INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

PENERAPAN ANALISIS SENTIMEN UJARAN KEBENCIAN TERHADAP VAKSINASI COVID-19 PADA TWEET BERBAHASA INDONESIA MENGGUNAKAN ALGORITME *K-NEAREST NEIGHBOR*

Juan Kalyzta[1], Muhammad Ardi Willdan[2], Selfiana Halfiani[3], Indra[4]

[1,2,3,4] Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Email: [1]1811500071@student.budiluhur.ac.id , [2]1811500402@student.budiluhur.ac.id,
[3]1811500899@student.budiluhur.ac.id, [4*]indra@budiluhur.ac.id
(* : coressponding author)

Abstrak-Sistem vaksinasi dengan memanfaatkan media komunikasi dan informasi, tanpa dibatasi oleh kendala waktu, ruang dan tempat serta keterbatasan sistem vaksinasi. Kurangnya kesiapan dalam menerapkan sistem vaksinasi baru tersebut memaksa banyak pihak untuk dapat beradaptasi dalam waktu yang cepat. Sistem vaksinasi yang semula dianggap sebagai solusi mulai menuai beragam pendapat dari masyarakat. Penelitian ini bertujuan untuk melakukan analisis ujaran kebencian pada masa vaksinasi pada media sosial Twitter. Metode yang digunakan adalah dengan melakukan analisis sentimen melalui pendekatan machine learning dengan pelabelan secara manual oleh pakar, dengan ekstraksi fitur menggunakan CountVectorizer dan Algoritme klasifikasi K-Nearest Neighbor. Nilai pengujian dan evaluasi tertinggi yang diperoleh sebesar: akurasi 35%, presisi 20% dan *recall* 100% menggunakan nilai K=3.

Kata Kunci: SuperDecisions, AHP, Kluster, Kriteria, Alternatif

Abstract-*The vaccination system utilizes communication and information media, without being limited by the constraints of time, space and place as well as the limitations of the vaccination system. The lack of readiness to implement the new vaccination system has forced many parties to adapt quickly. The vaccination system, which was originally considered a solution, has begun to reap various opinions from the public. This study aims to analyze the public's view of the vaccination system on Twitter social media. The method used is to perform sentiment analysis through a machine learning approach with a sentiment dictionary feature, with feature extraction using the CountVectorizer and the K-Nearest Neighbor classification algorithm. The highest test and evaluation values obtained were: 35% accuracy, 20% precision and 100% recall using the value of K=3*

Keywords*: sentiment analysis, twitter, vaccination, countvectorizer, k- nearest neighbor*

1. PENDAHULUAN

Virus Corona mulai menjadi wabah pada bulan November - Desember 2019 di kota Wuhan, China. Virus ini merupakan salah satu virus yang sangat berbahaya karena tingkat penyebarannya yang tinggi sehingga meluas dengan cepat ke seluruh dunia. Menurut catatan WHO, pada tahun 2020 sudah banyak laporan dari berbagai negara yang mengonfirmasi terjangkit virus corona atau COVID-19. Di Indonesia pertama kali terdeteksi adanya warga yang terjangkit virus corona pada tanggal 2 maret 2020, yang terjadi pada dua orang warga Depok, Jawa Barat. Semenjak saat itu, dari catatan satgas pemulihan COVID-19, diketahui semakin banyak kasus yang terkonfirmasi *Hatespeech* dari bulan ke bulan.

Media sosial Twitter menjadi salah satu tempat masyarakat dapat dengan bebas menyampaikan pendapat. Banyak metode analisis yang dapat digunakan untuk menganalisis pendapat masyarakat berdasarkan informasi yang ada pada media sosial semacam Twitter. Salah satu di antaranya adalah metode analisis sentimen. Metode analisis sentimen merupakan salah satu metode untuk menganalisis Data yang didapatkan dari internet sehingga dapat diketahui polaritas dari Data tersebut. Dengan menggunakan analisis sentimen, polaritas dari opini yang ada dapat dikumpulkan, sehingga akan dapat digunakan untuk memprediksi suasana publik atau gambaran perasaan netizen bersifat *Non Hatespeech* atau *Hatespeech*. Beberapa penelitian tentang klasifikasi sentimen pada konten media sudah dilakukan sebelumnya. Dalam sebuah penelitian, dilakukan prediksi penyebaran COVID-19 di Indonesia dengan menggunakan algoritme KNN.

K-Nearest Neighbor, menyatakan bahwa Algoritme *K-Nearest Neighbor* (KNN) mampu memperoleh nilai akurasi 32% dengan nilai K=3 [1]. Penelitian lain yang berjudul Perbandingan Metode *Naive Bayes*, KNN dan *Decision Tree* Terhadap Analisis Sentimen Transportasi KRL *Commuter Line*, menyatakan bahwa KNN dapat digunakan untuk analisis sentimen dengan nilai akurasi sebesar 80% terhadap 127 Data dan mampu mengimbangi Algoritme Naive Bayes Classifier [2]. KNN juga digunakan oleh Novelty dan Adiwijaya dalam penelitian yang berjudul Sentiment Analysis on Movie Reviews Using Information Gain and *K-Nearest Neighbor*, untuk melakukan analisis sentimen terhadap Data*set* review film dengan total 2000 Data, memperoleh hasil yang baik pada K=3

INDONESIA JOURNAL INFORMATION SYSTEM (IDEALIS)
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

dengan nilai akurasi sebesar 35% [3]. Penelitian deteksi Hate Speech pada dataset berbahasa Indonesia menggunakan metode Naïve Bayes, Support Vector Machine, Bayesian Logistic Regression, dan Random Forest Decision Tree (RFDT), dengan hasil metode RFDT memiliki nilai F-Measure tertinggi 93,5% [4]. Pelabelan data tweet menggunakan 3 relawan berumur 17-24 tahun dengan beragam suku, agama dan jenis kelamin yang berbeda. Jika 3 relawan menyatakan tweet tersebut adalah *hate speech* maka dilabelkan *hatespeech*, jika salah satu menyatakan bukan *hatespeech* maka tidak dilabelkan *hate speech*.

