e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 5 No. 2, Desember, 2021, **Hal. 167-176**
DOI: 10.29408/edumatic.v5i2.3664

Perbandingan Performansi Model pada Algoritma K-NN terhadap Klasifikasi Berita Fakta Hoaks Tentang Covid-19

Wahyu Hidayat[*1], Ema Utami [2], Ahmad Fikri Iskandar [3], Anggit Dwi Hartanto [4], Agung Budi Prasetio [5]
[1,2,3,4,5] Program Studi Teknik Informatika, Univeristas Amikom Yogyakarta
email: wahyuhublaa@gmail.com*[1], ema.u@amikom.ac.id[2], andfikri@gmail.com[3], anggit@amikom.ac.id[4], agung.bp@excelindo.co.id[5]

(Received: 15 Juni 2021 / Accepted: 14 Juli 2021/ Published Online: 20 Desember 2021)

Abstrak

Selama pandemi Covid-19 terdapat berbagai berita hoaks mengenai Covid-19. Terdapat platform klarifikasi fakta pemberitaan hoaks mengenai Covid-19 seperti Jala Hoaks dan Saber Hoaks yang mengkategorikan berita kedalam misinformasi dan disinformasi. Tujuan penelitian ini untuk membandingkan performa model dissimilarity measure dari K-NN terhadap klasifikasi berita fakta hoaks Covid-19. Klasifikasi metode *supervised learning* diterapkan untuk melakukan pembelajaran dari label fakta. Dataset yang digunakan diambil dari Jala Hoaks dan Saber Hoak sebanyak 559 data yang dibuat kedalam Class 1 (*Misleading Content, Satire or Parody, False Connection*), Class 2 (*False Context, Imposter Content*), Class 3 (*Fabricated Content, Manipulated Content*) yang kemudian dibagi 80% untuk data latih dan 20% untuk data uji. Algoritma K-Nearest Neighbor (K-NN) digunakan untuk mengklasifikasi kategori fakta misinformasi dan disinformasi. Model *dissimilarity measure* Jaccard Distance dibandingkan dengan berbagai model Euclidean, Manhattan, dan Minkowski serta menggunakan varian nilai *k* pada algoritma K-NN untuk mengetahui hasil perbandingan performasi setiap pengujiannya. Hasil pengujian model Jaccard Distance pada nilai *k* = 4 mendapatkan nilai lebih tinggi dibandingkan dengan hasil pengujian model lainnya dengan nilai *accuracy* 0,696, *precision* 0,710, *recall* 0,572 dan 0,599 untuk *F1-Score*. Hasil maksimal cenderung pada label *class* data terbanyak yaitu pada *Class* 1 dengan total 58 data benar dari 61 data uji.

Kata kunci: K-Nearest Neighbor, Euclidean, Manhattan, Minkowski, Jaccard.

Abstract

During Covid-19 pandemic, there was various hoax news about Covid-19. There are truth-clarification platforms for hoax news about Covid-19 such as Jala Hoax and Saber Hoax which categorize into misinformation and disinformation. Classification of supervised learning methods is applied to carry out learning from fact labels. Dataset is taken from Jala Hoax and Saber Hoax as many as 559 data which are made into Class 1 (Misleading Content, Satire/Parody, False Connection), Class 2 (False Context, Imposter Content), Class 3 (Fabricated and Manipulated Content). K-Nearest Neighbor (K-NN) is used to classify categories of misinformation and disinformation. Dissimilarity measure Jaccard Distance is compared with Euclidean, Manhattan, and Minkowski and uses k-value variance in the K-NN to determine the performance comparison results for each test. Results of Jaccard Distance at the value of k = 4 get a higher value than other model with an accuracy 0.696, precision 0.710, recall 0.572, and F1-Score. Maximum Results tend to be on the label of the most data class in Class 1 (Misleading Content, Satire or Parody, False Connection) with a total of 58 correct data from 61 test data.

Keywords: K-Nearest Neighbor, Euclidean, Manhattan, Minkowski, Jaccard

PENDAHULUAN

Setelah World Health Organization (WHO) menyatakan Corona Virus Disease of 2019 (Covid-19) sebagai wabah pandemi pada awal tahun 2020, penggunaan dan penerapan aplikasi daring bertambah digunakan dalam berbagai bidang untuk mencegah penyebaran Covid-19 karena dapat menjaga jarak atau mengurangi kerumunan. Selama pandemi Covid-19 berlangsung terdapat penyebaran informasi tidak tepat (Mathur, Kubde, & Vaidya, 2020). Berdasarkan informasi palsu yang disampaikan, terdapat pengaruh dari penyampaian berita hoaks tersebut mengenai virus Covid-19 yang terdapat pada internet terhadap pembentukan opini masyarakat (Roy & Junaidi, 2020) dengan persentase sebesar 58,7% yang dapat mempengaruhi pengguna internet dan 41,3% lainnya dipengaruhi oleh faktor lain.

