding* akan mengkonversi teks ke dalam bentuk *lowercase.*
3. *Tokenization* bertujuan untuk memecah setiap kalimat (data) menjadi kata-kata.

DOI : https://doi.org/ 10.26623/f4k55e04

4. Normalisasi merupakan tahap untuk mengubah kata0kata yang tidak baku menjadi kata baku.
5. *Stemming* merupakan tahap untuk mendapatkan kata dasar dari suatu kata dengan cara memisahkan kata dasar dan imbuhan dari kata tersebut, baik awalan (prefiks) maupun akhiran (sufiks).
6. *Convert negation* merupakan proses konversi kata-kata negasi. *Convert negation* dilakukan dengan cara menambahkan penanda pada kata yang mengandung negasi. *Convert negation* dilakukan karena kata negasi mempunyai pengaruh dalam mengubah nilai sentimen pada suatu ulasan.

D. Pembobotan Kata

Pembobotan kata atau *weighting word* merupakan sebuah mekanisme untuk memberikan nilai pada suatu kata berdasarkan pada frekuensi kemunculan kata tersebut dalam dokumen. TF-IDF merupakan sebuah proses transformasi data teks ke dalam data numerik dengan tujuan melakukan pembobotan pada setiap kata [6]. Pembobotan TF-IDF menggabungkan konsep TF (*term frequency*) dan IDF (*index document frequency*). Tahap awal yang perlu dilakukan dalam TF-IDF yaitu menghitung nilai *term frequency* (TF) yaitu menghitung seberapa sering sebuah kata muncul pada dokumen. Sedangkan IDF merupakan konsep yang digunakan untuk menghitung seberapa penting sebuah kata dalam keseluruhan dokumen. Pada perhitungan nilai IDF ditambahkan angka 1 dengan tujuan menghindari pembagian angka nol.

Berdasarkan perhitungan TF dan IDF maka akan diperoleh nilai TF-IDF dengan menggabungkan nilai TF dan IDF menggunakan persamaan 1.

$$TF\_IDF_{(t,d)} = \frac{T}{D} \times \ln\left(\frac{N+1}{1+df_t}\right) + 1 \quad \text{.....(1)}$$

Hasil TF-IDF kemudian akan dinormalisasi dengan menggunakan *euclidean norm* menggunakan persamaan 2.

$$TF\_IDF_{norm(t,d)} = \frac{TF\_IDF_{t,d}}{\sqrt{TF\_IDF_{1,d}^2 + TF\_IDF_{2,d}^2 \ldots\ldots + TF\_IDF_{m,d}^2}} \quad \text{.....(2)}$$

Keterangan:

d = Dokumen ke-d
t = Kata ke-t dari kata kunci
m = Kata ke-m dari dokumen d
T = Jumlah frekuensi kemunculan kata pada dokumen d
D = Jumlah total keseluruhan kata pada dokumen d
N = Jumlah total dokumen
df_t = Jumlah dokumen yang mengandung kata t

E. *K-Fold Cross Validation*

K-fold cross validation diawali dengan membagi data sejumlah *K-fold* yang diinginkan. Dalam proses *cross validation* data akan dibagi dalam *K* buah partisi dengan ukuran yang sama $Fold_1$, $Fold_2$,……, $Fold_n$ selanjutnya proses *testing* dan *training* dilakukan sebanyak *K* kali. Dalam iterasi ke-*i*, $fold_i$ akan dipartisi menjadi data testing dan sisanya menjadi data *training*[12].

F. *Naïve Bayes*

Naïve Bayes merupakan sebuah metode klasifikasi yang dapat memprediksi probabilitas keanggotaan suatu data dalam kelas tertentu berdasarkan perhitungan probabilitas. Metode ini digunakan dalam kasus *supervised learning* dimana terdapat label, kelas, atau target yang menjadi acuan[13]. Dalam klasifikasi analisis sentimen menggunakan *naïve bayes*, tahap awal yang perlu dilakukan yaitu menghitung probabilitas prior pada persamaan 3.