Penelitian [4] menjadi dasar dalam pelabelan tweet *hate speech* atau *non hate speech*. Namun, pada penelitian [4], belum menggunakan metode K-NN dalam mengklasifikasi *tweet hate speech* tersebut. Oleh karena itu, pada penelitian ini bertujuan untuk melakukan analisis sentimen ujaran kebencian di masyarakat pada masa vaksinasi di Juli 2021 menggunakan metode K-Nearest Neighbor. Data*set* yang digunakan berupa teks kicauan (*Tweet*) yang bersumber pada media sosial Twitter dengan kata kunci 'vaksinasi', 'vaksin','vaksinasicovid19','#vaksinasi', '#vaksin', '#vaksinasicovid19, "#jokowi", dan "#ppkm".

2. METODE PENELITIAN

2.1 Twitter

Twitter merupakan sebuah situs media sosial yang mulai dikembangkan pada tahun 2006. Situs ini pertama kali ditemukan oleh Jack Dorsey dan Evan Williams. Twitter merupakan *social networking* dimana memungkinkan penggunanya dapat saling berkomunikasi satu sama lain melalui fitur yang bernama *Tweet*.

2.2 *Hatespeech*

Hate Speech yaitu ucapan atau tulisan yang dibuat seseorang di muka umum untuk menyebarkan dan menyulut kebencian suatu kelompok terhadap kelompok lain yang berbeda ras, agama, keyakinan, gender, etnisitas, kecacatan, dan orientasi seksual [5]. Sedangkan menurut Margareth Brown Sica dan Jeffrey Beall menyatakan bentuk hate speech atau ujaran kebencian seperti menghina, merendahkan kelompok minoritas tertentu, dengan berbagai latar belakang dan sebab baik berdasarkan ras, gender, etnis, kecacatan, kebangsaan, agama, orientasi seksual atau karakteristik lain [6].

Dalam dunia hukum *hate speech* merupakan perkataan, perilaku, tulisan, dan pertunjukan yang dilarang karena dapat menimbulkan terjadinya aksi tindakan kekerasan dan sikap prasangka buruk dari pelaku pernyataan tersebut ataupun korban dari tindakan tersebut. Sedangkan penggunaan dan penerapan ujaran kebencian dalam dunia internet disebut *Hate Site*, kebanyakan dari situs ini menggunakan Forum Internet dan berita untuk mempertegas suatu sudut pandang tertentu. Berdasarkan hukum yang berlaku Hatespeech terdapat Udang-Undang menurut UU ITE Pasal 27 Ayat (3) UU ITE yaitu: Setiap Orang dengan sengaja dan tanpa hak mendistribusikan dan/atau mentransmisikan dan/atau membuat dapat diaksesnya Informasi Elektronik dan/atau Dokumen Elektronik yang memiliki muatan penghinaan dan/atau pencemaran nama baik. Menurut UU ITE Pasal 28 Ayat 2, setiap orang dilarang "dengan sengaja dan tanpa hak menyebarkan informasi yang ditujukan untuk menimbulkan rasa kebencian atau permusuhan individu dan/atau kelompok masyarakat tertentu berdasarkan atas suku, agama, ras, dan antargolongan (SARA). Dalam Kitab Undang-undang Hukum Pidana (KUHP) dijelaskan bahwa *Hatespeech* dapat dikategorikan sebagai berikut:

a) Penghinaan
   Proses, cara, perbuatan menghina(kan).
b) Pencemaran nama baik
   Perbuatan menyerang kehormatan atau nama baik seseorang dengan menuduhkan sesuatu hal yang maksudnya terang supaya hal itu diketahui umum.
c) Penistaan
   Proses, cara, perbuatan menistakan.
d) Perbutan tidak menyenangkan
   Memaksa orang lain untuk melakukan sesuatu dengan disertai ancaman baik verbal maupun fisik.
e) Meprovokasi
   Perbuatan untuk membangkitkan kemarahan, tindakan menghasut, penghasutan, pancingan.
f) Menghasut
   Membangkitkan hati orang supaya marah (melawan, memberontak, dan sebagainya).
g) Penyeberan berita bohong
   Usaha untuk menipu atau mengakali pembaca/pendengarnya untuk mempercayai sesuatu, padahal sang pencipta berita palsu tersebut tahu bahwa berita tersebut palsu.

INDONESIA JOURNAL INFORMATION SYSTEM (IDEALIS)
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

2.3 Data *Mining*

Data *mining* merupakan serangkaian proses untuk menggali informasi dengan melakukan analisa Data untuk menemukan suatu pola dari kumpulan Data tersebut. Data *mining* mampu menganalisa Data yang besar menjadi ekstraksi berupa pola yang mempunyai arti bagi pendukung keputusan [7]. Data mining juga bisa disebut knowledge discovery adalah proses pengambilan pola pada Data yang akan di proses lalu output tersebut berupa informasi yang sangat penting. Proses yang dilakukan untuk mengekstrak pengetahuan dalam Data mining adalah pengenalan pola, clustering, asosiasi, prediksi dan klasifikasi [8]. Data *mining* memiliki variasi untuk menemukan pola dari ekstraksi sebuah kumpulan sekumpulan Data tekstual yang disebut dengan *text mining*. *Text mining* memiliki fokus pada pengolahan Data berupa kata atau teks.

2.4 *Text Mining*

Text mining merupakan bagian dari Data *mining*, yang mana digunakan untuk mendapatkan informasi dari sebuah Data atau dokumen berupa sekumpulan teks yang memiliki format yang terstruktur ataupun tidak terstruktur dengan jumlah yang besar. Dalam *text mining* memiliki tugas khusus yaitu klasifikasi dan klasterisasi. Sedangkan dalam penerapannya, *text mining* berfungsi untuk mencari pola dalam teks, meganalisa teks agar bisa menghasilkan keluaran berupa informasi yang bermanfaat pada tujuan tertentu. Dikarenakan Data yang diproses pada *text mining* merupakan sebuah teks yang tidak terstruktur, maka diperlukan pemilihan teks sebelum dilakukan proses selanjutnya, pada tahap ini dikenal dengan prapemrosesan (*Preprocessing*).

2.5 Analisis Sentimen

Analisis Sentimen merupakan kajian tentang cara menyelesaikan dan memecahkan masalah dari berdasarkan opini masyarakat, sikap serta emosi suatu entitas, dimana entitas tersebut dapat mewakili individu. Analisis Sentimen atau yang juga disebut *opinion mining* merupakan proses memahami, mengekstrak serta mengolah Data tekstual secara otomatis guna mendapatkan informasi yang terkandung dalam suatu kalimat opini. Dilakukannya analisis sentimen ini bertujuan untuk melihat pendapat atau kecenderungan opini terhadap suatu masalah ataupun objek oleh seseorang, apa memiliki kecenderungan positif atau negatif Hatespeech [9].