Terdapat beberapa platform mengenai fakta berita hoaks tentang Covid-19 seperti Jala Hoaks (https://data.jakarta.go.id/jalahoaks) dan Saber Hoaks (https://saberhoaks.jabarprov.go.id/v2) yang memverifikasi berita hoaks yang beredar pada masyarakat dan difakta berdasarkan kategori misinformasi dan disinformasi. Sehingga tersedia data kebenaran terhadap hoaks tentang Covid-19 yang beredar. Supervised learning dengan metode klasifikasi dapat melakukan pembelajaran pada data latih untuk mendapatkan output label yang sama (Kristiawan, Somali, Linggan jaya, & Widjaja, 2020). Salah satu algoritma dalam klasifikasi yaitu K-Nearest Neighbor (K-NN) yang melakukan klasifikasi berdasarkan data latih dan dihitung berdasarkan tetangga terdekatnya (Satrian & Gusrianty, 2020).

Performa algoritma K-NN dibandingkan dengan algoritma Naïve Bayes pada kasus klasifikasi ketepatan waktu lulus mahasiswa (Sabilla & Putri, 2017) dengan menggunakan 159 data latih yang terdiri dari 82 data lulus tepat waktu dan 77 data lulus tidak tepat waktu yang kemudian dibagi menjadi 80%:20%;60%:40%;50%:50% untuk data latih dan data uji. Didapatkan pengujian *accuracy* sebesar 98.7% untuk algoritma K-NN dengan nilai $k = 2$ sedangkan untuk algoritma Naïve Bayes didapatkan hasil *accuracy* 80,9% sehingga algoritma K-NN memiliki performa pengujian yang lebih baik.

Penerapan varian nilai *k* pada algoritma K-NN mempengaruhi hasil dari pengujian yang dilakukan (Dinata, Akbar, & Hasdyna, 2020). Dalam penelitian untuk pengklasifikasian transportasi bus, menggunakan 176 data training dan 75 data testing serta menggunakan nilai varian *k* yaitu $k = 1$, $k = 2$, dan $k = 3$ dengan perhitungan model menggunakan Euclidean Distance dan Manhattan Distance. Penerapan K-NN dengan Euclidean Distance mendapatkan nilai *accuracy* 81.96% sedangkan untuk Manhattan Distance mendapatkan hasil *accuracy* 84.00% dengan hasil tertinggi dari nilai $k = 3$. Penelitian mengenai perbandingan model Euclidean, Minkowski dan Manhattan (Jedari, Wu, Rashidzadeh, & Saif, 2015) mendapatkan hasil pengujian yang sama untuk model Euclidean dan Minkowski yaitu dengan jarak terbaik dengan nilai 64,26, sedangkan model Manhattan mempunyai nilai terbaik 76,26.

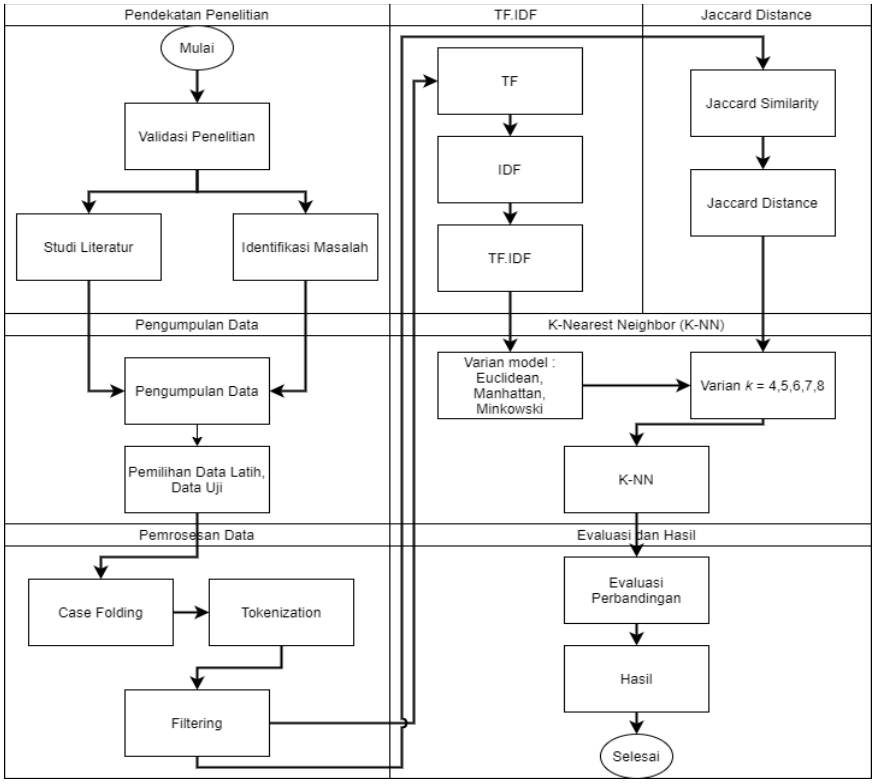
Model *dissimilarity measure* merupakan hasil dari jarak kesamaan dokumen yang dihitung dari Jaccard Similarity yang dibandingkan dengan jarak antara kemiripan *neighbor* terdekat dihitung menggunakan Jaccard Distance pada K-NN (Badhani & Muttoo, 2019) penerapan model Jaccard tersebut menghasilkan pengujian yang maksimal dari permasalahan mendeteksi malware mendapatkan akurasi 98% dari 3240 dataset aplikasi yang terdiri dari 1620 aplikasi biasa dan 1620 aplikasi berbahaya. Namun model *dissimilarity measure* menggunakan Jaccard Distance belum pernah dibandingkan karena memiliki langkah yang berbeda dibandingkan dengan model lainnya.

