(2025), 3 (3): 229–248 **Scientica** 3021-8209

Jurnal Ilmiah Sain dan Teknologi

PERBANDINGAN KLASIFIKASI DENGAN ALGORITMA K-NEAREST NEIGHBOR (KNN) DAN DECISION TREE PADA STRUKTUR KALIMAT BAHASA JAKSEL (CAMPUR KODE BAHASA INDONESIA-INGGRIS)

Alda Shafira [1], Arif Bijaksana Putra Negara [2], Hafiz Muhardi [3]
Fakultas Teknik Universitas Tanjungpura Pontianak
Aldashafira261299@gmail.com

Abstrak
Komunikasi menjadi media yang digunakan untuk bertukar informasi antara individu dan kelompok. Dalam berkomunikasi masyarakat menggunakan bahasa untuk dapat memudahkan pemahaman antara pembicara dan pendengar. Terdapat tren bahasa yang digunakan oleh masyarakat saat ini yaitu Bahasa Jaksel atau bahasa campuran yang biasa disebut sebagai campur kode. Bahasa Jaksel merupakan bahasa modifikasi yang digunakan masyarakat untuk saling berkomunikasi secara langsung maupun pada media sosial khususnya Twitter. Dalam realitasnya, struktur Bahasa Jaksel masih sulit untuk dipahami karena istilah dan kosa kata yang sangat beragam. Penelitian ini bertujuan untuk melakukan klasifikasi teks terhadap data teks yang mengandung Bahasa Jaksel yang telah dikumpulkan dari media sosial Twitter dan melalui proses *text preprocessing* menggunakan Algoritma *Decision Tree* dan *K-Nearest Neighbor*. Dari hasil penelitian ini didapatkan bahwa penggunaan Algoritma *Decision Tree* menghasilkan nilai akurasi tertinggi pada model klasifikasi menggunakan N-Gram = 2 dengan nilai sebesar 79,02%. Adapun model klasifikasi Algoritma *K-Nearest Neighbor* menghasilkan nilai sebesar 75,29% dengan menggunakan N-Gram = 3.

Sejarah Artikel
Submitted: 3 Januari 2025
Accepted: 8 Januari 2025
Published: 9 Januari 2025

Kata Kunci
Decision Tree, *K-Nearest Neighbor*, Bahasa Jaksel, Campur Kode, Twitter.

PENDAHULUAN

Komunikasi merupakan satu hal penting yang tidak dapat lepas dari kehidupan manusia. Manusia selalu berkeinginan untuk berbicara, bertukar pendapat, mengirim dan menerima informasi agar memenuhi kebutuhan pribadi. Berkomunikasi untuk mengetahui individualitas satu sama lain merupakan prasyarat untuk membangun hubungan sosial antar manusia. Setiap individu sering membuat kesalahan dengan menyangka bahwa komunikasi adalah kemampuan alami yang dimilikinya sejak lahir sehingga merasa tidak perlu untuk mempelajarinya. Komunikasi dapat terjadi secara pribadi atau dua orang, bisa juga dengan beberapa orang atau dengan kelompok orang banyak dan massa (Hardjana, 2003).

Menurut (Cangara, 2007) di dalam (Hadiwijaya, 2018) komunikasi juga dapat diartikan sebagai sebuah transaksi simbolik yang mengharuskan orang mengatur lingkungannya dengan membangun hubungan antar sesama manusia melalui pertukaran informasi dalam upaya menguatkan sikap dan tingkah laku orang lain serta merubah sikap perilaku tersebut. Terdapat dua perbedaan dalam komunikasi, yaitu komunikasi formal dan informal.

Menurut (Haryadi, 2009) di dalam (Cristanto, 2018) komunikasi secara umum hanya ada 2 (dua), yaitu yang bersifat formal dan informal. Komunikasi formal adalah komunikasi yang wajib mengikuti prosedur atau standar yang baku dan disesuaikan dengan tempat serta waktunya. Komunikasi informal adalah komunikasi yang terjadi tanpa direncanakan dan berlangsung dalam suasana santai. Komunikasi juga menggunakan berbagai bahasa agar mudah dipahami berdasarkan lingkungan tempat tinggal dan pergaulan. Saat ini di era milenial dan Gen Z, terdapat fenomena mencampurkan bahasa yaitu Bahasa Indonesia dan Bahasa Inggris atau sering disebut dengan Bahasa Jaksel. Masyarakat kaum muda sering berkomunikasi menggunakan bahasa tersebut. Adapun perkembangan Bahasa Jaksel muncul

di lingkungan masyarakat yang tinggal di kawasan Jakarta Selatan. Ciri khas Bahasa Jaksel sendiri yaitu selain mencampurkan bahasa inggris ke percakapan, kerap juga menggunakan *slang words* atau istilah gaul yang hanya dipahami oleh masyarakat yang paham menggunakan Bahasa Jaksel. Adapun contoh penggunaannya seperti "*Normally* kebanyakan orang sekarang *fomo* banget". Namun penggunaan bahasa seperti ini membuat sebagian besar masyarakat masih kesulitan untuk memahami dan berkomunikasi menggunakan Bahasa Jaksel karena hanya dapat dipahami oleh sebagian kalangan.

Berbagai cara masyarakat berkomunikasi menggunakan Bahasa Jaksel dalam kesehariannya maupun dalam bertukar informasi dan berkomunikasi melalui *platform digital* atau *social media*, salah satunya Twitter. Twitter adalah salah satu media sosial yang sering digunakan oleh penggunanya untuk mencari informasi, bersosialisasi dan saling bertukar pesan melalui *tweets*. Teruntuk generasi milenial atau Gen Z sering menggunakan Twitter sebagai media untuk berkomunikasi maupun berkeluh kesah menggunakan tata cara bahasa, salah satunya yaitu Bahasa Jaksel.

Penggunaan Bahasa Jaksel yang marak di Twitter menghasilkan banyak pula *tweets* atau cuitan masyarakat yang mengandung Bahasa Jaksel. Berbagai informasi dapat didapatkan namun tak menutup kemungkinan juga banyak masyarakat yang masih kesulitan untuk memahami kosa-kata Bahasa Jaksel tersebut. Hal tersebut dikarenakan data text dari Twitter tersebut masih belum memiliki struktur bahasa yang belum cukup baik dan perlu untuk dilakukan analisis terlebih dahulu untuk menghasilkan informasi dari data tersebut. Beberapa cara dapat dilakukan untuk mengolah data text tersebut salah satunya menggunakan *Text Mining*. *Text Mining* merupakan sebuah teknik metode pendekatan algoritmik berbasis komputer untuk mendapatkan suatu pengetahuan baru yang tersembunyi dari sekumpulan teks (Priyanto & Ma'arif, 2018).