$$P(c_i) = \frac{N(c_i)}{l} \quad \text{.....(3)}$$

Keterangan:
$P(c_i)$ = Probabilitas prior dari kelas semtimen
c_i = kelas sentimen
$N(c_i)$ = Jumlah data pada kelas sentimen c_i
l = Jumlah data latih

Selanjutnya menghitung probabilitas fitur atau probabilitas kondisional pada kelas sentimen menggunakan persamaan 4.

$$P(f_j|c_i) = \frac{v+1}{e+w} \quad \text{.....(4)}$$

Keterangan:
$P(f_i|c_i)$ = Probabilitas fitur f_j terhadap kelas c_i
f_j = Fitur yang dihitung, dimana $f_1, f_2, \ldots\ldots f_n$ merupakan fitur dalam data
v = Jumlah keseluruhan kata yang muncul pada kelas sentimen
w = Jumlah kosakata unik pada semua data latih

Angka 1 yang ada dalam persamaan 2 merupakan teknik laplace smoothing. Teknik ini dilakukan dengan menambahkan angka 1 pada setiap fitur untuk menghindari nilai probabilitas nol. Tahap akhir dari metode *naïve bayes* merupakan perhitungan probabilitas tertinggi atau *Maximum a Pasterior* (MAP) untuk mendapatkan probabilitas sentimen pada data uji. Perhitungan probabilitas tertinggi menggunakan persamaan 5.

$$C_{MAP} = \underset{\{positif, netral, negatif\}}{argmax} P(c_i) \times \prod P(f_j|c_i) \quad \text{.....(5)}$$

Perkalian nilai bilangan desimal yang terlalu besar dapat menghasilkan terlalu banyak angka dibelakang koma (*floating point underflow*). Maka dari itu, digunakan penambahan logaritma dalam perhitungan seperti persamaan 6.

$$C_{MAP} = \underset{\{positif, netral, negatif\}}{argmax} \log P(c_i) \times \sum_{i=1}^{n} \log P(f_j|c_i) \quad \text{.....(6)}$$

Keterangan:
C_{MAP} = Kelas probabilitas tertinggi
$P(c_i)$ = Probabilitas kelas sentimen
$P(c_i|f_j)$ =Probabilitas fitur f_j terhadap c_i

G. *k-Nearest Neighbor* (k-NN)

k-Nearest Neighbor merupakan salah satu algoritma klasifikasi yang biasa disebut juga metode berbasis jarak. Cara kerja dari metode ini adalah dengan menghafal semua data latih yang tersedia selama tahap pelatihan. Kemudian pada tahap pengujian, data yang telah diklasifikasi dibandingkan dengan data latih berdasarkan ukuran jarak yang telah ditentukan. Data yang paling mirip akan disebut sebagai "tetangga terdekat", penentuan jumlah tetangga terdekat ditentukan melalui jumlah nilai *k*[6]. Adapun perhitungan jarak untuk menentukan tingkat kesamaan data menggunakan *cosine similarity*. *Cosine similarity* merupakan sebuah metode yang digunakan untuk menghitung kemiripan antara semua data latih dengan data uji. Dimana nilai yang dihasilkan dari perhitungan *cosine similarity* akan berada pada rentang 0 sampai 1[14]. *Cosine similarity* dihitung dengan menggunakan persamaan 7.

$$CosSim\ (x, y) = \frac{\sum_{q=1}^{m} x_q y_q}{\sqrt{(\sum_{q=1}^{m} x_q)^2}\sqrt{(\sum_{q=1}^{m} y_q)^2}} \quad \text{.....(7)}$$

Keterangan:
x = Data uji
y = Data latih
x_q = Bobot kata yang terdapat pada data uji
y_q = Bobot kata yang terdapat pada data latih

DOI : https://doi.org/ 10.26623/f4k55e04

q = Data D_1, D_2,……,D_m

H. *Confusion Matrix*

Confusion matrix merupakan salah satu cara yang dapat digunakan untuk melihat performa dari metode yang digunakan. *Confusion matrix* digambarkan dalam bentuk tabel yang menyatakan jumlah data uji yang berhasil diklasifikasi dengan benar dan jumlah data uji yang salah diklasifikasi[15]. Tabel *confusion matrix* dapat dilihat pada tabel 1.