Kami tertarik untuk menguji perbandingan beberapa model *distance* dengan model Jaccard Distance beserta varian nilai *k* terhadap algoritma klasifikasi K-NN, karena dalam penelitian sebelumnya algoritma tersebut lebih unggul dibandingkan algoritma lainnya dalam menentukan label dan mempunyai berbagai model yang mempunyai hasil pengujian performa yang berbeda. Tujuan penelitian ini untuk membandingkan performa model *dissimilarity measure* menggunakan Jaccard Distance dan model Euclidean, Minkowski dan Manhattan dari

K-NN serta membandingkan performa varian nilai *k* pada pengujian hasil evaluasi sehingga dengan pelaksanaan pengujian yang akan dilakukan dapat diketahui perbandingan performa dari penerapan model terhadap algoritma klasifikasi K-NN dalam menentukan pengelompokkan fakta berita hoaks mengenai Covid-19 berdasarkan kategori misinformasi dan disinformasinya.

METODE

Berdasarkan studi literatur yang telah dilakukan, dibuatkan alur penelitian yang bertujuan untuk menyelesaikan permasalahan sehingga penelitian yang dilakukan dapat tertuju dan fokus terhadap penyelesaian masalah. Berikut gambar 1 memuat diagram alir penelitian.



Gambar 1. Digram Alir Penelitian

Berdasarkan diagram alir pada Gambar 1 berikut ditampilkan penjelasan dari beberapa proses sehingga memperjelas dari diagram alir tersebut.

Pengumpulan Data

Data yang digunakan berasal dari Jala Hoaks (https://data.jakarta.go.id/jalahoaks/) dan Saber Hoaks (https://saberhoaks.jabarprov.go.id/v2/). Data yang diambil mencakup ranah permasalahan berita Covid-19, data yang didapatkan sebanyak 559 data. Data mempunyai rentang waktu dari Januari 2020 sampai akhir bulan April 2021.

Pemrosesan Data

Data mentah yang telah didapatkan akan dibersihkan terlebih dahulu (Takdirillah, 2020). Proses ini terkait dalam pengolahan kata yang didapatkan dari data fakta berita hoaks, proses tersebut mencakup *case folding*, *tokenizing*, *filtering* atau *stopword removal* (Wibawa, Nasrun, & Setianingsih, 2018).

Pembobotan / Term Frequency - Invers Document Frequency (TF-IDF)

TF-IDF merupakan metode untuk merubah dokumen dalam bentuk susunan kata yang ditransformasikan dalam bobot dokumen berupa angka. Berikut perhitungan dari TF-IDF (Wang, Lu, Chow, & Zhu, 2020) yang dapat dilihat pada persamaan 1.

$$\begin{cases} tf(t,d) = \log(1 + freq(t,d)) \\ idf(t,D) = \log(\frac{N}{count(d \in D : t \in d)}) \\ tfidf(t,d,D) = tf(t,d) \times idf(t,D) \end{cases} \quad (1)$$

Keterangan :

t = *Term* pada dokumen koresponden
d = Dokumen koresponden
D = Jumlah semua dokumen
tf = Frekuensi kemunculan kata (*term*) terhadap setiap dokumen
idf = Hubungan pendistribusian kata pada dokumen yang bersangkutan

Jaccard Distance

Jaccard digunakan untuk menghitung seberapa mirip sepasang sample dengan membandingkan dua sampel data. Jaccard Similarity untuk menghitung indeks kemiripan himpunan ditampilkan sebagai berikut (Le & Phuong, 2020) pada persamaan 2.

$$J(A,B) = def \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Keterangan :

J(A,B) = Kemiripan dari dokumen yang dibandingkan
A, B = Dokumen koresponden

K-Nearest Neighbor (K-NN)