Dalam melakukan *Text Mining* terdapat pula berbagai metode yang dapat digunakan dalam mengklasifikasikan *data text* tersebut. Algoritma *K-Nearest Neighbor* merupakan salah satu dari beberapa algoritma yang dapat digunakan untuk mengklasifikasikan teks. Algoritma *K-Nearest Neighbor* sendiri merupakan algoritma dalam klasifikasi yang melakukan klasifikasi terhadap objek berdasarkan data latih (*training*) menggunakan jarak terdekat atau kemiripan terhadap objek tersebut (Wahyono, et al., 2020). Berdasarkan penelitian dengan judul "Klasifikasi Teks Dengan Menggunakan Algoritma *K-Nearest Neighbor* Pada Kasus Kinerja Pemerintah di Twitter" menghasilkan akurasi pada pengujian data set menggunakan Algoritma KNN menghasilkan nilai akurasi sebesar 90,00% (Prakasa & Lhaksamana, 2018). Penelitian terkait klasifikasi teks menggunakan Algoritma *K-Nearest Neighbor* juga dilakukan pada penelitian berjudul "Perbandingan Penghitungan Jarak pada *K-Nearest Neighbour* dalam Klasifikasi Data Tekstual", penelitian tersebut menghasilkan nilai akurasi yang diperoleh oleh algoritma *K-Nearest Neighbor* sebesar 85,50% menggunakan nilai K = 3 (Wahyono, et al., 2020). Selain itu, terdapat pula algoritma yang dapat digunakan dalam mengklasifikasikan teks yaitu Algoritma *Decision Tree*. Algoritma *Decision Tree* merupakan algoritma klasifikasi dengan membangun sebuah model prediksi terhadap suatu keputusan menggunakan struktur hirarki atau pohon keputusan (Sartika & Sensuse, 2017). Algoritma *Decision Tree* digunakan dalam penelitian yang berjudul "Klasifikasi Opini Terhadap Resesi Indonesia 2023 Pada Twitter Menggunakan Algoritma Decision Tree", penelitian ini melakukan pengujian terhadap performa Algoritma *Decision Tree* dengan menghasilkan nilai akurasi sebesar 89.86%, *recall* 82.64%, *precision* 84.22%, *f1-score* 83.32% ini membuat Algoritma *Decision Tree* dapat bekerja dengan sangat baik pada klasifikasi teks (Trianto, et al., 2023). Penelitian terhadap klasifikasi teks menggunakan Algoritma *Decision Tree* juga dilakukan pada penelitian yang berjudul "Perbandingan Akurasi Metode Deteksi Ujaran Kebencian dalam Postingan Twitter Menggunakan Metode SVM dan *Decision Trees* yang Dioptimalkan dengan *Adaboost*", dari penelitian yang dilakukan menghasilkan hasil nilai performa

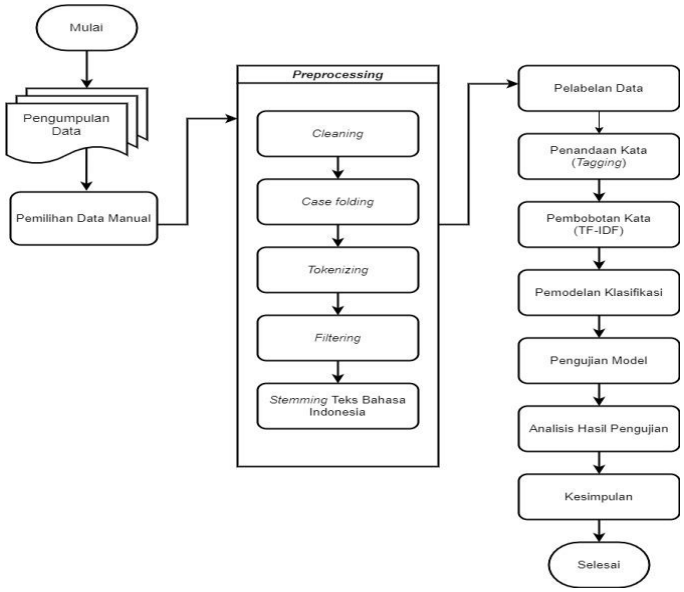
penggunaan Algoritma SVM sebesar 90,03% dan Algoritma *Decision Tree* sebesar 89,53% (Septiawan & Chairani, 2023). Berdasarkan hal ini, Algoritma *K-Nearest Neighbor* dan *Decision Tree* akan dipilih sebagai algoritma klasifikasi teks dalam penelitian ini.

Berdasarkan uraian di atas, pada penelitian ini akan dilakukan pengklasifikasian teks dari Bahasa Jaksel melalui tingkat keacakan struktur kalimat dari data teks yang digunakan. Data teks yang digunakan merupakan data yang dihasilkan melalui media social Twitter. Dalam penelitian ini akan mengimplementasikan penggunaan algoritma *K-Nearest Neighbor* dan *Decision Tree*, sehingga dari hasil tersebut dapat membandingkan nilai performa masing-masing algoritma untuk mengklasifikasikan data yang tersedia. Pada pengklasifikasian dari struktur kalimat, penelitian ini akan mengadopsi tiga aturan model dari penelitian yang berjudul "*Bilingual Speech: A Typology Of Code-Mixing*" yaitu, *insertion* adalah melakukan penyisipan kata dari suatu bahasa ke kalimat utama dalam bahasa lain, *alternation* adalah aturan yang melakukan penggantian dalam pencampuran bahasa dengan kondisi dua bahasa hadir dalam sebuah klausa, *congurent lexicalization* hasil menggabungkan dua unsur frasa dalam bahasa yang berbeda namun dengan konsep yang sama secara paralel dalam sebuah kalimat (Muysken, 2000). Berdasarkan hal di atas maka akan dihasilkan dari penelitian ini informasi terhadap klasifikasi teks Bahasa Jaksel dengan keacakan struktur kalimat dan menghasilkan nilai performa dari Algoritma *K-Nearest Neighbor* dan Algoritma *Decision Tree*.

