Jurnal Tugas Akhir Fakultas Informatika Februari 2024

Identifikasi Bahasa Daerah di Indonesia Dengan Multinomial Naïve Bayes

Maulana, Muhammad Ardhianda[1], Suryani, Arie Ardiyanti[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]ardhiandam@student.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id

Abstrak

Saat ini telah banyak penelitian yang melakukan identifikasi bahasa, namun untuk identifikasi bahasa daerah di Indonesia belum banyak hasil yang diberikan. Untuk itu penelitian ini akan membahas tentang identifikasi bahasa daerah di Indonesia dengan menggunakan tujuh bahasa yaitu, Bahasa Indonesia, Jawa, Sunda, Minang, Muna, Bugis, dan Madura. Pendekatan yang digunakan untuk mengidentifikasi bahasa di penelitian ini menggunakan metode Multinomial Naïve Bayes. Pendekatan ini dilakukan menghitung probabilitas setiap pola kata maupun deretan kata yang muncul pada kalimat berlabel. Model probabilitas yang dihasilkan kemudian digunakan untuk menentukan kelas kalimat baru yang akan ditentukan bahasanya. Performansi metode identifikasi bahasa ini diukur dengan melakukan dua skenario pengujian. Pengujian pertama dilakukan untuk melihat pengaruh pola n-gram terhadap *F-measure*, sedangkan pengujian kedua dilakukan untuk mengobservasi pengaruh jumlah data latih terhadap *F-measure*. Hasil pengujian menunjukkan bahwa pola unigram dan bigram memberikan hasil akurasi tertinggi sebesar 98,86%. Adapun untuk jumlah data latih sebesar 1500 kalimat pada setiap bahasa menunjukkan akurasi sebesar 98%.

Kata kunci : identifikasi bahasa, bahasa daerah, multinomial naïve bayes

Abstract

Currently, there has been a lot of research that has carried out language identification, but not many results have been provided for identifying regional languages in Indonesia. For this reason, this research will discuss the identification of local languages in Indonesia using seven languages, namely, Indonesian, Javanese, Sundanese, Minang, Muna, Bugis and Madurese. The approach used to identify languages in this research uses the Multinomial Naïve Bayes method. This approach is used to calculate the probability of each word pattern or row of words appearing in a labeled sentence. The resulting probability model is then used to determine the class of new sentences for which the language will be determined. The performance of this language identification method is measured by conducting two test scenarios. The first test was to find out the effect of n-gram pattern on the F-measure, while the second test was to observe the effect of the amount of training data on the F-measure. The test results show that the unigram and bigram patterns provide the highest accuracy results of 98.86%. As for the amount of training data of 1500 sentences for each language, it shows an accuracy of 98%.

Keywords: language identification, local languages, multinomial naïve bayes

1. Pendahuluan

Latar Belakang

Identifikasi Bahasa (*Language Identification*) adalah salah satu *task* dalam Pemrosesan Bahasa Alami (*Natural Language Processing*) yang bertujuan mengenali sebuah bahasa yang ada di dalam sebuah teks atau dokumen.[1] Identifikasi Bahasa dapat diaplikasikan di berbagai area, proses, atau *task* seperti *document resume, text classification, machine translation,* dan lainnya. Salah satu contoh penggunaan dari sistem identifikasi bahasa adalah pengembangan lebih lanjut dari *chatbot* dimana sistem secara otomatis mengidentifikasi bahasa dari lawan bicara sebelum memberi respon, analisis sentimen pada kalimat.

Saat ini penelitian terhadap *Language Identification* mengalami perkembangan yang pesat. Terdapat beberapa hasil penelitian yang membahas identifikasi Bahasa diantaranya *labeling hate speech*,[2] dan mengidentifikasi bahasa isyarat.[3] Namun untuk bahasa dengan korpus yang terbatas atau bahasa yang minor digunakan dalam penelitian, relatif sulit untuk menemukan hasil penelitian yang membahas tentang identifikasi bahasa tersebut, salah satu bahasa yang dimaksud adalah bahasa daerah yang terdapat pada di Indonesia, adapun saat ini penelitian identifikasi bahasa terhadap bahasa daerah yang dilakukan adalah *spoken language identification* dan identifikasi bahasa daerah di Indonesia dengan berbagai metode *machine learning*.

Topik dan Batasannya

Masalah yang diangkat dari Tugas Akhir ini adalah bagaimana mengenali sebuah bahasa dari input kalimat yang berupa teks dengan bahasa daerah tertentu. Adapun bahasa daerah yang ada di Indonesia seperti bahasa Jawa, Sunda, Banjar, Bali, dan lainnya. Untuk Batasan pada Tugas Akhir ini Bahasa yang digunakan dalam eksperimen adalah Bahasa Indonesia, Bahasa Jawa, Bahasa Sunda, Bahasa Minang, Bahasa Muna, Bahasa Madura, dan Bahasa Bugis.

Tujuan

Tujuan dari Identifikasi Bahasa daerah pada Tugas Akhir ini adalah untuk membuat identifikasi Bahasa daerah dengan menggunakan metode multinomial naïve bayes. Aplikasi ini lebih lanjut dapat digunakan untuk *user profiling*, mesin translasi, dan lainnya.

Organisasi Tulisan

Laporan tugas akhir ini dituliskan dengan sistematika sebagai berikut: Pendahuluan untuk menjelaskan latar belakang serta batasan penelitian dan tujuan dari tugas akhir ini. Selanjutnya dijelaskan studi terkait materi yang mendukung tugas akhir yakni studi tentang *language identification* dengan perkembangannya terhadap bahasa daerah, Naïve bayes dan. Setelahnya Sistem Identifikasi bahasa daerah, analisis hasil, serta kesimpulan.

