(2025), 4 (1): 8–24

Sabda

2962-5351

Jurnal Sastra dan Bahasa

KLASIFIKASI STRUKTUR KALIMAT BAHASA JAKSEL (CAMPUR KODE BAHASA INDONESIA-INGGRIS)

Farhan Syafaat [1], Arif Bijaksana Putra Negara [2], Hafiz Muhardi [3]
Fakultas Teknik Universitas Tanjungpura Pontianak
Farhansyafaat6699@gmail.com

Submitted 3 Januari 2025	Accepted 8 Januari 2025	Published 9 Januari 2025

ABSTRAK

Dinamika sosial dalam hubungan antar manusia tercermin dari kompleksitas komunikasi dalam kehidupan bermasyarakat. Proses komunikasi sebagai landasan utama interaksi sosial telah berkembang secara signifikan dengan bantuan teknologi *modern*. Di kalangan Milenial dan Gen Z, pola interaksinya ditandai dengan penggunaan bahasa gaul atau campur kode. Fenomena kebahasaan campur kode ini bermunculan di Jakarta dan menyebar ke media sosial, terutama *twitter*. Istilah bahasa JakSel ini merupakan sebuah modifikasi dari bahasa Indonesia dan Inggris, menampilkan keunikan dalam penggunaan campuran kode dan istilah *slang*. Popularitasnya di media sosial menimbulkan tantangan dalam pemahaman, terutama karena struktur bahasa JakSel tidak didefinisikan dengan jelas dalam data teks *twitter*. Penelitian ini bertujuan untuk mengklasifikasikan teks yang mengandung bahasa JakSel berdasarkan tingkat keacakan struktur kalimat pada cuitan *twitter* menggunakan *naive bayes classifier* dan algoritma *neural network*. Tahap pengujian menggunakan skenario *n-gram* mengimplementasikan *k-fold cross validation* dengan nilai K=10 dan *confusion matrix.* Berdasarkan hasil pengujian diperoleh rata-rata nilai akurasi optimal menggunakan algoritma *neural network* model *trigram* sebesar 70,30% dibandingkan menggunakan algoritma *naive bayes* sebesar 57,20%.

Kata kunci: *naive bayes classifier, neural network,* campur kode, bahasa Jaksel, *twitter*

PENDAHULUAN

Hubungan antar manusia selalu dipengaruhi oleh dinamika sosial yang terbentuk melalui komunikasi dalam kehidupan bersosial dan bermasyarakat. Komunikasi adalah suatu proses sosial dimana individu menggunakan simbol-simbol untuk menciptakan dan menafsirkan pesan baik secara verbal dan nonverbal (West dan Turner, 2007). Pesan yang disampaikan bisa bersifat abstrak dan konkrit. Pesan konkrit dapat berupa suara, ekspresi wajah, gerak tubuh, bahasa lisan, dan tulisan. Transmisi pesan lisan dari pengirim ke penerima melibatkan penggunaan bahasa. Bahasa digunakan untuk menyampaikan suatu ide, gagasan, dan pikiran. Saat ini, perkembangan bahasa sangat kompleks dikarenakan adanya dukungan teknologi sebagai sarana komunikasi yang cepat dan handal, khususnya dikalangan masyarakat modern.

Masyarakat di era modern ini dapat dikategorikan menjadi generasi milenial atau generasi Z. Pada umumnya pola interaksi sebagian besar kalangan generasi milenial atau generasi Z menggunakan penuturan campur kode atau lebih sering dikenal dengan sebutan bahasa gaul. Bahasa gaul merupakan pengembangan atau modifikasi dari banyak bahasa yang berbeda termasuk bahasa Indonesia, sehingga bahasa gaul tidak mempunyai struktur gaya bahasa yang pasti (Gunawan, 2008). Selain itu, bahasa gaul mempunyai ciri-ciri khusus bersifat pendek, lincah, unik, ringkas, dan kreatif. Fenomena bahasa gaul mulai digunakan oleh remaja di Jakarta, dan semakin populer di berbagai media sosial sehingga dikenal sebagai bahasa Jaksel. Keunikan dari bahasa Jaksel yaitu mengkombinasikan penggunaan bahasa Indonesia tidak baku dan bahasa Inggris baku dalam berkomunikasi, juga menggunakan istilah singkatan gaul atau kata-kata *slang*. Adapun contoh kalimat yang mengandung bahasa Jaksel berbunyi ”*Today* kita tidak perlu *too much overanalyzed* konsep rapat”. Perkembangan bahasa Jaksel sering kali ditemukan di media sosial khususnya *platform* Twitter. Twitter merupakan *platform* yang memudahkan pengguna dalam bertukar informasi secara luas, menyampaikan pendapat, mencari fenomena dari tren masa kini melalui kiriman cuitan atau *tweets*, hastag, dan balasan komentar. Meluasnya penggunaan bahasa Jaksel dalam *tweet* membuat banyak orang

kesulitan memahami kosakata bahasa tersebut. Hal ini dikarenakan data teks Twitter belum memiliki struktur kebahasaan yang memadai sehingga harus dianalisis terlebih dahulu untuk mengekstrak informasi dari data yang sedang tren tersebut. Cara yang dapat digunakan untuk mengolah data teks dalam bidang keilmuan *data mining* yaitu *text mining*.