Tabel 1. *Confusion Matrix*

Confusion Matrix	***Predicted Positive***	***Predicted Neutral***	***Predicted Negative***
Actual Positive	TP	FNt	FN
Actual Neutral	FP	TNt	FN
Actual Negative	FP	FNt	TN

Dalam evaluasi kinerja menggunakan *confusion matrix*, terdapat istilah-istilah yang mencerminkan hasil dari proses klasifikasi. Istilah-istilah tersebut adalah *True Positive* (TP) merupakan data *positive* yang diklasifikasi dengan benar sebagai *positive*, *True Negative* (TN) merupakan data *negative* yang berhasil diklasifikasi dengan benar sebagai negative, *True Neutral* (TNt) merupakan data *neutral* yang berhasil diklasifikasi dengan benar sebagai *neutral*, *False Positive* (FP) merupakan data *neutral* atau *negative* namun diklasifikasi sebagai *positive*, *False Negative* (FN) merupakan data *positive* atau neutral namun diklasifikasi sebagai *negative*, *False Neutral* (FNt) merupakan data *positive* atau *negative* yang diklasifikasi sebagai *neutral*. *Confusion matrix* dapat dilakukan perhitungan *accuracy* menggunakan persamaan 8, *precision* menggunakan persamaan 9, dan *recall* menggunakan persamaan 10.

$$Accuracy = \frac{TP+TNt+TN}{TP+FP+TN+FN+TNt+FNt} \quad \text{.....(8)}$$

$$Precision = \frac{TP}{TP+FP+FP} \quad \text{.....(9)}$$

$$Recall = \frac{TP}{TP+FNt+FN} \quad \text{.....(10)}$$

RESULTS

Hasil pengumpulan data yang dilakukan melalui teknik *web scraping*, diperoleh 8723 data ulasan dari rentang waktu 26 Agustus 2023 hingga 26 November 2023. Data tersebut kemudian diberi label oleh pakar pertama dengan jumlah sentimen positif 1598 data, sentimen netral 439 data, dan sentimen negatif 6686 data. Selanjutnya data ulasan akan akan melalui tahap persiapan data yaitu *text preprocessing* sebelum dilakukan proses klasifikasi. Hasil *text preprocessing* dapat dilihat dalam gambar 2.

	Ulasan sebelum text preprocessing	Ulasan setelah text preprocessing
0	Lebih seneng pake apk yg lama	lebih senang pakai aplikasi yang lama
1	Ribet masa bisa di kunci twiter akun saya.trus...	rumit masa bisa di kunci twitter akun saya ter...
2	Bercanda	canda
3	Tapi	tapi
4	Semakin lama aplokasi ini semakin matang.	makin lama aplikasi ini makin matang
...	...	...
8718	Bagus	bagus
8719	Kek ntut	seperti ntut
8720	Bague	bagus
8721	Semenjak dibeli Elon Musk, isi Twitter langsun...	semenjak beli elon musk isi twitter langsung u...
8722	Nama X	nama x

Gambar 2. Hasil *Text Preprocessing*

Selanjutnya untuk memberikan bobot pada setiap fitur dalam data, maka digunakan metode pembobotan kata TF-IDF yang dapat dilihat pada gambar 3.

Dokumen	Fitur	TF-IDF
0	aplikasi	0.226088
0	lama	0.394723
0	lebih	0.382244
0	pakai	0.408166
0	senang	0.625702
...	...	...
8721	twitter	0.265723
8721	ubah	0.282963
8721	yang	0.179054
8722	nama	0.746489
8722	x	0.665398

Gambar 3. Pembobotan TF-IDF

Setelah nilai TF-IDF dihasilkan maka tahap yang berikut adalah membagi data menjadi data latih dan data uji menggunakan metode *K-fold cross validation*. Hasil pembagian data tersebut kemudian digunakan untuk klasifikasi metode *naïve bayes* dan *k-Nearest Neighbor*. dimana pengujian menggunakan *K-fold cross validation* ini dilakukan sebanyak 9 kali yaitu menggunakan nilai K=2, K=3, K=4, K=5, K=6, K=7, K=8, K=9, K=10. Contoh hasil klasifikasi metode *naïve bayes* dan *k-Nearest Neighbor* pada data uji dapat dilihat pada gambar 4 dan gambar 5.

