

## ANALISIS TENDENSI PORTAL BERITA ONLINE TERHADAP VAKSINASI COVID-19 DI INDONESIA MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

Yana Adharani<sup>1\*</sup>, Ambar Dwi Saputra<sup>2</sup>, Nurvelly Rosanti<sup>3</sup>, Retnani Latifah<sup>4</sup>

Jurusan Teknik Informatika Universitas Muhammadiyah Jakarta

\*yana.adharani@umj.ac.id

Di Indonesia per 27 Juli 2021 ada 4.170.155 kasus terkonfirmasi COVID-19. Salahsatu cara untuk menekan peningkatan dan penyebaran kasus COVID-19 adalah dengan melakukan vaksinasi. Terkait dengan hal tersebut, ada beberapa berita tentang vaksin COVID-19 di portal berita. Pemberitaan yang diberikan memiliki tendensi positif, negatif maupun netral yang dapat memengaruhi pandangan masyarakat terhadap pemberian vaksin COVID-19. Dalam penelitian ini dilakukan klasifikasi terhadap tendensi pemberitaan portal berita terhadap vaksin COVID-19 menggunakan metode *K-Nearest Neighbor* (KNN). Pengujian dilakukan terhadap 1000 data dengan 5 kombinasi data *training* dan *testing* sebanyak 50 kali dengan nilai *k* berbeda. Pelabelan data dilakukan secara manual dan otomatis menggunakan *sentistrenth\_id*. Hasil Pengujian pelabelan manual menunjukkan tingkat *accuracy* tertinggi sebesar 78,50% pada *k* = 6. Untuk pelabelan otomatis diperoleh *accuracy* tertinggi sebesar 93% pada *k* = 6. *Accuracy* tertinggi diperoleh untuk penggunaan data *training* sebesar 80% dan data *testing* 20%. Tingkat *accuracy* yang belum optimal diakibatkan karena jumlah data pada setiap kelas tidak berimbang dan terdapat pelabelan pada data *training* maupun *testing* yang tidak akurat. Aplikasi dapat mengidentifikasi kata-kata dominan, baik dalam pemberitaan positif, negatif maupun netral untuk kelas yang teridentifikasi.

**Kata Kunci:** *klasifikasi, COVID-19, K-Nearest Neighbor, portal berita, vaksinasi*

### Abstract

In Indonesia as of July 27, 2021 there were 4,170,155 confirmed cases of COVID-19. One way to suppress the increase and spread of COVID-19 cases is to vaccinate. Related to this, there is some news about the COVID-19 vaccine on the news portal. The news provided hasa positive, negative and neutral tendency that can affect people's views on the COVID-19 vaccine. In this study, a classification of news portal news tendencies against the COVID-19 vaccine using the K-Nearest Neighbor (KNN) method. The test was conducted on 1000 data with 5 combinations of training data and testing as much as 50 times with different *k* values. Data labeling is done manually and automatically using *sentistrenth\_id*. Manual labeling test results showed the highest accuracy rate of 78.50% at *k* = 6. For automatic labeling obtained the highest accuracy of 93% at *k* = 6. The highest accuracy is obtained for the use of training data by 80% and data testing by 20%. The level of accuracy is not optimal because the amount of data in each class is not balanced and there are labels on the training and testing data that are not accurate. The application can identify dominant words, both in positive, negative and neutral news for the identified class.

**Keywords:** *classification, COVID-19, K-Nearest Neighbor, news portal, vaccination*

## 1. Pendahuluan

Pandemi Coronavirus atau COVID-19 yang terjadi di Indonesia sejak ditemukannya kasus positif di masyarakat pertama kali pada Maret 2020 tahun lalu belum ada tanda bahwa pandemi ini akan segera berakhir. Menurut data pada website [www.covid19.go.id](http://www.covid19.go.id) sampai dengan 27 Juli 2021 tercatat di Indonesia kasus yang terkonfirmasi positif sebanyak 3.239.936 dengan korban meninggal 86.835 jiwa. Virus Corona dapat menular melalui kontak fisik dengan orang yang tertular COVID-19. Namun, penularan juga bisa terjadi tanpa kontak fisik seperti dengan menularkan tetesan kecil air yang mengandung virus kepada orang lain saat batuk, bersin atau bernafas.

Melihat pesatnya penyebaran COVID-19 dan bertambahnya kasus harian orang yang positif virus corona maka diperlukan suatu cara untuk mencegah penyebaran virus ini agar tidak meluas, cara tersebut yaitu dengan melakukan vaksinasi. Pemberian vaksin dapat mencegah angka penyebaran virus corona karena vaksin dapat meningkatkan sistem imunitas tubuh untuk melawan virus corona, dengan demikian diharapkan angka kematian akibat COVID-19 juga akan menurun. Dalam pelaksanaan vaksinasi untuk mencegah virus corona Presiden Joko Widodo meresmikan Peraturan Presiden (PP) Nomor 99 Tahun 2020 tentang pengadaan vaksin dan pelaksanaan vaksinasi dalam rangka penanggulangan pandemi corona virus disease 2019 (COVID-19). Untuk tahap 1 pemerintah menggunakan vaksin yang di produksi oleh perusahaan biofarmasi asal China, yaitu vaksin corona Sinovac, uji klinis dilakukan oleh tim dari Fakultas Kedokteran Universitas Padjajaran. Dari hasil uji klinis tersebut menunjukkan efikasi vaksin Sinovac sebesar 65,3 persen, hasil tersebut sudah sesuai dengan persyaratan WHO di mana minimal efikasi vaksin adalah 50 persen (Farisa, 2021).