Klasifikasi K-NN merupakan metode yang melakukan perhitungan klasifikasi terhadap label tertentu berdasarkan pembelajaran terhadap jarak paling dekat dengan objek. Perhitungan K-NN dapat dilihat pada persamaan 3 (Walid & Darmawan, 2017).

$$d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (3)$$

Keterangan :

d = Jarak berdasarkan Euclidean Distance
a, b = Jarak antara dataset

Evaluasi

Penelitian ini menggunakan pengujian *accuracy*, *precision*, *recall* dan *F1-Score* untuk pengujian algoritma (Guillet & Hamilton, 2007), serta direpresentasikan dengan Confusion Matrix didasarkan pada evaluasi perkiraan benar atau salah (Riefky & Pramesti, 2020). Pengukuran atau pengujian algoritma digunakan untuk mengetahui performa dan keoptimalan algoritma (Sari, Firdausi, & Azhar, 2020).

HASIL DAN PEMBAHASAN

Hasil

Data mentah yang digunakan diambil dari *scraping* pada website https://data.jakarta.go.id/jalahoaks/ dan https://saberhoaks.jabarprov.go.id/v2/ pada ruang lingkup pembahasan hoaks pada berita Covid-19. Total data yang didapatkan sebanyak 559 data, setiap dokumen data tersebut memiliki fakta fakta berdasarkan kategori misinformasi dan

disinformasi, kategori tersebut yaitu *Misleading Content* atau penggunaan informasi yang sesat untuk membuat sebuah isu, *Satire or Parody* atau tidak ada niat untuk merugikan namun berpotensi untuk mengelabui, *False Connection* yang berarti ketika judul dan gambar tidak mendukung konten untuk membuat sebuah isu, *False Context* yaitu ketika konten yang asli dipadankan dengan konteks informasi yang salah, *Imposter Content* yaitu ketika sebuah sumber asli ditiru, *Fabricated Content* yang berarti konten baru yang sengaja di buat dan di desain untuk menipu dan merugikan, dan *Manipulated Content* yang ketika sebuah informasi di manipulasi untuk merusak atau menipu.

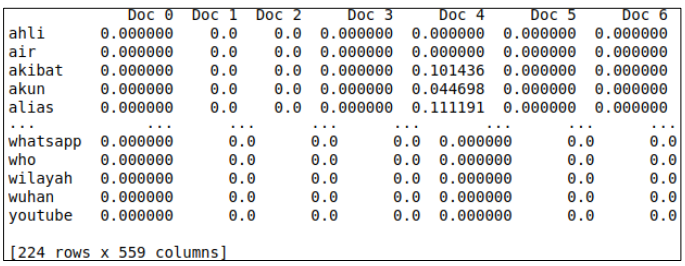
Berdasarkan kategori fakta yang ada, data tersebut dibagi menjadi 3 label *class* dikelompokkan berdasarkan makna yang menyerupai, yaitu pada label *Class* 1 untuk kategori konten menyesatkan (*Misleading Content*), konten satir (*Satire or Parody*) dan konten salah sambung (*False Connection*) sedangkan untuk label *Class* 2 terdiri dari kategori konten salah (*False Context*) dan konten penipuan (*Imposter Content*) serta pada label *Class* 3 terdiri dari kategori konten fabrikasi (*Fabricated Content*) dan konten manipulasi (*Manipulated Content*). Berdasarkan pembagian 3 label yang ada telah dipisah berdasarkan kategori fakta yang ada, terdapat sebaran data pada label *Class*1 sebanyak 321 data, pada label *Class* 2 sebanyak 139 data dan pada label *Class* 3 sebanyak 99 data. Serta pembagian data latih sebesar 80% dan data uji sebesar 20%. Data mentah ditampilkan pada tabel 1.

Tabel 1. Data Mentah

No	Berita	Cek_Fakta	Kesimpulan	*Class*
0	Beredar pesan berantai … 3M.	Berdasarkan hasil …8470.	Informasi tentang … Disinformasi	3
1	Beredar informasi di …Covid-19.	Berdasarkan hasil …Kusnandi.	Informasi bahwa …Disinformasi	2
…	…	…	…	…
558	JOKOWI: VIRUS CORONA …DISINFORMASI	Setelah ditelusuri …Dextromethorphan.	NaN	2

Berdasarkan tabel 1, perlu diterapkan teknik *case folding*, *tokenizing*, dan *stopword removal* untuk mebersihkan teks yang terdapat pada atribut selanjutnya pada atribut berita, cek_fakta dan kesimpulan dilakukan penyatuan menjadi 1 kelompok dokumen dan dilakukan pembobotan. Hasil dari sebaran *term frequency* terhadap semua dokumen yang ada ditampilkan pada gambar 2.

