Klasifikasi Teks Artikel Berita Hoaks Covid-19 dengan Menggunakan Algotrima K-Nearest Neighbor

Berlian Kaida Palma[1], Danang Triantoro Murdiansyah[2], Widi Astuti[3]
[1,2,3] Universitas Telkom, Bandung
[1]berliankaidap@students.telkomuniversity.ac.id, [2]danangtri@telkomuniversity.ac.id, [3]widiwdu@telkomuniversity.ac.id

Abstrak

Peran internet dan pertumbuhan informasi yang diberitakan di media sosial membuat perkembangan dan penyebaran berita semakin mudah, begitu pun dalam mengaksesnya. Pada masa pandemi Covid-19 saat ini banyak sekali berita yang tersebar sehingga masyarakat luas mencari atau mendapat informasi tentang virus ini. Berita berjudul Covid-19 ini banyak berisi informasi tidak penting bahkan memberitakan informasi hoaks. Hal ini membuat masyarakat internasional khususnya Indonesia resah akan berita yang beredar selama masa Covid-19. Oleh karena itu, penulis membuat sebuah model sistem untuk melakukan klasifikasi berita yang sesuai terjadi di lapangan. Informasi yang tersebar di media sosial sangat variatif sehingga banyak berita yang tidak penting bahkan berisikan informasi hoaks. Klasifikasi berita akan dilakukan dengan *K-Nearest Neighbor* (*KNN*). Berita yang ada dibagi menjadi beberapa kelas berdasarkan kategorinya, kemudian berita dilakukan klasifikasi teks dengan metode *K-Nearest Neighbor* (*KNN*) dan *k-fold cross validation* sebagai validasi model yang dibuat. Proses klasifikasi dilakukan dengan skema menggunakan 80% data train dan 20% data test serta mengubah parameter nilai k pada K-Nearest Neighbor dengan k = 3, k = 5, k = 7, k = 9, dan pada *k-fold cross validation* sebanyak k = 5 dan k = 10. Untuk evaluasi digunakan *confusion matrix*. Akhirnya, dari setiap model yang dilakukan dengan mengubah nilai k pada *K-Nearest Neighbor* didapatkan hasil akurasi terbaik dengan *F1-Score* sebesar 48% dari nilai k = 5, hasil validasi dari *k-fold cross validation* k = 5 sebesar 42% dan k = 10 sebesar 45%.

Kata Kunci : Covid-19, *K-Nearest Neighbor*, Hoax, intenet, klasifikasi.

Abstract

The role of the internet and the growth of information that is reported on social media makes the development and dissemination of news easier, as well as in accessing it. During the current Covid-19 pandemic, a lot of news is spreading so that the wider community is looking for or getting information about this virus. This news entitled Covid-19 contains a lot of unimportant information and even reports hoax information. This makes the international community, especially Indonesia nervous about the news circulating during the Covid-19 period. Therefore, the author makes a system model to classify news according to what happens in the field. The information spread on social media is so varied that many unimportant news even contain hoax information. News classification will be done with K-Nearest Neighbor (KNN). The existing news is divided into several classes based on the category, then the news is classified as text using the K-Nearest Neighbor (KNN) method and k-fold cross validation as validation of the model created. The classification process is carried out using a scheme using 80% train data and 20% test data and changing the parameter value of k on K-Nearest Neighbor with k = 3, k = 5, k = 7, k = 9, and on k-fold cross validation as much as k = 5 and k = 10. Confusion matrix is used for evaluation. Finally, from each model that was carried out by changing the value of k on the K-Nearest Neighbor, the best accuracy results were obtained with an F1-Score of 48% from the value of k = 5, the validation results of the k-fold cross validation k = 5 of 42% and k = 10 by 45%.

Keywords: Covid-19, K-Nearest Neighbor, Hoax, internet, classification.

1. Pendahuluan

1.1 Latar Belakang

Berita adalah informasi hangat dan menarik para pembaca yang disampaikan melalui media sosial. Dengan cepatnya perkembangan teknologi dan informasi, internet menjadi sumber utama persebaran informasi. Informasi yang tersebar tidak tersaring sehingga siapa pun dapat menyebarkan berita yang tidak jelas isinya bahkan berisi informasi hoaks yang memiliki arti "berita bohong". Berita yang tidak jelas dan hoaks dapat membuat kegaduhan dalam masyarakat apalagi konten yang tersebar merupakan konten yang sensitif yang menyangkut nyawa manusia, terutama selama masa pandemi Covid-19. Oleh karena itu penulis membuat sebuah model sistem klasifikasi dengan menggunakan metode *K-Nearest Neighbor* yang merupakan algoritma supervised learning. Algoritma *K-Nearest Neighbor* dapat mengklasifikasikan berita pada artikel berdasarkan atribut dan sample yang ada pada data. *K-Nearest Neighbor* (*KNN*) merupakan salah satu metode klasifikasi yang sering digunakan 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 dimasa pandemi Covid-19 secara cepat[18]. Model sistem klasifikasi yang dibuat oleh penulis ditujukan untuk menguji metode *K-Nearest Neighbor* pada studi kasus persebaran berita dimasa Covid-19 yang sudah tersebar di internet, pengujian juga akan disempurkan dengan *k-fold cross validation* sebagai validasi dan optimasi performansi sistem klasifikasi yang telah dibuat[11]. Alasan lain mengapa penulis mengunakan *K-Nearest Neighbor* (*KNN*) karena merupakan algoritma klasifikasi dengan performa tinggi namun sederhana[15].

1.2 Topik dan Batasan

Berdasarkan latar belakang yang telah disampaikan sebelumnya, maka rumusan masalah ialah bagaimana mengimplementasikan algoritma *K-Nearest Neighbor* (*KNN*) untuk mengklasifikasikan artikel berita hoaks, serta bagaimana performa sistem yang dibangun untuk mengklasifikasikan artikel berita hoaks dengan menggunakan *K-Nearest Neighbor* (*KNN*).

1.3 Tujuan

Tujuan yang dicapai dalam tugas akhir ini ialah membuat sebuah sistem yang dapat mengklasifikan sebuah data berita artikel yang tidak jelas isinya atau terdapat informasi hoaks didalamnya tentang Covid-19 dengan metode *K-Nearest Neighbor* serta mengukur performansi dari sistem yang telah dibuat.

2. Studi Terkait

Penelitian terkait klasifikasi berita hoax tentang Covid-19 masih jarang dilakukan karena wabah yang masih tergolong baru terjadi, akan tetapi penelitian yang terkait dengan berita hoax sudah cukup banyak. Salah satunya adalah [1], melakukan deteksi hoax dengan metode SVM dan SGD dengan kombinasi TF-IDF. Berdasarkan hasil penelitian yang didapat terbukti penggunaan SGD lebih baik 4%-20% daripada SVM.

Selanjutnya penelitian terkait berita hoax pilpres dengan menggunakan metode Modified K-Nearest Neighbor dan pembobotan dengan TF-IDF [4]. Pada penelitian tersebut, dapat disimpulkan dengan hasil yang tidak begitu baik disebabkan beberapa faktor yaitu: data yang diuji masih sedikit, sumberdaya yang masih sedikit dan kurangnya optimasi pada algoritma. Oleh karena itu, hasil pengujian bisa dikatakan kurang maksimal dengan nilai precission 93.75%, recall 90.90%, accuracy 92.31% dan f-measure 92.31%.

Berikutnya [2], dimana penulis melakukan penelitian tentang klasifikasi berita hoax dengan menggunakan klasifikasi naïve bayes dalam bahasa Indonesia. Dalam penelitiannya, data diambil melalui web crawler php-ml. Dengan melakukan 3 kali percobaan dengan pembagian data 70:30, 80:20 dan 60:40 didapatkan hasil yang berbeda, akan tetapi rata-rata hasil terbaik didapatkan dari pembagian data train/test pada 70:30, dengan accuracy 78,6%, hoax precission 67.1%, valid precission 91.6%, hoax recall 89.4% dan valid recall 71.4%.