DOI : https://doi.org/ 10.26623/f4k55e04

	Data Uji	Label Aktual	Label Prediksi Naive Bayes
0	aplikasi lama lebih pakai senang yang	negatif	negatif
1	akun bisa buka di foto grab harus kunci lagi m...	negatif	negatif
2	di pakai tidak_bisa	negatif	negatif
3	benar gila iya kamu nya susah x	negatif	negatif
4	buruk	negatif	negatif
...	...	...	...
4356	deh lah	netral	negatif
4357	selingkuh	netral	negatif
4358	hitam	netral	negatif
4359	ah au gelap	netral	negatif
4360	nama x	netral	negatif

Gambar 4. Klasifikasi Metode *Naïve Bayes* Pada Data Uji

	Data Uji	Label Aktual	Label Prediksi KNN
0	aplikasi lama lebih pakai senang yang	negatif	negatif
1	akun bisa buka di foto grab harus kunci lagi m...	negatif	negatif
2	di pakai tidak_bisa	negatif	negatif
3	benar gila iya kamu nya susah x	negatif	negatif
4	buruk	negatif	negatif
...	...	...	...
4356	deh lah	netral	positif
4357	selingkuh	netral	negatif
4358	hitam	netral	negatif
4359	ah au gelap	netral	negatif
4360	nama x	netral	negatif

Gambar 5. Klasifikasi Metode *K-Nearest Neighbor* Pada Data Uji

DISCUSSION

Hasil klasifikasi metode *naïve bayes* dan *k-Nearest Neighbor* kemudian akan dievaluasi menggunakan metode *confusion matrix* yang memberikan nilai akurasi, presisi dan *recall*. Berdasarkan nilai-nilai yang diperoleh tersebut akan disajikan dalam tiga skenario yaitu skenario untuk menguji metode *naïve bayes*, skenario untuk menguji metode *k-Nearest Neighbor* dan skenario untuk membandingkan hasil klasifikasi metode *naïve bayes* dan *k-Nearest Neighbor* dengan hasil pelabelan dari pakar kedua.

1) Skenario pengujian metode *naïve bayes*

Tabel 2. Hasil Pengujian Metode *Naïve Bayes*

K-Fold	**Akurasi**	**Presisi**			***Recall***		
		Negatif	Netral	Positif	Negatif	Netral	Positif
K=2	87.84%	86.90%	0.00%	94.46%	99.30%	0.00%	64.02%
K=3	88.40%	87.65%	0.00%	93.29%	99.13%	0.00%	67.77%
K=4	88.70%	88.09%	0.00%	92.43%	98.92%	0.00%	70.28%
K=5	88.69%	88.14%	0.00%	92.00%	98.85%	0.00%	70.53%
K=6	88.80%	88.29%	0.00%	91.91%	98.85%	0.00%	71.15%
K=7	88.78%	88.28%	0.00%	91.77%	98.82%	0.00%	71.15%
K=8	88.67%	88.22%	0.00%	91.50%	98.76%	0.00%	70.84%
K=9	88.82%	88.37%	0.00%	91.55%	98.77%	0.00%	71.59%
***K*=10**	**88.87%**	**88.39%**	**0.00%**	**91.78%**	**98.79%**	**0.00%**	**71.78%**

Berdasarkan tabel 2 dapat dilihat bahwa metode *naïve bayes* menghasilkan akurasi tertinggi pada nilai *K*=10 dengan 88.87%, presisi negatif 88.39%, presisi netral 0%, presisi positif 91.78%, serta nilai *recall* negatif 98.79%, *recall* netral 0%, *recall* positif 71.78%. Adapun nilai 0% pada presisi dan recall untuk kelas netral diakibatkan karena metode *naïve bayes* tidak dapat mengklasifikasi satupun kelas netral dengan benar. Hal ini dikarenakan metode *naïve bayes* sendiri mengklasifikasi data dengan mencari nilai tertinggi dari masing-masing kelas pada hasil