Selanjutnya dari perubahan data teks menjadi angka tersebut selanjutnya diproses pada tahap klasifikasi menggunakan algoritma K-NN dengan membandingkan model euclidean, manhattan dan Minkowski dengan membagi data latih sebesar 80% dan data uji sebesar 20%. Berikut hasil pengujian K-NN menggunakan model Euclidean ditampilkan pada tabel 2.



	Doc 0	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
ahli	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
air	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000
akibat	0.000000	0.0	0.0	0.000000	0.101436	0.000000	0.000000
akun	0.000000	0.0	0.0	0.000000	0.044698	0.000000	0.000000
alias	0.000000	0.0	0.0	0.000000	0.111191	0.000000	0.000000
...	...	...	...	...	...	...	...
whatsapp	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
who	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
wilayah	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
wuhan	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
youtube	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0

[224 rows x 559 columns]

Gambar 2. Sebaran *Term Frequency*

Hasil pengujian pada tabel 2, didapatkan nilai tertinggi pada keseluruhan varian nilai *k* dan beberapa metode pengujian mendapatkan nilai tertinggi pada hasil dari pengujian *precision* menggunakan nilai *k* = 5 dengan nilai 0,623. Sehingga berdasarkan nilai *precision* yang paling tinggi dalam mengklasifikasi berita fakta hoaks Covid-19 tersebut maka model Euclidean cenderung memiliki hasil klasifikasi yang mendapatkan *true positive* dan tidak cenderung pada hasil klasifikasi *false positive*.

Tabel 2. Hasil K-NN dengan Model Euclidean

k	**K-NN Euclidean**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
4	0,563	0,551	0,426	0,428
5	0,580	0,623	0,453	0,465
6	0,598	0,594	0,459	0,464
7	0,589	0,540	0,449	0,441
8	0,554	0,513	0,411	0,392

Hasil dari pengujian Tabel 2 menunjukkan bahwa penerapan model Euclidean mendapatkan rata-rata dari keseluruhan varian nilai *k* dengan rasio prediksi benar dari keseluruhan data dengan nilai *accuracy* 0,577, sedangkan untuk rasio prediksi benar *positive* dari keseluruhan hasil prediksi *positive* dengan nilai *precision* 0,564, serta rasio prediksi benar *positive* dari keseluruhan data yang benar *positive* memiliki nilai *recall* 0,440 dan memiliki perbandingan rata-rata *precision* dan *recall* yang dibobotkan dengan nilai *F1-Score* 0,438. Sehingga model Euclidean pada algoritma K-NN dalam mengklasifikasi berita fakta hoaks Covid-19 memiliki hasil performa dibawah 0,6 dari keseluruhan metode pengujian yang dilakukan yang berdasarkan dari rata-rata keseluruhan varian nilai *k*. Selanjutnya berikut hasil pengujian K-NN dengan model Manhattan pada tabel 3.

Hasil pengujian tertinggi pada tabel 3, didapatkan pada pengujian *precision* dengan nilai varian *k* = 8 yang memiliki hasil 0,684. Berdasarkan nilai pengujian yang paling tinggi dalam mengklasifikasi berita fakta hoaks Covid-19 tersebut maka model Manhattan cenderung memiliki hasil klasifikasi yang mendapatkan *true positive* dan tidak cenderung pada hasil klasifikasi *false positive*.

Tabel 3. Hasil K-NN dengan Model Manhattan

k	**K-NN Manhattan**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
4	0,589	0,494	0,465	0,467
5	0,607	0,594	0,488	0,489
6	0,643	0,646	0,520	0,531
7	0,634	0,603	0,525	0,538
8	0,643	0,684	0,514	0,527

Berdasarkan tabel 3, hasil rata-rata pengujian pada seluruh varian nilai *k* memiliki rasio prediksi benar dari keseluruhan data dengan nilai *accuracy* 0,632, sedangkan untuk rasio prediksi benar *positive* dari keseluruhan hasil prediksi *positive* dengan nilai *precision* 0,632, serta rasio prediksi benar *positive* dari keseluruhan data yang benar *positive* memiliki nilai *recall* 0,512 dan memiliki perbandingan rata-rata *precision* dan *recall* yang dibobotkan dengan nilai *F1-Score* 0.521. Sehingga model Manhattan pada algoritma K-NN dalam mengklasifikasi berita fakta hoaks Covid-19 memiliki hasil performa diatas 0,6 pada pengujain *accuracy* dan *precision*, sedangkan untuk pengujian *recall* dan *F1-Score* memiliki hasil performa dibawah

0,6 yang didapatkan dari rata-rata keseluruhan varian nilai *k*. Selanjutnya berikut ditampilkan hasil pengujian K-NN dengan model Minkowski pada tabel 4.