METODOLOGI PENELITIAN

Metode Penelitian

Berikut ini adalah metode penelitian dilakukan dalam penelitian yang diilustrasikan ke dalam bentuk *flowchart* pada **Gambar 3.1**.



Gambar 3. 1 Flowchart Penelitian

Alat dan Data Penelitian

Alat penelitian digunakan untuk mendukung pengerjaan pada pelitian dan data-data yang valid digunakan dalam menyusun laporan penelitian. Berikut ini merupakan alat dan data yang penulis gunakan dalam penelitian.

a. **Perangkat Keras**
   Perangkat keras utama yang digunakan dalam penelitian ini adalah ASUS ROG CPU AMD Ryzen™ 9 5900HX Mobile Processor dan GPU AMD Radeon™ RX 6800M.
b. **Perangkat Lunak**
   Perangkat lunak yang digunakan pada penelitian ini adalah Sistem Operasi Microsoft Windows 10, Aplikasi RapidMiner, Google Colaboratory.
c. **Data Penelitian**
   Data yang digunakan adalah hasil *scraping* yang telah dikumpulkan melalui media sosial Twitter. Data tersebut dikumpulkan berdasarkan 158 kata kunci berbahasa Inggris yang sering digunakan pada bahasa Jaksel dan pemilihan bahasa Indonesia pada *tweet* secara penuh. Data penelitian yang digunakan merupakan data *tweets* yang mengandung Bahasa Jaksel. Adapun karakteristik data *tweets* yang diambil yaitu merupakan *tweet* masyarakat yang mengandung campuran Bahasa Inggris dan Bahasa Indonesia. Dari data mentah tersebut, data masuk ke tahapan proses pengolahan data sehingga menghasilkan data bersih yang digunakan sebagai data penelitian.

Pengumpulan Data

Pada penelitian ini tahapan pengumpulan data dilakukan dengan berbagai tahapan. Pertama, melakukan pengumpulan data tweets yang dikumpulkan berdasarkan fenomena campur kode di tahun 2021. Fenomena campur kode ini adalah penggunaan Bahasa Jaksel yang berkembang dimasyarakat. Bahasa Jaksel sendiri merupakan penggabungan Bahasa Indonesia tidak baku dan Bahasa Inggris baku. Adapun Bahasa Jaksel juga terpengaruh oleh bahasa daerah yang digunakan pada wilayah DKI Jakarta yaitu Bahasa Betawi. Kemudian pengumpulan data dilanjutkan dengan melakukan *scraping* data *tweets* menggunakan API Twitter melalui aplikasi *Rapid Miner* dengan menginputkan parameter teksnya berupa Bahasa Inggris dan bahasa tweetsnya adalah Bahasa Indonesia serta *limit* 300 data tweets setiap penarikan data. Dari *scraping* data tersebut menghasilkan sebanyak 28.564 data tweets yang mengandung campur kode Bahasa Jaksel yang digunakan untuk penentuan kata kunci sebanyak 158 kata yang ditampilkan pada **Tabel 4.1**. dalam mengidentifikasi karakteristik data yang mengandung campur kode dari Bahasa Jaksel. Dari proses yang telah dilakukan maka akan menghasilkan data mentah yang kemudian akan proses ke tahapan selanjutnya, yaitu proses pemilihan secara manual berdasarkan pengetahuan konseptual dari penulis agar terhindar dari redudansi data maupun data yang bersifat spam.

Pemilihan Data

Tahapan pemilihan data dalam penelitian ini dilakukan secara manual. Adapun data yang telah dihasilkan melalui proses *scraping* diseleksi secara manual oleh peneliti agar dapat mengurangi redudansi data dan data *tweet* spam. Berdasarkan dari tahapan proses tersebut, maka dihasilkan sebanyak 5099 data *tweets* bersih yang digunakan pada penelitian ini.

Text Preprocessing

Berdasarkan pengumpulan data yang telah dilakukan, data yang dihasilkan merupakan data mentah yang belum dapat digunakan untuk pemodelan klasifikasi karena data yang dihasilkan masih tidak terstruktur. Sehingga diperlukan sebuah proses untuk menjadikan data tersebut menjadi data yang terlah terstruktur agar dapat digunakan ke tahapan pemodelan. Tahapan proses yang dilakukan yaitu *text preprocessing*. Dalam penelitian ini tahapan *text*

preprocessing meliputi *cleaning, case folding, tokenizing,* dan *filtering*, adapun proses masing-masing tahapan *text preprocessing* sebagai berikut.

a. *Cleaning*: Proses menghilangkan tanda baca, karakter, atau symbol yang tidak diperlukan.

Sebelum *Cleaning*: [Mine too!!! Actually aku juga lagi ada program gini terus belum dapet info juga huhu:(hoping we get good news dan hal hal baik lain semangat kak!!]

Setelah *Cleaning*: [Mine too Actually aku juga lagi ada program gini terus belum dapet info juga huhu hoping we get good news dan hal hal baik lain semangat kak]

b. *Case Folding:* Proses mengubah seluruh data menjadi *lowercase*.

Sebelum *Case Folding*: [Mine too Actually aku juga lagi ada program gini terus belum dapet info juga huhu hoping we get good news dan hal hal baik lain semangat kak]

Setelah *Case Folding*: [mine too actually aku juga lagi ada program gini terus belum dapet info juga huhu hoping we get good news dan hal hal baik lain semangat kak]

c. Tokenizing: Proses memecah data kalimat menjadi token-token kata.

Sebelum *Tokenizing*: [mine too actually aku juga lagi ada program gini terus belum dapet info juga huhu hoping we get good news dan hal hal baik lain semangat kak]

Setelah *Tokenizing*: [‘mine’, ‘too’, ‘actually’, ‘aku’, ‘juga’, ‘lagi’, ‘ada’, ‘program’, ‘gini’, ‘terus’, ‘belum’, ‘dapet’, ‘info’, ‘juga’, ‘huhu’. ‘hoping’, ’we’. ‘get,’ ‘good’, ‘news’, ‘dan’, ‘hal’, ‘hal’, ‘baik’, lain’, ‘semangat’, ‘kak’]

d. *Filtering*: Proses menghilangkan kata-kata tidak penting atau tidak mengandung makna atau *stopwords removal*.