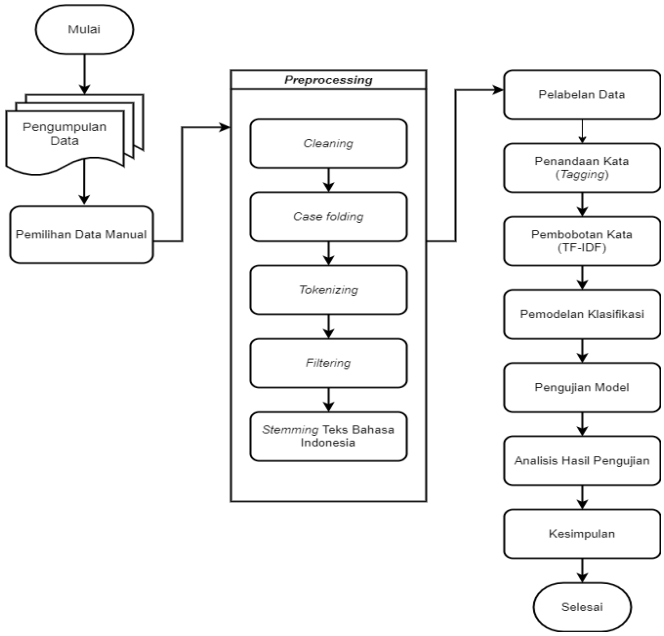
Text mining merupakan metode yang digunakan untuk menganalisis teks dan mengekstrak informasi berguna dari dokumen teks. Salah satu proses yang dapat dilakukan *text mining* adalah klasifikasi teks. Klasifikasi teks dapat diartikan sebagai proses pengklasifikasian dokumen teks ke dalam kelas-kelas tertentu (Prakoso et al., 2018). Metode yang dapat digunakan untuk membangun model klasifikasi data teks adalah algoritma *naive bayes classifier* dan algoritma *neural network*.

Naive bayes classifier menggambarkan perhitungan probabilitas untuk setiap kelas keputusan dengan syarat kelas keputusan tersebut benar, dengan adanya vektor informasi objek dan atribut objek bersifat independen (Olson dan Delen, 2008). Selain itu, algoritma jaringan saraf atau *neural network* adalah kumpulan unit masukan atau keluaran yang terhubung, dimana setiap koneksi memiliki bobot. Selama tahap pembelajaran, jaringan saraf menyesuaikan bobot untuk dapat memprediksi kelas yang benar dari kumpulan data (Han dan Kamber, 2006). Beberapa penelitian memperoleh hasil yang berbeda-beda dengan menggunakan algoritma klasifikasi *naive bayes* dan *neural network*, antara lain penelitian yang dilakukan oleh Abiodun (2023) berjudul "*Comparing the performance of three text classification algorithms: Convolutional Neural Networks, Multinomial Naive Bayes, and SVM*" menginformasikan performa algoritma klasifikasi teks *convolutional neural network* lebih unggul dalam semua parameter evaluasi dengan nilai akurasi sebesar 77%, diikuti SVM dengan akurasi 76%, dan *multinomial naive bayes* dengan akurasi terendah sebesar 69%. Selain itu, penelitian yang dilakukan oleh Chaturanga dan Ranathunga (2021) dengan judul "*Classification of Code-Mixed Text Using Capsule Networks*" memperoleh hasil akurasi menggunakan kombinasi *neural network* sebesar 89,8%, *precision* 83,7%, *recall* 81,1%, dan skor F1 80,6% dibandingkan tidak menggunakan kombinasi *neural network*.

Berdasarkan uraian di atas, maka dilakukan penelitian pengklasifikasian teks yang mengandung bahasa Jaksel berdasarkan tingkat keacakan struktur kalimat pada data cuitan Twitter dengan membandingkan dua metode algoritma, yaitu algoritma *naive bayes* dan algoritma *neural network* untuk membandingkan nilai presisi setiap algoritma. Untuk pengklasifikasian struktur kalimat bahasa Jaksel penelitian ini menggunakan penelitian Muysken (2000) berjudul "*Bilingual Speech: A Typology Of Code-Mixing*" yang menerapkan tiga aturan model klasifikasi yaitu *insertion, alternation,* dan *congurent lexicalization*. Aturan *insertion* mengacu pada penambahan kata dari suatu bahasa ke kalimat utama dalam bahasa lain. Dilanjutkan pada aturan model *alternation* melibatkan peralihan bahasa secara keseluruhan atau sebagian antar frasa pada masing-masing bahasa. Selanjutnya, model *congurent lexicalization* menyoroti hasil kombinasi dua unsur linguistik secara paralel dalam sebuah kalimat. Sehingga penelitian ini memberikan hasil pengklasifikasian keacakan stuktur kalimat yang mengandung bahasa Jaksel dengan mengimplementasikan performa dari algoritma *naive bayes* dan *neural network*.

METODOLOGI PENELITIAN

Ada beberapa langkah penelitian dengan berbagai metode yang dilakukan pada penelitian ini, dapat dilihat pada Gambar 3.1



Gambar 3. 1 *Flowchart* Langkah 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 yang digunakan dalam penelitian ini adalah Asus ROG GL553VD (Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz 2.81 GHz, RAM 16,0 GB.

B. Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini adalah Sistem Operasi Microsoft Windows 10 dan Google Colab.

C. Data Penelitian

Sumber data dari penelitian ini berasal dari data *tweet* melalui aplikasi Twitter. Data *tweet* yang dikumpulkan adalah fenomena campur kode tahun 2021 yang dikenal dengan bahasa Jaksel, yaitu menggabungkan bahasa Indonesia tidak baku dan bahasa Inggris baku. Metode pengumpulan data menggunakan teknik *scrapin*g terintegrasi Twitter API dengan aplikasi *rapid miner* berdasarkan 158 kata kunci bahasa Inggris yang umum digunakan dalam campur kode bahasa Jaksel dan kata-kata bahasa Indonesia dipilih dari keseluruhan *tweet*. Selain itu, data penelitian berjumlah 5099 dilakukan pembagian data menggunakan *cross validation* agar data uji dan data latih digunakan secara bergantian.

Pengumpulan Data

Pengumpulan data penelitian bersumber dari data *tweet* melalui aplikasi Twitter. Data *tweet* yang dikumpulkan adalah fenomena campur kode tahun 2021 yang dikenal dengan bahasa Jaksel, yaitu menggabungkan bahasa Indonesia tidak baku dan bahasa Inggris baku, serta bahasa Betawi. Metode pengumpulan data ini terarah yaitu dengan melakukan *scraping* data *tweet* melalui API Twitter pada aplikasi *rapid miner* dengan memasukan parameter teks dalam bahasa Inggris dan bahasa *tweets* adalah bahasa Indonesia serta limit 300 data *tweets* setiap penarikan data. Dari hasil *scraping* memperoleh 28.564 data *tweets* mengandung campur kode kalimat dalam bahasa Jaksel yang digunakan untuk penentuan kata kunci berjumlah 158

kata yang ditampilkan pada **Tabel 4.1** dalam menentukan karateristik data yang mengandung campur kode bahasa Jaksel. Dari tahapan yang sudah dilakukan memperoleh data mentah, selanjutnya dilakukan pembersihan secara manual berdasarkan konseptual pengetahuan penulis untuk menghindari cuitan data *tweets* mengandung spam dan duplikasi.