Masyarakat yang mendapatkan informasi tentang kegiatan vaksinasi tersebut melalui portal berita memberikan berbagai macam tanggapannya terhadap rencana pemerintah melakukan vaksinasi. Informasi yang diberikan portal berita ke masyarakat dapat membuat berbagai macam opini dan respon terhadap vaksin. Seperti yang ungkapkan oleh Dr dr Kohar Hari Santoso, Direktur RSSA Malang dan Ketua Tim Tracing Gugus Tugas COVID-19 Jawa Timur, Media sangat membantu program imunisasi MR di tahun 2017 dengan menyebarkan sosialisasi dan edukasi positif, upaya ini sangat efektif membuat masyarakat bersedia untuk diimunisasi (Suteja, 2021). Dengan banyaknya pemberitaan positif dan edukasi atau kampanye melalui portal berita kepada masyarakat tentang vaksinasi COVID-19 diharapkan agar target pemberian vaksin COVID-19 tercapai, sedangkan dengan banyaknya pemberitaan negatif dapat menghambat percepatan dalam pemberian vaksin COVID-19. Berdasarkan hal tersebut maka perlu dilakukan analisis terhadap tendensi atau kecenderungan pemberitaan portal berita online mengenai vaksin COVID-19, karena dapat mempengaruhi persepsi masyarakat dan tingkat keberhasilan program pemberian vaksin

*K-Nearest Neighbor* (KNN) adalah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasi sebelumnya, metode ini termasuk dalam supervised learning, dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam KNN (Informatikalogi, 2017). Berdasarkan latar belakang tersebut, maka dalam penelitian ini diimplementasikan metode *K-Nearest Neighbor* untuk mengetahui tendensi atau

kecenderungan pemberitaan portal berita terhadap vaksinasi COVID-19

## 2. Landasan Teori

### A. Text Mining

*Text mining* merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen (Soepomo, 2014). *Text mining* mencoba memecahkan masalah kelebihan informasi (*information overload*) dengan menggunakan teknik dari bidang ilmu terkait, *Text mining* dapat dipandang sebagai perluasan dari data mining atau *Knowledge Discovery in Database* (KDD), yang bertujuan untuk menemukan pola-pola menarik dari basis data berskala besar (Yulian, 2018). *Text mining* bertujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik. seperti terlihat pada contoh ditabel 1.

### B. Text Preprocessing

*Text Preprocessing* adalah mengolah sumber data baik yang terstruktur, terstruktur sebagian dan yang tidak terstruktur dari beberapa sumber. *Text Preprocessing* bertujuan untuk mengolah data yang masih bermacam – macam untuk dijadikan sebuah data teratur yang dapat dikenai atau diterapkan beberapa metode *text mining* yang ada (H, 2015). Pada proses *text preprocessing* terdapat 4 tahap yaitu meliputi *case folding*, *tokenizing*, *filtering*, dan *stemming* (Nugroho, 2011).

### C. Term Frequency-Inverse Document Frequency (TF-IDF)

Pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah suatu proses untuk melakukan transformasi data dari bentuk teks ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur. TF-IDF ini adalah ukuran statistik yang digunakan untuk menentukan kata-kata penting dalam dokumen. *Term Frequency* (TF) adalah frekuensi kemunculan kata atau term pada suatu dokumen. *Inverse Document*

*Frequency* (IDF) adalah inverse dari nilai DF. Hasil dari pembobotan kata menggunakan TF-IDF menghasilkan suatu nilai yang apabila bobot nilai tersebut besar maka semakin banyak kata yang muncul pada suatu dokumen, apabila bobot nilai tersebut kecil maka semakin kecil kata yang muncul di suatu dokumen (Septian et al., 2018).