Sebelum *Filtering*: [‘mine’, ‘too’, ‘actually’, ‘aku’, ‘juga’, ‘lagi’, ‘ada’, ‘program’, ‘gini’, ‘terus’, ‘belum’, ‘dapet’, ‘info’, ‘juga’, ‘huhu’. ‘hoping’, ’we’. ‘get,’ ‘good’, ‘news’, ‘dan’, ‘hal’, ‘hal’, ‘baik’, lain’, ‘semangat’, ‘kak’]

Setelah *Filtering*: [‘mine’, ‘too’, ‘actually’, ‘program’, ‘info’, ‘hoping’, ’we’, ‘get,’ ‘good’, ‘news’, ‘baik’, ‘semangat’]

e. *Stemming*: Proses mengubah kata menjadi akar kata atau kata dasar.

Sebelum *Stemming*: [‘mine’, ‘too’, ‘actually’, ‘program’, ‘menginfokan’, ‘hoping’, ’we’. ‘get,’ ‘good’, ‘news’, ‘baik’, ‘semangatnya’]

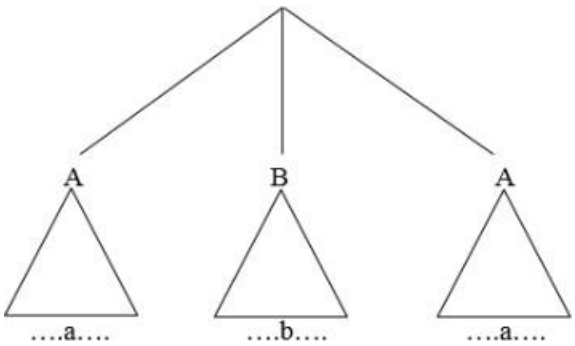
Setelah *Stemming*: [‘mine’, ‘too’, ‘actually’, ‘program’, ‘info’, ‘hoping’, ’we’. ‘get,’ ‘good’, ‘news’, ‘baik’, ‘semangat’]

Pelabelan Data

Pelabelan yang dilakukan pada penelitian ini adalah kategori pada campur kode yang diantaranya adalah (Muysken, 2000).

a. *Insertion*

Pada **Gambar 3.2** unsur ‘a’ mewakili frasa pada bahasa utama. Sedangkan unsur ‘b’

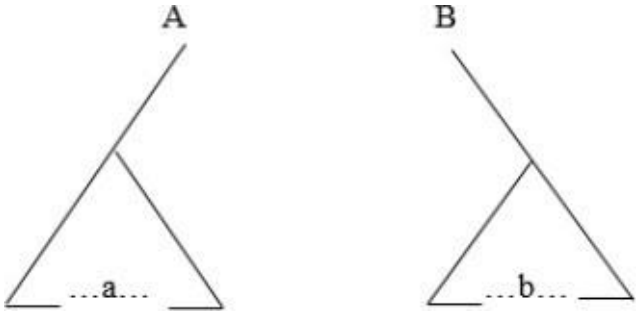


Gambar 3. 2 Ilustrasi Pada Insertion Dalam Campur Kode

mewakili frasa pada bahasa kedua yang telah disisipkan oleh penutur. Contoh pada *insertion*

yaitu “Mereka setiap akhir pekan selalu *self healing* bersama tanpa sepengetahuan saya” yang di mana frasa “*self healing*” merupakan unsur yang disisipkan oleh penutur.

b. Alternation

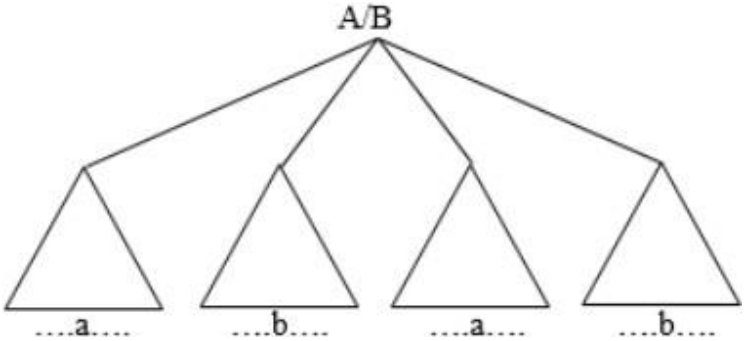


Gambar 3. 3 Ilustrasi Pada Alternation Dalam Campur Kode

Pada **Gambar 3.3** unsur ‘A’ dan ‘B’ merupakan representasi dari kedua bahasa yang berbeda yang di mana unsur tersebut mewakili pergantian bahasa pada campur kode dalam bentuk ujaran yang dihasilkan oleh penutur. Contoh pada *alternation* yaitu “*I think I can*, soalnya setiap aku menyanyikan suatu lagu, penonton terhibur”.

c. Congruent Lexicalization

Pada **Gambar 3.4** unsur ‘A’ dan ‘B’ merupakan representasi dari kedua bahasa yang



Gambar 3. 4 Ilustrasi Pada Congruent Dalam Campur Kode

berbeda. Pada kategori ini, penutur lebih cenderung untuk menggabungkan dua bahasa ketika dilihat dari struktur gramatikalnya sehingga penggunaan kategori ini menggunakan pengisian frasa secara leksikal dengan item leksikal dari kedua bahasa. Contoh pada *congruent lexicalization* yaitu “*Meeting* hari ini akan membahas tentang *urgent agenda* yang akan dilakukan *this week*”.

Berikut ini adalah hasil pelabelan kelas yang dilakukan secara manual ditampilkan pada

Tabel 3. 1 Hasil Pelabelan Kelas

Data Teks	Label Kelas
dari semalem sampe tadi pagi bangun tidur anxious sama kerja nyata pas udah dikerjain oh udah ya haha yaallah kenapa ya gue	INSERTION
please be better lupakan semua yang sudah berlalu	ALTERNATION
deep down aku nyata gampang anxious kalo ketemu tatap muka yang related to academics things padahal kalo onlen kayak gapunya takut	CONGRUENT

Penentuan Penandaan Kata

Pada penelitian ini, terdapat proses penentuan penandaan kata yang diterapkan pada setiap kata dalam kalimat Bahasa Jaksel atau campur kode. Penandaan kata dilakukan berdasarkan posisi dan jenis kata dalam kalimat campur kode. Proses penandaan kata ini bertujuan agar dapat menunjukkan tipe bahasa dan letak atau posisi kata tersebut dalam sebuah kalimat campur kode. Adapun dalam realitasnya, bahasa campur kode atau Bahasa Jaksel merupakan fenomena bahasa yang terjadi di masyarakat yang menggunakan Bahasa Indonesia tidak baku yang dicampur menggunakan Bahasa Inggris yang merupakan Bahasa Inggris baku. Maka dari itu, proses penandaan kata menjadi penting agar dapat memproses data lebih mudah dan menentukan letak posisi dan jenis kata pada kalimat dari data bahasa campur kode yang digunakan.