Lalu pada penelitian [6], dimana penulis melakukan penelitian tentang pendeteksi hoax dari berita Indonesia menggunakan KNN dan TF-TDM. Dari 74 berita dan 74 hoax yang di deteksi, akurasi yang didapatkan mencapai 83.6%. Akan tetapi ini bukanlah hasil maksimal karena adanya kesalahan hal ini karena ada ada false-positive detection yang mempengaruhi akurasi.

2.1. Berita & Hoaks

Menurut *Kamus Besar Bahasa Indonesia* (*KBBI*), berita merupakan informasi mengenai suatu kejadian atau peristiwa yang hangat[14]. Hoaks menurut *Kamus Besar Bahasa Indonesia* (*KBBI*) adalah berita bohong[16]. Hoaks adalah berita yang sengaja di manipulasi dengan tujuan untuk memberikan pemahaman yang salah (Dahlan,2017) [7]. Secara garis besar, hoaks merupakan berita bohong yang sengaja disebar di media sosial untuk mencari suatu sensasi atau masalah lainnya sehingga dapat meresahkan masyarakat.

2.2. TF-IDF

Term-Frequency atau *Invers Document Frequency* (TF-IDF) [4] adalah metode statistik numerik yang merepresentasikan pentingnya suatu kata dalam sebuah dokumen pada corpus. *Term-Frequency* merupakan jumlah frekuensi yang teradapat pada suatu dokumen sedangkan *Invers Document Frequency* merupakan invers jumlah dokumen yang memuat suatu term atau kata. Rumus pembobotan TF-IDF dapat dilihat dengan persamaan (3) yang diperoleh dari[4]:

- TF :

$$W(d,t) = TF(d,t) \quad (1)$$

$TF(d,t)$ adalah frekuensi dari t kata yang ada pada teks d.

- IDF:

$$idft = \log 10 \frac{N}{dft} \quad (2)$$

dft merupakan banyaknya dokumen yang memuat t dan N adalah jumlah total dokumen.

- Pembobotan TF-IDF
  Pembobotan disini merupakan perkalian antara tft dan $idft$, diketahui rumus dari TD-IDF ialah (3) :

$$Wt,d = \frac{Wt,d}{\sqrt{\sum_{t=1}^{n} Wt,d^2}} (3)$$

2.3. K-Nearest Neighbor

K-Nearest Neighbor (*KNN*) merupakan metode klasifikasi pada sekumpulan data berdasarkan pembelajaran data. *K-Nearest Neighbor* supervised learning, dan pada alogoritma ini akan dicari sejumlah k objek terdekat dengan data yang akan dilakukan klasifikasi, lalu data akan digolongan kedalam suatu kategori dengan voting berdasarkan probabilitas yang paling tinggi [5]. Dalam melakukan perhitungan jarak antar data menggunakan metode Euclidean distance dengan persamaan (4) yang diperoleh dari [17]:

$$dist = \sum_{i=1}^{p} \sqrt{(x2 - x1)^2} \quad (4)$$

- Euclidean dsitance:

  $dist$: Jarak

  $x1$: Data Latih

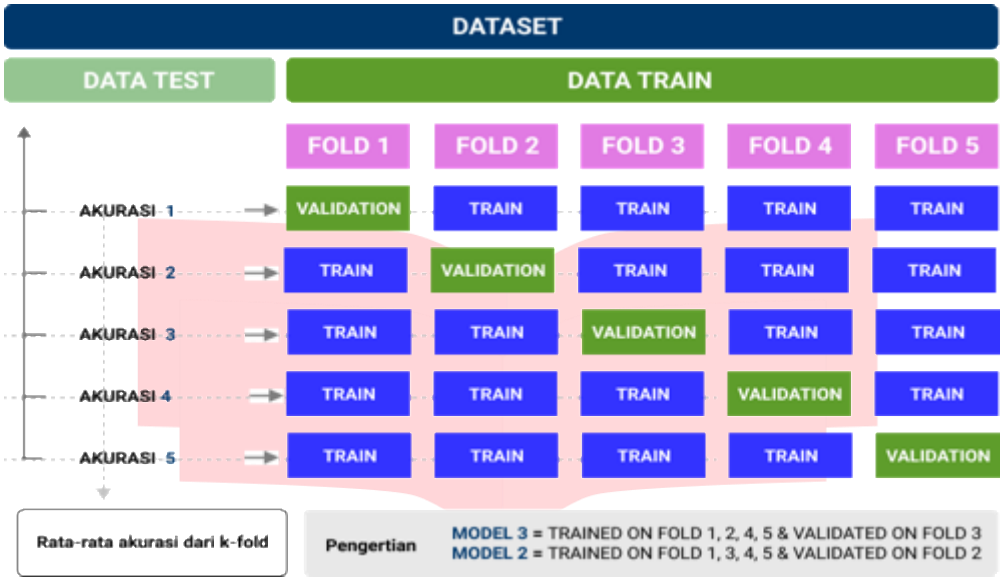
  $x2$: Data Test

  i : Indeks Penjumlahan

  p : Jumlah Atribut

2.4. Cross Validation

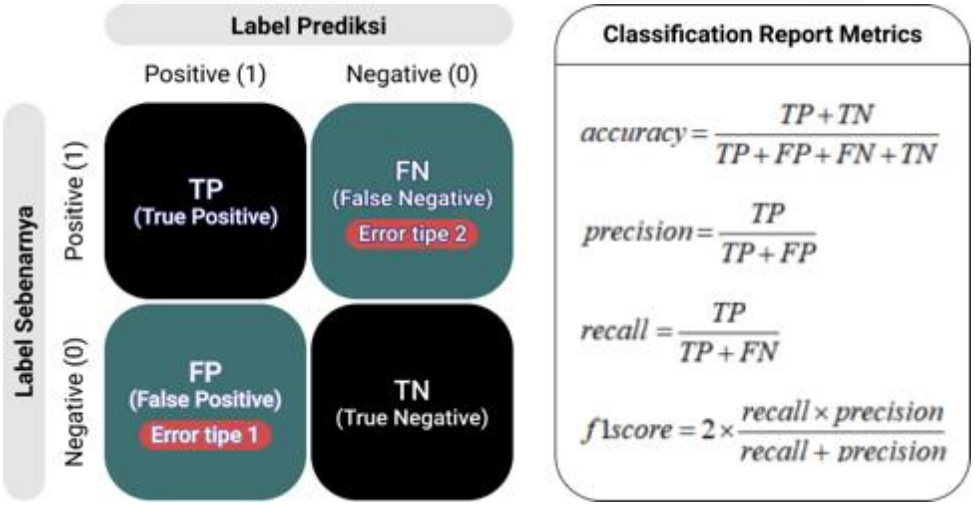
Dengan metode validasi *k-fold cross validation*, metode yang mainstream untuk melakukan optimasi performansi dari suatu algoritma klasifikasi maupun untuk dibandingkan dengan sebuah dataset. Pada tahap ini, data akan dibagi menjadi dua subset yaitu data train dan data vailidasi, *k-fold cross validation* membagi data sebanyak k dan akan melakukan perulangan sebanyak k, sehingga model yang dihasilkan akan sebanyak k dimana biasanya k = 5 dan k = 10 [11]. Berikut contoh proses *k-fold cross validation* pada Gambar 1:



Gambar 1 K-fold Cross Validation

2.5. Confusion Matrix dan Classification Report

Klasifikasi merupakan bagian dari *supervised learning*. Evaluasi performa merupakan langkah penting dalam Machine Learning atau Data Learning. Dalam melakukan evaluasi model klasifikasi penulis menggunakan *confusion matrix* dan *classification report*. Berikut penjelasannya pada Gambar 2:



Gambar 2 Confusion Matrix &Classification Report

1. Confusion Matrix
   Merupakan tabel N x N (N jumlah kelas/label/kategori) berisi nilai True dan False dari model klasifikasi yang dibuat, tujuannya adalah membandingkan nilai aktual dengan nilai prediksinya. Pada tabel matrix baris merupakan nilai kelas yang sebenarnya, sedangkan kolom adalah nilai prediksinya. Dalam *confusion matrix* terdapat 4 kategori nilai:

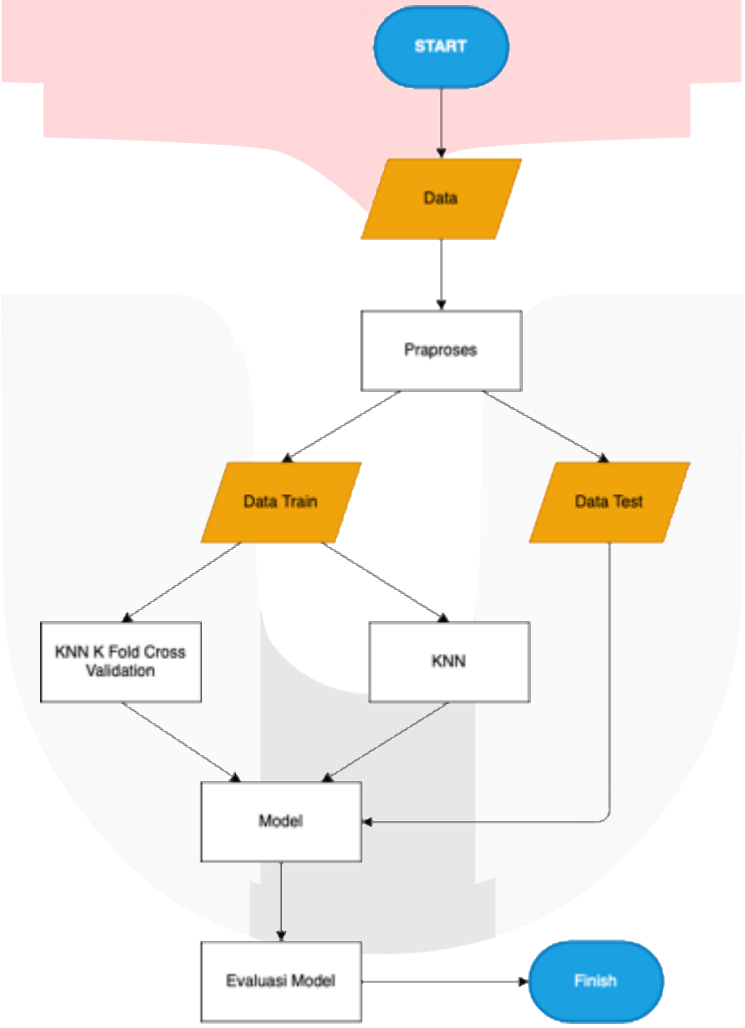
- True Positif (TP) : Prediksi Positif dan nilai aktualnya memang Positif.
- True Negatif (TN) : Prediksi Negatif dan nilai aktualnya Negatif.
- False Positif (FP) : Prediksi Positif dan nilai aktualnya Negatif.
- False Negatif (FN) : Prediksi Negatif dan nilai aktualnya Positif

2. Classification Report
   Hasil evaluasi yang dihasilkan sulit untuk diukur seberapa akurat model klasifikasi yang telah dibuat. Oleh karena itu data dari Confusion Matrix akan digunakan untuk menghitung nilai-nilai yang disimpan pada *classification report* (nilai kinerja dari model klasifikasi yang dibuat)[12]. Berikut nilai-nilai *classification report* yang dugunakan:

- *Accuracy* : kesesuaian hasil prediksi pengujian dengan nilai yang sebenarnya.
- *Precissio*n : ketepatan antara data yang user minta dengan hasil yang diberikan sistem.
- *Recall* : jumlah kesesuaian data dari hasil percobaan berdasarkan sudut pandang kelas yang digunakan.
- *F1-Score* : Gambaran rata-rata atau harmonic mean pada *Precission* dan *Recall.*

3. Sistem yang Dibangun

Pada bagian ini akan menjelaskan bagaimana perancangan sistem yang diterapkan, dapat dilihat pada flowchart dimana secara garis besar sistem berjalan dari input data, preprocessing pembobotan TF-IDF,training dengan *KNN*, *KNN* dengan *cross fold validation*, dan evaluasi yang telah dibuat oleh penulis. Berikut alur perancangan sistem pada Gambar 3:



Gambar 3 Alur Sistem

3.1 . Dataset

Dataset yang digunakan ialah Relevance & Stance Detection merupakan dataset model yang dibangun untuk mendapatkan analisis relevansi antara kueri dengan beberapa dokumen pembanding. Sebuah dokumen dikatakan relevan apabila memiliki kesamaan makna secara semantic dengan kueri yang akan dilakukan validasi kebenaran atau kebohongannya, Terdapat 4 label untuk kueri utama yaitu ***Fact, Hoax, Unknown, dan Unimportant,*** dan 4 juga untuk

label pembanding ***yaitu Support, Oppose, NEI,*** Irrelevant. Dengan total 15025 row data awal dan 9517 row data yang diproses. Berikut aturan dalam penggunaannya :

Tabel 1. Tabel Contoh Relevansi Kueri dan Dokumen Pembanding

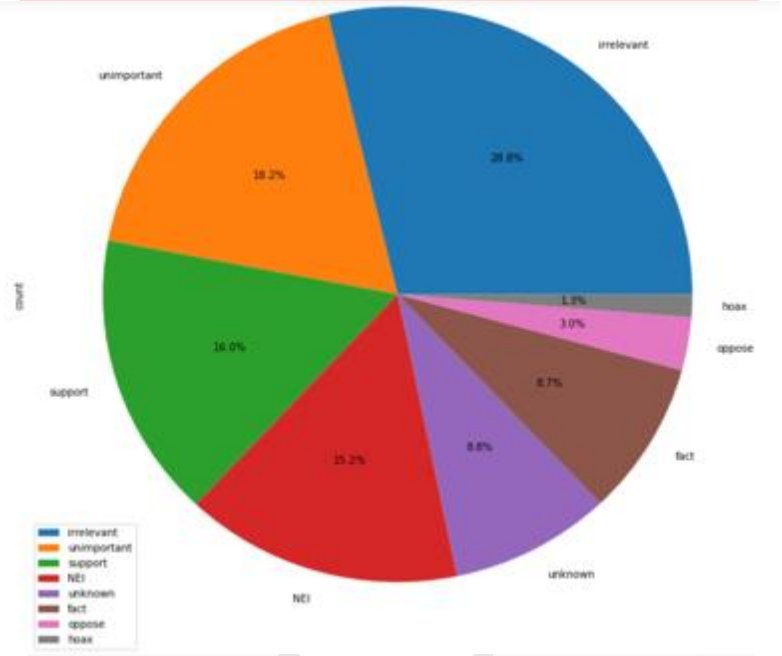
Label Kueri Utama	Penjelasan	Contoh
Fact	Fact adalah label yang digunakan untuk menandai kueri utama (kueri nomor 1) pada sebuah dokumen yang bernilai fakta.	- Bima Arya Positif Corona
Hoax	Hoax adalah label yang digunakan untuk menandai kueri utama (nomor 1) pada sebuah dokumen yang bernilai hoaks/bohong.	- Pasien Cimahi kabur
Unknown	Unkown adalah label yang digunakan untuk menandai kueri utama (nomor 1) pada sebuah dokumen yang tidak dikethaui nilai fakta atau bohongnya sekalipun sudah selancar di internet.	- Ramuan jahe dan lada bisa mencegah Corona
Unimportant	Unimportant adalah label yang digunakan untuk menandai kueri utama yang mengandung informasi tidak penting, kueri tidak penting diantaranya bisa berupa kueri yang tidak memiliki gagasan utama, hanya berisi subjek, kalimat harapan, dan ujaran kebencian. Pada kueri utama berlabel *unimportan*t, artikel-artikel pembanding yang ada tidak perlu diberikan label apa pun.	- Corona pergi ke laut aja

