Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

Penerapan Algoritma *K-Nearest Neighbor* dalam Klasifikasi Judul Berita *Hoax*

Muhammad Diki Hendriyanto[a], *Betha Nurina Sari*[b]

[a]*Universitas Singaperbangsa Karawang, Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Timur, Karawang, Jawa Barat 41361*
[b]*Universitas Singaperbangsa Karawang, Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Timur, Karawang, Jawa Barat 41361*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 22 Februari 2022
Revisi Akhir: 25 Juni 2022
Diterbitkan *Online*: 15 September 2022

KATA KUNCI

Klasifikasi

Hoax

KNN

KORESPONDENSI

E-mail: muhammad.diki18199@student.unsika.ac.id

A B S T R A C T

With the rapid development of information technology, especially in Indonesia, information is more easily obtained through online media. Therefore, the dissemination of information in online media becomes uncontrollable and a lot of information is not in accordance with the facts or can be said to be a hoax. Readers should be more careful when reading news headlines to avoid hoaxes. The purpose of this research is to find out how to apply the K-Nearest Neighbor (KNN) algorithm in classifying news including hoaxes or not hoaxes. In the process, the classification of hoaxes or non-hoaxes uses the KDD method in text mining and goes through several stages, namely preprocessing, word weighting with TF-IDF and classification using the KNN algorithm. There are 3 scenarios in the data split process, namely 90:10, 80:20, and 70:30. Evaluation is done by using a confusion matrix. The results of this study obtained the highest accuracy of 93.33% with a k value of 3 in the 90:10 scenario. So, the K-Nearest Neighbor algorithm is suitable for classifying hoax news titles.

1. PENDAHULUAN

Internet disebut sebagai bentuk peningkatan teknologi yang sekarang menjadi kebutuhan manusia karena sudah merambah ke prospek budaya masyarakat, dari gaya hidup, bidang edukasi, kajian, hingga ke bidang komersial. Salah satu manfaat dengan adanya kehadiran internet ini yaitu pertumbuhan akan ketersediaan pesat dan sederhana ketika digunakan oleh semua kalangan, dimanapun dan kapanpun [1]. Kumpulan informasi yang menyajikan tentang suatu tentang insiden, pandangan, kegemaran, keadaan, suasana, ulasan yang esensial, atraktif, sedang ramai di perbincangkan dan akan disebar ke masyarakat yang biasanya kita sebut sebagai berita. Berita berisi pemberitaan yang konsisten, relevan, dan akurat yang memberikan pengetahuan kepada penerimanya. Seringkali, berita disajikan kepada pihak ketiga atau kelompok orang dengan bentuk media cetak, dunia maya, atau dari mulut ke mulut [2]. Penyebaran informasi tersebut tidak tersaring, dan siapapun dapat menyebarkan berita yang tidak jelas, bahkan mengandung informasi *hoax* yang dikenal dengan istilah "berita palsu". Berita yang tidak jelas dan *hoax* dapat menimbulkan kegemparan di masyarakat, apalagi jika konten yang beredar sensitif terhadap kehidupan manusia [3].

Hoax adalah berita dan substansial yang bisa saja menyimpang dari penerimaan seseorang dengan menyebarkan berita palsu sebagai validitas. *Hoax* dapat merusak citra dan kredibilitas serta mempengaruhi banyak orang. *Hoax* seringkali menjadi berita yang mengecoh karena tidak adanya asal usul terpercaya dan indikasi yang jelas. Berita *hoax* memang telah direncanakan publikasinya oleh beberapa oknum demi keuntungan sendiri. Saat ini, berita *hoax* dapat meluas dengan internet. Pesatnya perkembangan telekomunikasi menimbulkan peredaran berita di internet yang tidak dapat ditanggulangi, contohnya seperti keterangan dokumen yang berisikan *hoax* [4]. Oleh karena itu upaya yang dapat segera direalisasikan dan dimengerti penerima berita tersebut adalah melakukan pengelompokan berita menurut kategori berita *hoax* atau tidak. Tujuan dari pengelompokan ini adalah untuk memperhitungkan tingkatan dari suatu target yang kelas dan karakteristiknya tidak ditemukan. Dalam memudahkan klasifikasi suatu berita, sebagai tahapan pertama peneliti akan menguji dalam pengklasifikasian judul berita dengan algoritma *K-Nearest Neighbor*.