HASIL DAN ANALISIS PENELITIAN

Hasil Pengumpulan Data

Dalam melakukan pengumpulan data, peneliti melakukan observasi pada media sosial Twitter terhadap kata atau frasa yang digunakan dalam tweets di komunitas Twitter yang mengandung campuran dari Bahasa Indonesia tidak baku dan Bahasa Inggris Baku. Kemudian, kata atau frasa tersebut disusun sebagai *corpus* yang digunakan sebagai parameter atau kata kunci pada pengumpulan data penelitian. Berikut parameter atau kata kunci yang ditunjukkan pada

Pemilihan parameter pada penelitian ini yaitu dengan mengobservasi penggunaan Bahasa Jaksel pada *tweets* masyarakat Twitter yang mengandung campuran bahasa Indonesia dan bahasa Inggris. Setelah melakukan proses observasi maka didapatkan kata kunci sebanyak 158 kata yang ditampilkan pada

Tabel 4. 1 Parameter atau Kata Kunci *Tweets*

No.	Parameter	No.	Parameter
1	actually	23	clingy
2	af	24	cmiiw
3	after	25	confirm
4	anhedonia	26	crush
5	anxious	27	couple goals
6	attitude	28	corporate
7	assess	29	confused
8	as fuck	30	confuse
9	around	31	deep talk
10	anyway	32	degrading
11	basically	33	denial
12	because why	34	detox sosmed
13	be like	35	discuss
14	bestie	36	fancy
15	better	37	eye catching
16	brain storming	38	ever
17	body shaming	39	end up
18	body goals	40	emotional abuse
19	bit	41	fashionable
20	bipolar	42	fear of missing out
21	bro mens	43	feminis
22	burnout	44	figuratively
No.	Parameter	No.	Parameter

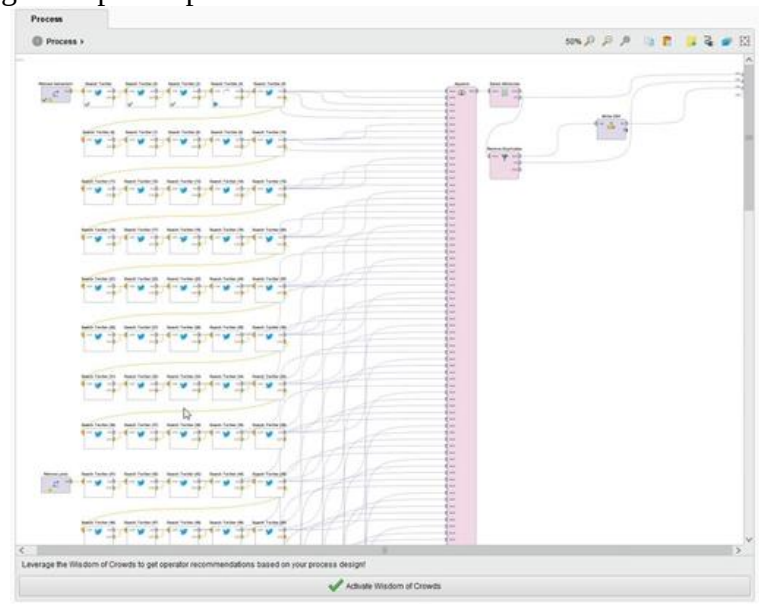
No.	Parameter
45	financial freedom
46	for your information
47	fomo
48	follow up
49	flirtatious behavior
50	flexing
51	for your page
52	fu
53	fuck
54	fyi
55	fyp
56	glow up
57	ghosting
58	get back
59	gate keeping
60	gaslighting
61	good looking
62	guilt tripping
63	healing
64	healthy relationship
65	hectic
66	idc
67	i dont know
68	i dont care
69	honestly
70	hence
71	idk
72	imo
73	in my opinion
74	income
75	inner child
76	let's say
77	judgemental
78	invasion of privacy
79	internal only
80	insecure
81	like
82	literally
83	little
84	living together
85	lowkey
86	money oriented
No.	Parameter
129	sleep call
130	silent treatment
131	spill
132	spill bill
133	start up
134	staycation
135	strict parent
136	surely

No.	Parameter
87	mindfullness
88	mental health
89	me time
90	make sure
91	morning person
92	mostly
93	negative vibes
94	normally
95	not yet
96	otw
97	open minded
98	once
99	on the way
100	noted
101	oversharing
102	overthinking
103	overwhelm
104	overwork
105	panic attack
106	possibility
107	positive vibes
108	personal space
109	perhaps later
110	pap
111	post a picture
112	prefer
113	price
114	probably
115	quarter life crysis
116	i don't know
117	sandwich generation
118	salty
119	roaming
120	revenue
121	second family
122	seldom
123	selflove
124	selfreward
125	sexist
126	somehow
127	socially awkward
128	social butterfly
No.	Parameter
144	too much information
145	toxic masculinity
146	verbally abuse
147	unlike
148	trust issue
149	toxic relationship
150	toxic positivity
151	wbk

No.	Parameter
137	supposed
138	support system
139	sugar daddy
140	sugar coating
141	that is
142	that's why
143	the point is

No.	Parameter
152	we been knew
153	well
154	whatever
155	whereas
156	i don't care
157	you know
158	work life balance

Kemudian tahapan pengumpulan data dilanjutkan dengan menarik data *tweets* menggunakan teknik *scraping* dari aplikasi Twitter menggunakan Twitter API melalui *Rapid Miner*. *Scraping* data dilakukan menggunakan spesifik *keywords* yang telah disusun sebelumnya sebagai *corpus*. Adapun data yang diambil merupakan *tweets* kombinasi bahasa Indonesia dan bahasa Inggris dari tahun 2021. Berikut ini hasil *scraping* data menggunakan *Rapid Miner* yang ditampilkan pada **Gambar 4.3**.