Pemilihan Data Manual

Tahap pemilihan data ini dilakukan dengan cara melakukan *filtering* data secara manual. Data yang telah dihasilkan melalui proses *scraping* diseleksi untuk menghindari cuitan data *tweets* mengandung spam dan duplikasi.

Text Preprocessing

Tahap *text preprocessing* bertujuan untuk mengubah data teks tidak terstruktur menjadi vektor data teks terstruktur. Pada saat melakukan tahap *preprocessing*, beberapa prosedur yang digunakan dalam penelitian ini antara lain:

A. *Cleaning* merupakan proses menghilangkan karakter maupun tanda baca yang tidak diperlukan, seperti contoh berikut ini.
   Sebelum *Cleaning* :”@#Anxi!Ous BAnget mAu liat wHAtsap mengheRANkan!245@”
   Setelah *Cleaning* :”AnxiOus BAnget mAu liat wHAtsap mengheRANkan”

B. *Case folding* meliputi penggunaan huruf kapital pada huruf kecil dan huruf besar. Karena data *tweet* seringkali menggunakan huruf kecil, maka metode *case folding* digunakan untuk mengubah ukuran huruf atau karakter menjadi *lowercase*, seperti contoh berikut ini.
   Sebelum *Case Folding* :”AnxiOus BAnget mAu liat wHAtsap mengheRANkan”
   Setelah *Case Folding* :”anxious banget mau liat whatsap mengherankan”

C. *Tokenizing* yaitu proses memecah data cuitan hasil pembersihan sebelumnya menjadi *token*, seperti contoh berikut ini.
   Sebelum *Tokenizing* :”anxious banget mau liat whatsap mengherankan”
   Setelah *Tokenizing* :”anxious”, ”banget”, ”mau”, ”liat”, ”whatsap”, ”mengherankan”

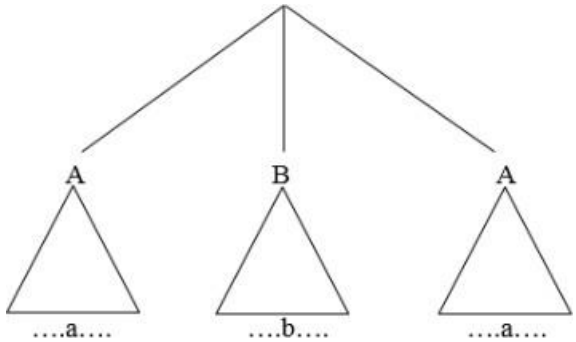
D. *Filtering* merupakan proses pemilihan kata-kata penting dari hasil pengkodean, khususnya kata-kata yang dapat digunakan untuk mewakili isi suatu teks atau dokumen, seperti contoh berikut ini.
   Sebelum *Filtering* : ”anxious banget mau liat whatsap mengherankan”
   Setelah *Filtering* : ”anxious banget liat whatsap mengherankan”

E. *Stemming* merupakan proses pencarian kata dasar untuk setiap kata pada suatu dokumen dengan cara menghilangkan imbuhan baik prefiks maupun sufiks dalam bahasa Indonesia, seperti contoh berikut ini.
   Sebelum *Stemming* : ”anxious banget liat whatsap mengherankan”
   Setelah *Stemming* : ”anxious banget lihat whatsap heran”

Pelabelan Data

Pelabelan data dilakukan secara manual setelah memperoleh data bersih dari tahapan sebelumnya. Penelitian ini menerapkan tiga jenis kelas pada campur kode yaitu *insertion*, *alternation*, dan *congruent lexicalization* (Muysken, 2000).

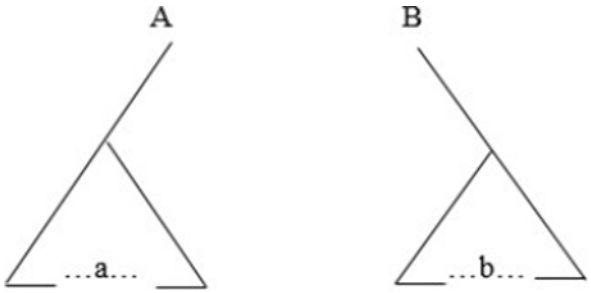
a. *Insertion*



Gambar 3. 2 Model *Insertion* pada Campur Kode

Pada Gambar 3.2 unsur "a" mewakili kalimat bahasa utama dan unsur "b" mewakili kalimat bahasa kedua yang disisipkan oleh penutur. Contoh penyisipannya adalah "Mereka itu merupakan *hardworker* walaupun hari ini adalah hari libur nasional" dimana frasa "*hardworker*" merupakan unsur yang disisipkan oleh penutur.

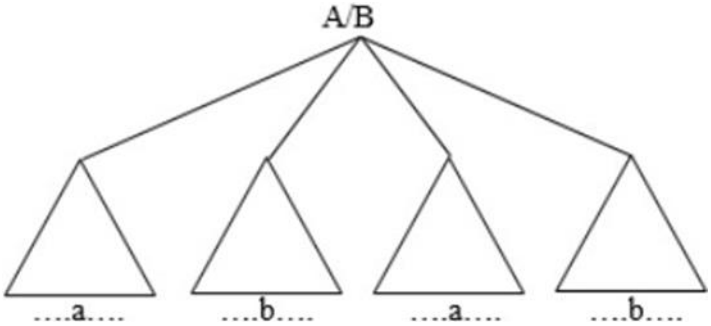
Alternation



Gambar 3. 3 Model *Alternation* pada Campur Kode

Pada Gambar 3.3 unsur "A" dan "B" merupakan representasi dari dua bahasa yang berbeda dimana unsur tersebut mewakili variasi kebahasaan dalam campuran kode-kode berupa tuturan yang diciptakan oleh penutur. Contoh dalam *Alternation* adalah "*I will get it*, karena setiap yang sudah aku recanakan tentu memerlukan tindakan yang efektif".