Algoritma *K-Nearest Neighbor* merupakan algoritma *supervised learning*. *K-Nearest Neighbor* (KNN) merupakan salah satu metode klasifikasi yang yang umumnya dipergunakan dalam data mining, dengan keunggulan mengeksekusi data dan waktu lebih singkat, dimana sangat berguna untuk mengeksekusi persebaran

berita yang tidak jelas isinya bahkan berita yang mengandung unsur hoaks [3]. Penerapan algoritma KNN ditemukan pada sejumlah penelitian yang telah dilakukan yaitu oleh Siti Ernawati dan Risa Wati mengenai analisis sentimen ulasan agen perjalanan menggunakan algoritma *K-Nearest Neighbor* diperoleh akurasi dengan hasil mencapai 87.00% [5]. Dari hasil penelitian lain oleh Pratama Dwi Nugraha dkk mengenai klasifikasi dokumen dengan algoritma KNN dan seleksi fitur *Information Gain* menunjukkan bahwa algoritma *K-Nearest Neighbor* tanpa seleksi fitur *Information Gain* untuk semua dokumen latih menggunakan berbagai parameter mempunyai akurasi tingkat tinggi sebesar 93.94438% [6].

Berdasarkan paparan di atas, diketahui bahwa klasifikasi dengan menerapkan algoritma KNN memberikan akurasi dan performa yang tinggi namun sederhana. Oleh karena itu, algoritma *K-Nearest Neighbor (KNN)* akan diterapkan pada penelitian ini dalam klasifikasi judul berita *hoax*. Tujuannya adalah untuk mengetahui tingkat performa algoritma KNN dalam melakukan klasifikasi judul berita *hoax* dengan bahasa Indonesia. Sehingga diharapkan dapat membantu pembaca berita untuk membedakan berita *hoax* dan fakta.

2. TINJAUAN PUSTAKA

2.1. Berita

Berita adalah informasi terkini yang menjadi ketertarikan pembaca kemudian disebarluaskan melalui surat kabar, televisi, dunia maya, social media dan juga media lainnya. Suatu berita semestinya tercantum komponen 5W+1H yaitu (siapa, apa, kapan, mengapa, dan di mana, serta bagaimana). Berita adalah pemberitahuan yang menerangkan kembali peristiwa yang sebelumnya terjadi. Berita harus autentik dan mengungkapkan fakta peristiwa. Pendapat dapat ditambahkan ke cerita, tetapi hanya jika itu adalah pendapat penulis berita atau pendapat orang lain tentang peristiwa tersebut. Di era teknologi canggih ini, proses penyebaran berita tidak lagi sulit. Tidak seperti di masa lalu, ketika berita menyebar, orang harus mengirim surat atau menulisnya di koran, mencetaknya, dan mendistribusikannya ke publik pada hari berikutnya. Ini akan membuat berita berhenti menjadi fakta. Sekarang, orang cukup menulis pesan singkat di perangkat selulernya, kemudian membagikannya di media sosial. Tanpa menunggu waktu lama, siapapun dapat membaca berita yang telah disebarkan [7].

2.2. *Hoax*

Hoax mengandung arti berita palsu, yaitu berita yang tidak memiliki sumber [8]. *Hoax* merupakan berita dan substansial yang bisa saja menyimpang dari penerimaan seseorang dengan menyebarkan berita palsu sebagai validitas [9]. *Hoax* adalah penyalahgunaan suatu berita yang telah terencana untuk menyampaikan berita palsu. *Hoax* merupakan berita dan substansial yang bisa saja menyimpang dari penerimaan seseorang dengan menyebarkan berita palsu sebagai validitas. *Hoax* itu sendiri bisa dirancang agar memberi pengaruh kepada penerima informasi sehingga si penerima dapat bertindak atas isi *hoax* tersebut [4].

2.3. *Text Mining*

Text mining mendefinisikan data mining dalam bentuk teks. Sumber data umumnya diperoleh dari dokumen. Tujuannya yaitu untuk menemukan istilah yang dapat menempati isi dari dokumen, sehingga dapat menganalisis konektivitas antar dokumen [5]. Konsep dalam *text mining* digunakan sebagai klasifikasi dokumen teks, dimana dokumen yang ada akan diklasifikasikan menurut dokumen yang akan diproses [10].