### D. Analisis Sentimen

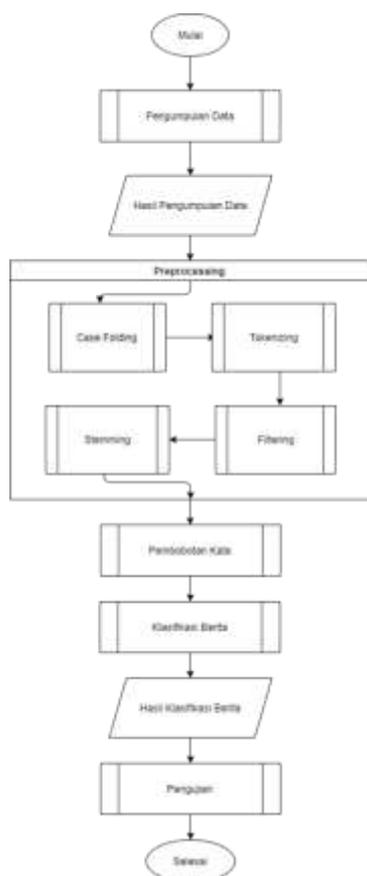
Analisis Sentimen adalah studi yang menganalisis suatu pendapat atau opini-opini orang untuk memberikan penilaian terhadap suatu entitas produk, layanan, peristiwa dan yang lainnya (Liu, 2012). Analisis sentimen telah berkembang menjadi salah satu penelitian aktif dalam *natural language processing*. Dalam beberapa tahun terakhir, penggunaan analisis sentimen juga berkembang dengan digunakannya pada kegiatan industri (Liu, 2012). Analisis sentimen dapat memberikan klasifikasi atas kalimat atau dokumen terhadap sesuatu, nilai yang biasa digunakan pada analisis sentimen adalah positif dan negatif, dimana dengan nilai tersebut dapat dijadikan parameter dalam pengambilan keputusan (Haryanto et al., 2018)

### E. K-Nearest Neighbor (KNN)

*K-Nearest Neighbor* (KNN) adalah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasi sebelumnya, metode ini termasuk dalam supervised learning, dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam KNN (Informatikalogi, 2017).

## 3. Metode Penelitian

Dalam penelitian ini adapun tahapan yang dibahas pada bab ini yaitu tahap pengumpulan data, tahap *text preprocessing*, tahap pembobotan kata dan tahap implementasi metode yang digunakan yaitu metode *K-Nearest Neighbor* (KNN). Untuk diagram alir proses untuk metode penelitian seperti Gambar 1 sebagai berikut :



Gambar. 1. Metode Penelitian

### A. Pengumpulan Data atau Informasi

Pengumpulan data yang digunakan untuk menganalisis sentimen berupa informasi yang terdapat pada portal berita tentang vaksinasi COVID-19 di Indonesia. Data yang dikumpulkan dari berita vaksinasi COVID-19 di Indonesia yaitu dicari dengan berbagai keyword yaitu vaksin covid-19, vaksin corona, vaksin dan vaksinasi. Website portal berita yang digunakan di antaranya detiknews.com, antranews.com, kompas.com, Liputan6.com dan cnbcindonesia.com. Data didapatkan dengan pengambilan judul berita pada web yang digunakan untuk analisis tendensi. Hasil data yang dikumpulkan yaitu berjumlah 1000 data. Dari data yang dikumpulkan akan diberi label secara manual yaitu label tendensi positif, negatif dan netral.

### B. Preprocessing

Tahap ini merupakan hal yang penting untuk tahap selanjutnya. Pada tahap ini adalah menghilangkan atau mengurangi atribut yang kurang berpengaruh terhadap proses

klasifikasi. Adapun tahap *preprocessing* yaitu meliputi *case folding*, *tokenizing*, *filtering* dan *stemming*.

*Case folding* adalah tahap perubahan huruf kapital pada suatu dokumen menjadi huruf kecil. Tujuannya untuk menghilangkan pengulangan data yang berbeda pada hurufnya saja.

*Tokenizing* merupakan pemisahan kata berdasarkan tiap kata yang menyusunnya menjadi potongan tunggal. Kata dalam dokumen yang dimaksud adalah kata yang dipisah oleh spasi.

*Filtering* adalah tahap pengambilan kata-kata yang penting setelah hasil dari tokenizing, atau biasa disebut pengeliminasian sebuah kata-kata sesuai dengan kaidahnya. Pada proses ini terdapat penghapusan stopword, yaitu kata yang tidak deskriptif.

*Stemming* merupakan pemecahan suatu varian-varian kata atau kata berimbuhan menjadi kata dasar sesuai dengan kata yang di proses.

### C. Pembobotan Term Frequency-Inverse Document Frequency (TF-IDF)

Pada tahap ini diberikan pembobotan tiap kata atau term sebelum melakukan klasifikasi dengan dokumen hasil *stemming*. Pembobotan TF-IDF penelitian kali ini akan digunakan library python sklearn `TfidfVectorizer`. Langkah pertama yang dilakukan adalah menentukan *Term Frequency* (TF) yaitu menghitung kemunculan term pada dokumen. Setelah menghitung *Term Frequency* (TF) tahap selanjutnya yaitu menghitung *Inverse Document Frequency* (IDF). Adapun rumus dari pembobotan TF-IDF adalah:

$$W = TF \times IDF \quad (1)$$

W = Bobot term atau kata.

TF = Banyaknya kemunculan suatu kata dalam suatu dokumen.

IDF = inverse frekuensi dokumen, dimana

Menghitung nilai IDF dengan menggunakan library python sklearn `TfidfVectorizer`, yaitu

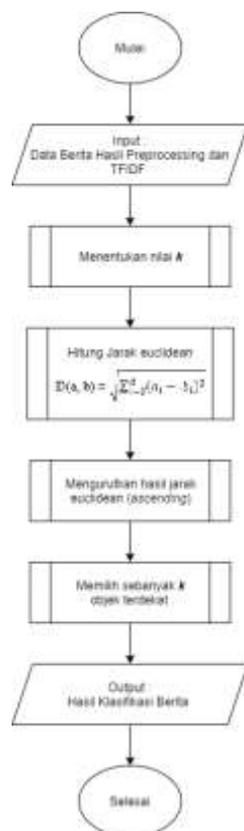
$$IDF = \ln \left( \frac{N}{n_t} \right) + 1 \quad (2)$$

N = Jumlah semua dokumen.

$n_t$  = Jumlah dokumen yang mengandung term

#### D. Klasifikasi K-Nearest Neighbor (KNN)

Data yang telah melalui *preprocessing* dan sudah dilakukan pembobotan kata atau term akan di klasifikasi sesuai kelasnya untuk menentukan polaritas dari text tersebut apakah text tersebut termasuk dalam berita yang positif, negatif maupun netral terhadap vaksin COVID-19. Adapun Proses klasifikasi menggunakan metode KNN dapat dilihat pada Gambar 2.



Gambar. 2. Tahapan Implementasi metode KNN

Penjelasan secara proses secara mendetail implementasi klasifikasi dengan menggunakan metode KNN pada Gambar 2 adalah sebagai berikut:

1. Input atau masukkan data berita hasil preprocessing dan pembobotan *Term*

*Frequency-Inverse Document Frequency* (TF-IDF).

2. Menentukan jumlah tetangga terdekat ( $k$ ) yang digunakan untuk membuat klasifikasi berita.
3. Menghitung kuadrat jarak *euclidean* objek dengan rumus:

$$D(a, b) = \sqrt{\sum_{i=1}^d (a_i - b_i)^2} \quad (3)$$

4. Mengurutkan hasil perhitungan kuadrat *Euclidean Distance* dari kecil ke besar (ascending).
5. Memilih sebanyak  $k$  objek terdekat dan melakukan pengklasifikasian berita berdasarkan jumlah tetangga terdekat.
6. Hasil output berupa tampilan klasifikasi berita berdasarkan jumlah tetanggaterdekat

#### E. Pengujian

Untuk menghasilkan performa model yang maksimal pada pengujian penelitian ini dilakukan pengukuran tingkat ketepatan atau *accuracy* yang dihasilkan oleh *metode K-Nearest Neighbor* (KNN). Pengujian akan dilakukan dengan 5 skenario pembagian data training dan data testing yang yaitu 60%:40%, 65%:35%, 70%:30%, 75%:25% dan 80%:20%, setiap skenario pembagian data akan dilakukan percobaan sebanyak 50 kali dengan nilai  $k$  yang berbeda yaitu 1 sampai 50, kemudian akan di uji menggunakan *Confusion Matrix* pada setiap percobaannya. *Confusion matrix* adalah suatu metode yang biasa digunakan untuk melakukan perhitungan tingkat *accuracy* pada data mining (Edukasi et al., 2018).

Setelah mendapatkan model pembagian data yang paling baik dengan melihat nilai *accuracy* tertinggi menggunakan hasil labeling secara manual, selanjutnya dilakukan pengujian pelabelan otomatis menggunakan program *pyhton sentistrengt\_id* yang berdasarkan *list* atau daftar kata positif dan negatif. Pengujian dilakukan untuk membandingkan dan melihat hasil labeling yang digunakan model dalam memberikan klasifikasi.

#### 4. Hasil dan Pembahasan

##### A. Tampilan Hasil

Tampilan hasil merupakan wujud dari rancangan web aplikasi yang telah dibangun. Pengkodean dilakukan untuk membuat tampilan yang bertujuan memberikan visualisasi atau gambaran jelas yang mudah dipahami terhadap klasifikasi tendensi portal berita.

- Tampilan *Dashboard*

*Dashboard* berfungsi sebagai halaman utama dan informasi hasil klasifikasi meliputi jumlah data, jumlah sentimen positif, jumlah sentimen netral, dan jumlah sentimen negatif, tampilan hasil dapat terlihat seperti pada Gambar 3.



Gambar. 3. Tampilan Dashboard

- Tampilan Data Master

Data Master yang berfungsi memberikan informasi isi semua data yang digunakan, pada tampilan data master akan dibagi menjad 2 menu, yaitu menu tampilan Hasil *Crawling* dan Hasil *Labeling*



Gambar. 4. Tampilan Hasil Crawling



Gambar. 5. Tampilan Hasil Labeling

- Tampilan Data Master

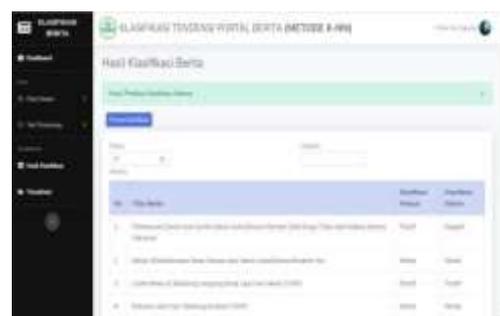
Hasil aplikasi web menu *Text Preprocessing* menampilkan 4 menu yang terdiri dari case folding, tokenizing, filterin, dan stemming. Berikut contoh hasil *text processing* pada tahap *case folding*.