Gambar 4. 1 Proses Pengumpulan Data Menggunakan Rapid Miner

Dari data yang telah diperoleh melalui tahapan *scraping tweets* menggunakan Rapid Miner maka didapatkan sebanyak 28.864 data tweets. Berikut ini adalah hasil data yang diperoleh yang ditampilkan pada **Tabel 4.2**.

Tabel 4. 2 Data Teks Hasil Rapid Miner

Data Teks Raw
after all, endingnya gue suka kok! semua pertanyaan yang ada di benak pembaca terjawab, jadi bukan buku yang bikin kecewa, tapi justru harus dibaca minimal COBA DAHHHH SEKALI AJA PASTI SUKA!
d kira yg lu buat ini konten positif apa? how about actually make some positive content? Kek maybe make some memes? So we can actually laugh ?
ada 2 tipe org di dunia ini. yg satu bilang gak sopan kaya gitu gapunya attitude, yg satu lg 'YA ALLAH LUCU BGT WKWKWK WKAKAKAK HAHAHAHA' ALIAS BEDA BGT REAKSI ORG' PDHL INIBLUCUUU https://t.co/KAZL0XNTn0
Gimana kalo sebenernya si Dr Strange ini memang menganggap kalo diri nya itu punya ilmu tenaga dalam? Alias memang sakit jiwa/halu?? Which is the most fucked up scenario we had
RT @justagudgurl: Bayangkan cantik-cantik, tapi tindas sesama kaum. Which is kita tau as perempuan, kita lemah, soft and etc..
besok uda agustus which is bulan dimana papi pergi hehe, kangen banget?

Hasil Pemilihan Data

Tahapan pemilihan data merupakan proses dimana peneliti melakukan pemilihan data secara manual, dimana proses ini peneliti menggunakan *spreadsheet* untuk dapat memilah data yang digunakan ke tahapan berikutnya. Pemilihan data ini bertujuan agar tidak terdapat data yang mengandung spam maupun duplikasi data. Dari hasil pemilihan data ini dihasilkan sebanyak 5099 data tweet yang digunakan dalam tahapan berikutnya. Berikut ini tahapan pemilihan data yang ditampilkan pada **Gambar 4.4**.

Data tweet
selalu siapkan satu stack brilliance kalo mau roaming
fokus dulu aja farming kasihan dia harus roaming ke lane lain
eh gue inget pernah baca ini di akun belah kayak haha lu ahli manip gaslighting lu mah verifikasi
benci banget ini yang di indosiar cowonya gaslighting
terus akun pra nikah itu gak tau cerita di balik sih laki laki kayak gitu ada malah nge judge kalo laki laki gak serius padahal udah juang demikian rupa gajelas banget a
lu bingung gak sakit bas lingkung ketemu sama dosen yang gaslighting lagi itu
abis gaslighting orang terus enggak tanggungjawab
ada butuh lain malah dimarahin emak gue sampe suruh pinjemin duit untuk adeknya bahkan gue kena gaslighting sama emak eh pas gue pinjemin malah adek mam
hadeh stop gaslighting kayak gini lah beneran deh jalan hidup manusia tuh macem macem banget ada yang udah nyiapin dana kuliah eh tiba tiba bapak sakit kena p
latian gaslighting buat nanti kalo jawab tanya pas sidang
gak ada nama bawa suasana nder emang cowo lu aja yang gatel kalo dia ngerhargain lu hubung lu dia gak akan mau kenal terus pergi orang lain apalagi sampe kissin
mungkin gak cuman pembully sih yang suka gaslighting aja rada susah mau maafinnya
atau masalah sama kayak di dalam suatu hubung asmara mana salah satu minor jadi gampang di gaslighting gitu juga kah

Gambar 4. 2 Hasil Pemilihan Data

Hasil Pelabelan Data

Berdasarkan hasil dari pemilihan data maka dihasilkan sebanyak 5099 data tweets yang digunakan untuk ketahapan berikutnya yaitu pelabelan data. Dalam proses pelabelan data, penulis melakukan pelabelan secara manual. Pelabelan dalam penelitian ini adalah memberikan kelas atau label pada setiap data. Adapun terdapat 3 (tiga) kelas atau label data yaitu, *Insertion*, *Congruent*, *Alternation*. Berikut ini hasil pelabelan data yang ditampilkan pada **Gambar 4.5**.

onboarding material yang katanya kurang	DENG TENG DIDN TIDN TIDN TIDN	ALTERNATION
sampai sekarang stop	DIDN TIDN DENG	ALTERNATION
buat bahan gaslighting that are better thar	DIDN TIDN DENG TENG TENG TENG TENG TENG	ALTERNATION
terus abis itu gaslighting pas gua mau tingg	DIDN TIDN TIDN DENG DIDN TIDN TIDN TIDN TIDN	CONGRUENT
gimana cara henti self gaslighting ya tapi al	DIDN TIDN TIDN DENG TENG DIDN TIDN TIDN TIDN	CONGRUENT
salah satu tipe manusia yang bahaya dan h	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	ALTERNATION
tipe tipe cowok kayak gini bentar lagi pasti	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	CONGRUENT
dah lama males nonton tiba tiba tarik nont	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	CONGRUENT
baru kali ini deh ngerasain punya temen b	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	INSERTION
asli kaget banget sama dialog reza rahardia	DIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG DIDN	CONGRUENT
adek gue ada bibit bibit gaslighting anjeeee	DIDN TIDN TIDN TIDN DENG DIDN	INSERTION
loh dan mereka tau itu mereka tau betapa	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	CONGRUENT
sedih banget kalo baca kasus begini padah	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	CONGRUENT
dia ini punya cowok yang manipulatif gasli	DIDN TIDN TIDN TIDN TIDN DENG DIDN TIDN TIDN	INSERTION
selalu gitu diajarin siapa sih gaslighting beg	DIDN TIDN TIDN TIDN TIDN DENG DIDN	INSERTION
mereka bisa ngelakuin apa aja ke kita kalau	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	INSERTION

Gambar 4. 3 Hasil Pelabelan Data

PENUTUP

Kesimpulan

Berdasarkan penelitian yang berjudul perbandingan klasifikasi dengan algoritma K-Nearest Neighbor (KNN) dan Decision Tree pada struktur kalimat bahasa jaksel (campur kode bahasa Indonesia-Inggris) penulis dapat mengambil kesimpulan sebagai berikut.