Congurent lexicalization



Gambar 3. 4 Model *Congurent lexicalization* pada Campur Kode

Pada Gambar 3.4 model *congurent lexicalization* dalam campur kode, unsur "A" dan "B" merupakan representasi dari dua bahasa yang berbeda. Pada frasa ini, penutur lebih cenderung menggabungkan dua bahasa dengan melihat struktur gramatikalnya, sehingga penggunaan frasa ini melibatkan pengisian kalimat secara leksikal dengan item leksikal dari kedua bahasa tersebut. Contoh leksikalisasi yang tepat adalah "*Today*, pimpinan kita akan melakukan *discuss* terkait *manufacture project* yang akan dibangun pada tahun 2024". Berikut ini adalah hasil pelabelan data secara manual pada penelitian ini.

Tabel 3. 1 Hasil Pelabelan Kelas

Data Teks	Label Kelas
gua kalo nunggu jawab chat penting tuh jadi gabisa nyambi ngerjain macem macem karena anxious nunggunya	INSERTION
anxious banget mau liat whatsap heran	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

Dalam penelitian ini dilakukan proses penentuan penandaan kata yang diterapkan pada setiap kata dalam suatu kalimat campur kode bahasa Jaksel. Aturan penandaan kata yang digunakan berdasarkan posisi dan tipe kata dalam kalimat campur kode. Menurut Muysken (2000) campur kode termasuk bidang linguistik sosiolinguistik yaitu data yang berasal dari pengucapan dalam suatu komunitas atau wilayah tertentu atau lebih dikenal sebagai bahasa masyarakat. Fenomena campur kode ini dikenal dengan sebutan bahasa Jaksel, yaitu memadukan struktur kalimat bahasa Indonesia tidak baku dan bahasa Inggris baku. Contoh aturan penandaan kata digunakan dalam penelitian ini memerlukan penandaan untuk menunjukkan kata mana mengandung kata dalam bahasa Indonesia dan bahasa Inggris. Dalam kalimat bahasa campur kode, setiap kata yang digunakan mempunyai kedudukan: depan dan tengah. Lalu, penandaan tersebut dibagi menjadi empat posisi kata dan jenis bahasa kata: DENG (Depan *English*), TENG (Tengah *English*), DIDN (Depan Indonesia), dan TIDN (Tengah Indonesia), seperti yang ditunjukkan pada Tabel 3.2 dan hasil penandaan kata pada Lampiran A data penelitian.

Tabel 3. 2 **Aturan Penandaan Kata**

Data Teks	Penandaan Kata
gua kalo nunggu jawab chat penting tuh jadi gabisa nyambi ngerjain macem macem karena anxious nunggunya	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG DIDN
anxious banget mau liat whatsap mengherankan	DENG DIDN TIDN TIDN TIDN TIDN
deep down aku nyata gampang anxious kalo ketemu tatap muka yang related to academics things padahal kalo onlen kayak gapunya takut	DENG TENG DIDN TIDN TIDN DENG DIDN TIDN TIDN TIDN TIDN DENG TENG TENG TENG DIDN TIDN TIDN TIDN TIDN TIDN

Pembobotan Kata (TF-IDF)

Pembobotan kata dilakukan menggunakan algoritma frekuensi dokumen invers frekuensi istilah yang menentukan bobot kata atau istilah berdasarkan frekuensinya di dalam dan di seluruh dokumen. Penggunaan *n-gram* diperlukan untuk memudahkan perhitungan dan menghindari kesalahan dalam ekstraksi *Term Frekuensi Inverse Document Frekuensi* (Nur et al., 2021). *N-gram* adalah subset dari *n* karakter dari sebuah *string.* Penggunaan *n-gram* memiliki keunggulan yaitu hasil yang diperoleh lebih akurat dan efektif (Sulis et al., 2018). *N-gram* adalah subset dari *n* karakter dari sebuah *string*, spasi ditambahkan pada awal dan akhir *strin*g untuk menentukan batas antara awal dan akhir *string* (Zaman et al., 2015). *N-Gra*m memiliki tiga fungsi: *Unigram* (*n-gram* = 1), *Bigram* (*n-gram* = 2), dan *Trigram* (*n-gram* = 3).

Contoh penggunaan *n-gram* dengan *string* "" Jika menambahkan awal dan akhir dengan "_" sebagai pengganti spasi lalu menambahkan "anxious banget mau liat whatsap mengherankan" akan memperoleh hasil seperti berikut ini.

Unigram: [’anxious’, ’banget’, ’mau’, ’liat’, ’whatsap’, ’mengherankan’]

Bigram: [’anxious banget’, ’banget mau’, ’mau liat’, ’liat whatsap’, ’whatsap mengherankan’]

Trigram: [’anxious banget mau’, ’banget mau liat’, ’mau liat whatsap’, ’liat whatsap mengherankan’]

Hyperparameter Tuning

Penyetelan *hyperparameter* adalah proses pemilihan parameter optimal untuk membangun model klasifikasi. Penelitian ini menggunakan optimasi parameter otomatis dengan fungsi *GridSearchCV* dari perpustakaan *Scikit Learn.* Optimasi parameter digunakan untuk mencari parameter yang menghasilkan nilai akurasi model terbaik dari distribusi rentang nilai parameter yang ditentukan dalam penelitian ini. Adapun parameter yang diujikan pada setiap masing-masing algoritma yaitu algoritma *naive bayes* adalah nilai *alpha* serta algoritma *neural network* adalah nilai *alpha, hidden layers size, activation, solver,learning rate*, serta *learning rate init*. Selanjutnya rentang parameter yang digunakan dalam penelitian ini yaitu nilai *alpha* mulai dari rentang 0,1 hingga 10. Serta rentang nilai *alpha neural network* dimulai dari 0,0001 hingga 0,001, kriteria aktivasi ReLu, Tanh, *hidden layers size* terdiri dari dua lapisan masing-masing berukuran 10, *solver* optimasi yang digunakan SGD, Adam, *learning rate init* dimulai dari 0,01 hingga 0,1, serta *learning rate* untuk menentukan kecepatan pembelajaran model yaitu *constant*, *adaptive*.