Dengan adanya konsep ini membuat artikel yang diteliti akan diperjelas kategori jenisnya melalui katakata yang akan muncul dari artikel yang ada. Kata-kata yang terkandung dalam artikel dapat dicocokkan dan dianalisis dengan kata kunci yang telah ditentukan sebelumnya sehingga proses dapat mengklasifikasikan dokumen secara efisien [11]. *Text mining* memiliki tujuan memperoleh informasi yang bermanfaat dan penting dari sekelompok dokumen [12].

2.4. *K-Nearest Neighbor*

KNN merupakan metode yang mengklasifikasikan objek didasarkan pada nilai k dengan melihat jarak terdekat suatu objek berdasarkan data latih atau data uji. Metode tersebut bertujuan untuk mengklasifikasikan objek baru berdasarkan sampel pelatihan dan atribut. Nilai prediksi akan ditentukan berdasarkan hasil klasifikasi tetangga terdekat [6]. Sebelum menggunakan algoritma *K-Nearest Neighbor*, langkah awal yaitu tentukan data latih dan data uji. Setelah itu dilanjutkan dengan proses perhitungan untuk menghitung jarak dengan rumus *Euclidean Distance*. Algoritma ini sangat simpel dan mudah diterapkan serta menyerupai metode *clustering*, yaitu penggolongan data baru berdasarkan jaraknya ke beberapa data atau tetangga terdekat. Pertama, tentukan nilai K tetangga sebelum menghitung jarak data ke tetangga. Kemudian, tentukan jarak antara dua titik, satu titik dalam data latih dan satu titik dalam data uji, menggunakan rumus *Euclidean Distance*. Berikut ini adalah tahapan penggunaan algoritma KNN [13].

1. Menetapkan nilai k.
2. Hitunglah jarak antara data baru dengan seluruh data *training* menggunakan rumus *Euclidean Distance.*

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

   Ket:
   d(x,y) = Jarak data uji dengan data latih
   n = Banyaknya data latih
   x = Data latih
   y = Data uji
3. Urutkan jarak berdasarkan nilai terdekat
4. Lihat kelas k dari tetangga dengan jarak terdekat
5. Kelas pada data baru diambil berdasarkan kelas terbanyak tetangga terdekat.

2.5. Pembobotan Kata

Pembobotan kata bergantung dari total kemunculan setiap token dalam dokumen. Metode untuk memboboti kata yang sangat sering digunakan yaitu *Term Frequency-Inverse Document Frequency (TF-IDF)*. *Term frequency* merupakan banyaknya kemunculan sebuah *term* dalam sebuah dokumen. *Inverse Document Frequency* adalah probabilitas sebuah kata muncul

dalam sekelompok dokumen [4]. Berikut adalah langkah-langkah memboboti kata dengan TF-IDF [14]:

1. Menghitung nilai TF dengan melihat jumlah kata pada sebuah dokumen.
2. Menghitung nilai IDF, rumus IDF = $log\ N/DF$
3. Selanjutnya melakukan proses TF-IDF dengan persamaan 6.

$$tfidf_{(a,b)} = tf_{(a,b)} \times idf_{(a)} \quad (6)$$

Keterangan:

$tf_{(a,b)}$ = Nilai TF kata a dalam dokumen b

$idf_{(a)}$ = Nilai IDF kata a terhadap keseluruhan dokumen

2.6. Evaluasi

Dalam melakukan evaluasi terhadap kinerja saat klasifikasi, khususnya klasifikasi teks, umumnya dilakukan dengan *Confusion Matrix* berupa hasil dengan nilai *accuracy*, *precision*, *recall*, dan *f1-score*. Nilai akurasi menunjukkan seberapa baik seluruh dokumen diklasifikasikan dengan benar. Semakin tinggi nilai akurasinya, maka model tersebut semakin baik dan akurat dalam klasifikasi. [4]. Untuk menghitung *accuracy*, *precision*, *recall*, dan *f1-score* dapat menggunakan rumus dibawah ini.