Model pengujian tertinggi Tabel 4, didapatkan pada pengujian *precision* dengan nilai k=5 dengan nilai 0,623. Sehingga berdasarkan nilai precision yang paling tinggi dalam mengklasifikasi berita fakta hoaks Covid-19 tersebut maka model Minkowski cenderung mendapatkan *true positive* dan tidak cenderung pada hasil klasifikasi *false positive*.

Tabel 4. Hasil K-NN dengan Model Minkowski

k	**K-NN Minkowski**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
4	0,563	0,551	0,426	0,428
5	0,580	0,623	0,453	0,465
6	0,598	0,594	0,459	0,464
7	0,589	0,540	0,449	0,441
8	0,554	0,513	0,411	0,392

Berdasarkan hasil pengujian tabel 4, dengan model Minkowski mendapatkan rata-rata dari keseluruhan varian nilai *k* dengan rasio prediksi benar dari keseluruhan data dengan nilai *accuracy* 0,577, sedangkan untuk rasio prediksi benar *positive* dari keseluruhan hasil prediksi *positive* dengan nilai *precision* 0,564, serta rasio prediksi benar *positive* dari keseluruhan data yang benar *positive* memiliki nilai *recall* 0,440 dan memiliki perbandingan rata-rata *precision* dan *recall* yang dibobotkan dengan nilai *F1-Score* 0,438. Sehingga model Minkowski pada algoritma K-NN dalam mengklasifikasi berita fakta hoaks Covid-19 memiliki hasil performa dibawah 0,6 dari keseluruhan metode pengujian yang dilakukan yang berdasarkan dari rata-rata keseluruhan varian nilai *k*.

Selanjutnya dilakukan proses untuk mengolah data dokumen yang belum dirubah pada TF.IDF untuk dipotong menjadi per 2 karakter atau 2-gram, selanjutnya potongan tersebut akan dihitung jaraknya menggunakan rumus Jaccard Distance untuk mengetahui jarak antara dokumen satu dengan dokumen lainnya. Berikut hasil dari pengujian K-NN menggunakan Jaccard Distance ditampilkan pada tabel 5.

Tabel 5. Hasil K-NN dengan Model Jaccard Distance

k	**K-NN Jaccard Distance**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
4	0.696	0.710	0.572	0.599
5	0.652	0.632	0.513	0.528
6	0.625	0.658	0.490	0.502
7	0.634	0.644	0.496	0.506
8	0.634	0.637	0.480	0.483

Hasil pengujian *precision* pada tabel 5, memiliki hasil tertinggi pada nilai varian $k = 4$ dengan nilai *precision* 0,710. Sehingga berdasarkan nilai *precision* yang paling tinggi dalam mengklasifikasi berita fakta hoaks Covid-19 tersebut maka model Jaccard Distance cenderung memiliki hasil klasifikasi yang mendapatkan *true positive* dan tidak cenderung pada hasil klasifikasi *false positive*. Keseluruhan hasil pengujian tabel 5, menunjukkan bahwa penerapan model Jaccard Distance mendapatkan rata-rata dari keseluruhan varian nilai *k* dengan rasio prediksi benar dari keseluruhan data dengan nilai *accuracy* 0,648, sedangkan untuk rasio prediksi benar *positive* dari keseluruhan hasil prediksi *positive* dengan nilai *precision* 0,656,

serta rasio prediksi benar *positive* dari keseluruhan data yang benar *positive* memiliki nilai *recall* 0,510 dan memiliki perbandingan rata-rata precision dan recall yang dibobotkan dengan nilai *F1-Score* 0,524. Sehingga model Jaccard Distance pada algoritma K-NN dalam mengklasifikasi berita fakta hoaks Covid-19 memiliki hasil performa diatas 0,6 pada pengujain *accuracy* dan *precision*, sedangkan untuk pengujian *recall* dan *F1-Score* memiliki hasil performa dibawah 0,6 yang didapatkan dari rata-rata keseluruhan varian nilai *k*.

Pembahasan

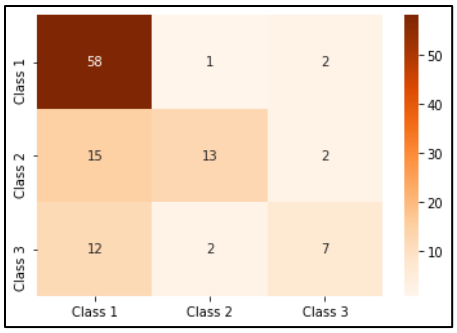
Berdasarkan hasil pengujian setiap model dapat ditampilkan perbandingan setiap model dengan hasil terbaik pada setiap pengujian sebagai berikut ditampilkan pada Tabel 6. Perbandingan tabel 6, *accuracy* K-NN dengan Jaccard Distance memiliki hasil *accuracy* paling tinggi dengan nilai 0.696 pada nilai *k* = 4 dibandingkan dengan pengujian K-NN dengan model lainnya. Serta pada pengujian *precision* model Jaccard Distance mendapatkan hasil paling tinggi dengan nilai 0,710 dibandingkan dengan model lainnya yaitu 0,623 (Euclidean), 0,684 (Manhattan), dan 0,623 (Minkowski). Pengujian *Recall* didapatkan nilai tertinggi yaitu dengan nilai 0,572 dengan model Jaccard Distance hasil tersebut lebih tinggi dibandingkan dari pengujian model lainnya dengan nilai 0,459 (Euclidean), 0,25 (Manhattan) dan 0,459 (Minkowski)