Gambar. 6. Tampilan Case Folding

- Tampilan Hasil Klasifikasi

Tampilan web menu Hasil Klasifikasi Berita yang sudah melalui proses klasifikasi menggunakan metode *K-Nearest Neighbor*.



Gambar. 7. Tampilan Hasil Klasifikasi

TABEL 1

HASIL SKENARIO PERCOBAAN MENGGUNAKAN PELABELAN MANUAL

Data Training	Data Testing	Nilai k terbaik	Accuracy
60%	40%	12	74,00
65%	35%	6	76,00
		8	76,00
70%	30%	6	76,33
75%	25%	14	78,00
80%	20%	6	78.50

- Tampilan Hasil Visualisasi  
Tampilan menu Visualisasi pada saat memberikan informasi gambaran yang berbentuk pie chart dan *word cloud* dari hasil proses klasifikasi berita menggunakan metode K-Nearest Neighbor.



Gambar. 8. Tampilan Hasil Visualisasi

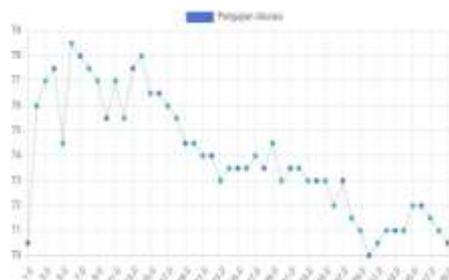
B. Hasil Pengujian

Untuk mengetahui model yang dibangun mempunyai accuracy dan tingkat ketepatan maka dibutuhkan suatu pengujian dalam implementasi klasifikasi berita menggunakan metode *K-Nearest Neighbor*. Pada pengujian penelitian ini adalah jumlah data yang digunakan sebesar 1000 data yang didapatkan dari hasil *scrapping* web pada portal berita online. Data yang dikumpulkan kemudian diberi label secara manual dan otomatis untuk melihat perbandingan hasil labeling serta klasifikasi berita.

- Pelabelan Manual

Setelah melakukan percobaan nilai *k* sebanyak 50 kali dengan pembagian data yang berbeda didapatkan hasil terbaik yaitu yang mempunyai nilai *accuracy* tertinggi pada setiap skenario percobaan seperti pada Tabel 1 sebagai berikut.

Berdasarkan skenario pembagian data yang dilakukan didapatkan hasil untuk performa model yang paling baik dengan melihat nilai *accuracy* tertinggi yaitu pada pembagian jumlah data *training* 80% dan data *testing* 20% dengan nilai *accuracy* sebesar 78,50% pada *k* = 6. Berikut grafik plot data nilai *accuracy* pada tiap nilai *k*.



Gambar. 9. Grafik Nilai k Pelabelan Manual

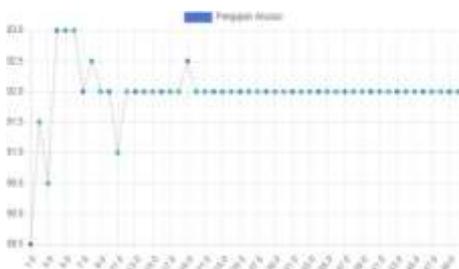
- Pelabelan Otomatis *sentistrength\_id*  
Pada percobaan sebelumnya pelabelan dataset dilakukan secara manual kemudian diujikan untuk melihat performa model yang paling baik, selanjutnya model kembali diujikan dengan pelabelan secara otomatis menggunakan program *pyhton* *sentistrength\_id*, *sentistrength\_id* adalah program deteksi untuk menghitung kekuatan sentimen pada kalimat atau kata Bahasa Indonesia. Pengujian ini dilakukan sebagai perbandingan hasil klasifikasi suatu dataset yang diberi label manual dengan hasil klasifikasi yang diberi label otomatis berdasarkan daftar kata positif dan negatif.

Setelah melakukan percobaan nilai *k* sebanyak 50 kali dengan pembagian data yang berbeda hasil terbaik yaitu yang mempunyai nilai *accuracy* tertinggi pada setiap skenario percobaan seperti pada Tabel 2 sebagai berikut.

TABEL 2  
HASIL SKENARIO PERCOBAAN MENGGUNAKAN  
PELABELAN OTOMATIS

Data Training	Data Testing	Nilai k terbaik	Accuracy
60%	40%	3	89,25
		4	89,25
65%	35%	4	90,57
		5	90,57
70%	30%	5	91,00
75%	25%	5	92,00
80%	20%	6	93,00
		8	93,00
		9	93,00

Berdasarkan skenario pembagian data yang dilakukan didapatkan hasil untuk performa model yang paling baik dengan melihat nilai *accuracy* tertinggi yaitu pada pembagian jumlah data training 80% dan data testing 20% dengan nilai *accuracy* sebesar 93,00% pada  $k = 6$ ,  $k = 8$  dan  $k = 9$ , karena memiliki nilai *accuracy* tertinggi yang sama pada  $k$  berbeda maka dipilih salah satu nilai  $k$  untuk dijadikan pembahasan yaitu pada nilai  $k = 6$ . Berikut grafik plot data nilai *accuracy* pada tiap nilai  $k$ .



Gambar. 10. Grafik Nilai  $k$  Pelabelan Otomatis

### C. Word Cloud

*Word Cloud* berguna untuk melihat frekuensi dari sebuah kata yang sering muncul. *Word Cloud* memberikan hasil gambaran yang jelas dan informatif sehingga mudah dipahami. Pada pengujian kali ini akan ditampilkan hasil dari *word cloud* berdasarkan klasifikasi tendensi portal berita dengan model pembagian 80% data *training* dan 20% data *testing* yang memiliki *accuracy* tertinggi yaitu  $k = 6$ .

- *Word Cloud* Pelabelan Manual

Kata-kata yang muncul pada tendensi berita positif menggunakan pelabelan manual adalah “astrazeneca”, “sinovac”, “pfizer”, “persen”, “antibodi”, “efektif”, ”halal” dan lainnya, untuk detail hasil *word cloud* tendensi kata positif dapat dilihat pada Gambar 11.



Gambar. 11. *Word Cloud* Tendensi Positif Pelabelan Manual

Kata-kata yang sering muncul pada tendensi berita negatif menggunakan pelabelan manual adalah “kanker”, “positif”, “beku”, “darah”, “efek”, “samping”, “astrazeneca”, “suntik” dan lainnya. Untuk detail hasil *word cloud* tendensi kata positif dapat dilihat pada Gambar 12.



Gambar. 12. *Word Cloud* Tendensi Negatif Pelabelan Manual

Kata-kata yang sering muncul pada tendensi berita netral menggunakan pelabelan manual adalah “astrazeneca”, “sinovac”, “suntik”, “lasnia”, “mulai”, “indonesia”, “terima”, “kemenkes” dan lainnya, untuk detail hasil *word cloud* tendensi kata netral dapat dilihat pada Gambar 13.



terbalik dengan netral yang mempunyai nilai *precision* dan *recall* yang cukup tinggi. Hal tersebut mengindikasikan tingkat *accuracy* untuk prediksi kelas positif dan negatif masih kurang. Kurangnya tingkat *accuracy* dapat diakibatkan karena tidak seimbang kelas atau *imbalanced data*, dari total data yang dikumpulkan dan diberi label manual, dimana jumlah kelas positif 222 data, kelas negatif 150 data sedangkan netral 628 data. Perbedaan jumlah label antar kelas yang signifikan membuat model sulit untuk memprediksi hasil yang diujikan. Disamping itu kurangnya tingkat *accuracy* dapat disebabkan karena data yang di klasifikasikan hanya berdasarkan judul tanpa mengikutsertakan isi berita.

Berbeda dari hasil pelabelan manual, untuk dataset yang diberi label secara otomatis menggunakan program *sentistrength\_id* berdasarkan *list* atau daftar kata yang mengandung positif dan negatif menghasilkan nilai *accuracy* yang lebih tinggi yaitu sebesar 93% pada  $k = 6$  dengan pembagian data 80% training dan 20% testing, walaupun memiliki nilai *accuracy* yang cukup tinggi namun hasil labeling otomatis dengan menggunakan *sentistrength\_id* masih kurang akurat. Labeling otomatis menggunakan *sentistrength\_id* menghasilkan label dominan netral. Dari Hasil labeling yang dominan netral tersebut apabila model melakukan *training* dan *testing* akan didapatkan hasil klasifikasi yang dominan netral. Penjelasan lebih detail untuk evaluasi confusion matrix hasil klasifikasi menggunakan pelabelan otomatis dapat dilihat pada Gambar 17 berikut.

	precision	recall	f1-score	support
Negatif	0.67	0.52	0.58	27
Netral	0.82	0.96	0.89	136
Positif	0.63	0.32	0.43	37
accuracy			0.79	200
macro avg	0.71	0.60	0.63	200
weighted avg	0.76	0.79	0.76	200

Gambar. 17. Confusion Matrix  $k = 6$  Pelabelan Otomatis

Terlihat pada Gambar 9 nilai positif maupun negatif untuk *precision* dan *recall* rata-rata sangat kecil, bahkan untuk nilai

*precision* dan *recall* untuk kelas negatif hanya 0 karena tidak ada hasil prediksi yang masuk kedalam kelas negatif. Dari hasil pelabelan otomatis yang dominan netral terlihat jelas bahwa nilai *precision* dan *recall* untuk kelas netral sangat tinggi. Hal tersebut mengindikasikan prediksi yang diberikan model cenderung dominan netral. Hal ini dapat diakibatkan karena tidak seimbang kelas atau *imbalanced data*, dimana dari total data yang dikumpulkan dan diberi label otomatis jumlah kelas positif 90 data, kelas negatif 20 data sedangkan netral 890 data. Dari hasil analisis secara manual masih banyak pelabelan *sentistrength\_id* yang tidak akurat sehingga dapat mengakibatkan ketidakakuratan dalam memprediksi data. Dengan demikian tingkat *accuracy* tinggi pada hasil klasifikasi data dengan pelabelan otomatis tidak merepresentasikan hasil yang sebenarnya dikarenakan tidak akuratnya hasil labeling secara otomatis pada data *training* maupun *testing*.

Hasil visualisasi text dalam bentuk *word cloud* memberikan sebuah gambaran kata yang sering atau paling banyak muncul dalam pemberitaan di portal berita online tentang vaksinasi COVID-19. Dari hasil *word cloud* kata tendensi berita positif pada Gambar 3 kata yang sering muncul di judul portal berita online ditandai dengan kata yang ukurannya lebih besar diantaranya yaitu "astrazeneca", "sinovac", "pfizer", "efektif", "lawan", "antibodi", "persen", "halal", "kebal", "varian", "sembuh" dan "baru". Hal ini menunjukkan portal berita telah memberikan informasi tentang vaksin COVID-19 mulai dari nama vaksin yang beredar di masyarakat hingga efektifitas penggunaan vaksin terhadap kekebalan tubuh. Disamping itu informasi kehalalan vaksin COVID-19 juga menjadi salah satu isu yang merupakan tendensi positif dalam pemberitaan.

*Word cloud* juga menampilkan visualisasi kata yang sering muncul pada tendensi berita negatif tentang vaksinasi COVID-19. Dari hasil *word cloud* kata tendensi berita negatif pada Gambar 4 kata yang sering muncul di title atau judul portal berita online meliputi "beku", "darah", "efek",

“samping”, “kanker”, “usai”, “positif”, “suntik”, “astrazeneca”. Hal ini menunjukkan bahwa vaksin yang beredar dan digunakan saat ini belum sepenuhnya berkerja dengan baik, munculnya kata “efek”, “samping”, “beku”, “darah” juga menjelaskan bahwa vaksin dapat menimbulkan efek negatif yang akan membuat masyarakat mengalami kejadian ikutan pasca imunisasi (KIPI) setelah melakukan vaksinasi COVID-19.

Hasil *word cloud* untuk tendensi berita netral menunjukkan bahwa kata yang sering muncul “indonesia”, “menkes”, “distribusi”, “juta”, “mulai”, “terima”. Jika dilihat dari kata yang muncul pada tendensi berita netral terlihat kata-kata tersebut tidak terindikasi sebagai berita yang bertendensi positif maupun negatif.

Berdasarkan hasil *word cloud* menggunakan pelabelan manual kata “astrazeneca” selalu muncul di setiap tendensi berita positif, negatif maupun netral. Hal tersebut mengindikasikan bahwa pemberitaan mengenai vaksin astrazeneca memiliki pemberitaan yang berimbang.

*Word cloud* pada pelabelan otomatis untuk tendensi kelas positif didapatkan hasil kata yang sering muncul yaitu “kriteria”, “pfizer”, “populer”, “kuat”, “lawan”. Adanya kata “kuat” dan “lawan” mengindikasikan bahwa salah satu isu berita yaitu tentang efektifitas vaksin yang kuat dan dapat melawan COVID-19. Berbeda dari hasil *word cloud* untuk tendensi negatif, pada pelabelan otomatis tidak ada *word cloud* yang dihasilkan karena tidak adanya hasil klasifikasi yang masuk kedalam kelas negatif. Sementara itu *word cloud* untuk tendensi netral menghasilkan kata yang sering muncul yaitu “astrazeneca”, “sinovac”, “menkes”, “lansia”, “mulai”, “indonesia”, “terima”, “kemenkes”, “aman”, “halal”, “beku darah”, “efek samping”, “resiko” dan lainnya. Jika dilihat dari kata yang muncul pada tendensi berita netral terlihat kata-kata tersebut ada yang sebenarnya lebih cocok untuk masuk kedalam kata pada kelas positif, seperti “aman” dan “halal”, serta ada yang lebih cocok untuk masuk kedalam kata pada kelas negatif,

seperti “beku darah”, “efek samping”, “resiko”.

Berdasarkan hasil ujicoba dapat disimpulkan bahwa metode *K-Nearest Neighbor* dapat digunakan untuk mengklasifikasikan tendensi berita tentang vaksinasi COVID-19, dengan tingkat *accuracy* yang berbeda antara pelabelan manual dan pelabelan otomatis menggunakan *sentistrength\_id*. Disamping itu aplikasi juga dapat menampilkan *word cloud* dari setiap kelas yang teridentifikasi.

## 5. Kesimpulan dan Saran

### A. Kesimpulan

Berdasarkan hasil penelitian analisis tendensi portal berita menggunakan metode *K-Nearest Neighbor*, dapat ditarik kesimpulan sebagai berikut :

1. Metode *K-Nearest Neighbor* dapat digunakan untuk mengklasifikasikan tendensi pada portal berita tentang vaksinasi COVID-19 dengan tingkat *accuracy* tertinggi sebesar 78,50% pada  $k = 6$  menggunakan pelabelan manual. Sementara itu pada data dengan pelabelan otomatis dengan *sentistrength\_id* diperoleh *accuracy* sebesar 93% pada  $k = 6$ . Adapun data ujicoba yang digunakan sebanyak 1000 data dengan 80% data *training* dan 20% data *testing*.
2. Tingkat *accuracy* berada di kisaran 78,50% pada pelabelan manual dapat diakibatkan karena adanya *imbalanced* data yang digunakan serta berita yang di klasifikasikan hanya berdasarkan judul tanpa mengikutsertakan isi berita.
3. Walaupun tingkat *accuracy* pada pelabelan otomatis mencapai 93% tetapi tidak merepresentasikan hasil yang sebenarnya dikarenakan tidak akuratnya hasil labeling secara otomatis pada data *training* maupun *testing*. Kata yang muncul pada tendensi berita netral ada yang lebih cocok untuk masuk kedalam kata pada kelas positif, seperti “aman” dan “halal”, serta ada yang lebih cocok untuk masuk kedalam kata pada kelas negatif, seperti “beku darah”, “efek samping”, “resiko”.

4. Pada pelabelan otomatis model tidak dapat mengidentifikasi kelas negatif dikarenakan jumlah data yang *imbalanced* untuk setiap kelas.
5. Dari hasil klasifikasi yang dilakukan dalam penelitian ini tentang tendensi portal berita terhadap pemberitaan vaksinasi COVID-19, metode ini dapat mengidentifikasi kata-kata yang dominan, baik dalam pemberitaan positif, negatif maupun netral untuk kelas yang teridentifikasi.
6. Berdasarkan hasil *word cloud* pada pelabelan manual terdapat kata yang muncul, baik di pemberitaan positif, negatif maupun netral yaitu kata "astrazeneca". Hal tersebut mengindikasikan bahwa pemberitaan mengenai vaksin astrazeneca adalah berimbang

#### B. Saran

Pada penelitian ini masih memiliki beberapa kekurangan dan keterbatasan. Oleh karena itu, ada beberapa hal yang perlu diperhatikan untuk pengembangan selanjutnya agar bisa memberikan informasi lebih baik :

1. Pelabelan manual dilakukan oleh pakar atau ahli bahasa agar dapat memberikan labeling lebih baik.
2. Untuk meningkatkan nilai *accuracy* dari hasil klasifikasi sebaiknya menggunakan data dengan jumlah yang berimbang untuk setiap kelas.
3. Klasifikasi dilakukan tidak hanya berdasarkan judul akan tetapi memperhitungkan isi berita.
4. Adanya penelitian terkait dengan menggunakan metode yang berbeda atau dibandingkan dengan metode klasifikasi lainnya. ntuk mengetahui model yang dibangun mempunyai *accuracy* dan tingkat ketepatan maka dibutuhkan

#### Daftar Pustaka

Edukasi, J., Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). *Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes*. 4(2), 113–118.

Farisa, F. C. (2021). *Jalan Panjang Vaksin Covid-19 Sinovac di RI: Dari Uji Klinis hingga Dapat Izin Penggunaan*. Www.Kompas.Com.

<https://nasional.kompas.com/read/2021/01/12/08364271/jalan-panjang-vaksin-covid-19-sinovac-di-ri-dari-uji-klinis-hingga-dapat?page=all>

Haryanto, D. J., Muflikhah, L., & Fauzi, M. A. (2018). Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 2(9), 2909–2916.

Informatikalogi. (2017). *Algoritma K-Nearest Neighbor (K-NN)*. Www.Informatikalogi.Com. <https://informatikalogi.com/algoritma-k-nn-k-nearest-neighbor/>

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Nugroho, E. (2011). *Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp*. Universitas Brawijaya Malang.

Septian, J. A., Fahrudin, T. M., & Nugroho, A. (2018). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF - IDF dan K - Nearest Neighbor. *Journal Of Intelligent Systems And Computaion*, 43–49.

Soepomo, P. (2014). Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes. *Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes*, 2(3), 73–83. <https://doi.org/10.12928/jstie.v2i3.2877>

Suteja, J. (2021). *Pemerintah Rangkul Media Massa Tingkatkan Kesadaran Masyarakat soal Vaksin Covid-19*. Www.Beritasatu.Com.

<https://www.beritasatu.com/kesehatan/701843/pemerintah-rangkul-media-massa-tingkatkan-kesadaran-masyarakat-soal-vaksin-covid19>

- Yulian, E. (2018). Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung. *Jurnal Matematika "MANTIK,"* 4(1), 53–58. <https://doi.org/10.15642/mantik.2018.4.1.53-58>