perhitungan probabilitas, kelas netral yang memiliki jumlah lebih sedikit dari antara dua kelas yang lain mengakibatkan hasil perhitungan probabilitasnya lebih rendah, sehingga *naïve bayes* cenderung mengklasifikasi kelas netral kedalam kelas negatif dan positif. Hasil pengujian untuk metode *naïve bayes* dalam tabel 1, akurasi yang dihasilkan semakin meningkat seiring dengan peningkatan jumlah nilai *K-fold* yang digunakan. Semakin tinggi nilai *K* maka semakin banyak pula data latih yang digunakan untuk melatih metode *naive bayes* dalam setiap *fold*. Hal ini menunjukan bahwa metode *naïve bayes* akan bekerja lebih baik dengan menggunakan data latih yang banyak.

2) Skenario pengujian metode *k-Nearest Neighbor*

Pada pengujian metode *k-Nearest Neighbor* digunakan variasi nilai *k* dari *k*=5 hingga *k*=20. Penggunaan variasi nilai *k* yang dibatasi sampai *k*=20 untuk menghindari *k-NN* memprediksi data ke dalam kelas mayoritas.

Tabel 3. Hasil Akurasi K-Fold

K-Fold	**Nilai *k* Optimal**	**Akurasi**	**Presisi**			***Recall***		
			Negatif	Netral	Positif	Negatif	Netral	Positif
***K*=2**	***k*=12**	**90.32%**	**91.98%**	**36.67%**	**84.61%**	**96.75%**	**3.19%**	**87.36%**
K=3	*k*=14	90.24%	92.07%	47.85%	83.99%	96.43%	5.92%	87.55%
K=4	*k*=10	90.09%	92.56%	37.96%	83.78%	95.75%	10.48%	88.30%
K=5	*k*=17	90.18%	92.23%	50.99%	82.96%	96.19%	6.15%	88.11%
K=6	*k*=12	90.16%	92.62%	37.52%	83.67%	95.87%	9.79%	88.36%
K=7	*k*=12	90.19%	92.67%	40.94%	83.54%	95.86%	11.16%	88.17%
K=8	*k*=13	90.17%	92.69%	42.15%	83.17%	95.77%	10.48%	88.67%
K=9	*k*=12	90.19%	92.62%	40.62%	83.84%	95.87%	10.94%	88.17%
K=10	*k*=14	90.27%	92.60%	42.86%	83.68%	95.96%	10.02%	88.48%

Hasil pengujian metode *k-Nearest neighbor* pada tabel 3, menghasilkan akurasi tertinggi pada nilai *K*=2 dengan nilai *k* optimal untuk tetangga terdekat *k-NN* adalah *k*=12 yaitu sebesar 90.32%, sedangkan nilai presisi negatif 91.98%, presisi netral 36.67%, presisi positif 84.61, dan nilai *recall* negatif 96.75%, *recall* netral 3.19%, *recall* positif 87.36%. Terlepas dari distribusi data yang tidak seimbang, metode *k-Nearest Neighbor* dapat memprediksi beberapa kelas netral dengan benar yang dapat dilihat pada nilai presisi dan *recall* untuk kelas netral. Peningkatan nilai *K-fold* yang digunakan, jumlah data latih juga meningkat namun akurasi yang dihasilkan tidak mengalami peningkatan secara konsisten.

3) Skenario Perbandingan

Sebagai kelanjutan dari klasifikasi sebelumnya, selanjutnya dilakukan evaluasi terhadap hasil klasifikasi dari metode *naïve bayes* dan *k-Nearest Neighbor* berdasarkan pelabelan dari pakar kedua. Pelabelan oleh pakar kedua ini dilakukan terhadap data yang telah melalui tahap *text preprocessing* yang tentunya mengalami perubahan bentuk kata dan mengubah struktur kalimat, sehingga berpengaruh terhadap sentimen yang terkandung dalam data. Perbandingan ini dilakukan dengan tujuan menjadikan label pakar kedua sebagai data acuan dalam analisis, untuk menentukan metode mana yang paling efektif dalam mencerminkan sentimen asli dalam data.