2.6. *Crawling*

Crawling merupakan proses mengumpulkan Data dari sebuah laman dan menyimpannya untuk diatur dan dianalisis lebih lanjut. Dalam penelitian ini proses *Crawling* dilakukan menggunakan standard search API Twitter dengan pustaka Tweepy. Penggunaan pustaka Tweepy bertujuan untuk memperoleh Data *Tweet* pada Twitter dengan akses menggunakan API *Key* yang didapatkan melalui akun *developer* Twitter. Kata kunci untuk penarikan data adalah ‘vaksinasi’, ‘vaksin’,‘vaksinasicovid19’,‘#vaksinasi’, ‘#vaksin’, ‘#vaksinasicovid19, “#jokowi”, dan “#ppkm”. Kata kunci ditentukan berdasarkan nama event terkait Hashtag Vaksinasi, tokoh yang berpengaruh di masa Vaksinasi dan organisasi yang telibat dalam Vaksinasi [10].

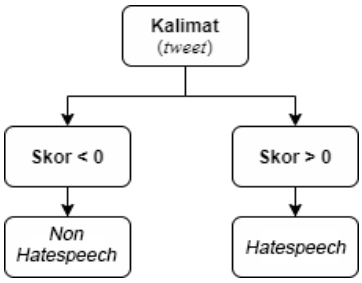
2.7. *Preprocessing*

Preprocessing merupakan bagian dari *text mining* yang dilakukan untuk menghapus *noise* pada dokumen atau kalimat. Selain itu, proses ini bertujuan untuk menghindari Data yang kurang sempurna; gangguan pada Data; dan Data yang tidak konsisten [11]. Proses pengubahan Data teks yang tidak terstruktur menjadi Data teks yang terstruktur sangat diperlukan sehingga perlu adanya proses pra-pemrosesan Data [12]. Merujuk pada penelitian yang telah dilakukan [13]–[16] maka dalam penelitian ini akan dilakukan beberapa tahapan *Preprocessing* teks antara lain: *Case Folding*, *Cleansing*, mengubah *slang word*, menghapus *stop word*, dan *stemming*.

2.8. Pelabelan

Pelabelan merupakan proses pemberian kelas berdasarkan ciri atau karakteristik yang terkandung dalam sebuah dokumen atau kalimat. Performa pembagian kelas lebih baik terbagi menjadi dua (2) kelas, yakni sentimen *Hatespeech* dan sentimen *Non Hatespeech*. Dalam penelitian ini proses pelabelan akan memberikan kelas pada tiap *Tweet* dengan *Hatespeech* atau *Non Hatespeech* (2 kelas) yang dapat dilakukan dengan cara membentuk tim pelabelan yang berisi 3 orang (2 laki-laki – 1 perempuan) dengan umur 17-24 Tahun, setiap tweet ujaran kebencian harus berdasarkan UU ITE dan dilakukan pengecekan fakta sesuai berita yang ada, setiap anggota memberikan skor 1 pada setiap kalimat yang dianggap mengandung unsut *Hatespeech*, pada Gambar 1 adalah hasil akhir pelabelan skor yang bernilai 3 dianggap *Hatespeech* dan skor 0 dianggap *Non Hatespeech* [4]. Dalam pelabelan masing masing anggota memberikan skor berupa satu. Jumlah data training yang digunakan dalam penelitian ini sebanyak 317 *record*, dengan label *Hatespeech* sebanyak 151 dan label non *Hatespeech* 161 *record*.

INDONESIA JOURNAL INFORMATION SYSTEM (IDEALIS)
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 1. Pemberian Kelas Sentiment

2.9. *CountVectorizer*

CountVectorizer merupakan proses pengolahan dokumen atau teks menjadi bentuk vektor. *CountVectorizer* digunakan untuk menghitung frekuensi kata dalam dokumen atau kalimat kemudian direpresentasikan ke dalam bentuk vector [17].

2.10. Pemodelan

Pemodelan merupakan proses pembuatan pengetahuan berdasarkan Data latih yang telah tersedia. Data latih yang dijadikan *model* dipilih dengan teknik sampling kuota (*quota sampling*). *Quota Sampling* merupakan teknik *sampling* yang menentukan jumlah sampel dari populasi yang memiliki ciri atau kriteria tertentu hingga jumlah kuota yang diinginkan tercapai [11].

2.11. K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah Algoritme klasifikasi *supervised learning* berbasis jarak. Algoritme ini bekerja dengan cara membandingkan jarak antara Data uji dengan semua Data latih yang ada [1], [2]. Untuk menghitung jarak antara Data digunakan perhitungan *euclidean distance* dengan rumus:

$$\boldsymbol{d_{(x,y)}}\sqrt{\sum_{i=1}^{n}(\boldsymbol{x_{i-}}\,\boldsymbol{y_i})^2} \qquad (1)$$

Keterangan:
d(x,y) = Jarak antara Data Uji dengan Data Latih
n = Jumlah Fitur
Xi = Fitur ke-*i* dalam Data Uji
Yi = Fitur ke-*i* dalam Data Latih

2.12. Pengujian

Pengujian merupakan hal penting untuk memastikan bahwa suatu algoritme yang telah dirancang dapat berjalan sesuai dengan harapan. Pengujian klasifikasi sentimen dilakukan dengan menguji aplikasi yang telah dibangun dengan membandingkan antara Data prediksi dan Data aktual. Data prediksi berupa hasil klasifikasi yang dihasilkan oleh aplikasi yang dibangun, sedangkan Data aktual berupa yang didapatkan melalui proses pelabelan. Dalam penelitian ini, pengujian dilakukan pada sebuah *model* terhadap Data uji yang tersedia. Hasil dari pengujian tersebut akan dievaluasi menggunakan *confusion matrix* untuk mengukur tingkat akurasi, presisi dan *recall*, yang terlihat pada Tabel 1 berikut.

Tabel 1. *Confusion Matrix*

		Nilai Aktual	
		True (P)	False (N)
Nilai Prediksi	True (P)	TP	FP
	False (N)	FN	TN

3. HASIL DAN PEMBAHASAN

3.1 Tahap *Crawling*

Data*set* penelitian bersumber dari media sosial Twitter berupa Data teks kicauan (*Tweet)*. Data*set* tersebut diperoleh secara rutin melalui proses *Crawling* menggunakan pustaka Tweepy, dimulai pada tanggal 9 juli 2021 hingga 18 Juli 2021, sehingga diperoleh total Data*set* sejumlah 454 Data (*Tweet*). Kata kunci yang digunakan antara

INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

lain: 'vaksinasi, 'vaksin', 'vaksincovid19', '#vaksinasi', '#vaksin', '#vaksincovid19', '#vaksinasicovid19', "#jokowi", dan "#ppkm".