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

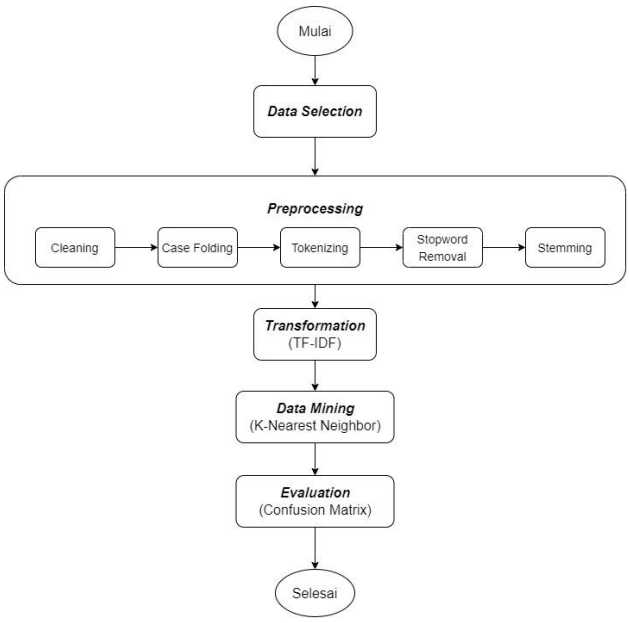
$$Precision = \frac{TP}{TP+FP} \quad (3)$$

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

$$F1-score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (5)$$

3. METODOLOGI

Metode yang diterapkan pada penelitian ini yaitu *Knowledge Discovery in Database*, dimana metode ini terdapat beberapa proses seperti pada alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

3.1. *Data Selection*

Data Selection adalah prosedur penentuan data yang akan digunakan dari sekelompok data. Data terpilih dari tahap ini nantinya akan dipergunakan untuk diolah pada tahap selanjutnya. Data tersebut akan disimpan dalam file dengan ekstensi csv.

3.2. *Preprocessing*

Tahapan *preprocessing* data adalah tahapan menyiapkan data mentah sebelum melakukan proses selanjutnya. Tahap ini membersihkan data yang bertujuan untuk menyatukan struktur kata dan mengurangi banyaknya kata dari sekumpulan dokumen. Pada tahap *preprocessing* akan dilakukan beberapa proses yaitu sebagai berikut.

1. *Cleaning*

   Di tahap ini akan membersihkan atribut yang tidak berkaitan dengan informasi yang terkandung dalam data seperti emoticon, *hashtag*, URL, dan *mention*.
2. *Case folding*

   Di tahapan ini dilakukan proses untuk mengubah keseluruhan karakter atau huruf dalam dokumen agar menjadi huruf kecil.
3. *Tokenizing*

   Tahap ini merupakan tahap setiap kata dalam kalimat dipisahkan dari kata yang membentuknya.
4. *Stopword Removal*

   Tahapan ini adalah tahapan dimana sejumlah kata yang tidak penting akan dihilangkan. Contohnya kata-kata seperti “ke”, “di”, “yang”, “dari”, “di”, “dengan”, “ada”, “akan”, “tidak”, “itu”, “mau”, dan lain-lain.
5. *Stemming*

   Di tahap ini merupakan tahapan mengubah seluruh kata menjadi kata dasar.

3.3. *Transformation*

Pada tahap *transformation* dilakukan proses perubahan pada data yang sebelumnya sudah terpilih sehingga data terpilih tersebut dapat digunakan dalam tahap *data mining*. Dalam tahap ini akan memboboti setiap kata dengan TF-IDF.

3.4. *Data Mining*

Di tahap ini akan dilakukan analisa dengan menggunakan algoritma klasifikasi *K-Nearest Neighbor*. Nilai k yang digunakan dalam proses klasifikasi yaitu 3, 5, 7, 9, dan 11. Dalam tahap ini akan dilakukan *split data* dengan 3 skenario yaitu 70:30, 80:20, dan 90:10. *Split data* menggunakan *train test split* dari *library scikit-learn* dengan parameter *random state*=42.

3.5. *Evaluation*

Penelitian ini menerapkan *Confusion Matrix* berupa hasil dengan nilai *accuracy*, *precision*, *recall* dan *F1-Score*.. Pengukuran atau pengujian algoritma digunakan untuk mengetahui performa dan keoptimalan algoritma.