1. Berdasarkan hasil validitas pengujian data menggunakan *K-Fold Cross Validation* dan *Confusion Matrix*, nilai *accuracy* tertinggi pada Algoritma K-Nearest Neighbor didapatkan pada fold ke-5 dan nilai N-Gram = 2 atau *Bigram* dengan nilai sebesar 75,29%. Kemudian dari hasil validitas pengujian data menggunakan *K-Fold Cross Validation* dan *Confusion Matrix*, nilai *accuracy* tertinggi pada Algoritma *Decision Tree* didapatkan pada fold ke-5, 6, 8 dan nilai N-Gram = 3 atau *Trigram* dengan nilai sebesar 79,21%.

2. Dari Skenario pengujian yang telah dilakukan membuktikan bahwa Algoritma *Decision Tree* menjadi lebih unggul daripada Algoritma *K-Nearest Neighbor* dalam mengklasifikasikan struktur kalimat yang mengandung Bahasa Jaksel atau Campur Kode Bahasa Inggris – Bahasa Indonesia.
3. Dalam melakukan pengumpulan data pada media sosial Twitter terdapat parameter yang digunakan sebagai kata kunci pencarian yaitu kosa kata yang sering digunakan dalam Bahasa Jaksel. Adapun kosa kata tersebut berdasarkan observasi yang telah dilakukan oleh peneliti dan disusun dalam sebuah tabel dengan *keywords* sebanyak 158 kata. Dari hasil melakukan *scraping* dan pemilihan data secara manual dihasilkan sebanyak 5099 data *tweets* yang digunakan pada penelitian.
4. Pengujian model dilakukan dengan 3 Skenario Pengujian pada Model Klasifikasi Algoritma *K-Nearest Neighbor* dan *Decision Tree*. Untuk Skenario 1 menggunakan Unigram, Skenario 2 menggunakan Bigram, dan Skenario 3 menggunakan Trigram. Parameter pengujian yang digunakan menggunakan parameter hasil dari *Grid Search*. Pada Algoritma *K-Nearest Neighbor* penyebaran *range* parameter K yaitu dari rentang 1 sampai 30. Kemudian, Algoritma *Decision Tree* rentang atau *range* parameter yang digunakan untuk *max depth* adalah 30 dan parameter *Criterion Gini* atau *Entropy*.

Saran

Berdasarkan dari penelitian yang telah dilakukan terdapat beberapa saran yang dapat dijadikan sebagai bahan pertimbangan untuk penelitian selanjutnya agar proses dan hasil penelitian lebih efektif dan baik sebagai berikut.

1. Pelabelan sebaiknya dilakukan secara otomatis, agar mempermudah dan mempersingkat waktu dalam pelabelan data.
2. Menambah dan memperluas jangkauan kosa kata Bahasa Jaksel agar data yang dihasilkan pada saat pengumpulan data lebih variatif, persebaran data yang merata, dan model dapat melakukan generalisasi data dengan baik. Adapun jangkauan perluasan untuk parameter untuk algoritma *K-Nearest Neighbor range* nilai K diperluas hingga 100. Kemudian untuk algoritma *Decision Tree* parameter *Criterion* ditambah lebih bervariasi dan *range maximal depth* diperluas menjadi 100.
3. Melakukan analisis lanjutan terhadap nilai dari performa masing-masing model klasifikasi algoritma.
4. Melakukan pembaharuan data pada rentang waktu pengambilan data agar menghasilkan data yang relevan dan terbaru.

DAFTAR PUSTAKA

Amin, F., 2012. Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. *JSINBIS (Jurnal Sistem Informasi Bisnis),* Volume Vol. 2 No. 2, pp. 78-83.

Amri, Y. K., 2019. *Alih Kode Dan Campur Kode Pada Media Sosial.* s.l., Prosiding Seminar Nasional PBSI II Tema: Guru dan Dosen Kreatif Abad XXI.

Berry, M. & Kogan, J., 2010. *Text Mining Aplication and theory.* United Kingdom: WILEY.

Cangara, H., 2007. *Pengantar Ilmu Komunikasi.* Jakarta: PT. Raja Grapindo Persada.

Cristanto, A. S., 2018. *Pola Komunikasi Internal Pada Sales dan Marketing PT Simnu.* [Online]
Available at: https://repository.uksw.edu/bitstream/123456789/21899/2/T0_102015004_Full%20text.pdf

Girnanfa, F. A. & Susilo, A., 2022. Studi Dramaturgi Pengelolaan Kesan Melalui Twitter Sebagai Sarana Eksistensi Diri Mahasiswa di Jakarta. *Journal of New Media and Communication*, Volume Vol. 1, pp. 58-73.

Guterres, A. & Santoso, J., 2019. Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based. *Teknika*, Volume Vol. 8 No. 2, pp. 142-147.

Hadi, C. F., Yasi, R. M. & Prasetyo, A., 2024. Model Decision Tree ForecastingBerbasis DHT22 pada Smart HydroponicMicrogreen. *Journal of Telecommunication, Electronics, and Control Engineering (JTECE)*, Volume Vol. 6 No. 1, pp. 29-38.

Hadiwijaya, H., 2018. Pengaruh Komunikasi Dan Kualitas Pelayanan Terhadap Kinerja. *International Journal of Social Science and Business.*, pp. 124-131.

Handoko, W. T. et al., 2022. Klasifikasi Opini Pengguna Media Sosial Twitter Terhadap JNT Di Indonesia dengan Algoritma Decision Tree. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, Volume Vol. 6 No. 2, pp. 790-799.

Hardjana, A. M., 2003. *Komunikasi intrapersonal & Komunikasi Interpersonal.* Yogyakarta: Kanisius.

Haryadi, 2009. *Administrasi Perkantoran untuk Manajer dan Staff.* Jakarta: Visimedia.

Jurafsky, D. & Martin, J., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Prentice Hall, Upper Saddle River, NJ.*

Munthe, C. J. E., Hasibuan, N. A. & Hutabarat, H., 2022. Penerapan Algoritma Text Mining Dan TF-RF Dalam Menentukan Promo Produk Pada Marketplace. *Resolusi: Rekayasa Teknik Informatika dan Informasi*, Volume Vol. 2 No. 3, pp. 110-115.

Muysken, P., 2000. *Bilingual Speech: A Typology of Code-Mixing.* Cambridge: Cambridge University.