Label Kueri Pembanding	Penjelasan	Contoh
Support	Support adalah label yang digunakan untuk menandai kueri pembanding yang memiliki nilai relevean dengan kueri utama (nomor 1) dan mendukung informasi dalam kueri utama.	- Bima Arya Positif corona ➔Walikota Bogor Bima Arya Sugiarto sudah beraktifitas seperti biasa, setelah pulih dari infeksi Covid-19.
Oppose	Oppose adalah label yang digunakan untuk menandai kueri pembanding yang memiliki nilai relevan dengan kueri utama (nomor 1), namun menentang informasi dalam kueri utama.	- Pasien di Cimahi kabur ➔beredar informasi melalui WhatsApp yang menyebutkan seorang pasien positif Corona atau Covid-19 kabur dari perawatan di RSUD Cibabat, Kota Cimahi, Jawa Barat Pemkot Cimahi buka suara, kepala Dinas Komunikasi Informasi Arsip dan Perspustakaan (Diskominfoarpus) Kota CImahi Harjono menyebutkan pasien positif kabur dari RSUD Cibabat itu tidak benar.
NEI	NEI atau Not Enough Information adalah label yang digunakan untiuk menandai kueri pembanding yang memiliki nilai relevan dengan kueri utama (nomor 1), namun tidak memiliki cukup informasi apakah kueri ini memiliki nilai mendukung atau menentang.	- Pasien di Cimahi kabur ➔Jumlah kasus terkonfirmasi positif Covid-19 di Kota Cimahi mencapai 88 orang. Jumlah ini terdiri atas 43 positif aktif, 42 sembuh, dan 3 orang meninggal dunia. Perkembangan ini dianggap cukup bagus apalagi ditambah data bahwa jumlah pasien yang sembuh tersebut sudah menyamai jumlah pasien positif. Lebih dari itu, serta tidak ada angka laporan kematian terbaru. “Sekarang pasien yang sembuh hampir menyamai pasien

		positif padahal awalhnya jauh dibawah.
irrelevant	Irrelevant adalah label yang digunakan untuk menandai kueri pembanding yang memiliki nilai tidak relevan dengan kueri utama (kueri nomor 1).	- Bima Arya Positif corona ➔ **Jakarta** – Wali Kota Bogor Bima Arya mengisi akhir pekan

Jumlah data berdasarkan kelasnya sebelum masuk ke tahap preprocessing :

Irrelevant : 2741
unimportant : 1732
support : 1524
NEI : 1449
unknown : 840
fact : 826
oppose : 284
hoax : 121

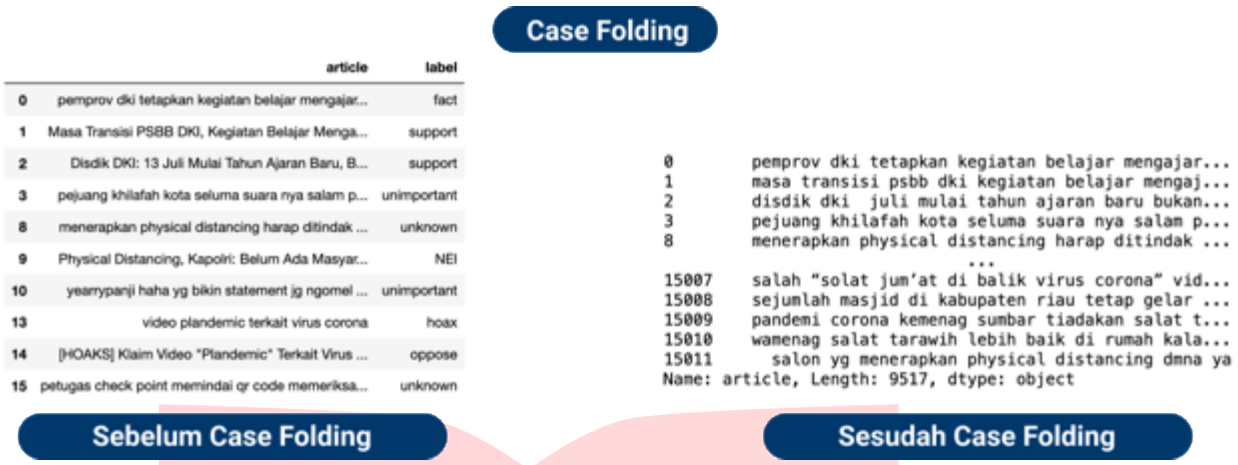


Gambar 4 Persebaran Data

3.2. Preprocessing

Text processing merupakan tahapan ekstraksi dokumen menjadi data yang akan digunakan pada tahapan selanjutnya. Pada bagian ini, data akan menjadi sebuah potongan-potongan atau token, selain itu akan dihilangkannya simbol-simbol angka dan yang tidak diperlukan [10]. Berikut tahapaan text processing :

- Case Folding : Tahap ini merupakan proses perubahan teks dimana semua huruf merupaakan huruf kecil. Berikut contohnya pada Gambar 5:

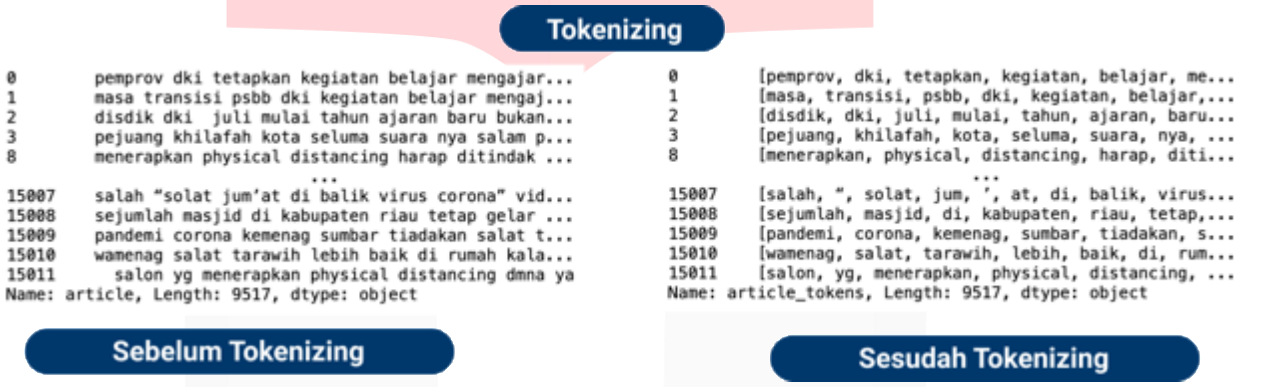


	article	label
0	pemprov dki tetapkan kegiatan belajar mengajar...	fact
1	Masa Transisi PSBB DKI, Kegiatan Belajar Menga...	support
2	Disdik DKI: 13 Juli Mulai Tahun Ajaran Baru, B...	support
3	pejuang khilafah kota seluma suara nya salam p...	unimportant
8	menerapkan physical distancing harap ditindak ...	unknown
9	Physical Distancing, Kapolri: Belum Ada Masyar...	NEI
10	yearrypanji haha yg bikin statement jg ngomel ...	unimportant
13	video plandemic terkait virus corona	hoax
14	[HOAKS] Klaim Video "Plandemic" Terkait Virus ...	oppose
15	petugas check point memindai qr code memeriksa...	unknown

```
0       pemprov dki tetapkan kegiatan belajar mengajar...
1       masa transisi psbb dki kegiatan belajar mengaj...
2       disdik dki  juli mulai tahun ajaran baru bukan...
3       pejuang khilafah kota seluma suara nya salam p...
8       menerapkan physical distancing harap ditindak ...
                              ...
15007   salah "solat jum'at di balik virus corona" vid...
15008   sejumlah masjid di kabupaten riau tetap gelar ...
15009   pandemi corona kemenag sumbar tiadakan salat t...
15010   wamenag salat tarawih lebih baik di rumah kala...
15011     salon yg menerapkan physical distancing dmna ya
Name: article, Length: 9517, dtype: object
```

Gambar 5 Case Folding

- Tokenizing : Melakukan pemenggalan kata dari data yang akan diolah. Berikut contohnya pada Gambar 6:

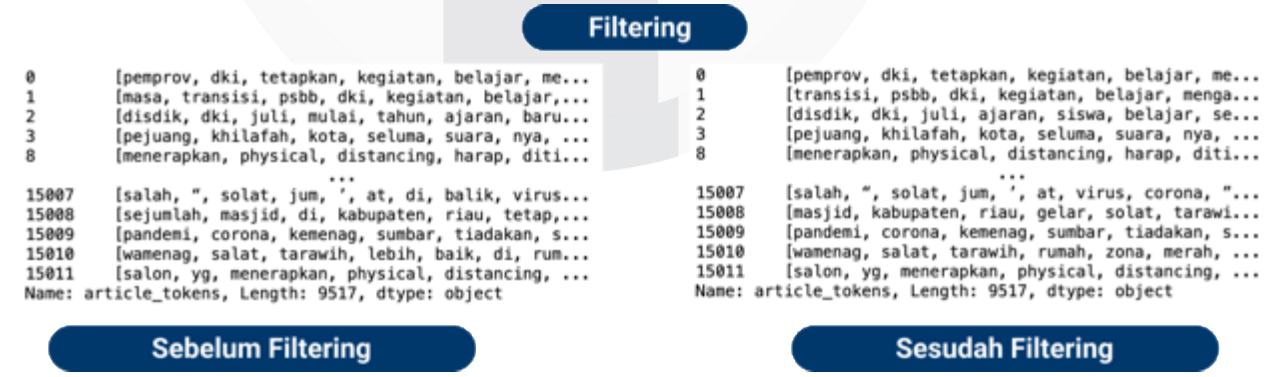



```
0       pemprov dki tetapkan kegiatan belajar mengajar...
1       masa transisi psbb dki kegiatan belajar mengaj...
2       disdik dki  juli mulai tahun ajaran baru bukan...
3       pejuang khilafah kota seluma suara nya salam p...
8       menerapkan physical distancing harap ditindak ...
                              ...
15007   salah "solat jum'at di balik virus corona" vid...
15008   sejumlah masjid di kabupaten riau tetap gelar ...
15009   pandemi corona kemenag sumbar tiadakan salat t...
15010   wamenag salat tarawih lebih baik di rumah kala...
15011     salon yg menerapkan physical distancing dmna ya
Name: article, Length: 9517, dtype: object
```

```
0       [pemprov, dki, tetapkan, kegiatan, belajar, me...
1       [masa, transisi, psbb, dki, kegiatan, belajar,...
2       [disdik, dki, juli, mulai, tahun, ajaran, baru...
3       [pejuang, khilafah, kota, seluma, suara, nya, ...
8       [menerapkan, physical, distancing, harap, diti...
                              ...
15007   [salah, ", solat, jum, ', at, di, balik, virus...
15008   [sejumlah, masjid, di, kabupaten, riau, tetap,...
15009   [pandemi, corona, kemenag, sumbar, tiadakan, s...
15010   [wamenag, salat, tarawih, lebih, baik, di, rum...
15011   [salon, yg, menerapkan, physical, distancing, ...
Name: article_tokens, Length: 9517, dtype: object
```

Gambar 6 Tokenizing

- Filtering / Stopword Removal : Pada proses ini akan dihapusnya kata yang tidak memiliki pengaruh begitu besar, seperti kata yang terlalu sering muncul bila dibandingkan dengan kata lainnya. Berikut contohnya pada Gambar 7:




```
0       [pemprov, dki, tetapkan, kegiatan, belajar, me...
1       [masa, transisi, psbb, dki, kegiatan, belajar,...
2       [disdik, dki, juli, mulai, tahun, ajaran, baru...
3       [pejuang, khilafah, kota, seluma, suara, nya, ...
8       [menerapkan, physical, distancing, harap, diti...
                              ...
15007   [salah, ", solat, jum, ', at, di, balik, virus...
15008   [sejumlah, masjid, di, kabupaten, riau, tetap,...
15009   [pandemi, corona, kemenag, sumbar, tiadakan, s...
15010   [wamenag, salat, tarawih, lebih, baik, di, rum...
15011   [salon, yg, menerapkan, physical, distancing, ...
Name: article_tokens, Length: 9517, dtype: object
```

```
0       [pemprov, dki, tetapkan, kegiatan, belajar, me...
1       [transisi, psbb, dki, kegiatan, belajar, menga...
2       [disdik, dki, juli, ajaran, siswa, belajar, se...
3       [pejuang, khilafah, kota, seluma, suara, nya, ...
8       [menerapkan, physical, distancing, harap, diti...
                              ...
15007   [salah, ", solat, jum, ', at, virus, corona, "...
15008   [masjid, kabupaten, riau, gelar, solat, tarawi...
15009   [pandemi, corona, kemenag, sumbar, tiadakan, s...
15010   [wamenag, salat, tarawih, rumah, zona, merah, ...
15011   [salon, yg, menerapkan, physical, distancing, ...
Name: article_tokens, Length: 9517, dtype: object
```

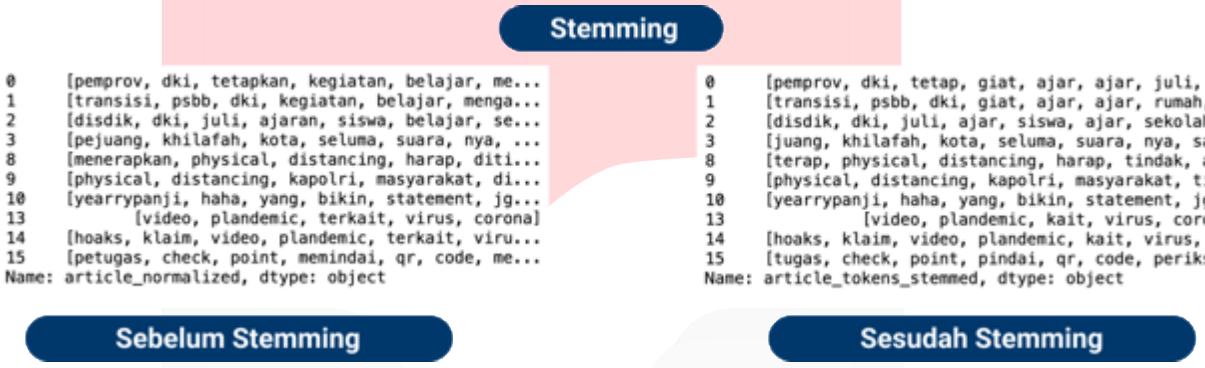
Gambar 7 Filtering

- Normalization : Merupakan tahapan modifikasi kata singkatan menjadi kata yang seharusnya, seperti ‘yg’ menjadi ‘yang’, ‘kmrn’ menjadi ‘kemarin’, ‘krn’ menjadi ‘karena’, dan masih sangat banyak contoh lainnya.

Berikut contoh pada Gambar 8:

Gambar 8 Normalization

- Stemming or Lemmatization : Stemming merupakan proses perubahan kata menjadi kata dasar, dimana seluruh kata berimbuhan seperti di, ke dan sebagainya akan diubah menjadi kata dasar. Berikut contoh pada Gambar 9:



Gambar 9 Stemming

3.3. TF-IDF

Pada tahap ini akan dilakukan pembobotan data, ini agar terjadi penyempitan antar masing-masing data, sistem akan melakukan pengujian terhadap data latih yang ada, sehingga hasil akan lebih variatif. Pada sub bab 2 sudah dijelaskan bagaimana perhitungan TF-IDF.

3.4. Training K-Nearest Neighbor

Pada bagian training *K-Nearest Neighbor* akan dilakukan klasifikasi sebanyak k, dimana nilai k yang akan diuji diantaranya k = 3, k = 5, k = 7, dan k = 9 yang akan dicari hasil terbaiknya. Untuk penjelasan *K-Nearest Neigbhor* sudah dijelaskan pada sub bab 2.

3.5. Training K-Nearest Neigbhor Cross Validation

Pada bagian ini akan dilakukan validasi KNN dengan menggunakan k-fold cross validation, pada tahap ini data dibagi menjadi 2 yaitu data validasi dan data train sebagai proses pembelajaran, pada proses ini, berdasarkan[12] penulis menggunakan fold sebanyak k = 5 dan k = 10.

4. Evaluasi

4.1. Skenario Pelatihan & Pengujian

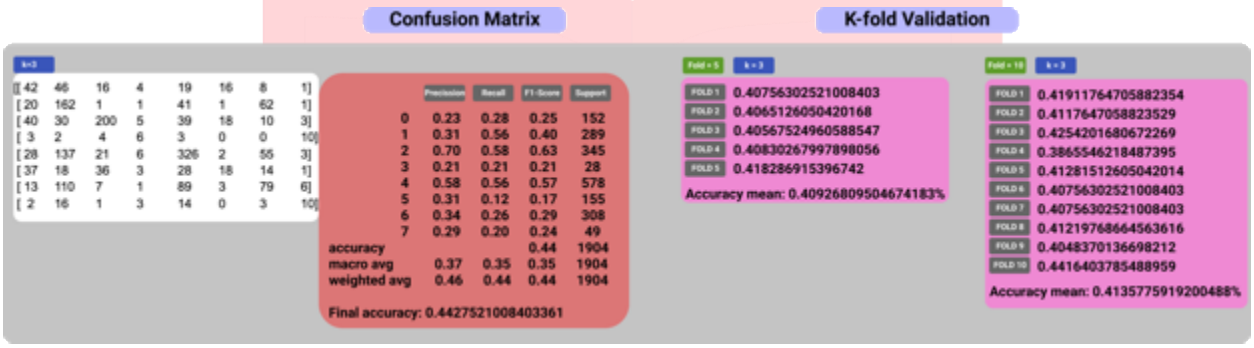
Terdapat enam skenario yang digunakan pada penelitian ini. Pada tahap awal dilakukan split data dengan 80% data training, dan 20% data testing. Nilai k pada *KNN* divariasikan menjadi: k =3 , k = 5, k = 7, dan k = 9 untuk mencari nilai k terbaik. Setelah mendapatkan nilai k terbaik, komposisi data train dan test diubah untuk melihat pengaruh persentase nilai tersebut. Untuk setiap skenario, pada data training digunakan k-fold cross validation dengan k=5 dan k=10.

Tabel 2. Tabel Skenario

Skenario ke-	data training		data testing	nilai *k* pada kNN
1	80%	Nilai k pada cross fold validation dengan k=5 dan k=10	20%	3
2				5
3				7
4				9
5	70%		30%	Sesuai yang terbaik dari skenario 1-4
6	60%		40%	

4.2. Analisis Hasil Pelatihan dan Pengujian

Dari skenario yang telah dibuat pada subbab sebelumnya. Pada *confusion matrix* inisiasi label 'fact' :0, 'support' :1, 'unimportant' :2, 'hoax' :3, 'irrelevant' :4, 'unknown' :5, 'NEI' :6, dan 'oppose' :7. Implementasi dari skenario sebagai berikut:

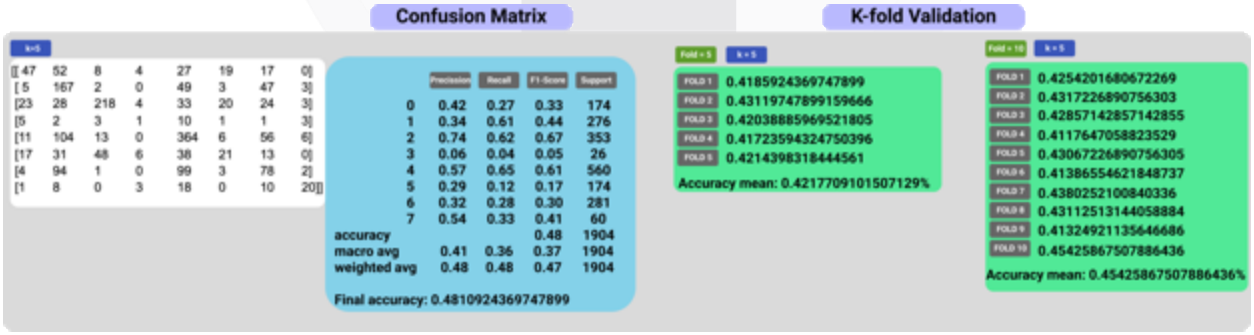


Gambar 10 Hasil Pengujian Skenario ke-1 (k =3)

Seperti yang terlihat pada Gambar 10, total akurasi model adalah 44% dengan komposisi sebagai berikut:

- Label 0 memiliki akurasi 28% dengan prediksi benar sebesar 42 dari 152 data.
- Label 1 memiliki akurasi 56% dengan prediksi benar sebesar 162 dari 289 data.
- Label 2 memiliki akurasi 58% dengan prediksi benar sebesar 200 dari 345 data.
- Label 3 memiliki akurasi 21% dengan prediksi benar sebesar 6 dari 28 data.
- Label 4 memiliki akurasi 56% dengan prediksi benar sebesar 326 dari 578 data.
- Label 5 memiliki akurasi 12% dengan prediksi benar sebesar 18 dari 155 data.
- Label 6 memiliki akurasi 26% dengan prediksi benar sebesar 79 dari 308 data.
- Label 7 memiliki akurasi 20% dengan prediksi benar sebesar 10 dari 49 data.

Label akurasi penulis divalidasi menggunakan k-fold cross validation dengan dua nilai k yaitu k = 5 dengan akurasi rata – rata 40% dan k = 10 dengan nilai rata-rata akurasi 41%.



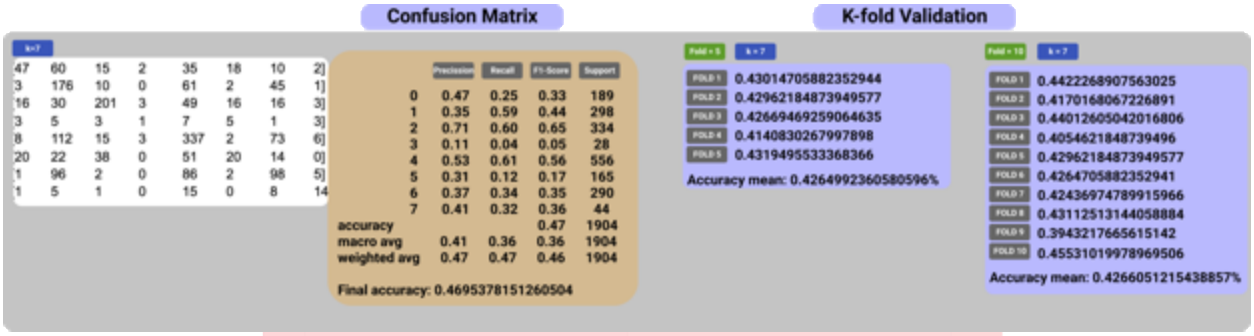
Gambar 11 Hasil Pengujian Skenario ke-2 (k =5)

Seperti yang terlihat pada Gambar 11, total akurasi model adalah 44% dengan komposisi sebagai berikut:

- Label 0 memiliki akurasi 27% dengan prediksi benar sebesar 47 dari 174 data.
- Label 1 memiliki akurasi 61% dengan prediksi benar sebesar 167 dari 276 data.
- Label 2 memiliki akurasi 62% dengan prediksi benar sebesar 218 dari 353 data.
- Label 3 memiliki akurasi 4% dengan prediksi benar sebesar 1 dari 26 data.
- Label 4 memiliki akurasi 65% dengan prediksi benar sebesar 364 dari 560 data.

- Label 5 memiliki akurasi 12% dengan prediksi benar sebesar 21 dari 174 data.
- Label 6 memiliki akurasi 28% dengan prediksi benar sebesar 78 dari 281 data.
- Label 7 memiliki akurasi 33% dengan prediksi benar sebesar 20 dari 60 data.

Label akurasi penulis divalidasi menggunakan k-fold cross validation dengan dua nilai k yaitu k = 5 dengan akurasi rata – rata 42% dan k = 10 dengan nilai rata-rata akurasi 45%.

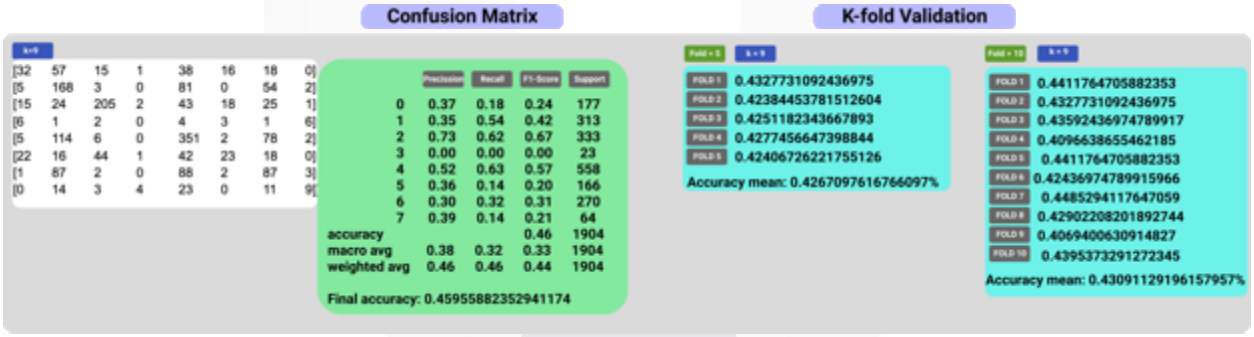


Gambar 12 Hasil Pengujian Skenario ke-3 (k =7)

Seperti yang terlihat pada Gambar 12, total akurasi model adalah 44% dengan komposisi sebagai berikut:

- Label 0 memiliki akurasi 25% dengan prediksi benar sebesar 47 dari 189 data.
- Label 1 memiliki akurasi 59% dengan prediksi benar sebesar 176 dari 298 data.
- Label 2 memiliki akurasi 60% dengan prediksi benar sebesar 201 dari 334 data.
- Label 3 memiliki akurasi 4% dengan prediksi benar sebesar 1 dari 28 data.
- Label 4 memiliki akurasi 61% dengan prediksi benar sebesar 337 dari 556 data.
- Label 5 memiliki akurasi 12% dengan prediksi benar sebesar 20 dari 165 data.
- Label 6 memiliki akurasi 34% dengan prediksi benar sebesar 98 dari 290 data.
- Label 7 memiliki akurasi 32% dengan prediksi benar sebesar 14 dari 44 data.

Label akurasi penulis divalidasi menggunakan k-fold cross validation dengan dua nilai k yaitu k = 5 dengan akurasi rata – rata 42% dan k = 10 dengan nilai rata-rata akurasi 42%.



G Gambar 13 Hasil Pengujian Skenario ke-4 (k =9)

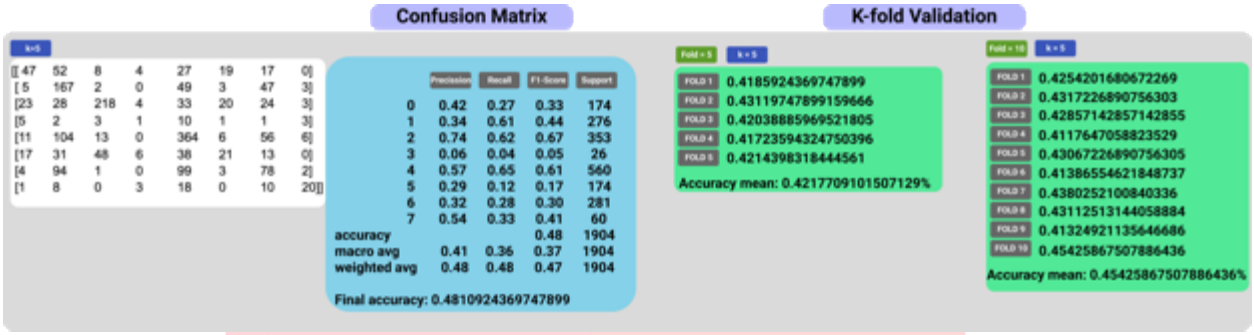
Seperti yang terlihat pada Gambar 13, total akurasi model adalah 44% dengan komposisi sebagai berikut:

- Label 0 memiliki akurasi 18% dengan prediksi benar sebesar 32 dari 177 data.
- Label 1 memiliki akurasi 54% dengan prediksi benar sebesar 168 dari 313 data.
- Label 2 memiliki akurasi 62% dengan prediksi benar sebesar 205 dari 333 data.
- Label 3 memiliki akurasi 0% dengan prediksi benar sebesar 0 dari 23 data.
- Label 4 memiliki akurasi 63% dengan prediksi benar sebesar 351 dari 558 data.
- Label 5 memiliki akurasi 14% dengan prediksi benar sebesar 23 dari 166 data.
- Label 6 memiliki akurasi 32% dengan prediksi benar sebesar 87 dari 270 data.
- Label 7 memiliki akurasi 14% dengan prediksi benar sebesar 9 dari 64 data.

Label akurasi penulis divalidasi menggunakan k-fold cross validation dengan dua nilai k yaitu k = 5 dengan akurasi rata – rata 42% dan k = 10 dengan nilai rata-rata akurasi 43%.

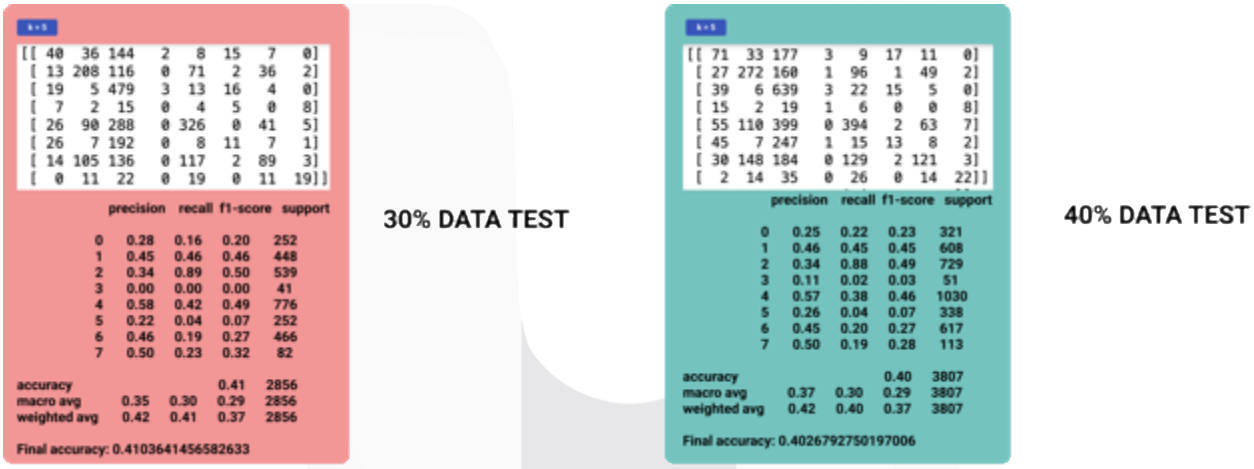
4.3. Analisis Model KNN Terbaik

Dari masing-masing skenario pengujian hasil akurasi dari model berbeda-beda dari traning KNN dengan nilai k = 3, k = 5, k = 7, dan k = 9 serta validasi pada k-fold = 5 dan k-fold = 10. Hasil terbaik didapat dari nilai k = 5 untuk akurasi KNN sebesar 48%, akurasi k-fold = 5 sebesar 42%, dan k-fold = 10 sebesar 45% menggunakan 20% data test. Jumlah data sangat berpegaruh terhadap akurasi model yang dibuat, hal ini dapat kita lihat pada classification *report* dimana jumlah data yang tinggi memiliki akurasi yang lebih tinggi dibandingkan dengan data dengan jumlah sedikit. Berikut model terbaik yang di dapat pada Gambar 14 :



Gambar 14 Hasil Pengujian Skenario Terbaik (k =5)

Karena pengujian pertama menghasilkan nilai akurasi terbaik dengan nilai k = 5 pada training KNN dan 20% data test, maka penulis mencoba mengubah parameter dengan menggunakan 30% data test dan 40% data test. Hasil pengujian mendapati bahwa menggubah parameter ke 30% dan 40% data test tidak membuat akurasi lebih baik. Dengan hasil 41% pada data test 30% dan 40% pada data test 40%. Dapat dilihat Gambar 15 :



Gambar 15 Hasil Pengujian Skenario ke-5 & ke-6 (k =5)

5. Kesimpulan & Saran

5.1. Kesimpulan

Dari hasil analisis dan pengujian dan pelatihan yang telah dilakukan, maka dapat disimpulkan bahwa penggunaan nilai k = 5 pada proses klasifikasi KNN lebih akurat dibanding menggunakan k = 3, k = 7, dan k = 9 dengan menggunakan 80% data train dan 20% data test, menghasilkan akurasi sebesar 48%, akurasi model di validasi dengan *k-fold cross validation*. Dengan nilai k-fold, k = 5 menghasilkan akurasi 42%, dan pada k-fold, k = 10 menghasilkan akurasi 45%. Pada hasil penelitian diketahui hasil yang tidak begitu baik dikarenakan masih adanya beberapa kelas yang memiliki jumlah data sedikit sehingga mengurangi kemampuan pembelajaran model dalam mengklasifikasikan kelas tersebut.

5.2. Saran

Saran dari penulis adanya peneliti yang menambah jumlah data sehingga hasil lebih bervariasi, serta implementasi ke dalam platform website atau mobile sehingga model penelitian dapat dirasakan oleh masyarakat luas.

REFERENSI

[1]. Agung B. Prasetijo, R. Rizal Isnanto, Dania Eridani, Yosua Alvin Adi Soetrisno, M. Arfan, Aghus Sofwan, "Hoax Detection System on Indonesian News Sites Based on Text Classification using SVM and SGD", in International Conference on Information Tech., Computer, and Electrical Engineering, (2017).

[2]. Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, Faisal Rahutomo, "Study of Hoax News Detection Using Naïve Bayes Calasifier in Indonesian Language", in International Conference on Information & Communication Technology and System, (2017).

[3]. I Made Riarta Prawira, Adiwijaya, Mohamad Syahrul Mubarok, " Klasifikasi *Multi-Label* Pada Topik Berita Berbahasa Indonesia Menggunakan *Multinomial Naïve Bayes*", in e-Proceeding of Engineering vol. 15, no. 3, (2018).

[4]. Faizal Nur Rozi, Dwi Hartini Sulistyawati, "Klasifikasi Berita Hoax Pilplres Menggunakan Metode Modified K-Nearest Neighbor dan Pembobotan Menggunakan TF-IDF", in konvergensi vol. 15, no.1, (2019).

[5]. Afrian Hanafi, Adiwijaya, Widi Astuti, "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor", in Jurnal Sistem Informasi dan Komputer vol. 09, no. 03, (2020).

[6]. Eri Zuliarso, Muchamad Taufiq Anwar, Kristophorus Hadiono, Iswatun Chasanah, "Detecting Hoaxes in Indonesian News Using TF/IDM and K Nearest Neighbor", in IOP Conference Series : Materials Science and Engineering, (2020).

[7]. Andre Rino Prasetyo, Indriati, Putra Pandu Adikara, "Klasifikasi *Hoax* Pada Berita Kesehatan Berbahasa Indonesia Denga Menggunakan Metode Modified *K-Nearest Neighbor*", in Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer vol. 2, no. 12, (2018).

[8]. Achmad Ridok, Retnani Latifah, "Klasifikasi Text Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN", in Konferensi Nasional Sistem & Informatika, (2015).

[9]. Nikmah Isnaini, Adiwijaya, Mohamad Syahrul Mubarok, Muhammad Yuslan Abu Bakar,"A multi-label classification on topics of Indonesian news using K-Neareast Neighbor", in The 2nd International Conference on Data and Information Science, (2019).

[10]. Putu Wira Buana, Sesaltina Jannet D.R.M, I Ketut Gede Darma Putra, "Combination of K-Nearest Neighbor and K-means based on Terms Re-Weighting for Classifiy Indonesian News", in International Jurnal of Computer Applications, (2012).

[11]. Juan Diego Rodriguez, Jose A. Lozano, Aritz Perez, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation", in IEE Transaction on Pattern Analysis and Machine Inteligence, vol. 32, no. 3, (March 2010).

[12]. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Stastistical Learning", Springer Texts in Statistics, (2013).

[13]. Chin-Min Ma, Wei-Shui Yang, Bor-Wen Cheng, "How the Parameters of K-Nearest Neighbor Algortihm Impact on the Best Classification Accruracy : In Case of Parkinson Dataset", Asian Network For Scientific Information, (2014).

[14]. KBBI, "KBBI" [2012-2021], Available: https://kbbi.web.id/berita. [Accessed 14 September 2021].

[15]. Anurag P. Jain, Mr. Vijay D. Katkar, "Sentiment Analysis Of Twitter Data Using Data Mining", 2015 International Conference on Information Processing (ICIP) Vishwakarma Institute of Technology. (2015).

[16]. KBBI, "KBBI Daring", Available: https://kbbi.kemdikbud.go.id/entri/hoaks. [Accessed 14 September 2021].

[17]. Octaryo Sakti Yudha Prakasa, Kemas Muslim Lhaksamana, "Klasifikasi Teks dengan Menggunakan Algoritma K-Nearest Neighbor Pada kasus Kinerja Pemerintah di twitter", e-Proceeding of Engineering : vol.5, No.3 (Desember 2018).