Pemodelan Klasifikasi Algoritma

Tahap pemodelan klasifikasi algoritma merupakan langkah penting dalam mengimplementasikan model algoritma. Pada penelitian ini menggunakan algoritma *naive bayes classifier* dan *neural network* sebagai teknik klasifikasi. Terdapat beberapa parameter yang dicari untuk menemukan parameter terbaik yang digunakan pada setiap algoritma.

Tabel 3. 3 **Parameter *Grid* yang dicari pada Algoritma *Naive Bayes***

No	Algoritma	Parameter yang digunakan	Nilai Parameter yang Dicari
1	*Naive Bayes*	Nilai *Alpha*	Menggunakan rentang nilai dari 0.1 hingga 10

Berdasarkan Tabel 3.3 parameter yang digunakan pada algoritma klasifikasi *naive bayes* adalah *Alpha* (Noto dan Saputro, 2022). *Alpha* sebagai parameter penghalusan untuk mengontrol jika suatu kata tidak ada dalam data pelatihan, dimana hasil probabilitasnya tidak bernilai nol. Dalam dokumentasi Sklearn (2022) penggunaan rentang *alpha* sebaiknya tidak terlalu mendekati 0.0000000001. Oleh karena itu, penelitian ini menggunakan rentang nilai parameter dimulai dari 0.1 hingga 10 untuk menghindari kesalahan numerik yang dapat terjadi jika *alpha* terlalu dekat dengan nol.

HASIL DAN ANALISIS

Hasil Pengumpulan Data

Tahap pengumpulan data ini dilakukan dengan cara *scraping* data dari aplikasi Twitter menggunakan Twitter API dengan *rapid miner*. *Scraping* data yang diambil menggunakan *filtering* yang spesifik yaitu mengambil data cuitan kombinasi bahasa Indonesia dan bahasa Inggris dari tahun 2021. Namun sebelumnya, diperlukan metode *observasi* terkait penggunaan kata dalam bahasa Jaksel pada cuitan komunitas Twitter yang mengandung campuran bahasa Indonesia tidak baku dan bahasa Inggris baku. Data yang diperoleh diproses menjadi data *tweet* tidak terstruktur yang digunakan dalam penelitian ini. Selanjutnya parameter yang digunakan merupakan hasil dari observasi sebanyak 158 kata kunci lalu dibuat menjadi *corpus* campur kode kalimat bahasa Jaksel yang biasa digunakan dalam cuitan komunitas Twitter, seperti pada Tabel 4.1.

Tabel 4. 1 Parameter atau Kata Kunci *Tweets*

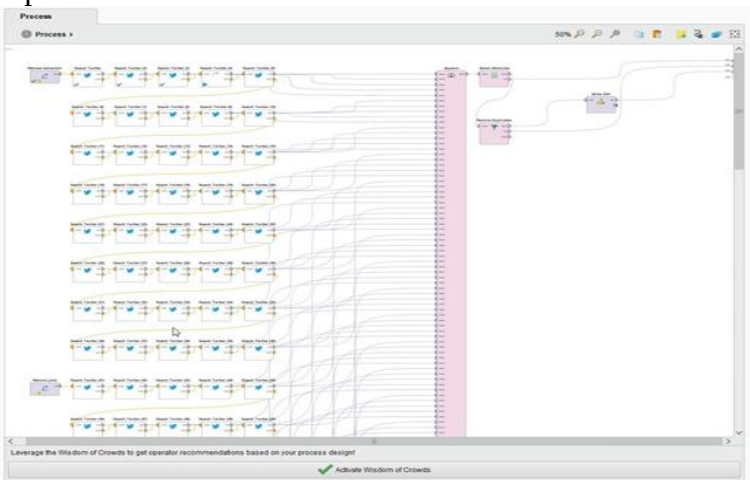
No.	Parameter
1	actually
2	af
3	after
4	anhedonia
5	anxious
6	attitude
7	assess
8	as fuck
9	around
10	anyway
11	basically
12	because why
13	be like
14	bestie
15	better
16	brain storming
17	body shaming
18	body goals
19	bit
20	bipolar
21	bro mens
22	burnout
No.	Parameter
45	financial freedom
46	for your 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

No.	Parameter
23	clingy
24	cmiiw
25	confirm
26	crush
27	couple goals
28	corporate
29	confused
30	confuse
31	deep talk
32	degrading
33	denial
34	detox sosmed
35	discuss
36	fancy
37	eye catching
38	ever
39	end up
40	emotional abuse
41	fashionable
42	fear of missing out
43	feminis
44	figuratively
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

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

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

Tahapan pengumpulan data kemudian dilanjutkan dengan menarik data *tweets* menggunakan teknik *scraping* dari aplikasi Twitter menggunakan Twitter API melalui *rapid miner*. *Scraping* data dilakukan dengan menggunakan kata kunci tertentu yang telah disusun sebelumnya sebagai *corpus*. Adapun hasil pengumpulan data menggunakan aplikasi *rapid miner*, ditunjukkan pada Gambar 4.1.



Gambar 4. 1 Hasil *Scraping* data menggunakan *Rapid Miner*

Data yang diperoleh dari hasil *scraping* menggunakan *rapid miner* berjumlah 28.564 data *tweets*, seperti ditunjukkan 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..

Selanjutnya dilakukan pembersihan secara manual berdasarkan konseptual pengetahuan peneliti untuk menghindari spam dan data cuitan yang duplikat. Sehingga diperoleh data bersih yang digunakan dalam penelitian sebanyak 5099 data *tweets* dan disimpan dalam format CSV.

Hasil Pemilihan Data Manual

Tahap pemilihan data ini dilakukan dengan cara melakukan *filtering* data secara manual berdasarkan konseptual penulis. Dalam proses ini peneliti menggunakan *tools spreadsheet* untuk memilah data yang digunakan pada tahap selanjutnya. Tujuan pemilihan data ini adalah untuk menghindari cuitan data *tweets* mengandung spam dan duplikasi, sehingga data bersih yang diperoleh dari tahap pemilihan data berjumlah 5099 data cuitan, ditunjukkan pada Gambar 4.2.