DOI : https://doi.org/ 10.26623/f4k55e04

Tabel 4. Perbandingan Klasifikasi *Naïve Bayes* dan *k-Nearest Neighbor* dengan Label Pakar Kedua

K-Fold		*K*=2	*K*=3	*K*=4	*K*=5	*K*=6	*K*=7	*K*=8	*K*=9	*K*=10
NB	Sama	92.30%	92.51%	92.56%	92.47%	92.55%	92.48%	92.40%	92.42%	92.56%
	Beda	7.70%	7.49%	7.44%	7.53%	7.45%	7.52%	7.60%	7.58%	7.44%
k-NN	Sama	91.75%	91.45%	90.93%	91.28%	91.04%	91.01%	90.92%	90.83%	91.12%
	beda	8.25%	8.55%	9.07%	8.72%	8.96%	9.02%	9.08%	9.17%	8.88%

Hasil perbandingan pada tabel 4, metode *naive bayes* menghasilkan persentase hasil klasifikasi yang sesuai dengan label pakar kedua yaitu berkisar antara 92.30% hingga yang tertinggi yaitu 92.56%, sedangkan untuk metode *k-Nearest Neighbor* berkisar antara 90.83% hingga tertinggi 91.73%. Sehingga dapat disimpulkan bahwa metode *naive bayes* memiliki tingkat kesesuaian yang lebih tinggi terhadap label pakar kedua.

CONCLUSIONS AND RECOMMENDATIONS

Dalam proses analisis, data yang digunakan sebanyak 8723 data yang terdiri dari 6686 kelas negatif, 1598 kelas positif, dan 439 kelas netral. Data kemudian dibagi menjadi data uji dan data latih dengan metode *K-fold cross validation* menggunakan nilai *K-fold K*=2 sampai dengan *K*=10. Berdasarkan hasil pengujian, diperoleh akurasi metode *naïve bayes* sebesar 88.87% pada nilai *K-fold K*=10, sementara metode *k-nearest neighbor* memperoleh akurasi tertinggi pada nilai *K-fold K*=2 dengan menggunakan *k* optimal untuk *k-Nearest Neighbor* adalah *k*=12 dengan akurasi sebesar 90.32%. Hasil akurasi metode *k-Nearest Neighbor* yang lebih tinggi dibanding *naïve bayes* menunjukan bahwa metode *k-Nearest Neighbor* lebih unggul jika digunakan pada data dengan distribusi kelas tidak seimbang. Sedangkan pada perbandingan hasil klasifikasi metode *naïve bayes* dan *k-Nearest Neighbor* dengan label dari pakar kedua, metode *naïve bayes* menghasilkan persentase kesesuaian yang lebih tinggi yaitu 92.56% dibandingkan metode *k-Nearest Neighbor* yang menghasilkan persentase tertinggi sebesar 91.73%.

Berdasarkan kesimpulan dari penelitian ini, maka saran yang dapat diberikan untuk penelitian selanjutnya adalah dapat memperbanyak daftar kata dalam kamus normalisasi yang digunakan karena dalam penelitian ini masih banyak kata yang tidak ternormalisasi karena adanya kesalahan penulisan maupun singkatan-singkatan.

REFERENCES

[1] I. A. Ratnamulyani and B. I. Maksudi, "PERAN MEDIA SOSIAL DALAM PENINGKATAN PARTISIPASI PEMILIH PEMULA DIKALANGAN PELAJAR DI KABUPATEN BOGOR," *Sosiohumaniora*, vol. 20, no. 2, Art. no. 2, Jul. 2018, doi: 10.24198/sosiohumaniora.v20i2.13965.