Tabel 6. Perbandingan Hasil K-NN pada Setiap Model

Model	*k*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
K-NN Euclidean	5	0,580	0,623	0,453	0,465
	6	0,598	0,594	0,459	0,464
K-NN Manhattan	6	0,643	0,646	0,520	0,531
	7	0,634	0,603	0,525	0,538
	8	0,643	0,684	0,514	0,527
K-NN Minkowski	5	0,580	0,623	0,453	0,465
	6	0,598	0,594	0,459	0,464
K-NN Jaccard Distance	4	0,696	0,710	0,572	0,599

Hasil pengujian *F1-Score* didapatkan nilai tertinggi yaitu pada pengujian model Jaccard Distance dengan nilai 0,599, hasil tersebut lebih tinggi jika dibandingkan dengan hasil pengujian nilai *F1-Score* dengan model lainnya yaitu dengan nilai 0,465 (Euclidean), 0,538 (Manhattan), dan 0,465 (Minkowski). Model Jaccard Distance dengan nilai varian *k* = 4 unggul pada pengujian *accuracy*, *precision*, *recall* dan *F1-Score* dibandingkan model lainnya, dengan nilai pengujian *accuracy* 0,696, nilai pengujian *precision* 0,710, nilai pengujian *recall* 0,572 dan nilai pengujian *F1-Score* 0,599.

Hasil dari Confusion Matrix untuk mengetahui sebaran data klasifikasi benar terhadap data aktual pada berita fakta hoaks tentang Covid-19 berdasarkan penerapan model Jaccard Distance dengan nilai varian *k* = 4, ditampilkan pada gambar 3. Berdasarkan gambar 3, hasil klasifikasi *true positive* terbaik pada *Class* 1 (*Misleading Content, Satire or Parody, False Connection*) dengan total 58 data benar dari 61 data uji, *Class* 2 (*False Context, Imposter Content*) mendapatkan *true positive* 13 data dari 30 data uji, serta untuk *Class* 3 (*Fabricated Content, Manipulated Content*) hanya mendapatkan *true positive* 7 data dari 21 data uji. Sehingga hasil klasifikasi yang didapatkan dari model Jaccard Distance terhadap data aktual pada berita fakta hoaks tentang Covid-19 cenderung memiliki hasil yang maksimal pada label *class* data uji yang memiliki mayor data.



Gambar 3. Hasil Confusion Matrix Model Jaccard Distance K-NN (k = 4)

Perbandingan dari penelitian sebelumnya (Widiyaningsih & Pertiwi, 2020) pada penelitian tersebut menggunakan nilai k = 1 dimana nilai jarak ditentukan hanya dari 1 tetangga saja sehingga pada penelitian yang telah dilaksanakan ini menggunakan varian nilai k untuk mengetahui perbandingan dari penerapan varian nilai k tersebut. Penelitian terkait (Dinata et al., 2020) mempunyai kesamaan hasil dengan karakteristik model Manhattan yang lebih unggul dibandingkan model Euclidean, namun pada penelitian yang telah dilakukan usulan model Jaccard Distance lebih unggul dibandingkan model lainnya pada nilai k = 4.

SIMPULAN

Pengklasifikasian menggunakan algoritma K-NN dengan varian nilai k pada berita fakta hoaks tentang Covid-19 memiliki hasil pengujian yang berbeda karena jumlah k mempengaruhi dari pemilihan nilai tetangga terdekat. Penerapan model Jaccard Distance pada algoritma K-NN dengan nilai k = 4 mendapatkan hasil pengujian dengan nilai 0,696 untuk *accuracy*, 0,710 untuk *precision*, 0,572 untuk *recall* dan 0,599 untuk *F1-Score*, hasil pengujian tersebut lebih tinggi dibandingkan dengan hasil pengujian dengan model Euclidean, Manhattan dan Minkowski. Ketidakseimbangan sebaran jumlah data terhadap setiap label *class* mempengaruhi hasil dari nilai pengujian. Penerapan model Jaccard Distance pada algoritma K-NN memiliki hasil yang paling baik pada klasifikasi label yang memiliki data uji paling banyak yaitu *Class* 1 (*Misleading Content*, *Satire or Parody*, *False Connection*), saran untuk penelitian selanjutnya adalah mengatasi ketidakseimbangan data yang digunakan.