Dataset yang berhasil dikumpulkan melalui proses *Crawling* dengan informasi antara lain: *Tweet id, full_text* (*Tweet*), *created_at, user.screen name* (*username*), akan disimpan ke dalam file *Excel (*.xlsx*)*, yang kemudian akan dimasukkan ke dalam basis Data (*Database*) MySQL untuk tahap selanjutnya (*Preprocessing*).

3.2 Tahap *Preprocessing*

Tahap *Preprocessing* merupakan tahapan yang hanya dapat dilakukan setelah tersedianya satu atau lebih Data*set* pada basis Data (*Database*) hasil dari tahapan pengumpulan Data tahapan ini terdiri atas lima (5) proses utama antara lain*: Case Folding*, *Cleansing*, mengubah *Slang Word*, menghapus *Stop Word*, dan *stemming*, yang terlihat pada Tabel 2 berikut.

Tabel 2. Tahapan Preprocessing

Tahapan *Preprocessing*	Hasil Tahapan *Preprocessing*
Tweet Awal	Mau vaksin gratis Atau vaksin berbayar Saya tetap #TOLAKVAKSIN
Case Folding	mau vaksin gratis atau vaksin berbayar saya tetap #tolakvaksin
Cleansing	mau vaksin gratisatau vaksin berbayarsaya tetap
Mengubah *Slang Word*	mau vaksin gratis atau vaksin berbayar saya tetap
Menghapus *Stop Word*	vaksin gratis atau vaksin berbayar saya tetap
Stemming	vaksin gratis atau vaksin berbayar saya tetap

3.3

3.3 Tahap Pelabelan

Pada Tabel 3 merupakan Tahap Pelabelan merupakan tahapan yang hanya dapat dilakukan setelah tersedianya satu atau lebih Data *clean text* pada basis Data (*Database*) hasil dari tahapan *Preprocessing*. Tahap Pelabelan utama dilakukan dengan membentuk satu kelompok yang beranggotakan 3 (tiga) orang yang terdiri dari 1 (satu) wanita dan 2 (dua) pria. *Dataset* yang sudah bersih dapat diberikan label. Label yang diberikan berupa *No* daan *Yes*, untuk menentukan *Hatespeech* dan *NonHatespeech* apabila dia *Hatespeech* maka *yes* kalau tidak Hatespeech maka *No*, sama dengan halnya *Hatespeech*. Kalau dia *NonHatespeech* maka *Yes*, apabila dia tidak *NonHatespeech* maka *No*. Cuitan yang dinilai sebagai *Hatespeech* harus memenuhi unsur provokasi, penghinaan, diskriminasi, ancaman kepada Suku, Agama, Ras, Pemerintahan dan Antar golongan yang bertujuan untuk memusuhi Subjek tertentu.

Tabel 3. Tahapan Pelabelan

Subjek	HS	NONHS
Gila gila Penguasa, udah ga ada mikir rakyat nih partai. Engga tau apa rakyat nya sudah susah masih ajh mikirin cuan cuan	*Yes*	*No*
Vaksinasi sdh mulai digodok dimana-mana, walaupun harus diakui (sayangnya) kl pelaksanaanya bs dikatakan amburadul di bbrp tempat????. Jepang salah satunya. Gmn g, prnh dgr ada org yg disuntik smpe 4x???????????????? Blog	*Yes*	*No*
Vaksin gratis msh tersedia banyak. Segera daftar dan datang ke gerai vaksin terdekat. Vaksin berbayar bagi yg mampu	*No*	*Yes*
Anggota DPR Mengkritisi kebijakan melalui PT Kimia Farma Tbk yang berencana memberlakukan vaksinasi berbayar.	*No*	*Yes*

3.3 Seleksi Data Latih

Seleksi Data latih dilakukan setelah Data melalui proses *Preprocessing*, Pelabelan, dan pembagian Data. Menggunakan teknik sampling kuota (*quota sampling*), tahap pertama dalam Pemodelan adalah pengambilan sampel dari populasi Data latih untuk dijadikan sebagai pengetahuan berdasarkan kriteria tertentu, kriteria yang dimaksud adalah dengan menyamakan jumlah antara Data berlabel *Hatespeech* dengan Data berlabel *Non Hatespeech*. Sampel Data latih yang diambil dapat dilihat pada Tabel 4 berikut:

INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Tabel 4. Sampel data latih

Tweet (*Tlatih-i*)	*Clean Text*	*Tipe*
latih-1	ppkm darurat tepat kondisi perintah lalu kemensos dukung cair bantu bst lain lain	*HS*
latih-2	alhamdulillah dapat bantu pemerinta biar rumah tangga tetap bisa jalan	*HS*
latih-3	hore bst cair buru informasi website resmi	*HS*
latih-4	mangsa turut dakwa marah minta madam rakaman video lepas dapat sunti benar	*NONHS*
latih-5	tahniah malaysia register dah baru je ko*Non* update ingatkan selit namaehhhh sama jeeeee nyahhh tetiba pulak tl penuh naik pangkat	*NONHS*
latih-6	register lambat dapat is real gaissss register julai julai dapat appointment tahniahyg daftar tidak reput haaaa	*NONHS*

3.4 Pembuatan List Kata

Kemudian akan dipisahkan menjadi satuan kata. Pemisahan menjadi kata dilakukan berdasarkan spasi (*whitespace*), kemudian hasilnya akan ditampung dalam sebuah wadah *list*. Hasil proses ini dapat dilihat pada Tabel 5 berikut:

Tabel 5. List kata

List Kata
[‘ppkm’, ‘darurat’, ‘tepat’, ‘kondisi’, ‘perintah’, ‘lalu’, ‘kemensos’, ‘dukung’, ‘cair’, ‘bantu’, ‘bst’, ‘lain’, ‘alhamdulillah’, ‘dapat’, ‘pemerinta’, ‘biar’, ‘rumah’, ‘tangga’, ‘tetap’, ‘bisa’, ‘jalan’, ‘horebst’, ‘buru’, ‘informasi’, ‘website’, ‘resmi’, ‘mangsa’, ‘turut’, ‘dakwa’, ‘marah’, ‘minta’, ‘madam’, ‘rakaman’, ‘video’, ‘lepas’, ‘sunti’, ‘benar’, ‘tahniah’, ‘malaysia ‘, ‘register’, ‘dahbaru’, ‘je’, ‘ko*Non*’, ‘update’, ‘ingatkan’, ‘selit’, ‘namaehhhh’, ‘sama’, ‘jeeeee’, ‘nyahhh’, ‘tetiba’, ‘pulak’, ‘tl’, ‘penuh’, ‘naik’, ‘pangkat’, ‘register’, ‘lambat’, ‘is’, ‘real’, ‘gaissss’, ‘julai’, ‘appointment’, ‘tahniahyg’, ‘daftar’, ‘tidak’, ‘reput’, ‘haaa’]

3.5 Membuat Vektor Kosong Latih

Membuat vektor kosong dimaksudkan untuk menyiapkan wadah berbentuk vektor dengan isian nilai awal yaitu angka (*integer*) nol (0). Wadah vektor tersebut dibentuk dengan panjang berdasarkan jumlah fitur kata. Maka vektor kosong yang dihasilkan seperti pada Tabel 6 berikut:

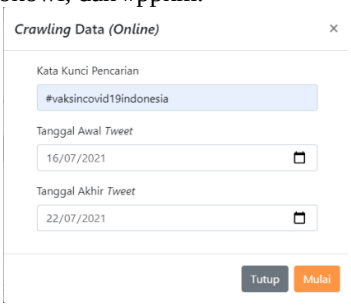
Tabel 6. Vektor Kosong

Tweet (*Tlatih-i*)	Vektor Kosong
latih-1	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
latih-2	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
latih-3	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
latih-4	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
latih-5	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
latih-6	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

3.7. Proses *Crawling*

Pada Gambar 1 akan dilakukan uji coba mengambil Data secara *online* melalui Twitter menggunakan *hashtag* #VaksinCovid19, #Vaksinasi, #vaksin, #jokowi, dan #ppkm.



Gambar 1. Data Crawling

Gambar 2 adalah hasil *Crawling* Data *Online* yang tersimpan ke dalam file Excel dengan format .xlsx.

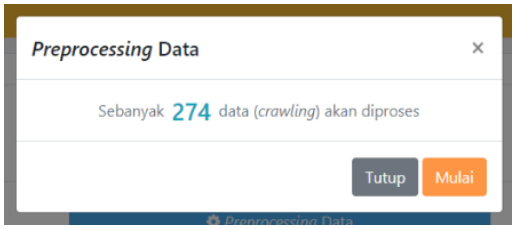
	A	B	C	D
1	id	text	username	created_at
2	1,41679E+18	Dalam rangka memperingati Hari Anak Nasional, Polres Metro Jakarta Barat bekerja sama dengan UNIFAM menyelenggarakan"V	Resjakbar	Sun Jul 18 15:50:24 +0000 2021
3	1,41678E+18	GERAI VAKSIN PRESISISilahkan bagi masyarakat yang belum melaksanakn vaksin bisa langsung datang ke Gerai Vaksin Presisi ya	Ujangsa96801330	Sun Jul 18 15:08:52 +0000 2021
4	1,41674E+18	[Video] Vaksin: Jururawat Teruja Dapat Suntik Nenek SendiriBerita penuh: https://t.co/F2uiDy6RRPhttps://t.co/PuyrPP9bPs@KKM	bernamadotcom	Sun Jul 18 12:47:37 +0000 2021
5	1,41674E+18	100 peratus dewasa di Malaysia dijangka lengkap vaksin pada Oktober https://t.co/qnQkJUaoMg#VaksinCOVID19 #MenangBers	PMR_Kuching	Sun Jul 18 12:47:26 +0000 2021
6	1,41674E+18	Gerai Vaksinasi Presisi mobile Polres Metro Jakarta BaratKehadiran nya tersebut sebagai bentuk kepedulian polri dalam rangka p	PolsekKbnJeruk	Sun Jul 18 12:39:19 +0000 2021
7	1,41674E+18	100 peratus dewasa di Malaysia dijangka lengkap vaksin pada Oktober https://t.co/orUY1u3zBL#VaksinCOVID19 #MenangBersa	wewantamar	Sun Jul 18 12:26:50 +0000 2021
8	1,41674E+18	100 peratus dewasa di Malaysia dijangka lengkap vaksin pada Oktober https://t.co/EPmgnsaJda#VaksinCOVID19 #MenangBersa	JapenSarawak	Sun Jul 18 12:26:49 +0000 2021
9	1,41674E+18	[3] #vaksinCovid19 maka akan digunakan terlebih dahulu untuk masyarakat umum yang memasuki jadwal vaksinasi #Covid19 ta	radiopatria	Sun Jul 18 12:23:21 +0000 2021
10	1,41673E+18	Gerai Vaksinasi Presisi mobile Polres Metro Jakarta BaratKehadiran nya tersebut sebagai bentuk kepedulian polri dalam rangka p	PolsekKbnJeruk	Sun Jul 18 12:18:24 +0000 2021
11	1,41673E+18	Gerai Vaksinasi Presisi mobile Polres Metro Jakarta BaratKehadiran nya tersebut sebagai bentuk kepedulian polri dalam rangka p	PolsekKbnJeruk	Sun Jul 18 12:18:23 +0000 2021
12	1,41673E+18	Gerai Vaksinasi Presisi mobile Polres Metro Jakarta BaratKehadiran nya tersebut sebagai bentuk kepedulian polri dalam rangka p	PolsekKbnJeruk	Sun Jul 18 12:13:35 +0000 2021
13	1,41673E+18	.Ayoo kita Vaksin untuk menuju Indonesia sehat dari Covid-19.#PolresJakbar #PolresMetroJakartabarat #pandemicovid19 #Covid1	HumasBar	Sun Jul 18 12:01:03 +0000 2021
14	1,41673E+18	.Ayoo kita Vaksin untuk menuju Indonesia sehat dari Covid-19.#PolresJakbar #PolresMetroJakartabarat #pandemicovid19 #Covid1	HumasBar	Sun Jul 18 12:00:48 +0000 2021
15	1,41673E+18	.Ayoo kita Vaksin untuk menuju Indonesia sehat dari Covid-19.#PolresJakbar #PolresMetroJakartabarat #pandemicovid19 #Covid1	HumasBar	Sun Jul 18 11:59:51 +0000 2021
16	1,41673E+18	.Ayoo kita Vaksin untuk menuju Indonesia sehat dari Covid-19.#PolresJakbar #PolresMetroJakartabarat #pandemicovid19 #Covid1	HumasBar	Sun Jul 18 11:59:47 +0000 2021
17	1,41673E+18	Yuk.. saksikan talkshow seputar vaksinasi bersama @radiomuara #vaksincovid19 https://t.co/al4wdgfgQa	mikirkriting	Sun Jul 18 11:43:11 +0000 2021
18	1,41672E+18	100 peratus dewasa di Malaysia dijangka lengkap vaksin pada Oktober https://t.co/xxkKDUCd6s#VaksinCOVID19 #MenangBersa	bernamadotcom	Sun Jul 18 11:03:51 +0000 2021
19	1,41671E+18	Done my first dose of sinovac. dah 3 hari tapi baru post hee, no side effect yang teruk kecuali lenguh tangan dan kuat makan	nrnhsya	Sun Jul 18 10:48:17 +0000 2021
20	1,41671E+18	Dah masuk cip pertama 5G Tunggu cip kedua pulak. Lepasni mesti saya dapat high speed connection #sinovac #vaksin	AdDien90	Sun Jul 18 10:36:42 +0000 2021
21	1,4167E+18	Sulitnya Perjuangan Indonesia Berebut Vaksin Covid-19 Dibeberkan Menlu Retno Marsudi - https://t.co/d6vhyP6enD https://t.co/	rmolbantencom	Sun Jul 18 10:12:03 +0000 2021
22	1,4167E+18	Alhamdulillah, saya sudah vaksin jenis Abdulla & Abdulla ! Secara rasminya saya boleh disambung melalui bluetooth ke sen	syahmie_14	Sun Jul 18 10:09:19 +0000 2021
23	1,4167E+18	Kapan jadwal kembali vaksin dosis ke 2 ??#vaksincovid19 #dinkeskotamalang #puskesmasbarengkotamalang @ Kota Malang - Ja	pkm_bareng_4a	Sun Jul 18 09:56:38 +0000 2021

Sheet1

Gambar 2. File Excel Data Crawling

3.8. Proses *Preprocessing*

Setelah selesai melakukan proses *Crawling* Data, maka selanjutnya adalah melakukan proses *Preprosessing* yang akan ditampilkan pada Gambar 3 berikut:



Gambar 3. Preprocessing Data

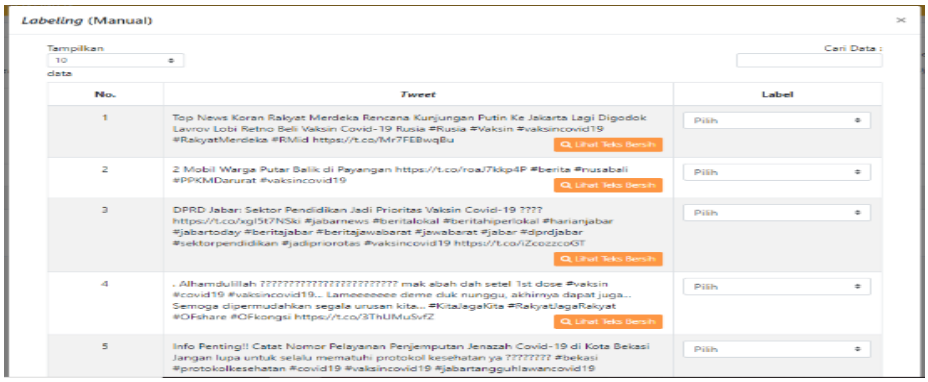
Gambar 4 adalah penjelasan detail tentang langkah-langkah proses *Preprocessing.*

INDONESIA JOURNAL INFORMATION SYSTEM (IDEALIS)
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Gambar 4. Proses Preprocessing

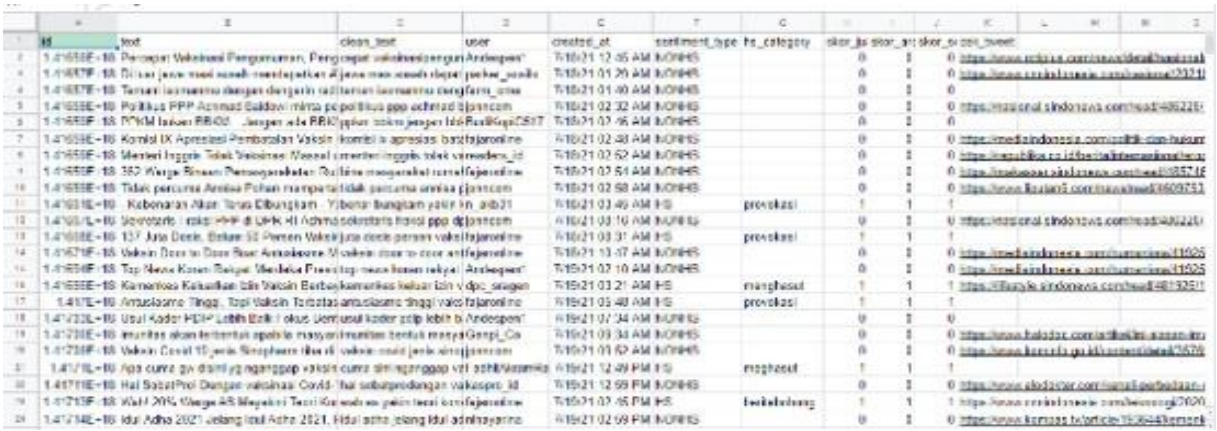
3.9. Proses Pelabelan

Berdasarkan hasil dari *Preprocessing* Data maka langkah selanjutnya adalah melakukan pelabelan pada tiap *Tweet* yang didapat, ada 2 cara untuk melakukan Pelabelan yaitu secara Manual dan Otomatis menggunakan Kamus. Gambar 5 adalah proses Pelabelan.



Gambar 5. Proses Pelabelan

Gambar 6 adalah proses pelabelan *manual* yang mana disini Data *Tweet* akan diolah lebih lanjut dimana setiap *Tweet* akan diberikan skor dengan 0 atau 1 dengan tujuan agar Tweet yang yang dilabelkan mutlak dan tidak ada yang bersifat ambigu, serta pada tahap ini setiap *Tweet* akan ditentukan sesuai golongannya apabila *Tweet* berlabel *Hatespeech* dan dicocokkan dengan berita yang terkait apabila *Tweet* berlabel *Non Hatespeech*.



Gambar 6. Proses Pelabelan Manual

INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

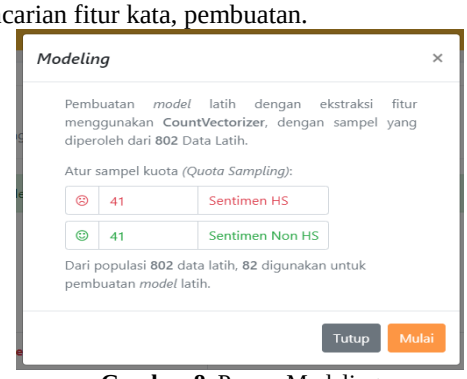
3.10. Proses Pembagian Data

Berdasarkan Data hasil Pelabelan, selanjutnya Data *Tweet* berlabel akan dibagi menjadi dua (2) bagian, yaitu Data Uji dan Data Latih. Pembagian Data dilakukan menggunakan rasio yang telah ditentukan yaitu 1:9 (Data Uji: Data Latih) atau 10% Data Uji dan 90% Data Latih. Yang terlihat pada Gambar 7.

Gambar 7. Proses Pembagian Data

3.11. Proses Pemodelan

Pada Gambar 8 dan Gambar 9 merupakan proses pemodelan menggunakan *CountVectorizer* dilakukan setelah *Tweet* melalui proses *Preprocessing*, Pelabelan dan pembagian Data. Proses ini bertujuan untuk memperoleh seleksi Data Latih, pembuatan list kata, pencarian fitur kata, pembuatan.



Gambar 8. Proses Modeling

Vektor hasil CountVectorizer:

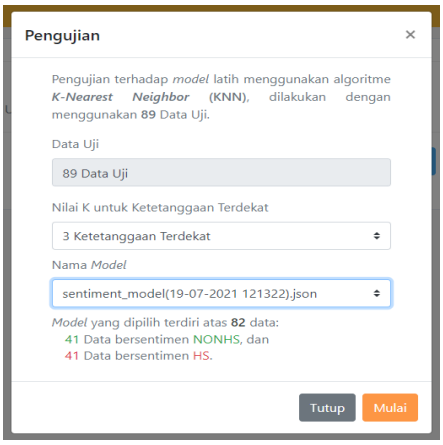
	breaking	sampai	maaf	kimia	farma	putus	tunda	jadwal	vaksinasi	bayar	covid	malinau	baru	capai	persen	satgas	usul	tambah	vaksin	sasar	salah	satu	momen	sejarah	dalam
Tweet ke-1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tweet ke-2	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Tweet ke-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Tweet ke-4	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
Tweet ke-5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tweet ke-6	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
.....																									

Gambar 9. Hasil CountVectorizer

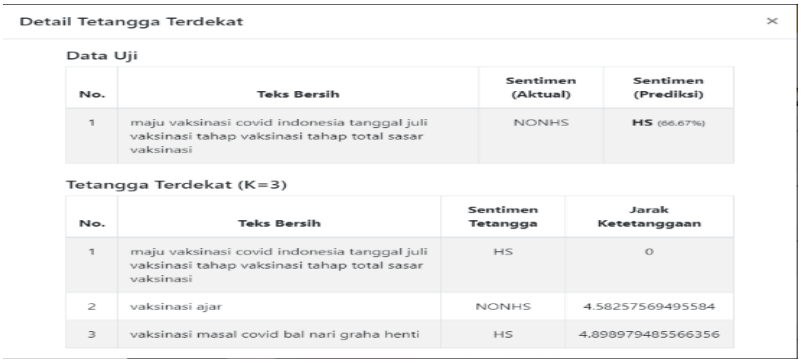
3.12. Proses Pengujian Data

Tahap Pengujian menggunakan metode klasifikasi algoritme *K-Nearest Neighbor* (KNN) merupakan tahapan yang dilakukan setelah tahap ekstraksi fitur (Pemodelan). Tahapan ini bertujuan untuk memprediksikan label untuk setiap Data Uji berdasarkan model latih yang dihasilkan pada tahap Pemodelan, Pengujian Data dilakukan menggunakan Nilai K=3, K=5, K=7, K=9, K=11. Gambar 10 menjelaskan pengaturan pengujian terhadap model latih menggunakan 3 data uji (K=3). Gambar 11 menampilkan hasil prediksi sentiment dan pelabelan manual (actual) nilai jarak ketetanggaan terdekat menggunakan K=3 dan data latih 82 dengan 41 tweet berlabel Non *Hatespeech* dan 41 berlabel *Hatespeech*.

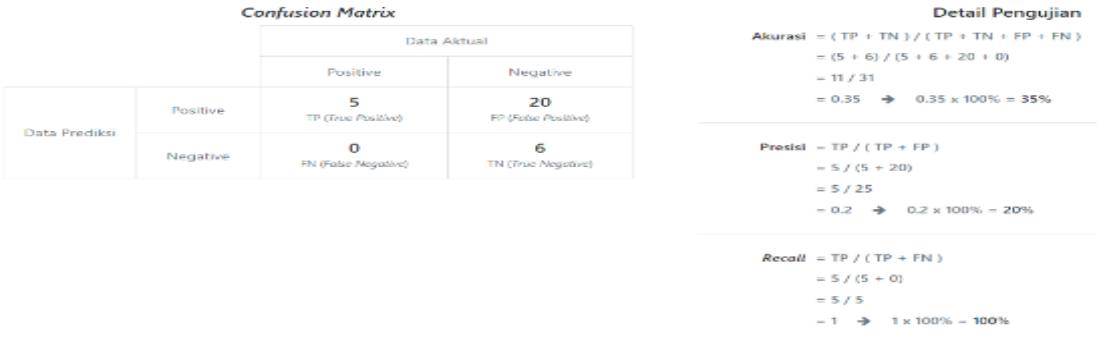
INDONESIA JOURNAL INFORMATION SYSTEM **(IDEALIS)**
Volume 5, Nomor 2, Juli 2022
ISSN 2684-7280 (*online*)
Halaman 87-97
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



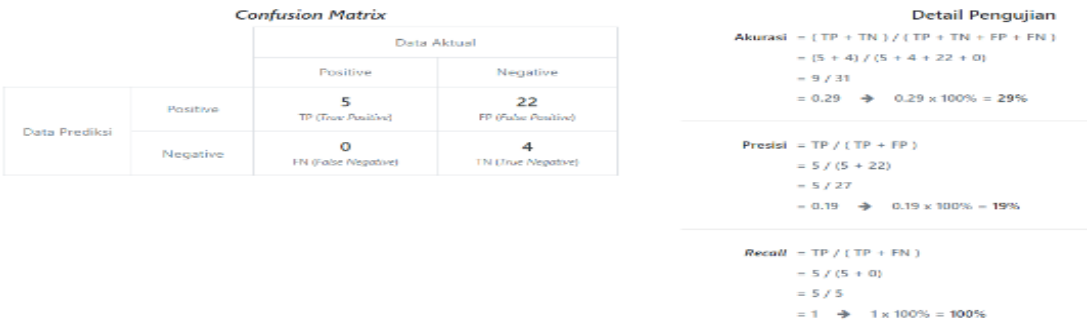
Gambar 10. Hasil Pengujian



Gambar 11. Data Uji



Gambar 12. Nilai K=3



Gambar 13. Nilai K=5

Gambar 12 dan 13 menjelaskan hasil pengujian data testing dalam memprediksi tweet dengan kategori Hatespeech (HS) atau Non Hatespeech (Non HS). Nilai pengujian dengan K=5 berturut-turut untuk nilai akurasi,

presisi dan recall adalah 29%, 19% dan 100%. Disisi lain, nilai pengujian dan evaluasi tertinggi yang diperoleh sebesar: akurasi 35%, presisi 20% dan *recall* 100% berturut-turut untuk nilai akurasi, presisi dan recall menggunakan nilai K=3.

4. KESIMPULAN

Berdasarkan hasil pengujian dan evaluasi dari aplikasi yang dibuat menggunakan *Dataset* dan Algoritme yang diusulkan, maka dapat disimpulkan bahwa:

a. Berdasarkan 312 *Tweet*, arah pandangan (sentimen) masyarakat Indonesia terhadap vaksinasi cenderung ke arah sentimen *Hatespeech* sebesar 76.56% pada periode Juli 2021.
b. Tahap utama yang terdapat dalam penelitian ini antara lain: *Crawling*, *Preprocessing*, Pelabelan, Pemodelan, Pembagian Data dan klasifikasi *K- Nearest Neighbor* (KNN). Tahap *Preprocessing* yang baik menjadi penentu dalam terbentuknya hasil yang optimal untuk tahap selanjutnya. Penggunaan kamus sentimen dapat membantu proses pemberian kelas atau *label* juga meminimalisir waktu dan usaha dalam melakukan proses Pelabelan.
c. Penggunaan ekstraksi fitur *CountVectorizer* dan Algoritme *K-Nearest Neighbor* (KNN) dalam melakukan analisis sentimen dapat berjalan dengan baik, dengan nilai pengujian dan evaluasi tertinggi yang diperoleh sebesar: akurasi 35%, presisi 20% dan *recall* 100% menggunakan nilai K=3.

DAFTAR PUSTAKA

[1] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Journal of Intelligent Systems and Computation 43," pp. 43–49, 2019, [Online]. Available: https://t.co/9WloaWpfD5.
[2] N. Tri Romadloni, I. Santoso, and S. Budilaksono, "Perbandingan Metode Naive Bayes, KNN, dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line," *J. IKRA-ITH Inform.*, vol. 3, no. 2, pp. 1–9, 2019.
[3] N. O. F. Daeli and Adiwijaya, "Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor," *J. Data Sci. Its Appl.*, vol. 3, no. 1, pp. 1–7, 2020, doi: 10.34818/JDSA.2020.3.22.
[4] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017, vol. 2018-January, no. October, pp. 233–237, 2018, doi: 10.1109/ICACSIS.2017.8355039.
[5] U. P. Harapan, ""Hate Speech, Kenapa di ributkan? Ujaran Kebencian (Hate Speech) di Indonesia"." .
[6] A. Masyhur Effendi, Dimensi/Dinamika Hak Asasi Manusia dalam Hukum Nasional dan Internasional. Ghalia Indonesia, 1994.
[7] G. Gunadi and D. I. Sensuse, "Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori Dan Frequent Pattern Growth (Fp-Growth) :," Telematika, vol. 4, no. 1, pp. 118–132, 2012.
[8] E. Fitri, N. Zola, and I. Ifdil, "Profil Kepercayaan Diri Remaja serta Faktor-Faktor yang Mempengaruhi," JPPI (Jurnal Penelit. Pendidik. Indones., vol. 4, no. 1, pp. 1–5, 2018, doi: 10.29210/02017182.
[9] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," J. EECCIS, vol. 6, no. 1, pp. 37–43, 2012.
[10] L. M. Aiello et al., "Sensing trending topics in twitter," IEEE Trans. Multimed., vol. 15, no. 6, pp. 1268–1282, 2013, doi: 10.1109/TMM.2013.2265080.
[11] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," J. SIMETRIS, vol. 10, no. 2, pp. 681–686, 2019.
[12] A. V. Sudiantoro et al., "Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier," Din. Inform., vol. 10, no. 2, pp. 398–401, 2018.
[13] T. K. Ronal Watrianthos, Muhammd Noor Hasan Siregar, Dewa Putu Yudhi Ardiana, Dyah Gandasari, Ramen A Purba, Yusra Fadhillah Nur Azizah Affandy, Janner Simarmata, Diena Dwidienawati Tjiptadi, Yo Ceng Giap, Oris Krianto Sulaiman, Noverita Sprinse Vinolina, Deddy, Belajar dari Covid-19 Perspektif Teknologi & Pertanian, vol. ق ث ڤ ف ق, no. ڤ ف ڤ ڤ ڤ ڤ. Yayasan Kita Menulis, 2020.
[14] E. B. Santoso and A. Nugroho, "Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook," Eksplora Inform., vol. 9, no. 1, pp. 60–69, 2019, doi: 10.30864/eksplora.v9i1.254.
[15] S. N. J. Fitriyyah, N. Safriadi, and E. E. Pratama, "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," J. Edukasi dan Penelit. Inform., vol. 5, no. 3, p. 279, 2019, doi: 10.26418/jp.v5i3.34368.
[16] P. Antinasari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 12, pp. 1718–1724, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.
[17] Munawar, "Sistem Pendeteksi Berita Palsu (Fake News) di Media Sosial dengan Teknik Data Mining Scikit Learn," 2019. [Online]. Available: https://digilib.esaunggul.ac.id/UEU-Research-16_0402/13322