JIKO (Jurnal Informatika dan Komputer)
Vol. 4, No. 2, April 2021, hlm. 105-111
DOI: 10.33387/jiko

Akreditasi KEMENRISTEKDIKTI, No.36/E/KPT/2019
p-ISSN: 2614-8897
e-ISSN: 2656-1948

PERBANDINGAN KLASIFIKASI BAHASA MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (NBC) DAN SUPPORT VECTOR MACHINE (SVM)

Deglorians Tuhenay[1], Evangs Mailoa[2]

[1,2]Fakultas Teknologi Informasi, Program Studi Teknik Informatika, Universitas Kristen Satya Wacana
Email: [1]lorituhenay12@gmail.com, [2]evangs.mailoa@uksw.edu

(Naskah masuk: 27 Maret 2021, diterima untuk diterbitkan: 30 April 2021)

Abstrak

Beragam macam suku dan budaya yang ada di Indonesia menjadikan Indonesia sebagai salah satu negara dengan bahasa daerah terbanyak di dunia, karena begitu banyak bahasa daerah yang ada di Indonesia maka seringkali menyebabkan kesulitan dalam berkomunikasi terutama pada proses penyampaian informasi atau percakapan berupa teks. Penelitian ini bertujuan untuk mengidentifikasi bahasa dalam bentuk teks. Melalui penulisan ini akan dipaparkan hasil identifikasi bahasa Indonesia, Ambon dan Jawa dengan menggunakan teknologi komputerisasi *Machine Learning* dan memakai metode *Naïve Bayes Classifier* dan *Support Vector Machine* untuk menghitung nilai *accuracy* pada kedua metode tersebut agar dapat mengidentifikasi bahasa sesuai dengan teks yang dimasukkan. Penelitian ini juga melakukan perbandingan dari kedua metode tersebut agar dapat mengetahui metode apa yang lebih efektif dipakai untuk mengidentifikasi bahasa. Hasil yang didapatkan dengan memakai metode NBC dan SVM adalah keduanya bagus dalam melakukan identifikasi bahasa karena memperoleh nilai *accuracy* di atas 0,90 hanya saja melalui perhitungan confusion matrix, metode SVM lebih efektif dengan nilai *accuracy* 0,9634 dibandingkan dengan nilai NBC 0,9378.

Kata kunci: *Bahasa, NBC, SVM, Identifikasi, Perbandingan*

COMPARISON OF LANGUAGE CLASSIFICATION USING NAIVE BAYES CLASSIFIER (NBC) AND SUPPORT VECTOR MACHINE (SVM) METHOD

Abstract

The variety of ethnicities and cultures in Indonesia makes Indonesia be the one as the countries with the most regional languages in the world because there are so many regional languages in Indonesia that often cause difficulties in communicating, especially in the process of delivering information or conversations in the form of text. This study aims to identify the language in the form of text. Through this writing, the identification results of Indonesian, Ambon, and Javanese languages will be presented using computerized machine learning technology and using the Naïve Bayes Classifier and Support Vector Machine methods to calculate the accuracy value in both methods to identify the language according to the entered text. This study also conducted a comparison of the two methods to find out which method was more effective in identifying language. The results obtained using the NBC and SVM methods are both good at identifying languages because getting an accuracy value above 0.90, it's just that through the calculation of confusion matrix, the SVM method is more effective with an accuracy value of 0.9634 compared to the NBC value of 0.9378.

Keywords: *Language, NBC, SVM, Identification, Comparison*

1. PENDAHULUAN

Bahasa mempunyai sifat yang arbitrer atau sewenang-wenang yang digunakan manusia pada zaman dahulu hingga sekarang dan mempunyai peran penting bagi kehidupan umat manusia, karena sifat alami manusia adalah sebagai makhluk sosial maka dengan adanya bahasa, manusia dapat berinteraksi dan berkomunikasi satu dengan lainnya[1]. Bahasa juga penting bagi kehidupan manusia karena manusia dapat dengan mudah menyampaikan informasi yang ingin disampaikan baik secara lisan maupun tulisan bahkan melalui simbol-simbol atau kode tertentu[2], bahasa juga bisa digunakan dalam penyampaian informasi secara satu arah maupun dua arah agar dapat berkomunikasi satu dengan lainnya.[3]

Beragam macam suku dan budaya di Indonesia menjadikan negara Indonesia menjadi salah satu negara dengan bahasa daerah terbanyak di dunia yaitu lebih dari 700 bahasa,[4] karena begitu banyak bahasa yang terdapat di Indonesia, maka seringkali ketika ada orang yang berbeda daerah di Indonesia melakukan percakapan atau menyampaikan

informasi secara lisan maupun tulisan, seringkali terjadi ketidakpahaman pada makna dan pemakaian bahasa yang tidak dimengerti dalam percakapan mereka, terlebih lagi jika percakapan atau penyampaian informasi secara tekstual atau tulisan melalui media-media tertentu contohnya media sosial, maka jika kurangnya pengetahuan tentang bahasa dari tiap daerah maka bisa saja menimbulkan berbagai masalah sosial akibat ketidakpahaman tentang bahasa daerah dari tiap daerah yang ada di Indonesia, maka dari itu dengan adanya kemajuan teknologi di era modern ini sehingga dibutuhkannya sebuah sistem yang bisa mengidentifikasi bahasa pada tiap daerah yang di Indonesia. Penelitian ini memanfaatkan teknologi komputerisasi *Artificial Inteligence* (AI) atau juga biasa disebut dengan *Machine Learning* dalam cabangnya yaitu *Natural Language Processing* (NLP) yang memungkinkan sistem mengetahui bahasa manusia dengan memakai pendekatan algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) untuk melakukan perhitungan secara komputerisasi serta melakukan perbandingan, agar memudahkan orang dalam mengidentifikasi apakah bahasa daerah yang dipakai adalah bahasa daerah dari daerah mana saja yang ada di Indonesia,

Bahasa yang menjadi dataset dalam penelitian ini menggunakan bahasa Indonesia sebagai bahasa Nasional dan bahasa daerah menggunakan bahasa Jawa dan Ambon sebagai sampel, maka perlu adanya sistem untuk mengidentifikasi bahasa Indonesia, Ambon, dan Jawa agar dapat diterjemahkan menurut bahasa daerah masing-masing dan dapat dipakai sebagai acuan untuk bahasa daerah lain yang ada di Indonesia[5].

Dalam dunia komputerisasi, *Natural Language Processing* (NLP) adalah cabang dari *Artificial Inteligence* (AI) yang mempelajari tentang pengelolaan bahasa alami atau bahasa yang biasa digunakan manusia untuk berkomunikasi satu dengan lainnya[6]. Proses penyampaian informasi seringkali mengalami kesalahpahaman artikulasi dan makna dari bahasa, apalagi penyampaian informasi disampaikan dalam bentuk tulisan dan menggunakan bahasa daerah masing-masing maka akan sangat berpotensi menimbulkan kesalahpahaman dalam proses pemaknaan kalimat, maka dari itu perlu adanya sistem untuk menerjemahkan bahasa daerah menurut pengelompokan bahasa daerah dari tiap-tiap daerah di Indonesia. Secara tidak langsung bahasa daerah sudah diperkenalkan melalui sistem ini dan jika dipakai untuk mengidentifikasi bahasa-bahasa daerah lainnya yang ada di Indonesia maka kemungkinan dapat melestarikan bahasa daerah lebih banyak lagi.

Proses Pengelompokan bahasa daerah menggunakan salah satu algoritma dari *Natural Language Processing* (NLP) yaitu *Classification Naïve Bayes* bertujuan untuk mencari tau nilai kemungkinan bahasa yang diterjemahkan sesuai dengan klasifikasi bahasa berdasarkan data yang dimasukkan sebelumnya[7].

Pentingnya kesadaran untuk melestarikan bahasa di zaman modern ini membuat Farel Fathurrahman, Mayanda Mega Santoni, dan Anita Muliawati menulis tentang "Penerapan *Artificial Neural Network* untuk Klasifikasi Citra Teks dalam Penerjemahan Bahasa Daerah" bertujuan agar bahasa daerah dapat dilestarikan melalui sistem penerjemah gambar yang berisikan teks bahasa Indonesia menjadi teks bahasa daerah menggunakan metode *Artificial Neural Network* (ANN) yaitu metode pengklasifikasian pada citra atau gambar[8]

Berkembangnya dunia teknologi membuat banyak orang seringkali dalam lingkup *social media* sering melontarkan opini pada suatu postingan tertentu dan terkadang opini tersebut ada yang aktual dan ada yang tidak maka dari itu penelitian yang di lakukan oleh Nico Munasatya dan Sendi Novianto [9] yaitu membahas tentang *Analysis Sentiment* dalam opini publik terhadap Presiden Jokowi yang juga menggunakan *Natural Language Processing* (NLP) untuk proses *preprocessing* dalam mengubah kumpulan struktur *text* menjadi token, karena dengan adanya NLP maka dapat membantu mesin mengenal bahasa alami manusia dengan mengolah sejumlah besar data agar mesin memiliki kemampuan untuk mengerti ucapan dan tulisan dalam bahasa tertentu.

2. METODE PENELITIAN

2.1. *Naïve Bayes Classifier*

Naïve Bayes Classifier (NBC) digunakan untuk menentukan nilai probabiblitas atau kemungkinan dalam memprediksi peluang berdasarkan data pada pengalaman sebelumnya atau memungkinkan untuk membuat pengelompokan pada suatu sistem. Pada penelitian ini, metode NBC digunakan untuk mengelompokan bahasa berdasarkan *label* yang akan dihitung sehingga menghasilkan nilai bobot pada tiap *label*, dan akan mengidentifikasi bahasa yang dimasukkan sesuai dengan *label*. Salah satu contoh *Naïve Bayes Classifier* (NBC) yang dipakai pada penelitian [10]yaitu mengoptimalkan algoritma NBC sebagai pembanding untuk mengetahui kinerja optimasi PSO pada klasifikasi teks E-Government.

Secara umum rumus Naïve Bayes pada dasarnya adalah sebagai berikut[11]:

$$P(H \mid X) = \frac{P(H \mid X)\, P(H)}{P(X)}$$

Dimana :
X = Data *class* yang belum diketahui
H = Data X merupakan hipotesa *class* yang spesifik
P(H | X) = Probabilitas hipotesa H berdasarkan kondisi X
P(H) = Probabilitas hipotesa H (*prior*)
P(X | H) = Probabilitas X berdasarkan kondisi
P (X) = Probabilitas dari X

Dalam proses ini metode NBC perlu mengambil petunjuk agar dapat menentukan kelas yang cocok bagi sampel data yang diuji coba maka dari itu rumus umum NBC diubah menjadi sebagai berikut :

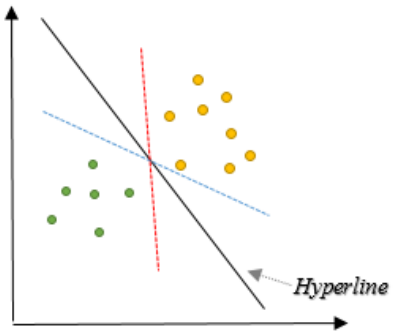
$$P(H \mid X_{1\ldots}X_n) = \frac{P(H)\ P\ (X1\ldots Xn|H)}{P(X1\ldots Xn)}$$

Dimana variabel H merepresentasikan *class* dan $X_1...X_n$ merepresentasikan class yang belum diketahui dan nantinya akan dibutuhkan untuk proses klasifikasi.

2.2. *Support Vector Machine*

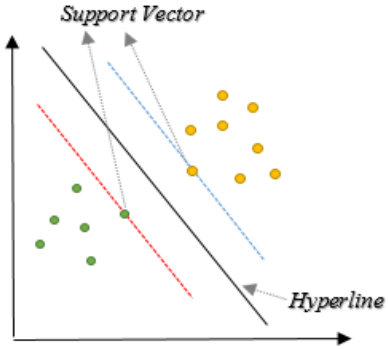
Support Vector Machine (SVM) digunakan untuk membagi data yang sudah diketahui atau sudah dibuat sebelumnya berdasarkan klasifikasinya agar dapat menguji keakuratan data pada sebuah sistem. SVM digunakan untuk pembagian data non linier dan akan membaginya dalam *hyperline* sebagai pemisah dari titik *vector*[12]. Dalam penelitian sebelumnya yang membahas tentang perbandingan pemilihan fitur dan pengklasifikasian klasifikasi teks pendek menunjukan hasil bahwa regresi *logistic* dan *Support Vector Machine* mencapai akurasi tertinggi dan paling stabil[13] namun pada penelitian ini masih menggunakan *Support Vector machine* (SVM) dan *Naïve Bayes Classifier* (NBC) untuk menjadi perbandingan keakuratan sistem.

Cara kerja *Support Vector Machine* (SVM) secara umum dapat dilihat pada Gambar 1 dan 2.



Gambar 1. Ilustrasi Pembagian SVM 1

Pada Gambar 1. Terdapat 3 garis pembatas, yang di tengah adalah garis *hyperline* dan 2 garis lainnya adalah pembagi antara titik *vector* yang ditandai dengan titik warna kuning dan hijau. Metode SVM melakukan pembagian pada kelompok titik *vector* berwarna kuning dan hijau dengan cara mencari titik *vector* terdekat dari kelompoknya yang mendekati garis *hyperline* agar dapat dicari tingkat keakuratannya.



Gambar 2. Ilustrasi Pembagian SVM 2

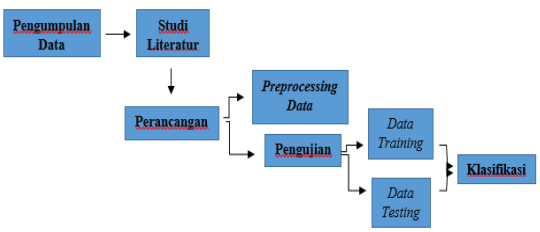
Pada Gambar 2. Titik *vector* yang terdekat dengan *hyperline* harus searah lurus dengan *hyperline*, jika dilihat pada gambar 2 titik *vector* ini sudah searah lurus dengan *hyperline*, 2 titik *vector* ini yang dinamakan dengan *Support Vector*.

2.3. *Performance Evaluation Measure*

Dalam pemrosesan data dari *data set* terdapat suatu tahapan yang dilakukan untuk mengukur evaluasi performa pada suatu sistem yang disebut sebagai *Perfomance Evaluation Measure* (PEM) perhitungan yang dipakai dalam sistem ini yaitu mencari informasi perbandingan hasil identifikasi dengan *confusion matrix* dengan menghitung nilai *accuracy*, *precision*, *recall* dan *F1-score*. *Accuracy* adalah hasil dari data yang terbaca oleh sistem dari keseluruhan informasi, *precision* adalah tingkat ketepatan antara permintaan pengguna dengan jawaban dari sistem, *recall* adalah perbandingan ketepatan membaca sistem dengan informasi yang pernah di *running* sebelumnya dan *F1-score* adalah *harmonic mean* atau nilai rata-rata dari *precision* dan *recall*. Pada perancangan sistem ini g*oogle colab* digunakan sebagai *coding environment* yang bahasa pemrogramannya adalah *python*, jika dipakai dalam proses perhitungan *machine learning* maka tidak harus menghitung tingkat *accuracy*, *precision*, *recall* dan *F-1-score* secara manual karena pada *codingan python* sudah bisa langsung menghitung secara otomatis.

2.4. Tahapan Penelitian

Proses identifikasi bahasa menggunakan beberapa tahapan dalam penelitian yaitu teknik pengumpulan data, studi literatur, tahap perancangan penelitian dan identifikasi sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. Alur Tahapan Penelitian

Teknik pengumpulan data yaitu dengan mencari sumber data berupa teks yang di ambil dari internet dan media cetak, adapun data yang sengaja dimasukkan berdasarkan percakapan keseharian dari bahasa Indonesia, Ambon, dan Jawa yang dibuat dalam bentuk teks dengan tujuan untuk memperbanyak kalimat agar dapat memenuhi target minimum data, karena dengan semakin banyak kalimat maka kemampuan menganalisa pada sistem komputerisasi akan semakin baik. Data haruslah berupa teks dan dibuat pada *M.Excel* dengan memasukan minimal 3.000 data pada masing-masing *label* dalam bentuk kalimat dengan menggunakan bahasa Indonesia dan bahasa daerah Ambon dan Jawa untuk menjadi data mentah agar ketika diproses menggunakan metode NBC dan SVM dapat mengetahui nilai probabilitas dan persamaan *hyperline* pada sistem agar dapat membandingkan keakuratan pada kedua metode tersebut.

Studi literatur memakai acuan dari penelitian sebelumnya dengan memanfaatkan metode *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier* (NBC) untuk proses pengklasifikasian maupun pengelompokan data [14]maka sudah semakin mudah bagi penelitian ini dalam membuat perbandingan klasifikasi bahasa dengan memanfaatkan kedua metode tersebut[15].

Tahapan perancangan pada sistem yang dibuat yaitu dengan dilakukannya *Preprocessing* data yaitu membuat data sesuai dengan model yang sudah ditentukan seperti, menghilangkan tanda baca pada *cleantext* dan juga kalimat dibuat lebih dari lima kata agar tidak terjadi kesalahan dalam proses *running* dalam sistem.

Contoh :

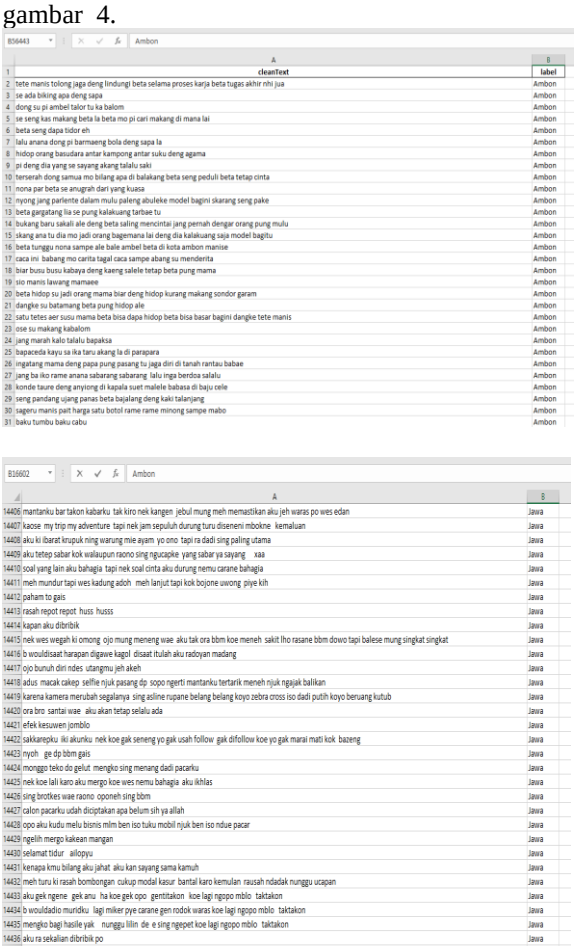
> Kalimat asli : “biar busu-busu kabaya deng kaeng salele, tetap beta pung mama”
> Preprocessing : “biar busu busu kabaya deng kaeng salele tetap beta pung mama”
> Tanda baca (-) dan (,) dihilangkan

Setelah data *cleantext* di buat pada *Ms.Excel* maka selanjutnya melakukan pengelompokan data berdasarkan *label* yaitu bahasa Indonesia, Jawa dan Ambon. Pengujian dilakukan dengan mengelompokan data untuk menentukan dan membedakan data berdasarkan bahasa daerah, dalam proses ini juga dilakukan *data set* dengan proses perhitungannya dilakukan secara komputerisasi yaitu dengan memasukan *codingan* pada sistem. Data dibagi menjadi dua yaitu *data training* 70% dan *data test* 30%. *Data training* dijadikan sebagai basis untuk menguji dan melatih apakah data yang kita pakai sudah bisa berjalan dan dapat melakukan penerjemahan sesuai dengan aturan yang ditentukan atau tidak, dan *data test* yaitu data baru yang digunakan untuk menguji keakuratan dalam sebuah sistem apakah sudah bekerja sesuai dengan standar yang diinginkan atau tidak, setelah itu masuk pada proses identifikasi yaitu dengan menghitung hasil uji *accuracy*, *precision*, *recall* dan *F1-score* pada sistem yang telah dibuat[5] sehingga dapat mengetahui hasil dari metode NBC dan SVM agar dapat dilakukan perbandingan pada kedua metode tersebut.

3. HASIL DAN PEMBAHASAN

3.1. Data Set

Data yang dipakai untuk pembuatan sistem ini menggunakan bahasa Indonesia, Ambon dan Jawa yang dimasukkan kedalam *Ms.Excel* yang dibagi menjadi *cleantext* sebagai isi dari kalimat dari tiap bahasa dan *label* sebagai identitas dari bahasa. Data yang dimasukkan di *Ms.Excel* dapat dilihat pada gambar 4.



	A	B
1	cleanText	label
2	tete manis tolong jaga deng lindungi beta selama proses karja beta tugas akhir nhi jua	Ambon
3	se ada biking apa deng sapa	Ambon
4	dong su pi ambel talor tu ka balom	Ambon
5	se seng kas makang beta la beta mo pi cari makang di mana lai	Ambon
6	beta seng dapa tidor eh	Ambon
7	lalu anana dong pi barmaeng bola deng sapa la	Ambon
8	hidop orang basudara antar kampong antar suku deng agama	Ambon
9	pi deng dia yang se sayang akang talalu saki	Ambon
10	tersenah dong samua mo bilang apa di balakang beta seng peduli beta tetap cinta	Ambon
11	nona par beta se anugrah dari yang kuasa	Ambon
12	nyong jang parliente dalam mulu paleng abuleke model bagini skarang seng pake	Ambon
13	beta gangatang lia se pung kalakuang tarbae tu	Ambon
14	bukang baru sakali ale deng beta saling mencintai jang pernah dengar orang pung mulu	Ambon
15	skang ana tu dia mo jadi orang bagemana lai deng dia kalakuang saja model bagitu	Ambon
16	beta tunggu nona sampe ale bale ambel beta di kota ambon manise	Ambon
17	caca ini babang mo carita tagal caca sampe abang su menderita	Ambon
18	biar busu busu kabaya deng kaeng salele tetap beta pung mama	Ambon
19	sio manis lawang mamaee	Ambon
20	beta hidop su jadi orang mama biar deng hidop kurang makang sondor garam	Ambon
21	dangke su batamang beta pung hidop ale	Ambon
22	satu tetes aer susu mama beta bisa dapa hidop beta bisa basar bagini dangke tete manis	Ambon
23	ose su makang kabalom	Ambon
24	jang marah kalo talalu bapaksa	Ambon
25	bapaceda kayu sa ika taru akang la di parapara	Ambon
26	ingatang mama deng papa pung pasang tu jaga diri di tanah rantau babae	Ambon
27	jang ba iko rame anana sabarang sabarang lalu inga berdoa salalu	Ambon
28	konde taure deng anylong di kapala suet makele babasa di baju eele	Ambon
29	seng pandang ujang panas beta bajalang deng kaki talanjang	Ambon
30	sageru manis pait harga satu botol rame rame minong sampe mabo	Ambon
31	baku tumbu baku cabu	Ambon

	A	B
14406	mantanku bar takon kabarku tak kiro nek kangen jebul mung meh memastikan aku jeh waras po wes edan	Jawa
14407	kaose my trip my adventure tapi nek jam sepuluh durung turu diseneni mbokne kemaluan	Jawa
14408	aku ki ibarat kirupuk ning warung mie ayam yo ono tapi ra dadi sing paling utama	Jawa
14409	aku tetep sabar kok walaupun raono sing ngucapke yang sabar ya sayang xaa	Jawa
14410	soal yang lain aku bahagia tapi nek soal cinta aku durung nemu carane bahagia	Jawa
14411	meh mundur tapi wes kadung adoh meh lanjut tapi kok bojone uwong piye kih	Jawa
14412	paham to gais	Jawa
14413	rasah repot repot huss husss	Jawa
14414	kapan aku dibribik	Jawa
14415	nek wes wegah ki omong ojo mung meneng wae aku tak ora bbm koe meneh sakit lho rasane bbm dowo tapi balese mung singkat singkat	Jawa
14416	b wouldisast harapan digawe kagol disaat itulah aku radoyan madang	Jawa
14417	ojo bunuh diri ndes utangmu jeh akeh	Jawa
14418	adus macak cakep selfie njuk pasang dp sopo ngerti mantanku tertarik meneh njuk ngajak balikan	Jawa
14419	karena kamera merubah segalanya sing asline rupane belang belang koyo zebra cross iso dadi putih koyo beruang kutub	Jawa
14420	ora bro santai wae aku akan tetap selalu ada	Jawa
14421	efek kesuwen jomblo	Jawa
14422	sakkarepku iki akunku nek koe gak seneng yo gak usah follow gak difollow koe yo gak marai mati kok bazeng	Jawa
14423	nyoh ge dp bbm gais	Jawa
14424	monggo teko do gelut mengko sing menang dadi pacarku	Jawa
14425	nek koe lali karo aku mergo koe wes nemu bahagia aku ikhlas	Jawa
14426	sing brotkes wae raono oponeh sing bbm	Jawa
14427	calon pacarku udah diciptakan apa belum sih ya allah	Jawa
14428	opo aku kudu melu bisnis mlm ben iso tuku mobil njuk ben iso ndue pacar	Jawa
14429	ngelih mergo kakean mangan	Jawa
14430	selamat tidur allopyu	Jawa
14431	kenapa kmu bilang aku jahat aku kan sayang sama kamuh	Jawa
14432	meh turu ki rasah bombongan cukup modal kasur bantal karo kemulan rausah ndadak nunggu ucapan	Jawa
14433	aku gek ngene gek anu ha koe gek opo gentitakon koe lagi ngopo mblo taktakon	Jawa
14434	b wouldadio muridku lagi miker pye carane gen rodok waras koe lagi ngopo mblo taktakon	Jawa
14435	mengko bagi hasile yak nunggu lillin de e sing ngepet koe lagi ngopo mblo taktakon	Jawa
14436	aku ra sekalian dibribik po	Jawa

Gambar 4. Data Set Bahasa Indonesia,Ambon dan Jawa

Proses penginputan data dilakukan secara manual dengan memasukan data ke *Ms.Excel* dengan membaginya berdasarkan *label* dan *cleantext* yaitu *label* adalah pengelompokan bahasa dan *cleantext* adalah isi dari kalimat sesuai bahasa. Data yang dimasukkan dalam penelitian ini tidak memakai tanda baca dan angka. Jumlah data dari keseluruhan data yang diolah adalah ≥ 3.000 bahasa Ambon, ≥ 10.000, bahasa Jawa, dan ≥ 3.500 bahasa Indonesia. Standar minimum data yang harus dimasukkan dalam sistem ini adalah minimal 3.000 data agar sistem dapat membaca peluang kemungkinan dan membagi data sesuai dengan rumus yang sudah ditentukan dalam penelitian ini yaitu *split* data uji sekitar 30%.

3.2. *Naïve Bayes Classifier*

Codingan Naïve Bayes Classifier untuk dapat mengetahui hasil uji tingkat keakuratan sistem pada *confusion matrix* dapat dilihat sebagai berikut:

```
#Multinomial Naive Bayes
pipeline_mnb = Pipeline([
    ('vect', CountVectorizer()),
    #('vect', CountVectorizer(ngram_range=(2,2))),
    ('tfidf', TfidfTransformer(use_idf=True, smooth_idf=True)),
    ('clf', MultinomialNB(alpha=1))
])
X_train, X_test, y_train, y_test = train_test_split(data['cleanText'], data['label'],  test_size=0.33, random_state = 0)
clf_mnb = pipeline_mnb.fit(X_train, y_train)
pred_mnb = pipeline_mnb.predict(X_test)
```

Dengan Perhitungan *count vector n-gram* 2,2 dan *split* data uji 0,33 atau sekitar 30% data.

3.3. *Naïve Bayes Classifier result*

Hasil perhitungan *confusion matrix* dapat dilihat pada tabel 1 dan heatmap *naïve bayes classifier* ditunjukkan pada Gambar 5.

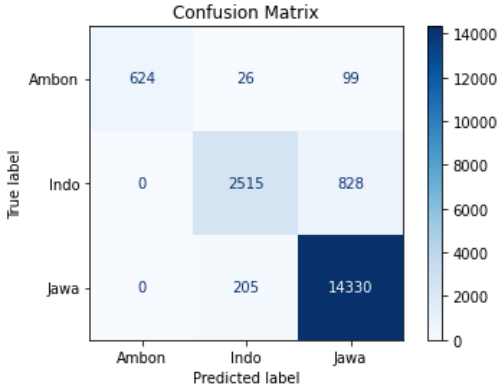
Accuracy: 0.9378

Confusion matrix:

[[624 26 99]
[0 2515 828]
[0 205 14330]]

Tabel 1. Hasil *Confusion Matrix* NBC

	Precision	Recall	F1-score	Support
Ambon	1.00	0.83	0.91	749
Indo	0.92	0.75	0.83	3343
Jawa	0.94	0.99	0.96	14535
Accuracy			0.94	18627
Macro avg	0.95	0.86	0.90	18627
Weighted avg	0.94	0.94	0.94	18627



Gambar 5. *Heatmap Confusion Matrix* NBC

Berdasarkan Tabel 1 dapat dilihat bahwa tingkat keseluruhan perhitungan *data test* dengan *confusion matrix* yang mencakup nilai *precision*, *recall* dan *F1-score* mendapatkan nilai *accuracy* 0,9378.

3.4. *Identification Naïve Bayes Classifier*

Penginputan kalimat dari tiap bahasa pada sistem akan diidentifikasi sesuai dengan bahasa daerah yang di input. Gambar 5 dapat dilihat bahwa sistem membaca dengan membedakan bahasa dari bahasa Indonesia, Ambon dan Jawa, yaitu kalimat "aku telah mendapatkannya", "maju deng kaeng berang tu saja", "sapa pung ana par se" diidentifikasikan sebagai bahasa "Jawa", "Ambon", "Ambon".

Gambar 6. *Prediction* Bahasa

3.5. *Support Vector Machine*

Codingan Support Vector Machine untuk menghitung nilai dari *confusion matrix* dapat dilihat sebagai berikut

```
#SVM
pipeline_svc = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer(use_idf=True, smooth_idf=True)),
    ('clf', LinearSVC())
])
X_train, X_test, y_train, y_test = train_test_split(data['cleanText'], data['label'],  test_size=0.33, random_state = 0)
clf_svc = pipeline_svc.fit(X_train, y_train)
pred_svc = pipeline_svc.predict(X_test)
```

Dengan memakai split data uji yaitu 0,33 atau sekitar 30% data.

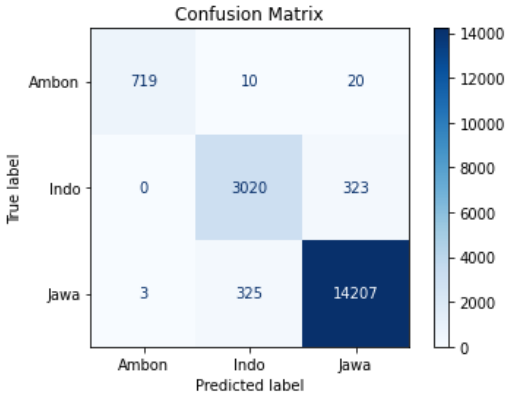
3.6. *Support Vector Machine Result*

Hasil perhitungan confusion matrix dapat dilihat pada tabel 2, sedangkan heatmap confusion matrix ditunjukkan pada Gambar 7.

Accuracy: 0.9634
Confusion matrix:
[[719 10 20]
[0 3020 323]
[3 325 14207]]

Tabel 2. Hasil *Confusion Matrix* SVM

	Precision	Recall	F1-score	Support
Ambon	1.00	0.96	0.98	749
Indo	0.90	0.90	0.90	3343
Jawa	0.98	0.98	0.98	14535
Accuracy			0.96	18627
Macro avg	0.96	0.95	0.95	18627
Weighted avg	0.96	0.96	0.96	18627

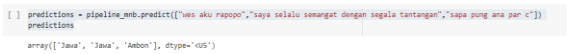


Gambar 7. *Heatmap Confusion Matrix* SVM

Perhitungan nilai keakuratan *data test* secara keseluruhan pada *confusion matrix* mendapat nilai *accuracy* sebesar 0,9634.

3.7. *Identification Support Vector Machine*

Kalimat dimasukkan kedalam sistem agar bisa diidentifikasi bahwa kalimat tersebut dapat diterjemahkan sesuai dengan bahasa yang sudah ditentukan atau tidak, dapat dilihat pada gambar 7, kalimat yang dimasukkan adalah kalimat "wes aku rapopo", "saya selalu semangat dengan segala tantangan", "sapa pung ana par c" diidentifikasikan sebagai bahasa "Jawa", "Jawa", "Ambon".



Gambar 8. *Prediction* Bahasa

3.8. Perbandingan nilai akurasi

Berdasarkan hasil analisis perbandingan nilai keakuratan sistem yang dihitung dari *confusion matrix* maka dapat dilihat hasil dari metode *Naïve Bayes Classifier* dan *Support Vector Machine* ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Nilai Akurasi

Metode	Nilai Akurasi
Naïve Bayes Classifier	0,9378
Support Vector Machine	0,9634

4. KESIMPULAN

Berdasarkan hasil pengujian dan perbandingan yang dilakukan dalam penelitian ini, kedua metode yang dipakai yaitu *Naïve Bayes Classifier* dan *Support Vector Machine* sama-sama bagus dalam melakukan penerjemahan bahasa karena memiliki tingkat keakuratan di atas 0,9 atau di atas 90%. Akan tetapi dalam proses perhitungan menggunakan *confusion matrix* pada keseluruhan nilai keakuratan yang mencakup nilai *precision, recall* dan *F1-score* maka dapat dilihat bahwa nilai *accuracy* pada metode *Support Vector Machine* lebih tinggi dibandingkan dengan nilai dari *Naïve Bayes Classifier* dengan nilai *accuracy* SVM adalah 0,9634 atau 96,34%. Maka dapat disimpulkan bahwa metode *Support Vector Machine* lebih efektif dibandingkan dengan metode *Naïve Bayes Classifier* dalam hal mengidentifikasi bahasa.

5. DAFTAR PUSTAKA

[1] Rina Devianty. 2017. "Bahasa Sebagai Cermin Kebudayaan," *J. Tarb.*, vol. 24, no. 2, pp. 226–245.

[2] L. Wicaksono. 2016. "Bahasa Dalam Komunikasi Pembelajaran," *J. Pembelajaran Prospektif*, vol. 1, no. 2, pp. 9–19.

[3] R. P. Suminar. 2016. "Pengaruh Bahasa Gaul terhadap Penggunaan Bahasa Indonesia Mahasiswa Unswagati," *Logika*, vol. 18, no. 3,pp.114–119[Online].Available: www.jurnal.unswagati.ac.id.

[4] T. Berlianty. 2018. "Penguatan Eksistensi Bahasa Tana dalam Upaya Perlindungan Hukum Bahasa Daerah sebagai Warisan Budaya Bangsa," *Kertha Patrika*, vol. 40, no. 2,p. 99, DOI: 10.24843/kp.2018.v40.i02.p04.

[5] A. A. Budiman. 2018. "Pendeteksi Bahasa Daerah Pada Twitter Dengan Machine Learning," p. 11523262.

[6] T. Lalwani, S. Bhalotia, A. Pal, S. Bisen, and V. Rathod. 2018. "Implementation of a Chat Bot

System using AI and NLP," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 6, no. 3, pp. 26–30, DOI: 10.21276/ijircst.6.3.2.

[7] P. H. Saputro, M. Aristin, and Dy. L. Tyas. 2017. "Berdasarkan Lirik Menggunakan Metode Tf-," *J. Teknoloi Inform. dan Terap.*, vol. 4, no. 1, pp. 45–50.

[8] F. Fathurrahman *et al.* 2020. "PENERAPAN ARTIFICIAL NEURAL NETWORK UNTUK KLASIFIKASI CITRA TEKS DALAM PENERJEMAHAN BAHASA," pp. 585–594.

[9] N. Munasatya and S. Novianto. 2020. "Natural Language Processing untuk Sentimen Analisis Presiden Jokowi Menggunakan Multi Layer Perceptron," *Techno.Com*, vol. 19, no. 3, pp. 237–244, DOI: 10.33633/tc.v19i3.3630.

[10] K. S. Nugroho, I. Istiadi, and F. Marisa. 2020. "Naive Bayes classifier optimization for text classification on e-government using particle swarm optimization," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 1, pp. 21–26, 2020, DOI: 10.14710/jtsiskom.8.1.21-26.

[11] S. Kasus and D. I. Iain. 2020. "DATA MINING UNTUK PENENTUAN MODEL TINGKAT KESUKSESAN KELULUSAN MURID SMA PADA PERGURUAN TINGGI NEGERI : DATA MINING FOR DETERMINATION OF HIGH SCHOOL STUDENT GRADUATION MODEL AT STATE UNIVERSITY ; CASE STUDY IN IAIN BONE," vol. 3, no. 2, pp. 113–118, DOI: 10.33387/jiko.

[12] E. Anindika Sari, M. Thereza Br. Saragih, I. Ali Shariati, S. Sofyan, R. Al Baihaqi, and R. Nooraeni. 2020. "Klasifikasi Kabupaten Tertinggal di Kawasan Timur Indonesia dengan Support Vector Machine," *JIKO (Jurnal Inform. dan Komputer)*, vol. 3, no. 3, pp. 188–195, DOI: 10.33387/jiko.v3i3.2364.

[13] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu. 2017. "Comparisons and Selections of Features and Classifiers for Short Text Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, no. 1, DOI: 10.1088/1757-899X/261/1/012018.

[14] L. Flek. 2020. "Returning the N to NLP: Towards Contextually Personalized Classification Models," pp. 7828–7838, DOI: 10.18653/v1/2020.acl-main.700.

[15] M. L. Laia and Y. Setyawan. 2020. "Perbandingan Hasil Klasifikasi Curah Hujan Menggunakan Metode SVM dan NBC," vol. 05, no. 2, pp. 51–61.