REFERENSI

Badhani, S., & Muttoo, S. K. (2019). Android Malware Detection Using Code Graphs. *System Performance and Management Analytics*, 203–215. https://doi.org/10.1007/978-981-10-7323-6_17

Dinata, R. K., Akbar, H., & Hasdyna, N. (2020). Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus. *ILKOM Jurnal Ilmiah*, *12*(2), 104–111. https://doi.org/10.33096/ilkom.v12i2.539.104-111

Guillet, F., & Hamilton, H. J. (2007). *Quality Measures in Data Mining*. New York: Springer. https://doi.org/10.1007/978-3-540-44918-8

Jedari, E., Wu, Z., Rashidzadeh, R., & Saif, M. (2015). Wi-Fi based indoor location positioning employing random forest classifier. *International Conference on Indoor Positioning and Indoor Navigation, IPIN 2015*, 13–16. IEEE. https://doi.org/10.1109/IPIN.2015.7346754

Kosub, S. (2019). A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters*, *120*, 36–38. https://doi.org/10.1016/j.patrec.2018.12.007

Kristiawan, K., Somali, D. D., Linggan jaya, T. A., & Widjaja, A. (2020). Deteksi Buah Menggunakan Supervised Learning dan Ekstraksi Fitur untuk Pemeriksa Harga. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(3), 541–548. https://doi.org/10.28932/jutisi.v6i3.3029

Le, T. T. N., & Phuong, T. V. X. (2020). Privacy Preserving Jaccard Similarity by Cloud-Assisted for Classification. *Wireless Personal Communications*, *112*(3), 1875–1892. https://doi.org/10.1007/s11277-020-07131-6

Mathur, A., Kubde, P., & Vaidya, S. (2020). Emotional analysis using twitter data during pandemic situation: Covid-19. *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020*, (Icces), 845–848. https://doi.org/10.1109/ICCES48766.2020.09138079

Riefky, M., & Pramesti, W. (2020). Sentiment Analysis of Southeast Asian Games (SEA Games) in Philippines 2019 Based on Opinion of Internet User of Social Media Twitter with K-Nearest Neighbor and Support Vector Machine. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(1), 26–41. https://doi.org/10.20956/jmsk.v17i1.9947

Roy, J., & Junaidi, A. (2020). Pengaruh Terpaan Media Berita Hoax di Instagram terhadap Opini Masyarakat Milenials Akan Sumber Berita. *Koneksi*, *4*(2), 280-285. https://doi.org/10.24912/kn.v4i2.8138

Sabilla, W. I., & Putri, T. E. (2017). Prediksi Ketepatan Waktu Lulus Mahasiswa dengan k-Nearest Neighbor dan Naïve Bayes Classifier (Studi Kasus Prodi D3 Sistem Informasi Universitas Airlangga). *Jurnal Komputer Terapan*, *3*(2), 233–240.

Sari, V., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 1–9. https://doi.org/10.29408/edumatic.v4i2.2202

Satrian, B., & Gusrianty. (2020). Penerapan Algoritma K-Nn untuk Klasifikasi Gamers Usia Sekolah. *Jurnal Mahasiswa Aplikasi Teknologi Komputer Dan Informasi*, *2*(1), 19–23.

Takdirillah, R. (2020). Penerapan Data Mining Menggunakan Algoritma Apriori Terhadap Data Transaksi Sebagai Pendukung Informasi Strategi Penjualan. *Edumatic: Jurnal Pendidikan Informatika*, *4*(1), 37–46. https://doi.org/10.29408/edumatic.v4i1.2081

Walid, M., & Darmawan, A. K. (2017). Pengenalan Ucapan Menggunakan Metode Linear Predictive Coding (LPC) dan K-Nearest Neighbor (KNN). *Energy, Universitas Panca Marga*, *7*(1), 13–22.

Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access*, *8*, 138162–138169. https://doi.org/10.1109/ACCESS.2020.3012595

Wibawa, D. W., Nasrun, M., & Setianingsih, C. (2018). Sentiment Analysis on User Satisfaction Level of Cellular Data Service Using the K-Nearest Neighbor (K-NN) Algorithm. *International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018*, 235–241. https://doi.org/10.1109/ICCEREC.2018.8711992

Widiyaningsih, S. D., & Pertiwi, A. (2020). Analysis of Ovo Application Sentiment Using Lexicon Based Method and K-Nearest Neighbor. *Jurnal Ilmiah Ekonomi Bisnis*, *25*(1), 14–28. https://doi.org/10.35760/eb.2020.v25i1.2416