[2] kenny, "Digital 2023 October Global Statshot Report," We Are Social Indonesia. Accessed: Jan. 28, 2024. [Online]. Available: https://wearesocial.com/id/blog/2023/10/digital-2023-october-global-statshot-report/

[3] T. I. Suari and D. Gustian, "SENTIMEN ANALISIS TERHADAP PENGGUNA APLIKASI TWITTER PADA GOOGLE PLAYSTORE MENGGUNAKAN METODE NAIVE BAYES," in *Prosiding Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra*, 2023, pp. 63–69. Accessed: Jan. 27, 2024. [Online]. Available: https://sismatik.nusaputra.ac.id/index.php/sismatik/article/view/204

[4] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *INTEGER: Journal of Information Technology*, vol. 2, no. 1, Art. no. 1, Mar. 2017, doi: 10.31284/j.integer.2017.v2i1.95.

[5] M. Yasid, "ANALISIS SENTIMEN MASKAPAI CITILINK PADA TWITTER DENGAN METODE NAÏVE BAYES," *JURNAL ILMIAH INFORMATIKA*, vol. 7, no. 02, Art. no. 02, Oct. 2019, doi: 10.33884/jif.v7i02.1329.

[6] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *INSYST: Journal of Intelligent System and Computation*, vol. 1, no. 1, Art. no. 1, Aug. 2019, doi: 10.52985/insyst.v1i1.36.

[7] R. Apriani and D. Gustian, "ANALISIS SENTIMEN DENGAN NAÏVE BAYES TERHADAP KOMENTAR APLIKASI TOKOPEDIA," *Jurnal Rekayasa Teknologi Nusa Putra*, vol. 6, no. 1, Art. no. 1, Sep. 2019, doi: 10.52005/rekayasa.v6i1.86.

[8] D. C. Hidayati, S. A. Faraby, and A. Adiwijaya, "Klasifikasi Topik Multi Label pada Hadis Shahih Bukhari Menggunakan K-Nearest Neighbor dan Latent Semantic Analysis," *JURIKOM (Jurnal Riset Komputer)*, vol. 7, no. 1, Art. no. 1, Feb. 2020, doi: 10.30865/jurikom.v7i1.2013.

[9] K. Sihotang and R. Ghaniy, "Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir Pada Website Perpustakaan STIKOM Binaniaga," vol. 9, 2019.

[10] B. Bijanto, Z. Abidin, and T. Tamrin, "PEMBELAJARAN ALGORITMA K-NN UNTUK BIG DATASET MENGGUNAKAN METODE SAMPLE BOOTSTRAP DAN WEIGHTED GINI INDEX," *JDPT*, vol. 12, no. 2, pp. 71–77, Jan. 2022, doi: 10.34001/jdpt.v12i2.2091.

[11] D. F. Sari, A. Kusjani, D. Kurniawati, and I. Setiawan, "PENCARIAN DATA QUICK COUNT PILPRES DENGAN TEKNIK WEB SCRAPING," *Journal of Innovation Research and Knowledge*, vol. 3, no. 5, Art. no. 5, Oct. 2023, Accessed: Nov. 24, 2023. [Online]. Available: https://bajangjournal.com/index.php/JIRK/article/view/6695

[12] E. S. Y. Pandie, "IMPLEMENTASI ALGORITMA DATA MINING NAIVE BAYES PADA KOPERASI," *J-Icon : Jurnal Komputer dan Informatika*, vol. 6, no. 1, Art. no. 1, Mar. 2018, doi: 10.35508/jicon.v6i1.350.

[13] A. Y. Simanjuntak, I. S. eptian S. Simatupang, and A. Anita, "IMPLEMENTASI DATA MINING MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER UNTUK DATA KENAIKAN PANGKAT DINAS KETENAGAKERJAAN KOTA MEDAN," *JOURNAL OF SCIENCE AND SOCIAL RESEARCH*, vol. 5, no. 1, Art. no. 1, Feb. 2022, doi: 10.54314/jssr.v5i1.804.

[14] Y. Luhulima, "Sentiment Analysis pada Review Barang Berbahasa Indonesia dengan Metode K-Nearest Neighbor (K-NN).," PhD Thesis, Universitas Brawijaya, 2013. Accessed: Mar. 13, 2024. [Online]. Available: http://repository.ub.ac.id/145885/

[15] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *Jurnal Informatika*, vol. 11, no. 1, p. 36, Jan. 2017, doi: 10.26555/jifo.v11i1.a5452.

DOI : https://doi.org/ 10.26623/f4k55e04