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 Tampilan Hasil Pemilihan Data Manual

Hasil Pelabelan Data

Berdasarkan hasil pemilihan data maka diperoleh data bersih berjumlah 5099 data cuitan yang digunakan ke tahap selanjutnya yaitu pelabelan data. Dalam proses ini peneliti melakukan pelabelan secara manual. Tahap pelabelan yang diterapkan pada penelitian ini adalah jenis kode campur yang terdiri dari 3 kelas meliputi *insertion*, *alternation*, dan *congruent lexicalization* (Muysken, 2000) yang ditampilkan seperti pada Gambar 4.3 dan Lampiran A.

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 TIDN	CONGRUENT
gimana cara henti self gaslighting ya tapi al	DIDN TIDN TIDN DENG TENG DIDN TIDN TIDN TIDN TIDN	CONGRUENT
salah satu tipe manusia yang bahaya dan k	DIDN TIDN 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 TIDN T	CONGRUENT
dah lama males nonton tiba tiba tarik non	DIDN TIDN 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 TIDN T	INSERTION
asli kaget banget sama dialog reza rahardi	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG DIDN	CONGRUENT
adek gue ada bibit bibit gaslighting anjeeee	DIDN TIDN TIDN TIDN TIDN DENG DIDN	INSERTION
loh dan mereka tau itu mereka tau betapa	DIDN TIDN 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 TIDN T	CONGRUENT
dia ini punya cowok yang manipulatif gasli	DIDN TIDN 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 kala	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T	INSERTION

Gambar 4. 3 Hasil Pelabelan Data

Hasil pelabelan data berdasarkan kelas penandaan kata memperoleh frekuensi yang berbeda. Dimana, jumlah penandaan kata untuk kelas *alternation* sebanyak 1608, kelas *congurent* sebanyak 1781, dan kelas *insertion* sebanyak 1710, seperti ditunjukkan pada Gambar 4.4.

	label
ALTERNATION	1608
CONGRUENT	1781
INSERTION	1710

Gambar 4. 4 Jumlah Persebaran Label Pada *Dataset*

Hasil *Preprocessing*

Dari hasil pengumpulan data cuitan dengan teknik *scraping*, proses pemilihan, dan pelabelan data maka akan dilakukan pengolahan data menjadi data terstuktur yang akan digunakan untuk pembuatan model klasifikasi. Dalam mengelola data cuitan yang diperoleh menggunakan *tools google colaboratory* dengan *import* data cuitan dalam bentuk file dokumen CSV, seperti pada Kode Program 4.1.

Kode Program 4. 1 *Import Data*

```
# Tambah data menggunakan pandas
# Corpus = pd.read_csv(r"datacsv.csv",encoding='latin-1')
# Corpus = pd.read_excel('data.xlsx')
sheet_id = "1WmfsZ-Ax_D9ob8pd3U5xm_ONbI1fiQzC9Q1lqOUVMgA"
sheet_name = "hasil"
url =
f"https://docs.google.com/spreadsheets/d/{sheet_id}/gviz/tq?tqx=out:csv
&sheet={sheet_name}", sep=","
Corpus = pd.read_csv(url)
#Ganti nama kolom
Corpus.columns = ['text', 'tag', 'label']
Corpus.head(10)
```

Adapun hasil dari tahapan *preprocessing* ditampilkan ke dalam bentuk data *frame* melalui *librariy* Pandas yang terdiri dari No, *text*, *tag*, dan *class label*, seperti ditunjukkan pada Gambar 4.5.

	text	tag	label
0	[sama, terus, ketemu, orang, baru, tuh, bikin,...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG DIDN T...	INSERTION
1	[sebenernya, yang, buat, anxious, itu, lu, sen...	DIDN TIDN TIDN DENG DIDN TIDN TIDN DENG DIDN T...	CONGRUENT
2	[di, internet, salah, satu, yang, bikin, aku, ...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG D...	INSERTION
3	[anxious, banget, mau, liat, whatsap, heran]	DENG DIDN TIDN TIDN TIDN TIDN	ALTERNATION
4	[tapi, beneran, ini, tiket, kereta, ke, bus, w...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T...	INSERTION
5	[aku, ngantuk, tapi, enggak, bisa, tidur, kare...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN DENG	ALTERNATION
6	[gua, kalo, nunggu, jawab, chat, penting, tuh,...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T...	INSERTION
7	[nanti, kalo, aku, udah, baik, baik, lagi, uda...	DIDN TIDN TIDN TIDN TIDN TIDN TIDN TIDN T...	INSERTION
8	[deep, down, aku, nyata, gampang, anxious, kal...	DENG TENG DIDN TIDN TIDN DENG DIDN TIDN TIDN T...	CONGRUENT
9	[pantes, aja, anjer, bangun, bangun, langsung,...	DIDN TIDN TIDN TIDN TIDN TIDN DENG DIDN TIDN T...	INSERTION

Gambar 4. 5 Hasil *Preprocessing*

Selanjutnya dilakukan tahap *preprocessing* bertujuan untuk mengubah data teks tidak terstruktur menjadi vektor data teks terstruktur. Dalam tahap *preprocessing* meliputi *case folding* digunakan untuk mengubah ukuran huruf atau karakter menjadi *lowercase*, seperti ditunjukkan pada Kode program 4.2.

Kode Program 4. 2 *Case Folding*

```
# Case folding mengubah teks menjadi huruf kecil semua
# Step - 1a : hapus baris kosong jika ada
Corpus['text'].dropna(inplace=True)
```

Kemudian dilakukan *tokenizing* untuk memotong data teks yang diperoleh menjadi *token-token*, seperti ditunjukkan pada Kode program 4.3.

Kode Program 4.3 *Tokenizing*

```
# Tokenizing memecah dokumen teks menjadi token kata
Corpus['text']= [word_tokenize(entry) for entry in Corpus['text']]
```

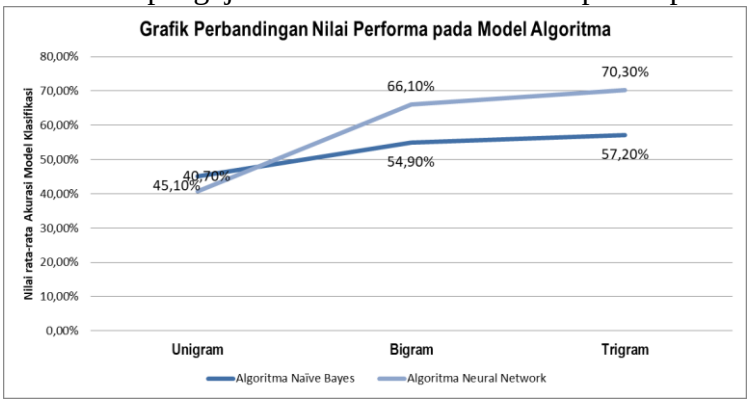
Selanjutnya *stemming* yaitu setiap kata terkait akan diubah menjadi kata dasar atau sufiks. Seperti Kode program 4.4.