2. Studi Terkait

Language Identification

Language Identification merupakan task yang menentukan bahasa alami dalam sebuah dokumen yang tertulis. *Language Identification* digunakan untuk task seperti normalisasi teks *code-mixed*, *spoken language identification*, *hate speech language detection*. Identifikasi bahasa merupakan task *text categorization* dan beberapa riset sebelumnya mengaplikasikan metode standar *text categorization* terhadap *Language Identification*.[4]

Penggunaan LI dapat digunakan dalam berbagai area seperti *Native Language identification*, mesin translasi, pembuatan *chatbot*, mendeteksi *hate speech*, hoaks dan sebagainya. Dalam perkembangannya saat ini adanya identifikasi bahasa untuk teks *code-mixed*[5] dan rekonstruksi hubungan *native language*.[6]

Language Identification Bahasa Daerah

Perkembangan *Language Identification* untuk bahasa daerah di Indonesia dengan adanya *Spoken Language Identification* untuk empat bahasa daerah yakni bahasa Jawa, Sunda, Minang, dan Bugis. dimana pada penelitian yang dilakukan oleh Wijonarko dan Zahra[7] menggunakan deep learning untuk *automatic speech recognition.* Selain itu adapun identifikasi 10 bahasa daerah dengan menggunakan berbagai metode *machine learning* yang dilakukan oleh Nugraha dan Romadhony yang menggunakan *Support Vector Machine*, *Naïve Bayes Classifier*, *Decision Tree*, *Rocchio Clasification*, *Logistic Regression*, dan *Random Forest*.[8]

Perkembangan pemrosesan bahasa alami pada bahasa daerah juga dilakukan dengan *crowdsourcing* korpus Bahasa Indonesia dan bahasa daerah yakni NusaCrowd[9] dengan bertujuan untuk membantu mengembangkan NLP terhadap bahasa Indonesia dan bahasa daerah. Namun untuk saat ini pengembangan NLP terhadap bahasa daerah lebih didominasi oleh bahasa Jawa dan Sunda.

Naïve Bayes

Naïve Bayes merupakan sebuah metode algoritma yang berdasarkan teorema Bayes dengan menghitung probabilitas asumsi diantara *feature*, dimana kemunculan suatu *feature* tertentu tidak tergantung dengan kemunculan *feature* lainnya. Metode ini kerap digunakan dalam ilmu pembelajaran mesin, statistika, dan bioinformatika.[10] Dalam Naïve Bayes persamaan yang digunakan berdasarkan teorema bayes dengan notasi P(A|B) adalah probabilitas terjadinya A dengan data atau bukti B, P(B|A) probabilitas B dengan data atau bukti A, dan P(A) & P(B) merupakan probabilitas A dan B. (1)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (1)$$

Naïve Bayes kerap juga merupakan sebuah klasifier linear yang sederhana namun efisien. Klasifier ini juga digunakan untuk berbagai task seperti mendeteksi memprediksi cuaca, sistem rekomendasi, dan klasifikasi teks, salah satunya adalah *language identification*. Dalam *language identification* klasifikasi Naïve Bayes adalah metode yang cukup baik karena strukturnya yang kokoh, mudah untuk diimplementasi, cepat, dan akurat.[11]

F-Score Evaluation

Evaluasi F-Score umumnya digunakan untuk mengukur akurasi pengujian pada pemrosesan bahasa alami dengan menghitung *precision* dan *recall* dimana *F-measure* merupakan *harmonic mean* antara *precision* dan *recall*.[12]

Precision (PR) merupakan hasil dari pengujian dimana label yang diprediksi dinyatakan positif, untuk menghitung presisi *PR* diberikan jumlah *true positive* t_p dan *false positive* f_p. Sementara *recall* merupakan hasil dimana model dapat memprediksi label yang merupakan positif nyata, *recall* dapat dikalkulasikan dengan menggunakan *true positive* t_p dan *false negative* f_n. (2)

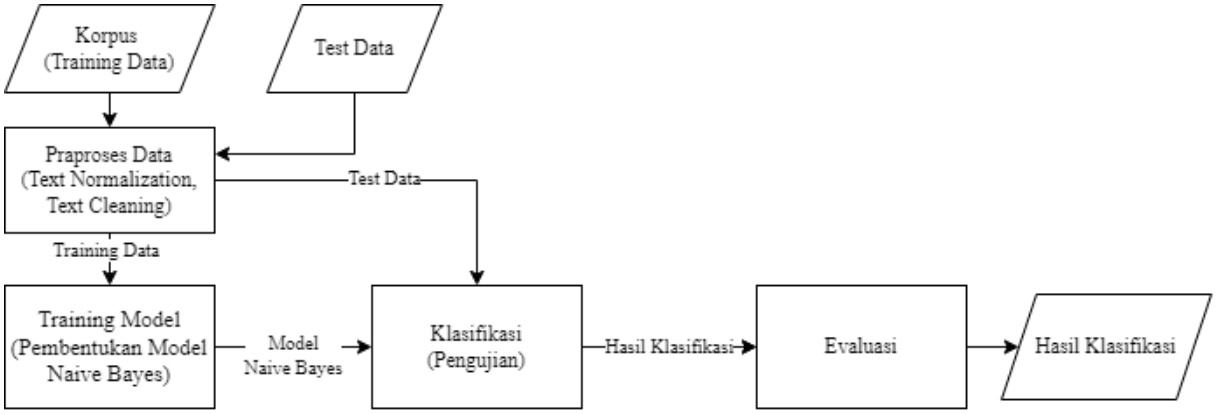
$$PR = \frac{|t_p|}{|t_p| + |f_p|} \quad R = \frac{|t_p|}{|t_p| + |f_n|} \quad (2)$$