Nata, G. N. M. & Yudiastra, P. P., 2017. Preprocessing Text Mining pada email box berbahasa Indonesia. *E-Proceedings KNS&I STIKOM Bali*, pp. 479-483.

Nikmatun, I. A. & Waspada, I., 2019. Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, Volume Vol. 10 No. 2, pp. 421-432.

Noveanto, M., 2022. *UJI AKURASI KLASIFIKASI EMOSI PADA LIRIK LAGU BAHASA INDONESIA.* Universitas Tanjungpura Pontianak: s.n.

Nur, A., U. E. & Nina, S., 2021. PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM UNTUK TEXT CLASSIFICATION. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, Volume Vol. 6 No. 2, pp. 129-136.

Nursyafitri, G. D., 2023. *DQ-LAB.* [Online] Available at: https://dqlab.id/machine-learning-model-and-hyperparameter-tuning [Accessed 20 Maret 2024].

Nurzahputra, A. & Muslim, M. A., 2016. *Analisis Sentimen pada Opini Mahasiswa Menggunakan Natural Language Processing.* Semarang, Seminar Nasional Ilmu Komputer (SNIK 2016).

Pamungkas, F. S. & Kharisudina, I., 2021. *Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter.* Semarang, Universitas Negeri Semarang, pp. 628-634.

Permana, A. P., Ainiyah, K. & Holle, K. F. H., 2021. Analisis Perbandingan Algoritma Decision Tree, KNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-Up. *JISKA (Jurnal Informatika Sunan Kalijaga)*, Volume Vol 6 No.3, pp. 178-188.

Prakasa, O. S. Y. & Lhaksamana, K. M., 2018. KLASIFIKASI TEKS DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR PADA KASUS

KINERJA PEMERINTAH DI TWITTER. *e-Proceeding of Engineering*, Vol.5 No.3(ISSN : 2355-9365), pp. 8237-8248.

Priyanto, A. & Ma'arif, M. R., 2018. Implementasi Web Scrapingdan Text Mininguntuk Akuisisi dan Kategorisasi Informasi Laman WebTentang Hidroponik. *Indonesian Journal of Information Systems (IJIS)*, pp. 25-33.

Putri, T. A. E., Widiharih, T. & Santoso, R., 2022. PENERAPAN TUNING HYPERPARAMETER RANDOMSEARCHCV PADA. *JURNAL GAUSSIAN*, Volume Vol.1 No.3, pp. 397-406.

Rahman, M. F., Alamsah, D., Darmawidjadja, M. I. & Nurma, I., 2017. Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). *urnal Informatika*, Volume Vol. 11 No. 1, p. 36.

Ridwana, Y., 2018. *Campur Kode Dalam Lirik Lagu Grup Band One Ok Rock Dalam Album ゼイタクビョウ (Zeitakubyou)*, s.l.: Doctoral dissertation, Universitas Komputer Indonesia.

Romzi, M. & Kurniawan, B., 2020. Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma. *JTIM: Jurnal Teknik Informatika Mahakarya*, Volume Vol. 3 No. 2, pp. 37-44.

Sa'adah, L., 2020. *Analisis Sentimen Review E-Commerce pada Twitter Menggunakan dan Metode Klasifikasi Support Vector Machine.* [Online] Available at: https:/ /repository.telkomuniversity. ac.id/pustaka/158487/analisis-sentimen-review-e-commerce-pada-twitte r-menggunakan-metode-klasifikasi-support-vector-machine.html [Accessed 2 Februari 2024].

Saragih, R. R., 2016. *Pemrograman dan bahasa Pemrograman. STMIK-STIE Mikroskil.* s.l.:s.n.

Setianingrum, A., Hindayanti, A., Cahya, D. M. & Purnia, D. S., 2021. Perbandingan Metode Algoritma K-NN & Metode Algoritma C45 Pada Analisa Kredit Macet (Studi Kasus PT Tungmung Textil Bintan). *EVOLUSI: Jurnal Sains Dan Manajemen*, Volume Vol. 9 No. 2.

Sivakumar, A. & Gunasundari, R., 2017. A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining. *International Journal of Pure and Applied Mathematics*, Volume 117, pp. 785-794.

Sulis, M., Muhammad Zidny, N. & Indra, H., 2018. *EKSTRAKSI TF-IDF N-GRAM DARI KOMENTAR PELANGGAN PRODUKSMARTPHONE PADA WEBSITE E-COMMERCE*. Yogyakarta, Seminar Nasional Teknologi Informasi dan Multimedia.

Surbakti, A. Q., Hayami, R. & Al-Amien, J., 2021. Analisa Tanggapan Terhadap PSBB di Indonesia dengan Algoritma Decision Tree pada Twitter. *Jurnal Computer Science and Information Technology (CoSciTech)*, Volume Vol. 2 No.2, pp. 91-97.

Susanto, E. B., Cahyanto, T. A. & Umilasari, R., 2019. ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN KNN UNTUK KLASIFIKASI MULTI DATASET. *Skripsi*, p. Universitas Muhammadiyah Jember.

Tuntun, R., Kusrini, K. & Kusnawi, K., 2022. Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation. *Jurnal Media Informatika Budidarma*, Volume Vol. 6 No. 4, pp. 2111-2119.

Ulfiyani, S., 2014. Alih Kode dan Campur Kode Dalam Tuturan Masyarakat. *CULTURE*, Volume Vol. 1 No.1, pp. 92-100.

Wicaksono, A. F. & Purwarianti, A., 2010. *HMM based part-of-speech tagger for Bahasa Indonesia.* Jakarta, In Fourth International MALINDO Workshop.

Wijaya, E. T., 2013. PERANCANGAN INFORMATION RETRIEVAL (IR) BERBASIS TERM FREQUENCYINVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK

PERINGKASAN TEKS TUGAS. *Jurnal Ilmiah Teknologi dan Informasi ASIA,* Volume Vol. 7 No.1, pp. 78-92.

Yunita, D., 2017. Perbandingan Algoritma K-Nearest Neighbor dan Decision Tree untuk Penentuan Risiko Kredit Kepemilikan Mobil. *Jurnal Informatika Universitas Pamulang,* Volume Vol. 2 No. 2, pp. 103-107.

Zaman, B., Hariyanti, E. & Purwanti, E., 2015. Sistem Deteksi Bahasa pada Dokumen. *JURNAL MULTINETICS,* Volume Vol. 1 No. 2, pp. 21-26.