PENUTUP

Kesimpulan

Berdasarkan penelitian yang berjudul klasifikasi struktur kalimat bahasa Jaksel (campur kode bahasa Indonesia – Inggris) penulis dapat mengambil beberapa kesimpulan diantaranya sebagai berikut.

1. Penelitian yang dilakukan ini membangun model klasifikasi menggunakan dua algoritma yaitu algoritma *naive bayes* dan *neural network.* Selanjutnya, untuk mengetahui perbandingan kinerja model terbaik pada masing-masing model klasifikasi menggunakan data teks bersumber dari data cuitan Twitter terhadap penggunaan bahasa campur kode yaitu bahasa Jaksel. Hasil pengujian model klasifikasi ditampilkan pada Gambar 5.1.



Gambar 5. 1 Perbandingan Nilai Performa Klasifikasi Algoritma

Berdasarkan Gambar 5.1 diperoleh nilai rata-rata kinerja tertinggi terdapat pada algoritma *neural network* model *trigram* sebesar **70,30%**. Selain itu, algoritma *naive bayes* juga memperoleh nilai rata-rata kinerja tertinggi pada model *trigram* sebesar **57,20%**.

2. Pemilihan parameter pada penelitian ini dilakukan oleh peneliti melalui observasi penggunaan 158 kata kunci yang sering digunakan dalam bahasa Jaksel sehingga memperoleh data penelitian berjumlah 5099 data cuitan.
3. Pengujian model dilakukan menggunakan 3 skenario pada algoritma *naive bayes* dan *neural network*. Pada skenario 1 menggunakan *unigram*, skenario 2 menggunakan *bigram*, dan skenario 3 menggunakan *trigram*. Adapun parameter pengujian yang digunakan merupakan parameter hasil dari *grid search*. Pada algoritma *naive bayes range* parameter *alpha* adalah dari 0,1 hingga 10. Selanjutnya, untuk algoritma *neural network range* parameter *alpha* adalah 0,0001 hingga 0,001, kriteria aktivasi ReLu, Tanh, *hidden layers size* terdiri dari dua lapisan masing-masing berukuran 10, *solver* optimasi yang digunakan SGD, Adam, *learning rate init* dimulai dari 0,01 hingga 0,1, serta *learning rate* untuk menentukan kecepatan pembelajaran model yaitu *constant, adaptive*.
4. Berdasarkan hasil skenario pengujian *k-fold cross validation* dan *confusion matrix* menggunakan algoritma *naive bayes* memperoleh nilai *accuracy* tertinggi di *fold* 5 pada model *trigram* dengan nilai sebesar **61,00%**. Sedangkan hasil skenario pengujian *k-fold cross validation* dan *confusion matrix* menggunakan algoritma *neural network* memperoleh nilai *accuracy* tertinggi di *fold* 4,5,6 pada model *trigram* dengan nilai sebesar **72,00%**.
5. Dari keseluruhan skenario pengujian yang sudah dilakukan pada penelitian ini membuktikan bahwa model klasifikasi algoritma *neural network* memiliki kinerja yang lebih unggul dibandingkan algoritma *naive bayes* dalam mengklasifikasikan struktur kalimat campur kode bahasa Inggris – bahasa Indonesia atau bahasa Jaksel.

Saran

Berdasarkan dari penelitian yang telah dilakukan untuk membangun model klasifikasi struktur kalimat bahasa Jaksel (campur kode bahasa Indonesia – Inggris) ada beberapa saran yang dapat dijadikan pertimbangan untuk penelitian selanjutnya sebagai berikut.

1. Menambahkan data korpus dan kosakata baru yang lebih luas dan beragam sehingga model dapat mengeneralisasi data dengan lebih baik.
2. Memperluas nilai rentang parameter *alpha* pada algoritma *naive bayes* yaitu 100. Nilai rentang parameter *alpha* pada *neural network* hingga 1.0, *hidden layers size* hingga 5 lapisan, *learning init* hingga 1.0 serta penambahan *solver* LBFGS dan fungsi aktivasi sigmoid pada model untuk mendapatkan nilai kombinasi parameter optimal yang lebih luas.
3. Melakukan analisis lanjutan terhadap nilai kinerja setiap model klasifikasi algoritma.
4. Melakukan *update* data pada rentang waktu *scraping* data agar menghasilkan data yang relevan dan terkini.
5. Adanya proses *language detection* untuk menentukan kata bahasa Inggris dan bahasa Indonesia

DAFTAR PUSTAKA

Abiodun, A. 2023. Comparing the performance of three text classification algorithms: Convolutional Neural Networks, Multinomial Naive Bayes, and SVM. Journal of Text Classification, 1(1), 1-10. https://doi.org/10.1234/jtc.2023.0001.

Amri, Yusni Khairul. 2019. *Alih Kode Dan Campur Kode Pada Media Sosial.* Prosiding Seminar Nasional Pendidikan Bahasa dan Sastra Indonesia II, 2. pp. 149-154. ISSN 978-623-92504-4-7.

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

Chaer, Abdul dan Leonie Agustina. 2010. *Sosiolinguistik Perkenalan Awal*. Jakarta: Rineka Cipta.

Chathuranga, S., dan Ranathunga, S. 2021. *Classification of Code-Mixed Text Using Capsule Networks*. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 256-263. Held Online. INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.30

D. L, Olson dan D. Delen. 2008. *Advanced Data Mining Techniques*. Verlag Berlin Heidelberg: Springer

D. R. Anamisa, 2021. *Rancang Bangun Metode OTSU untuk Deteksi Hemoglobin*. Jurnal Ilmu Komputer dan Sains Terapan, p. 5.