Setelah menghitung *precision* dan *recall*, maka dapat dilakukan penghitungan *F-score* yang merupakan rata-rata dari kedua perhitungan *precision* dan *recall*, dan persentase akurasi dapat dihitung dengan menghitung jumlah *true positive* t_p dan *true negative* t_n dengan total semua hasil uji (*true positive + true negative + false positive + false negative)*. (3)

$$F = 2\frac{PR.R}{PR + R} \quad Acc = \frac{(t_p + t_n)}{(t_p + t_n + f_p + f_n)} \quad (3)$$

3. Sistem Identifikasi Bahasa Daerah

Dalam Tugas Akhir ini metode yang digunakan adalah metode klasifier Naïve Bayes. Untuk alur pemodelan yang digagas dalam Tugas Akhir dijelaskan melalui diagram berikut.



Gambar 1. Diagram Sistem Identifikasi Bahasa Daerah

Dataset yang digunakan berupa korpus dengan tujuh bahasa yakni Bahasa Indonesia, Jawa, Sunda, Minang, Muna, Madura, dan Bugis dalam bentuk dataset sebagai data latih (*training data*) dengan data berupa kalimat dan label bahasa sebagai kelas. Untuk korpus Bahasa Indonesia, Sunda, Minang dan Muna korpus yang digunakan didapat dari repositori Dataverse Telkom University[13], Sedangkan korpus Bahasa Jawa, Madura dan Bugis menggunakan korpus Nusantara.[14] Adapun juga korpus tambahan menggunakan korpus NusaX.[15] Korpus yang digunakan berisi kalimat-kalimat dengan Panjang kalimat yang bervariasi untuk setiap Bahasa daerah yang digunakan.

Tabel 1. Data Latih

No.	Kalimat	Bahasa
1	"Panorama sakalilingnyo manjadi tujuan wisata."	Minang
2	"Maksudna nu salah téh rahayat anu milih pamingpin."	Sunda
3	"Di ahérat jaga saban manusa bakal dibingbing ku Alloh, masing-masing bakal di anteurkeun ku malaikat nepika lawang sawarga pikeun ahli surga jeung ahli naraka ogé bakal di anteur ku para malaikat nepi ka lawang naraka."	Sunda
…	…	…

Untuk *preprocessing* data dilakukan normalisasi teks dimana teks yang memiliki aksen seperti kata "nyaéta" menjadi "nyaeta", umumnya teks yang memiliki aksen ini muncul pada Bahasa Jawa dan Sunda. Seterusnya dilanjutkan dengan mengubah semua *string* pada dataset sehingga hanya terdiri dari kalimat saja, dan dilanjutkan dengan mengubah semua huruf menjadi huruf kecil. Setelah *preprocessing data*, data latih digunakan kedalam model yang digunakan untuk deteksi Bahasa.

Klasifier yang digunakan adalah model klasifikasi dengan metode Multinomial Naïve Bayes[16], data latih (*training data*) digunakan untuk klasifier dan dilakukan *training* pada model dengan menggunakan data latih yang berupa korpus dengan contoh pada Tabel 1. Pertama dilakukan mencari peluang c dengan jumlah data pada data latih N_{doc}. (4)

$$P(c) = \frac{N_c}{N_{Doc}} \quad (4)$$

Berdasarkan rumus pada teorema Bayes *training model* melakukan perhitungan dengan menggunakan persamaan Naïve Bayes dengan persamaan berikut. (5)

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)} \quad (5)$$

Setelah pembentukan model bahasa dan pelatihan model, selanjutnya model digunakan untuk klasifikasi atau *testing* dengan menggunakan *test data* dengan berupa korpus kumpulan kalimat dalam empat bahasa daerah yang serup. Dalam proses klasifikasi diberikan sebuah kalimat dalam *test data* yang dicari probabilitas per n-gram kalimat tersebut yang kemudian disamakan dengan korpus *training data* dan bahasa yang diklasifikasi pada data latih.

Tabel 2. Data Test

No.	Kalimat	Bahasa
1	Kanaekan intensitas sarta rupi kagiatan ekstrakurikuler	Jawa
2	Akses medal uga mlebet dhusun cekap ngenteni	Jawa
3	Pan lawang-ati kang puat nenggih panganjate mring wirit jalalah	Sunda
…	…	…

Berikut ini adalah simulasi klasifikasi bahasa untuk data yang kecil. Dalam klasifikasi pada teks pada naïve bayes, asumsi yang digunakan berupa *Bag-of-words* sebagai representasi pada dokumen dengan menghitung jumlah kata yang digunakan pada kalimat, sehingga jika teorema Bayes digunakan pada dokumen dan kelas maka diberikan rumus sebagai berikut. Dimana jika dokumen d direpresentasikan featurenya yaitu kata persamaan yang dapat di $c = argmax\ P(x1, x2, \dots, xn|c)P(c)$.

Namun rumus diatas masih belum bisa digunakan jika tanpa adanya asumsi, dan memperkirakan probabilitas setiap kombinasi fitur yang bisa digunakan karena memerlukan jumlah parameter dan *training set* yang besar, maka dari itu klasifier Naïve Bayes membuat dua asumsi. Pertama adalah *bag-of-words* dimana posisi sebuah kata tidak berpengaruh dan sebuah kata memiliki hasil yang sama pada klasifikasi dan fitur *f1,f2,…,fn* hanyalah encode identitas sebuah kata. Selanjutnya adalah *Conditional Independence / Naïve Bayes Assumption* dimana *feature* probabilitas $P(f_i|c)$ independen diberikan kelas c, dan dapat dirumuskan sebagai berikut. (6)

$$P(f_1, f_2, \dots, f_n|c) = P(f_1|c)\,P(f_2|c) \dots P(f_n|c) \quad (6)$$

Dimana persamaan akhirnya berupa $c_{NB} = argmax\ P(c) \prod_{f \in F} P(f|c)$ dan untuk mengaplikasikan Naïve Bayes untuk teks, posisi kata harus dipertimbangkan sehingga menjadi $c_{NB} = argmax\ P(c) \prod_{i \in Positions} P(w_i|c)$. untuk mencari probabilitas $P(w_i|c)$ semua dokumen dengan sebuah kelas c digabungkan menjadi satu mega-dokumen yang kemudian digunakan untuk menghitung probabilitas $P(w_i|\ c)$ yang menggunakan mega-dokumen tersebut dengan *maximum likelihood*. (7)

$$P(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)} \quad (7)$$

Namun Adapun masalah dengan menggunakan *maximum likelihood*. Jika sebuah kata tidak muncul pada sebuah kelas namun muncul di kelas lainnya, katakan sebuah kata tidak dikategorikan kelas A melainkan kelas B pada dokumen training, maka tidak ada *likelihood* pada sebuah kelas yang menghasilkan probabilitas kelas itu nol

walaupun di kelas lainnya ada. Solusi untuk mengatasi hal tersebut adalah *laplace smoothing* sehingga persamaannya sebagai berikut. (8)

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} count(w, c) + 1} \quad (8)$$

Sehingga untuk melakukan klasifikasi pada kalimat dilakukan penghitungan peluang w_i untuk setiap kata pada kalimat *w* terhadap kelas *c* , dan setiap kata *w* yang unik dimasukkan kedalam variabel kosakata V. (9)

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} count(w_i, c) + |V|} \quad (9)$$

Untuk masalah lainnya dalam melatih klasifier Naïve Bayes seperti kata yang muncul di *test data* tapi tidak ada di data latih maka kata tersebut dianggap *unknown word* yang dapat diatasi dengan menghapusnya dari *test data* atau tidak dimasukkan dalam probabilitas.

Untuk Contoh simulasi klasifikasi bahasa, diberikan training dan testing Naïve Bayes dengan menggunakan *add-one smoothing*, untuk simulasi ini menggunakan 1-gram atau unigram. Dalam contoh ini *task* yang dilakukan adalah klasifikasi bahasa Indonesia dan bahasa daerah, berikut *training data* dan juga *test data*.

Tabel 3. Simulasi klasifikasi bahasa

No.	Kalimat	Bahasa
1	Upaya kuwi memang cukup kanggo sekolah	Jawa
2	Mulane, ing Jawa Timur, contone, laporan ing wangun SPJ kudu digawe saben sasi	Jawa
3	Kanggo sakola anu jumlah siswa na seueur ,peunteun kanaekan penerimaan kasebat cekap signifikan	Sunda
4	Samentara eta penerimaan hiji sakola di Cilegon mudun kira-kira 15%	Sunda
5	liburan sikola lah tibo pulo	Minang
6	Sumawonten ,sababaraha SD di pilemburan kakurangan murid	Sunda
7	Amarga kekurangan pangerten sekolah, pembayaran pajak sing digawe dening sekolah beda-beda	Jawa
8	Kolot siswa terkesan henteu paduli kalawan urusan administrasi pamakean dana	Sunda
9	Biaya itu cukup untuk sekolah	Indonesia
10	Memang jumlah penerimaan siswa sekolah dasar tahun ini meningkat	Indonesia
K_1	Nanging panampa siswa sd ing jawa Tengah isih ngalami masalah	?
K_2	Namun penerimaan siswa SD di Jawa Tengah masih mengalami masalah	?

Untuk mencari klasifikasi dalam data test diatas, dihitung semua probabilitas kelas bahasanya terlebih dahulu.
$P(Jawa) = \frac{3}{10}$ $P(Sunda) = \frac{4}{10}$ $P(Minang) = \frac{1}{10}$ $P(Indonesia) = \frac{2}{10}$
Setelah itu dihitung peluangnya per kata pada data test. Karena kata “nanging”, “panampa”, “Tengah”, “isih”, “ngalami”, “masalah”, “namun”, “masih”, dan “mengalami” tidak ada pada training maka dapat diabaikan karena kata tidak diketahui atau *unknown word*.
Sehingga kalimat K_1 yang semula “Nanging panampa siswa sd ing jawa Tengah isih ngalami masalah” menjadi “siswa sd ing jawa” dan K_2 yang semula “Namun penerimaan siswa SD di Jawa Tengah masih mengalami masalah” berubah menjadi “penerimaan siswa SD di Jawa”. Selanjutnya menghitung seluruh kata unik yang ada di dataset sebagai variabel *V*, total kata unik yang ada dalam simulasi terdapat 72 kata unik, kemudian dihitung probabilitas per kata pada kalimat K untuk setiap kelas yang ada.

Penentuan kelas bahasa data tes K_1 adalah sebagai berikut:
$P("siswa\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}$ $P("siswa\ "|Sunda) = \frac{2+1}{39+72} = \frac{3}{111}$
$P("siswa\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("siswa"|Indonesia) = \frac{1+1}{14+72} = \frac{2}{86}$

$P("sd\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}, P("sd\ "|Sunda) = \frac{1+1}{39+72} = \frac{1}{111}$
$P("sd\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("sd\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

$P("ing\ "|Jawa) = \frac{2+1}{30+72} = \frac{3}{102}$ $P("ing\ "|Sunda) = \frac{0+1}{39+72} = \frac{1}{111}$

$P("ing\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("ing\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

$P("jawa\ "|Jawa) = \frac{1+1}{30+72} = \frac{2}{102}$ $P("jawa\ "|Sunda) = \frac{0+1}{39+72} = \frac{1}{111}$
$P("jawa\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("jawa\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

Setelah itu dilakukan perkalian Probabilitas kelas bahasa terhadap kalimat K_1
$P(Jawa)(K_1|Jawa) = \frac{3}{10} \times \frac{1\times1\times3\times2}{102^4} = \frac{3}{10} \times \frac{6}{102^4} = 1{,}66 \times 10^{-8}$
$P(Sunda)(K_1|Sunda) = \frac{4}{10} \times \frac{3\times1\times1\times1}{111^4} = \frac{4}{10} \times \frac{3}{111^4} = 7{,}9 \times 10^{-9}$
$P(Minang)(K_1|Minang) = \frac{1}{10} \times \frac{1\times1\times1\times1}{77^4} = \frac{1}{10} \times \frac{1}{77^4} = 2{,}84 \times 10^{-9}$
$P(Indonesia)(K_1|Indonesia) = \frac{2}{10} \times \frac{2\times1\times1\times1}{86^4} = \frac{1}{5} \times \frac{2}{86^4} = 7{,}31 \times 10^{-9}$

Dari hasil simulasi diatas kalimat K_1 dapat diklasifikasikan sebagai Kalimat dengan kelas bahasa Jawa.

Adapun untuk penentuan kelas bahasa data tes K_2 adalah sebagai berikut:
$P("penerimaan\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}$ $P("penerimaan\ "|Sunda) = \frac{2+1}{39+72} = \frac{3}{111}$
$P("penerimaan\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("penerimaan"|Indonesia) = \frac{1+1}{14+72} = \frac{2}{86}$

$P("siswa\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}$ $P("siswa\ "|Sunda) = \frac{2+1}{39+72} = \frac{3}{111}$
$P("siswa\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("siswa"|Indonesia) = \frac{1+1}{14+72} = \frac{2}{86}$

$P("sd\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}, P("sd\ "|Sunda) = \frac{1+1}{39+72} = \frac{2}{111}$
$P("sd\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("sd\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

$P("di\ "|Jawa) = \frac{0+1}{30+72} = \frac{1}{102}$ $P("di\ "|Sunda) = \frac{2+1}{39+72} = \frac{3}{111}$
$P("di\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("di\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

$P("jawa\ "|Jawa) = \frac{1+1}{30+72} = \frac{2}{102}$ $P("jawa\ "|Sunda) = \frac{0+1}{39+72} = \frac{1}{111}$
$P("jawa\ "|Minang) = \frac{0+1}{5+72} = \frac{1}{77}$ $P("jawa\ "|Indonesia) = \frac{0+1}{14+72} = \frac{1}{86}$

$P(Jawa)(K_2|Jawa) = \frac{3}{10} \times \frac{1\times1\times1\times1\times2}{102^5} = \frac{3}{10} \times \frac{2}{102^5} = 5{,}43 \times 10^{-11}$
$P(Sunda)(K_2|Sunda) = \frac{4}{10} \times \frac{3\times3\times2\times3\times1}{111^5} = \frac{4}{10} \times \frac{54}{111^5} = 1{,}28 \times 10^{-9}$
$P(Minang)(K_2|Minang) = \frac{1}{10} \times \frac{1\times1\times1\times1\times1}{77^5} = \frac{1}{10} \times \frac{1}{77^5} = 3{,}69 \times 10^{-11}$
$P(Indonesia)(K_2|Indonesia) = \frac{2}{10} \times \frac{2\times2\times1\times1\times1}{86^5} = \frac{1}{5} \times \frac{4}{86^5} = 1{,}70 \times 10^{-10}$

Dari hasil perhitungan diatas K_2 dikategorikan sebagai kelas bahasa Sunda. Untuk kasus terhadap K_2 mengalami *error* dikarenakan kata-kata yang terdapat pada K_2.

4. Evaluasi

Pada tahap evaluasi, dilakukan eksperimen pengujian dengan tiga skenario. Skenario pertama dilakukan dengan mengobservasi pengaruh n-gram pada model bahasa terhadap hasil klasifikasi. Adapun skenario kedua dilakukan dengan mengobservasi pengaruh perubahan jumlah data latih terhadap hasil klasifikasi, dan skenario ketiga untuk membandingkan akurasi metode multinomial naïve bayes dengan beberapa metode klasifikasi lainnya. Metrik evaluasi yang digunakan adalah evaluasi *F-Score* karena evaluasi metrik ini umumnya digunakan untuk *task* klasifikasi.

Untuk pengujian skenario pertama, korpus yang digunakan untuk pengujian ini dilakukan dengan menggunakan 1000 kalimat untuk 7 bahasa (total 7000 kalimat) menggunakan korpus paralel Dataverse Telkom University dan Korpus Nusantara. Adapun rasio *training / testing data* pada pengujian skenario pertama adalah 80:20 dan diuji dengan berbagai *range* n-gram.

Pengujian skenario kedua dilakukan untuk enam bahasa yang ada dikarenakan korpus bahasa muna yang ada saat ini jumlahnya sangat kecil, korpus yang digunakan sebagai *data training* adalah korpus paralel Dataverse Telkom University dan Korpus Nusantara untuk 1000 kalimat per bahasa, sedangkan 500 kalimat tambahan untuk setiap bahasa menggunakan korpus NusaX *data test* yang digunakan untuk skenario kedua dataset yang digunakan adalah

NusaX. Untuk pengujian skenario kedua *range* n-gram yang digunakan adalah 1-3-gram (unigram, bigram, trigram).
Pengujian skenario ketiga dilakukan dengan menggunakan pengaturan yang sama pada korpus pengujian pertama dan diujikan dengan menggunakan empat metode algoritma, Multinomial Naïve Bayes, *Support Vector Machine* (SVM) dengan *kernel Radial Basis Function* (RBF), dan *Random Forest*, dan *Decision Tree*. Pengujian ini juga menggunakan tiga *feature extraction* yaitu N-gram, TF-IDF, dan N-gram-TF-IDF. Untuk n-gram yang digunakan pada pengujian ini adalah *range* n-gram 1-3-gram (unigram, bigram, trigram).

4.1 Hasil Pengujian
Hasil pengujian skenario pertama yaitu pengubahan n-gram dapat dilihat pada tabel 1 berikut ini.

Tabel 4. Hasil Pengujian Skenario Pertama

N-Gram range	Accuracy	Precision	Recall	F1-score
1-2-gram	**98,86%**	**98,84%**	**98,9%**	**98,86%**
1-3-gram	98,64%	98,63%	98,68%	98,65%
1-4-gram	98,57%	98,56%	98,60%	98,57%
2-3-gram	86,36%	88,03%	86,38%	86,63%
2-4-gram	85,86%	87,56%	85,85%	86,14%
3-4-gram	49,5%	77,35%	49,97%	54,65%
Rata-Rata	86,29%	91,49%	86,39%	87,25%

Berdasarkan dari hasil pengujian yang ada di tabel, pengaturan 1-2-gram (unigram dan bigram) merupakan hasil terbaik yang dapat dicapai dengan rata-rata yang dicapai adalah 86,29%.

Tabel 5. Hasil Pengujian Skenario Kedua

Bahasa	Ukuran Data Latih	Precision	Recall	F1-Score
Indonesia	1000	79,67%	98,00%	87,89%
	1250	97,06%	99,00%	98,02%
	1500	**98,02%**	**99,00%**	**98,51%**
Jawa	1000	88,76%	79,00%	83,60%
	1250	97,94%	95%	96,45%
	1500	**98,97%**	**96,00%**	**97,46%**
Sunda	1000	76,23%	93,00%	83,78%
	1250	96,08%	98,00%	97,03%
	1500	**98,00%**	**98,00%**	**98,00%**
Minang	1000	70,15%	94,00%	80,34%
	1250	95,19%	99,00%	97,06%
	1500	**95,19%**	**99,00%**	**97,06%**
Bugis	1000	76,47%	39,00%	51,66%
	1250	97,98%	97,00%	97,49%
	1500	**98,02%**	**99%**	**98,51%**
Madura	1000	85,19%	69,00%	76,24%
	1250	100%	96,00%	97,96%
	1500	**100%**	**97%**	**98,48%**

Tabel 6. Hasil Prediksi Masing-Masing Bahasa Pada Pengujian Skenario Kedua

Bahasa Asli	Bahasa Yang Diprediksi	Persentase Prediksi		
		Jumlah Data Latih		
		1000	1250	1500
Indonesia	**Indonesia**	**98%**	**99%**	**99%**
	Sunda	0%	0%	0%

	Minang	2%	1%	1%
	Jawa	0%	0%	0%
	Madura	0%	0%	0%
	Bugis	0%	0%	0%
Sunda	Indonesia	1%	0%	0%
	Sunda	**92%**	**98%**	**98%**
	Minang	2%	1%	1%
	Jawa	4%	1%	1%
	Madura	1%	0%	0%
	Bugis	0%	0%	0%
Bugis	Indonesia	9%	1%	0%
	Sunda	13%	1%	0%
	Minang	27%	1%	1%
	Jawa	3%	0%	0%
	Madura	9%	0%	0%
	Bugis	**39%**	**97%**	**99%**
Jawa	Indonesia	5%	0%	0%
	Sunda	8%	2%	1%
	Minang	4%	1%	1%
	Jawa	**79%**	**95%**	**96%**
	Madura	2%	0%	0%
	Bugis	2%	2%	2%
Madura	Indonesia	6%	1%	1%
	Sunda	8%	1%	1%
	Minang	5%	1%	1%
	Jawa	2%	1%	0%
	Madura	**69%**	**96%**	**97%**
	Bugis	10%	0%	0%
Minang	Indonesia	5%	1%	1%
	Sunda	0%	0%	0%
	Minang	**94%**	**99%**	**99%**
	Jawa	1%	0%	0%
	Madura	0%	0%	0%
	Bugis	0%	0%	0%

Adapun untuk hasil pengujian skenario kedua yaitu pengubahan jumlah data latih dapat dilihat pada tabel 5 dan tabel 6 untuk validasi dari masing-masing pengujian dengan menggunakan jumlah data latih yang berbeda dengan garis bawah sebagai hasil prediksi yang benar untuk masing-masing bahasa. Adapun hasil dari pengujian skenario ketiga yang menggunakan berbagai metode algoritma untuk model bahasa didapatkan sebagai berikut.

Tabel 7. Hasil Pengujian Skenario Ketiga

Model	Feature Extraction	Accuracy	Precision	Recall	F-Score
MultinomialNB	N-Gram	98,64%	98,63%	98,66%	98,64%
	TF-IDF	98,64%	98,63%	98,68%	98,65%
	N-Gram + TF-IDF	**99,00%**	**98,99%**	**99,03%**	**99,00%**
SVM	N-Gram	95,21%	95,77%	95,17%	95,31%
	TF-IDF	**98,64%**	**98,66%**	**98,65%**	**98,65%**
	N-Gram + TF-IDF	**98,64%**	**98,66%**	**98,65%**	**98,65%**
Random Forest	N-Gram	97,21%	97,33%	97,23%	97,24%
	TF-IDF	97,21%	97,33%	97,23%	97,25%
	N-Gram + TF-IDF	**97,43%**	**97,52%**	**97,46%**	**97,46%**
Decision Tree	N-Gram	**90,36%**	**90,60%**	**90,50%**	**90,43%**
	TF-IDF	89,29%	89,32%	89,43%	89,28%
	N-Gram + TF-IDF	89,29%	89,32%	89,43%	89,28%

Dari hasil pengujian skenario ketiga, akurasi tertinggi yang dapat dicapai sebesar 99% dengan model bahasa Multinomial Naïve Bayes dan *feature extraction* N-Gram dan TF-IDF.

4.2 Analisis Hasil Pengujian

Berdasarkan hasil pengujian skenario pertama yang terdapat pada Tabel 4, diketahui bahwa pengubahan *range* n-gram berpengaruh terhadap akurasi klasifikasi bahasa. Semakin besar n-gram-nya tertinggi diperoleh dengan menggunakan model bahasa Multinomial Naïve Bayes *range* n-gram 1-2-gram. Pada pengujian dengan *range* n-gram yang lebih besar, terdapat penurunan terhadap akurasi, hal ini disebabkan karena frasa yang muncul sebagai fitur identifikasi bahasa bentuknya adalah frasa pendek yang terdiri dengan satu hingga tiga kata. Adapun faktor lain yang disebabkan karena korpus yang dimiliki terdapat campuran antara bahasa pada label dengan bahasa lain (*code-mixed*). Berikut merupakan beberapa kalimat *code-mixed* yang terdapat pada korpus.

Tabel 8. Data latih dengan kalimat *code-mixed*

Index	Kalimat	Bahasa
2081	Pisang goreng bisa dijadian kawan untuak minum teh talua.	Minang
2596	Pakaian asli Indonesia dari Sabang sampai Merauke lainnyo dapek dikataui dari ciri-cirinyo nan dipakaikan di satiok daerah.	Minang
3378	Ari geus ngarupa taman mah, bisa jadi tempat paniisan sarta tempat ulin anu murah meriah.	Sunda
4684	saking pikantuk FGD ing RT 1 RW 6 dipundeningaken sekawan klasifikasi kesejahteraan keluarga	Jawa
5344	Keluarga kasebbhut biyasahna keluarga oreng tuwah tunggal se mondhi kereman dari keluarga lain	Madura
6037	Instansi terkait ndek mempunyai kewajiban untok menindaklanjuti mengadu yang iletui'I LSM dan media lokal	Bugis

Sehingga pada klasifikasi dengan menggunakan tri-gram dan 4-gram semakin banyak kata yang digunakan untuk satu fitur namun fitur tersebut sulit untuk ditemukan pada klasifikasi saat pengujian, dimana pada kalimat "Pakaian asli Indonesia dari Sabang sampai Merauke lainnyo dapek dikataui dari ciri-cirinyo nan dipakaikan di satiok daerah" fitur yang didapat berupa "Pakaian asli Indonesia dari" sebagai 4-gram pertama pada kalimat tersebut. Selain itu kata-kata yang digaris bawahi pada tabel 8 merupakan kata-kata yang terdapat pada bahasa Indonesia.

Untuk hasil pengujian skenario kedua, berdasarkan Tabel 5 terjadi peningkatan akurasi yang sangat signifikan pada pengujian 1250 data latih yang meningkat hingga 19,16%. Pada pengujian menggunakan 1500 data latih terjadi peningkatan dengan persentase kecil. Berdasarkan Pengujian skenario kedua ini didapat bahwa nilai *recall* terendah didapatkan pada bahasa bugis dengan pengujian 1000 *training data* sebesar 39%. Pada Tabel 6 klasifikasi pada *data test* bahasa bugis mengalami kesalahan sebesar 27% bahasa minang dan 13% bahasa sunda, sehingga model tidak dapat membedakan bahasa bugis dan bahasa minang. Adapun juga peningkatan akurasi secara signifikan pada pengujian 1250 dan 1500 data latih terjadi karena data latih tambahan memiliki topik yang sama dengan data uji dan memiliki ciri khas kuat sebagai fitur bahasa tersebut.

Berdasarkan hasil pengujian yang didapat pada tabel 7 akurasi tertinggi yang dapat dicapai adalah 99% dengan menggunakan model Multinomial Naïve Bayes dengan *feature extraction* N-Gram dan TF-IDF. Selain itu hasil pengujian dengan model bahasa lainnya mendapatkan akurasi diatas 89%, dan juga untuk penggunaan *feature extraction* yang berbeda menghasilkan akurasi secara tidak signifikan. Adapun juga pengujian ini dilakukan untuk membandingkan terhadap hasil penelitian sebelumnya yang dilakukan oleh Nugraha dan Romadhony[8] dan didapatkan hasil perbandingan sebagai berikut.

Tabel 9. Perbandingan Antara Kedua Hasil Pengujian

Feature Extraction	Model Bahasa	Hasil Pengujian		Nugraha dan Romadhony	
		Accuracy	F-Score	Accuracy	F-Score
N-Gram	Naïve Bayes	98,64%	98,64%	**99,40%**	**99,50%**
	SVM	95,21%	95,31%	**97,20%**	**97,20%**
	Random Forest	**97,21%**	**97,24%**	96,80%	96,80%
	Decision Tree	**90,36%**	**90,43%**	89,40%	89,70%
TF-IDF	Naïve Bayes	98,64%	98,65%	**99,20%**	**99,30%**
	SVM	98,64%	98,65%	**98,80%**	**98,90%**
	Random Forest	**97,21%**	**97,25%**	96,70%	96,80%

	Decision Tree	**89,29%**	**89,28%**	87,80%	88,10%

Berdasarkan hasil perbandingan pada tabel 8, akurasi tertinggi terdapat pada penelitian sebelumnya yang memiliki akurasi sebesar 99,40% dengan model bahasa Naïve Bayes dan *feature extraction* N-Gram, adapun juga model bahasa SVM dengan akurasi 98,80% dengan *feature extraction* TF-IDF yang dicapai oleh penelitian sebelumnya. Untuk *Random Forest* dan *Decision Tree* pengujian yang dilakukan memiliki akurasi tertinggi dibanding penelitian sebelumnya sebesar 97,21% untuk *Random Forest* dan 90,36% untuk *Decision Tree* dengan *feature extraction* N-Gram. Didapatkan hasil pebandingan sebagai berikut dimana penelitian yang dilakukan Nugraha dan Romadhony menggunakan jumlah *data training* dan *data test* yang lebih besar dibandingkan pengujian yang dilakukan pada tugas akhir ini sehingga memiliki akurasi yang lebih tinggi dari seluruh hasil pengujian yang dilakukan.

5. Kesimpulan

Identifikasi bahasa daerah dengan menggunakan metode Multinomial Naïve Bayes dapat menghasilkan akurasi yang baik untuk menggunakan jumlah n-gram yang singkat yakni 1-gram (unigram) hingga 2-gram (bigram) dan akurasi model bahasa meningkat dengan bertambahnya jumlah data latih yang diberikan. Rata-rata akurasi yang dihasilkan pada Naïve Bayes dan F-measure sebesar 98,86%. Adapun untuk *range* n-gram 3-gram dan 4-gram nilai akurasi yang dihasilkan hanya sebesar 49,5% dikarenakan kalimat yang digunakan umunya dicampur dengan bahasa lain sehingga saat melakukan klasifikasi kata yang asalnya bahasa asing menjadi fitur dalam suatu tersebut. Adapun jumlah data latih mempengaruhi akurasi pada model bahasa, sehingga semakin banyak data latih maka akurasi pada model bahasa semakin tinggi. Dengan hasil perbandingan yang dilakukan model yang dibangun saat ini belum dapat mencapai hasil yang lebih tinggi dari penelitian sebelumnya,

Daftar Pustaka

[1] Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, *2*, 27-40.

[2] Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online (pp. 46-57)

[3] Moryossef, A., Tsochantaridis, I., Aharoni, R., Ebling, S., & Narayanan, S. (2020, August). Real-time sign language detection using human pose estimation. In European Conference on Computer Vision (pp. 237-248). Springer, Cham.

[4] Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research, 65, 675-782.

[5] Zhang, Y., Riesa, J., Gillick, D., Bakalov, A., Baldridge, J., & Weiss, D. (2018). A fast, compact, accurate model for language identification of codemixed text. *arXiv preprint arXiv:1810.04142*.

[6] Nishijima, M., & Liu, Y. (2021). Native Language Identification and Reconstruction of Native Language Relationship Using Japanese Learner Corpus. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (pp. 368-376).

[7] Wijonarko, P., & Zahra, A. (2022). Spoken language identification on 4 Indonesian local languages using deep learning. Bulletin of Electrical Engineering and Informatics, 11(6), 3288-3293

[8] Nugraha, A. B., & Romadhony, A. (2023). Identification of 10 Regional Indonesian Languages Using Machine Learning. *Sinkron: jurnal dan penelitian teknik informatika*, *8*(4), 2203-2214.

[9] Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

[10] Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.

[11] Baldwin, T., & Lui, M. (2010, June). Language identification: The long and the short of the matter. In Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 229-237).

[12] Sasaki, Y. (2007). The truth of the F-measure. *Teach tutor mater*, *1*(5), 1-5.

[13] ARDIYANTI SURYANI, ARIE; Widyantoro, Dwi Hendratmo; Purwarianti, Ayu; Sudaryat, Yayat, 2022, "Sundanese-Indonesian Parallel Corpus", https://doi.org/10.34820/FK2/HDYWXW, Telkom University Dataverse, V1

[14] Sujaini, H. (2020). Improving the role of language model in statistical machine translation (Indonesian-Javanese). *International Journal of Electrical and Computer Engineering*, *10*(2), 2102.

[15] Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., ... & Ruder, S. (2022). Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *arXiv preprint arXiv:2205.15960*.
[16] Jurafsky, Daniel & Martin, James. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.