4. HASIL DAN PEMBAHASAN

4.1. *Data Selection*

Pada tahap *data selection* ini dimulai dengan melakukan pengumpulan data. Data judul berita dikumpulkan secara manual oleh penulis dari beberapa situs online. Data judul berita yang fakta dikumpulkan melalui situs detikcom dan cnnindonesia, sedangkan data judul berita *hoax* dikumpulkan melalui situs turnbackhoax. Data yang diambil merupakan data pada tanggal 24 November 2021 – 22 Januari 2022 yang berjumlah 75 judul berita fakta dan 75 judul berita *hoax*. Pada awalnya data memiliki empat atribut yaitu tanggal, judul, sumber, dan label, tetapi hanya 2 atribut yang akan diolah yaitu judul dan label. Data yang sudah dikumpulkan ditunjukkan pada tabel 1.

Tabel 1. Data Judul Berita

No	Judul	Label
1	Jorok! Sungai di Kopo Sayati, Bandung Dipenuhi Sampah	Fakta
2	Booster Vaksin Sinovac Bisa Pakai AstraZeneca Pfizer, Ini Kombinasinya	Fakta
3	Ekstradisi Diteken, Bagaimana Kabar Tersangka KPK Tannos di Singapura	Fakta
4	Razia STNK, Kendaraan yang Terlambat Membayar Pajak akan Dikandangkan	Hoax
5	Atlet Tenis Dalila Jakupovic Mengalami Kesulitan Bernapas saat Babak Kualifikasi Australia Open karena Vaksin Covid-19	Hoax
6	Bill Gates akan Menarik Semua Peredaran Vaksin Covid-19	Hoax

4.2. *Preprocessing*

Data yang berhasil dikumpulkan masih berbentuk data mentah sehingga dibutuhkan tahap *preprocessing* agar data tersebut dapat diolah pada tahapan selanjutnya. Pada tahap *preprocessing* ini dilakukan 5 tahap sebagai berikut.

1. *Cleaning*

Tahap *cleaning* yaitu dilakukan penghapusan karakter yang tidak berpengaruh terhadap klasifikasi. Gambar 2 merupakan hasil data yang telah selesai melalui proses *cleaning*.

	Judul	Label
0	Jorok Sungai di Kopo Sayati Bandung Dipenuhi S...	Fakta
1	Booster Vaksin Sinovac Bisa Pakai AstraZeneca ...	Fakta
2	Ridawan Kamil Jabar Lampu Kuning Omicron	Fakta
3	Ekstradisi Diteken Bagaimana Kabar Tersangka K...	Fakta
4	Antisipasi Omicron Pemkab Garut Siagakan Rumah...	Fakta
...	...	...
145	Jepang Menggunakan Ivermectin dan Menghentikan...	Hoax
146	WHO Mengatakan bahwa Menyimpan Ponsel Dekat Ke...	Hoax
147	Aliansi Dokter Dunia Mengatakan Bahwa Varian D...	Hoax
148	Bawang Putih Ampuh Atasi Hidung Tersumbat	Hoax
149	Jalur Wisata Gunung Bromo Amblas	Hoax

Gambar 2. Hasil Proses *Cleaning*

2. *Case Folding*

Pada tahap ini mengubah keseluruhan huruf menjadi *lowercase* (huruf kecil). Gambar 3 menunjukkan hasil proses *case folding*.

	Judul	Label
0	jorok sungai di kopo sayati bandung dipenuhi s...	Fakta
1	booster vaksin sinovac bisa pakai astrazeneca ...	Fakta
2	ridawan kamil jabar lampu kuning omicron	Fakta
3	ekstradisi diteken bagaimana kabar tersangka k...	Fakta
4	antisipasi omicron pemkab garut siagakan rumah...	Fakta
...	...	...
145	jepang menggunakan ivermectin dan menghentikan...	Hoax
146	who mengatakan bahwa menyimpan ponsel dekat ke...	Hoax
147	aliansi dokter dunia mengatakan bahwa varian d...	Hoax
148	bawang putih ampuh atasi hidung tersumbat	Hoax
149	jalur wisata gunung bromo amblas	Hoax

Gambar 3. Hasil Proses *Case Folding*

3. *Tokenizing*

Pada tahap ini dilakukan pemisahan kalimat menjadi kata-kata penyusunnya. Gambar 4 menunjukkan hasil dari proses *tokenizing*.

	Judul	Label
0	[jorok, sungai, di, kopo, sayati, bandung, dip...	Fakta
1	[booster, vaksin, sinovac, bisa, pakai, astraz...	Fakta
2	[ridawan, kamil, jabar, lampu, kuning, omicron]	Fakta
3	[ekstradisi, diteken, bagaimana, kabar, tersan...	Fakta
4	[antisipasi, omicron, pemkab, garut, siagakan,...	Fakta
...	...	...
145	[jepang, menggunakan, ivermectin, dan, menghen...	Hoax
146	[who, mengatakan, bahwa, menyimpan, ponsel, de...	Hoax
147	[aliansi, dokter, dunia, mengatakan, bahwa, va...	Hoax
148	[bawang, putih, ampuh, atasi, hidung, tersumbat]	Hoax
149	[jalur, wisata, gunung, bromo, amblas]	Hoax

Gambar 4. Hasil Proses *Tokenizing*

4. *Stopword Removal*

Pada tahap *stopword removal* dilakukan penghapusan kata yang tidak penting menggunakan *stopword* Indonesia dari *library* NLTK. Gambar 5 merupakan hasil dari proses *stopword removal*.

	Judul	Label
0	[jorok, sungai, kopo, sayati, bandung, dipenuh...	Fakta
1	[booster, vaksin, sinovac, pakai, astrazeneca,...	Fakta
2	[ridawan, kamil, jabar, lampu, kuning, omicron]	Fakta
3	[ekstradisi, diteken, kabar, tersangka, kpk, t...	Fakta
4	[antisipasi, omicron, pemkab, garut, siagakan,...	Fakta
...	...	...
145	[jepang, ivermectin, menghentikan, vaksin, mem...	Hoax
146	[who, menyimpan, ponsel, kepala, menyebabkan, ...	Hoax
147	[aliansi, dokter, dunia, varian, delta]	Hoax
148	[bawang, putih, ampuh, atasi, hidung, tersumbat]	Hoax
149	[jalur, wisata, gunung, bromo, amblas]	Hoax

Gambar 5. Hasil Proses *Stopword Removal*

5. *Stemming*

Pada tahap *stemming* ini dilakukan pengubahan kata ke dalam bentuk kata dasar. Hasil dari proses *stemming* ditunjukkan pada Gambar 6.

	Judul	Label
0	[jorok, sungai, kopo, sayat, bandung, penuh, s...	Fakta
1	[booster, vaksin, sinovac, pakai, astrazeneca,...	Fakta
2	[ridawan, kamil, jabar, lampu, kuning, omicron]	Fakta
3	[ekstradisi, teken, kabar, sangka, kpk, tannos...	Fakta
4	[antisipasi, omicron, pemkab, garut, siaga, ru...	Fakta
...	...	...
145	[jepang, ivermectin, henti, vaksin, berantas, ...	Hoax
146	[who, simpan, ponsel, kepala, sebab, tumor]	Hoax
147	[aliansi, dokter, dunia, varian, delta]	Hoax
148	[bawang, putih, ampuh, atas, hidung, sumbat]	Hoax
149	[jalur, wisata, gunung, bromo, amblas]	Hoax

Gambar 6. Hasil Proses *Stemming*

4.3. *Transformation*

Dalam tahap *transformation* ini awalnya akan dilakukan split data sesuai dengan skenario yang telah ditentukan, setelah itu dilakukan pembobotan setiap kata menggunakan TF-IDF. Gambar 7 menunjukkan hasil dari pembobotan kata judul berita.

	abad	adam	aff	ahli	ahmad	air	airlines	akun	alam	alami	...	warna
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
130	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
131	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
132	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
133	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
134	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

Gambar 7. Hasil TF-IDF

4.4. *Data Mining*

Dalam tahap *data mining* dilakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor*. Dalam melakukan klasifikasi dengan KNN digunakan nilai k 1, 3, 5, 7, 9, dan 11. Tabel 2 menunjukkan hasil dari jumlah data *training* dan data testing untuk 3 skenario pembagian data.

Tabel 2. Skenario Pembagian Data

Skenario	Data Training	Data Testing
Skenario 1 (90:10)	135	15
Skenario 2 (80:20)	120	30
Skenario 3 (70:30)	105	45

Hasil dari klasifikasi dengan 3 skenario yaitu 90:10, 80:20, dan 70:30 dapat dilihat pada tabel 3.

Tabel 3. Hasil Klasifikasi KNN

Skenario	Nilai k	Akurasi
Skenario 1 (90:10)	1	80%
	3	**93,33%**
	5	86,67%
	7	86,67%
	9	66,67%
	11	66,67%
Skenario 2 (80:20)	1	70%
	3	80%
	5	80%
	7	73,33%
	9	73,33%
	11	66,67%
Skenario 3 (70:30)	1	75,56%
	3	73,33%
	5	71,11%
	7	75,56%
	9	73,33%
	11	71,11%

4.5. *Evaluation*

Evaluation dilakukan terhadap model dengan menggunakan *Confusion Matrix*. Hasil *accuracy, precision, recall*, dan *f1-score* dengan skenario 1 (90:10) ditunjukkan pada tabel 4.

Tabel 4. Hasil Evaluasi Skenario 90:10

Nilai k	Akurasi	Precision	Recall	F1-score
1	80%	66,67%	80%	72,73%
3	**93,33%**	**100%**	**80%**	**88,89%**
5	86,67%	80%	80%	80%
7	86,67%	80%	80%	80%
9	66,67%	50%	60%	54,55%
11	66,67%	50%	40%	44,44%

Hasil *accuracy, precision, recall*, dan *f1-score* dengan skenario 2 (80:20) ditunjukkan pada tabel 5.

Tabel 5. Hasil Evaluasi Skenario 80:20

Nilai k	Akurasi	Precision	Recall	F1-score
1	70%	68,75%	73,33%	70,97%
3	80%	90,91%	66,67%	76,92%
5	80%	84,62%	73,33%	78,57%
7	73,33%	70,59%	80%	75%
9	73,33%	68,42%	86,67%	76,47%
11	66,67%	63,16%	80%	70,59%

Hasil *accuracy, precision, recall*, dan *f1-score* dengan skenario 2 (70:30) ditunjukkan pada tabel 6.

Tabel 6. Hasil Evaluasi Skenario 70:30

Nilai k	Akurasi	Precision	Recall	F1-score
1	75,56%	72,73%	76,19%	74,42%
3	73,33%	66,67%	85,71%	75%
5	71,11%	66,67%	76,19%	71,11%
7	75,56%	72,73%	76,19%	74,42%
9	73,33%	66,67%	85,71%	75%
11	71,11%	66,67%	76,19%	71,11%

Berdasarkan tabel 4, tabel 5, dan tabel 6, dapat dilihat bahwa hasil terbaik diperoleh pada skenario pertama dengan rasio (90:10) menggunakan nilai k=3 dengan hasil akurasi sebesar 93,33 %, *precision* sebesar 100%, *recall* sebesar 80%, dan *f1-score* sebesar 88,89%. Sedangkan untuk hasil terendah didapatkan pada skenario 1 (90:10) menggunakan nilai k=11 dengan hasil akurasi sebesar 66,67%, *precision* sebesar 50%, *recall* sebesar 40%, dan *f1-score* sebesar 44,44 %.

5. KESIMPULAN DAN SARAN

Adapun beberapa hal yang bisa disimpulkan dari penelitian ini yaitu:

1. Implementasi algoritma *K-Nearest Neighbor* dalam klasifikasi judul berita hoax menghasilkan hasil yang sangat baik dengan akurasi tertinggi sebesar 93,33%, *precision* sebesar 100%, *recall* sebesar 80%, dan *f1-score* sebesar 88,89% sehingga dapat dikatakan bahwa algoritma *K-Nearest Neighbor* cocok untuk mengklasifikasikan judul berita *hoax*.
2. Penggunaan skenario dalam melakukan proses klasifikasi dengan skenario 1 (90:10), skenario 2 (80:20), dan skenario 3 (70:30) ternyata mempengaruhi hasil akurasi.

Untuk penelitian yang akan dilakukan selanjutnya, ada beberapa saran yaitu:

1. Menambah *dataset* yang digunakan untuk melakukan penelitian sehingga dapat menghasilkan tingkat akurasi yang lebih akurat.
2. Menggunakan algoritma atau metode klasifikasi yang lain untuk mengetahui apakah ada algoritma yang lebih baik dalam melakukan klasifikasi judul berita *hoax*.
3. Menerapkan model klasifikasi judul berita *hoax* ke dalam sebuah sistem.

DAFTAR PUSTAKA

[1] R. Sagita, U. Enri, and A. Primajaya, "Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN)," *JOINS (Journal Inf. Syst.*, vol. 5, no. 2, pp. 230–239, 2020, doi: 10.33633/joins.v5i2.3705.

[2] D. A. Fauziah, A. Maududie, and I. Nuritha, "Klasifikasi Berita Politik Menggunakan Algoritma K-nearst Neighbor," *Berk. Sainstek*, vol. 6, no. 2, p. 106, 2018, doi: 10.19184/bst.v6i2.9256.

[3] B. K. Palma, D. T. Murdiansyah, and W. Astuti, "Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K- Nearest Neighbor," vol. 8, no. 5, pp. 10637–10649, 2021.

[4] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo J. Inf. Technol.*, vol. 1, no. 1, p. 1, 2019, doi: 10.21580/wjit.2019.1.1.3915.

[5] S. Ernawati and R. Wati, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel," *J. Khatulistiwa Inform.*, vol. VI, no. 1, pp. 64–69, 2018, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/khatulistiwa/article/view/3802/2626.

[6] A. Nugraha, pratama dwi,, Said al faraby, "Klasifikasi Dokumen Menggunakan Metode k-Nearest Neighbor (kNN) Dengan Information Gain," *eProceedings Eng.*, vol. 5, no. 1, pp. 1541–1550, 2018.

[7] G. M and V. Mesyura, "Fake news detection using naive Bayes classifier," *First Ukr. Conf. Electr. Comput. Eng.*, 2017.

[8] A. Rahmadhany, A. Aldila Safitri, and I. Irwansyah, "Fenomena Penyebaran Hoax dan Hate Speech pada Media Sosial," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 3, no. 1, pp. 30–43, 2021, doi: 10.47233/jteksis.v3i1.182.

[9] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *J. Ilmu Komput. Agri-Informatika*, vol. 5, pp. 1–10, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jika.

[10] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter BMKG Nasional," *J. Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.

[11] C. Dhaneswara, Y. Azhar, and N. Hayatin, "Deteksi Berita Hoax Pada Dokumen Berbahasa Indonesia Menggunakan Metode Modified K - Nearest Neighbor," *Semin. Nas. Teknol. dan Rekayasa 2020*, pp. 165–170, 2020.

[12] D. Ariyanti and K. Iswardani, "Teks Mining untuk Klasifikasi Keluhan Masyarakat Pada Pemkot Probolinggo Menggunakan Algoritma Naïve Bayes," *J. IKRA-ITH Inform.*, vol. 4, no. 3, pp. 125–132, 2020.

[13] A. Rahmat Dian Nugraha, K. Auliasari, and Y. Agus Pranoto, "Implementasi Metode K-Nearest Neighbor (KNN) Untuk Seleksi Calon Karyawan Baru (Studi Kasus : BFI Finance Surabaya)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 4, no. 2, pp. 14–20, 2020, doi: 10.36040/jati.v4i2.2656.

[14] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," 2018, [Online]. Available: http://arxiv.org/abs/1806.06407.

[15] Simanjuntak, P., & Elisa, E. (2019). Data Mining Untuk Menentukan Pemilihan Celular Card Di Kota Batam. *Journal Information System Development (ISD)*, *4*(2).

BIODATA PENULIS

Muhammad Diki Hendriyanto
Mahasiswa Teknik Informatika, Universitas Singaperbangsa Karawang, Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Timur, Karawang, Jawa Barat 41361.

Betha Nurina Sari
Dosen Teknik Informatika, Universitas Singaperbangsa Karawang, Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Timur, Karawang, Jawa Barat 41361.