Dhande, L. L., dan Patnaik, P. G. K. (2014). Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(4), 313–320.*

Dobrow, T. 2016. Text-Dependent Speaker Verification via Neural Networks (Thesis). Middlebury College.

Ernawati, S. 2016. Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen Review Perusahaan Penjualan Online Menggunakan Naïve Bayes. Jurnal Evolusi, 4(1), 45–54.

Girnanfa, F. A. dan 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.

Gunawan, Agustin Wydia. 2008. *Pedoman Penyajian Karya Ilmiah*. Bogor:IPB Press.

Han, J., dan Kamber, M. 2006. *Data Mining Consepts and Technoloques Second Edition*. San Francisco:Diane Cerra.

Hulu, Sitefanus. 2020. Analisis Kinerja Metode Cross Validation dan K-Nearest Neighbor Dalam Klasifikasi Data [Online]. Available: https://repositori.usu.ac.id/bitstream/handle/123456789/29827/177038034.pdf?sequence=1&isAllowed=y

Indriani, A. 2014. Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. Seminar Nasional Aplikasi Teknologi Informasi (SNATI). ISSN: 1907-5022, pp. G5-G10.

Indrayuni, E. 2019. Klasifikasi *Text Mining Review* Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma *Naive Bayes*. JURNAL KHATULISTIWA INFORMATIKA, VOL. VII, NO. 1 JUNI 2019 p-ISSN: 2339-1928 & e-ISSN: 2579-633X.

Iskandar, D. dan Suprapto, Y. K. 2015. Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C 4.5 Dan Naïve Bayes. Jurnal Ilmiah NERO. Vol. 2, No.1, pp 37-43.

Manurung, R. 2016. Tutorial: Pengenalan terhadap POS Tagging dan Probabilistic Parsing, Workshop Nasional INACL, Jakarta.

Muhammad Sholeh Hudin, M. A. 2018. Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2, 11.

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

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

Noto, A. P., dan Saputro, D. R. S. (2022). *Classification data mining with Laplacian Smoothing on Naïve Bayes method*. International Conference of Mathematics and Mathematics Education (I-CMME) 2021 AIP Conf. Proc. 2566, 030004-1–030004-5; https://doi.org/10.1063/5.0116519.

Natasuwarna, A.P. 2019. Tantangan Menghadapi Era Revolusi 4.0 – Big Data dan Data Mining. pp. 23.

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

Nur, A., U. E. dan 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]

P. Antinasari, R. S. Perdana dan M. A. Fauzi, 2016. Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer , vol. 1, no. 12, pp. 1733-1741.

Pranckevicius, T., dan Marcinkevicius, V. 2017. Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. Baltic Journal of Modern Computing, 5(2), 221-232. http://dx.doi.org/10.22364/bjmc.2017.5.2.05

Prasanna, S., dan Rao, S. I. 2017. *An Overview of Wireless Sensor Networks , their Applications and Technical Challenges*. International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, 2(2), 538–540.

Puspitasari, D., Saputra, P. Y., dan Prakoso, I. A. 2018. *Analisa Sistem Klasifikasi Judul Skripsi Menggunakan Metode Naïve Bayes classifie*R. Jurnal Informatika Polinema, 5(1), 43-45.

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

Putu, Manik Prihatini. 2016. I*mplementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia*. Jurnal Jurusan Teknik Elektro, Vol. 6, No. 3.

Rahman, A., Wiranto, dan Doewes, A. 2017. *Online News Classification Using Multinomial Naive Bayes.* ITSMART: Jurnal Ilmiah Teknologi dan Informasi, 32-38.

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.

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

Sabok, M., dan Brown, P. F. 2016. *A Survey of Part-of-Speech Tagging Techniques*. Computational Linguistics, 42(3), 423-482.

Sajid, F., Putra Negara, A. B., dan Muhardi, H. 2021. *Perbandingan Algoritma Klasifikasi terhadap Emosi Tweet Berbahasa Indonesia*. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 7(2), pp. 242-249

Santosa, B. 2007. Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis, 1st ed. Yogyakarta: Graha Ilmu.

Santosa, B., Conway, T., & Trafalis, T. 2007. *A Hybrid Knowledge BasedClustering Multi-Class SVM Approach for Genes Expression Analysis*. Springer Optimization and Its Applications, 7, 231–274. https://doi.org/10.1007/978-0-387-69319-4_15/COVER/.

Scikit-learn. 2022. *Model Lasso Parameter*. Diakses dari https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Scikit-learn. 2022. *Multinomial Naive Bayes Documentation*. Diakses dari https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Scikit-learn. 2022. *Neural Network Documentation*. Diakses dari https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Septiani, A. 2021. *Klasifikasi Suara Paru Normal dan Abnormal dengan* Menggunakan Discrete Wavelet Transform dan Support Vector Machine. 8(1), 731–742.

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

Subroto, G., Sulistiyowati, N. &Ridha, A, A., 2022. *Klasifikasi Jenis Kekerasan Pada Perempuan dan Anak.* Journal of Information Technology and Computer Science (INTECOMS), vol. 5 No. 1, pp. 104-113.

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.

Turban, E,. 2005. *Decision Support Systems and Intelligent Systems* Edisi Bahasa Indonesia Jilid 1. Andi: Yogyakarta

West, R., dan Turner, L. H. 2007. *Introducing Communication Theory: Analysis and Application* (edisi ke-3). McGraw Hill. Jakarta:Salemba Humanika

Widjaya, A., Hiryanto, L., dan Handhayani, T. 2017. *Prediksi Masa Studi Mahasiswa Dengan Voting Feature Interval 5 Pada Aplikasi Konsultasi Akademik Online*. Journal of Computer Science and Information Systems. Vol.1, pp 25-33.

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

Zhang, J. 2020. *Sentiment Analysis of Movie Reviews in Chinese (Master's Thesis)*. Uppsala University, Department of Linguistics and Philology, Master Programme in Language Technology.

Zhang, X., Zhao, J., & LeCun, Y. 2015. *Character-level convolutional networks for text classification*. In Advances in Neural Information Processing Systems (pp. 649-657). Retrieved from https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification