

**KLASIFIKASI LIRIK LAGU DAERAH MENGGUNAKAN
METODE NAIVE BAYES**

SKRIPSI

**OLEH
TRIYANTI SIMBOLON
NIM 180535632505**



**UNIVERSITAS NEGERI MALANG
FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INFORMATIKA
AGUSTUS 2022**

**KLASIFIKASI LIRIK LAGU DAERAH MENGGUNAKAN
METODE NAIVE BAYES**

SKRIPSI
diajukan kepada
Universitas Negeri Malang
untuk memenuhi salah satu persyaratan
dalam menyelesaikan program Sarjana Teknik Informatika

OLEH
TRIYANTI SIMBOLON
NIM 180535632505

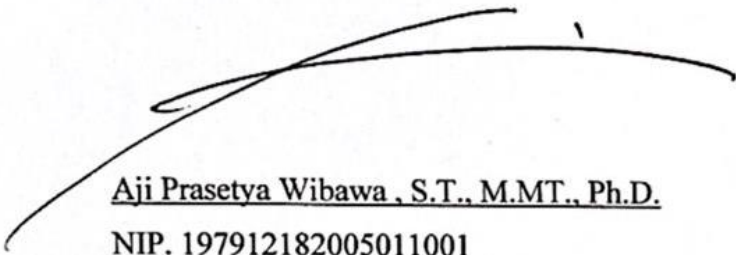
UNIVERSITAS NEGERI MALANG
FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INFORMATIKA
AGUSTUS 2022

LEMBAR PERSETUJUAN PEMBIMBING SKRIPSI

Skripsi oleh Triyanti Simbolon ini telah diperiksa dan disetujui untuk diujikan.

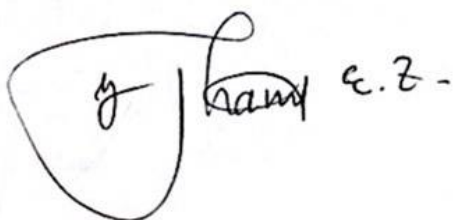
Malang, Agustus 2022.

Pembimbing I,



Aji Prasetya Wibawa, S.T., M.MT., Ph.D.
NIP. 197912182005011001

Pembimbing II,




Ilham Ari Elbaith Zaeni, S.T., M.T., Ph.D.
NIP. 198106262006041004

LEMBAR PENGESAHAN

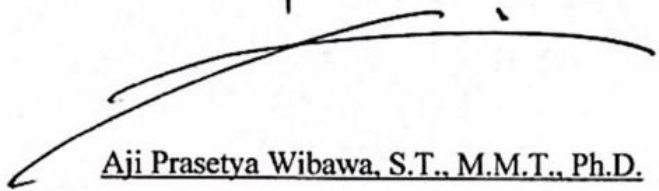
Skripsi oleh Triyanti Simbolon ini telah dipertahankan di depan dewan penguji pada tanggal Agustus 2022.

Dewan Penguji

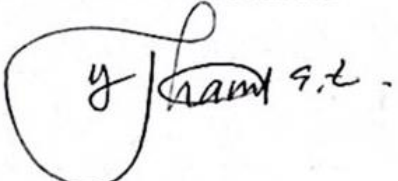


Utomo Pujianto, S.Kom., M.Kom.
NIP. 198206042012121001

Ketua



Aji Prasetya Wibawa, S.T., M.M.T., Ph.D. Anggota
NIP. 197912182005011001

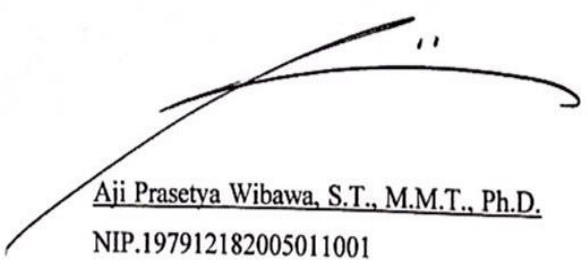


Ilham Ari Elbaith Zaeni, S.T., M.T., Ph.D. Anggota
NIP. 198106262006041004

Mengesahkan,
Dekan Fakultas Teknik

Mengetahui,
Ketua Departemen Teknik Elektro

Prof. Dr. Marji, M.Kes.
NIP. 195902031984031001



Aji Prasetya Wibawa, S.T., M.M.T., Ph.D.
NIP.197912182005011001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Triyanti Simbolon
NIM : 180535632505
Departemen/Program Studi : Teknik Elektro/S1 Teknik Informatika
Fakultas/Program : Fakultas Teknik/S1

Menyatakan dengan sesungguhnya bahwa skripsi yang saya tulis ini benar – benar tulisan saya, dan bukan merupakan plagiasi/falsifikasi/fabrikasi baik sebagian atau seluruhnya.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa skripsi ini hasil plagiasi/falsifikasi/fabrikasi, baik sebagian atau seluruhnya, maka saya bersedia menerima sanksi atas perbuatan tersebut sesuai dengan ketentuan yang berlaku.

Malang,

Yang membuat pernyataan

MATERAI

Triyanti Simbolon

RINGKASAN

S, Triyanti. 2022. *Klasifikasi Lirik Lagu Daerah Menggunakan Metode Naïve Bayes*. Skripsi, Program studi S1 Teknik Informatika, Departemen Teknik Elektro, Fakultas Teknik, Universitas Negeri Malang. Pembimbing: (I) Aji Prasetya Wibawa, S.T., M.MT., Ph.D. (II) Ilham Ari Elbaith Zaeni, S.T., M.T., Ph.D.

Kata Kunci: Lagu Daerah, Lagu Nasional, *Naive Bayes*, SMOTE.

Indonesia merupakan negara yang kaya akan bahasa dan budaya, salah satunya adalah lagu daerah dan lagu nasional. Dalam upaya mempermudah melakukan pencarian lagu daerah berdasarkan asal daerah maka dibuatlah proses klasifikasi teks. *Dataset* yang akan diuji terdiri dari 480 lagu daerah dan 90 lagu nasional. *Dataset* dibagi menjadi skenario 2 label, 4 label, dan 31 label dengan menggunakan dan tanpa menggunakan SMOTE. Sebelum dilakukan klasifikasi, *dataset* harus melalui beberapa tahapan *preprocessing text*. *Preprocessing text* terdiri dari *cleansing*, *case folding*, *tokenizing*, dan *indexing with term frequency*. Teknik yang digunakan untuk mengatasi *dataset* yang tidak seimbang adalah *upsampling* dengan teknik *Synthetic Minority Over-Sampling Technique* (SMOTE). *Dataset* akan dilakukan klasifikasi dengan metode *Naïve Bayes Multinomial* yang akan menghitung jumlah frekuensi kemunculan kata pada suatu dokumen. Evaluasi akan dilakukan dengan menggunakan metode *10-fold cross validation*. Kemudian ditampilkan dalam *confusion matrix* yang terdiri dari akurasi, presisi, dan *recall*. Berdasarkan hasil penelitian didapatkan skenario terbaik dengan menggunakan SMOTE pada 31 label dengan akurasi sebesar 98,1%.

SUMMARY

S, Triyanti. 2022. *Lyrics Classification of Indonesia's Folk Songs Using Naïve Bayes Method*. Skripsi, Program studi S1 Teknik Informatika, Departemen Teknik Elektro, Fakultas Teknik, Universitas Negeri Malang. Pembimbing: (I) Aji Prasetya Wibawa, S.T., M.MT., Ph.D. (II) Ilham Ari Elbaith Zaeni, S.T., M.T., Ph.D.

Key words: *Folk Songs, National Songs, Naïve Bayes, SMOTE.*

Indonesia is a country that is rich in language and culture, one of which is folk songs and national songs. To make it easier for people to search for regional songs based on the regional origin, a text classification process was made. The dataset to be tested consists of 480 folk songs and 90 national songs. The dataset is divided into scenarios of 2 labels, 4 labels, and 31 labels using and without using SMOTE. Before classification, the dataset must go through several stages of preprocessing text. Preprocessing text consists of cleansing, case folding, tokenizing, and indexing with term frequency. The technique used to overcome unbalanced datasets is up-sampling with the Synthetic Minority Over-Sampling Technique (SMOTE). The dataset will be classified using the Naïve Bayes Multinomial method which will calculate the number of occurrences of words in a document. The evaluation will be carried out using the 10-fold cross-validation method. Then it is displayed in a confusion matrix of accuracy, precision, and recall. Based on the study's results, the best scenario was obtained using SMOTE on 31 labels with an accuracy of 98.1%.

KATA PENGANTAR

Puji serta syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa karena dengan karunia – Nya penulis sampai saat ini telah menyelesaikan skripsi yang berjudul “*Klasifikasi Lirik Lagu Daerah Menggunakan Metode Naïve Bayes*”. Tujuan penulisan ini disusun guna untuk melengkapi salah satu syarat dalam menyelesaikan jenjang Strata 1 Program Studi Teknik Informatika di Universitas Negeri Malang. Dalam kesempatan ini juga penulis ingin menyampaikan ucapan terima kasih kepada semua pihak, sehingga penulis mampu menyelesaikan penulisan ini. Ucapan terima kasih tersebut khususnya kepada:

1. Prof. Dr. Marji M.Kes. selaku Dekan Fakultas Teknik Universitas Negeri Malang.
2. Aji Prasetya Wibawa, S.T., M.M.T., Ph.D. selaku Ketua Departemen Teknik Elektro sekaligus Pembimbing 1 yang telah memberikan pengarahan dan bimbingan hingga skripsi selesai serta membantu kelancaran dalam pengurusan skripsi.
3. Ilham Ari Elbaith Zaeni, S.T., M.T., Ph.D. selaku Koordinator Program Studi S1 Teknik Informatika sekaligus Pembimbing 2 yang telah banyak membantu dan membimbing penulis selama proses penyusunan skripsi.
4. Orang tua serta keluarga penulis yang telah memberikan dorongan moral dan material.

Dengan segala Keterbatasan pengetahuan yang saya miliki, penulis menyadari bahwa dalam penulisan ilmiah ini masih jauh dari kata sempurna. Penulis berharap agar kiranya tulisan ini dapat bermanfaat dan merupakan salah satu informasi yang berguna bagi pembaca, saran dan kritik sangat penulis harapkan.

Malang, 2022

Penulis

DAFTAR ISI

LEMBAR PERSETUJUAN	i
LEMBAR PENGESAHAN	ii
PERNYATAAN KEASLIAN TULISAN	iii
RINGKASAN	iv
SUMMARY	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	vii
DAFTAR TABEL	viii
DAFTAR GAMBAR.....	ix
DAFTAR LAMPIRAN	x
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	1
1.3 Tujuan Penelitian.....	2
1.4 Batasan Masalah	2
1.5 Manfaat Penelitian.....	2
1.6 Definisi Operasional	2
BAB II KAJIAN PUSTAKA	3
BAB III METODE PENELITIAN	5
3.1 Pengumpulan Data.....	5
3.2 <i>Preprocessing</i>	6
3.3 Klasifikasi.....	9
3.4 Evaluasi	9
BAB IV HASIL DAN PEMBAHASAN	11
BAB V KESIMPULAN DAN SARAN	17
5.1 Kesimpulan.....	17
5.2 Saran	17
DAFTAR PUSTAKA	18
LAMPIRAN.....	21

DAFTAR TABEL

Tabel 3.1 Pembagian Label <i>Dataset</i>	6
Tabel 3.2 Hasil Proses <i>Cleansing, Case Folding, Tokenizing</i>	7
Tabel 3.3 <i>Term Frequency</i>	7
Tabel 4.1 Skenario Pengujian	11
Tabel 4.2 Perbandingan Performa Klasifikasi Pada Skenario 1 dan 2	12
Tabel 4.3 Perbandingan Performa Klasifikasi Pada Skenario 3 dan 4	13
Tabel 4.4 Perbandingan Performa Klasifikasi Pada Skenario 5 dan 6	14
Tabel 4.5 Perbandingan Skenario Dengan Nilai Performa Terbaik	14
Tabel 4.6 Perhitungan Probabilitas Kalimat.....	15

DAFTAR GAMBAR

Gambar 3.1 Desain Penelitian	5
Gambar 3.2 Jumlah Data Setiap Kelas Sebelum SMOTE.....	8
Gambar 3.3 Jumlah Data Setiap Kelas Setelah SMOTE.....	9

DAFTAR LAMPIRAN

Lampiran	23
----------------	----

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Bangsa Indonesia terkenal dengan keragaman seni dan budaya. Kebudayaan menjadi ciri khas dan identitas bangsa Indonesia yang membedakannya dengan bangsa lain (Imam & Sismoro, 2015). Budaya terbentuk dari banyak unsur, seperti adat istiadat, pakaian, bangunan, karya seni dan bahasa (Setiowati, 2020). Indonesia memiliki bahasa daerah yang sangat beragam pada setiap provinsinya. Menurut data dari Kemendikbud RI tahun 2021, Indonesia memiliki 718 bahasa daerah (Mayasari & Indarti, 2022). Salah satu bentuk penerapan bahasa daerah dalam kebudayaan adalah lagu daerah.

Lagu daerah merupakan lagu yang ada pada daerah tertentu dengan bentuk yang sangat sederhana dan lirik yang menggunakan bahasa dari daerah tersebut (Setiadi, 2019). Lirik pada lagu daerah biasanya menceritakan kebiasaan masyarakat setempat, tradisi perjuangan suatu daerah, dan nilai-nilai budaya (Yati et al., 2020). Banyaknya bahasa daerah yang ada di Indonesia membuat masyarakat kesulitan dalam mengetahui asal daerah dari lirik lagu daerah. Permasalahan ini dapat diatasi dengan menggunakan klasifikasi teks yang bekerja dengan cara mengkategorikan objek berdasarkan ciri objek tersebut (Wibawa et al., 2018).

Klasifikasi pada penelitian ini dilakukan dengan mengkategorikan lagu daerah berdasarkan lirik dan asalnya. Klasifikasi dilakukan menggunakan metode *Multinomial Naïve Bayes* (MNB). MNB digunakan karena metode ini terbukti dapat menghasilkan performa baik pada *dataset* yang berupa teks (Setianingrum et al., 2018). *Dataset* yang digunakan dalam penelitian ini merupakan *dataset* lagu daerah dari beberapa provinsi di Indonesia.

1.2 Rumusan Masalah

Berdasarkan uraian dari latar belakang, didapatkan rumusan masalah penelitian antara lain:

1. Bagaimana kinerja algoritma *Multinomial Naive Bayes* dalam melakukan klasifikasi teks?

2. Bagaimana pengaruh metode *Synthetic Minority Over-Sampling Technique* (SMOTE) terhadap performa algoritma dalam melakukan klasifikasi teks lagu daerah dan lagu nasional?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang diuraikan, didapatkan tujuan penelitian sebagai berikut:

1. Mengetahui kinerja algoritma *Multinomial Naive Bayes* dalam melakukan klasifikasi teks lagu daerah dan lagu nasional.
2. Mengetahui pengaruh metode *Synthetic Minority Over-Sampling Technique* (SMOTE) dalam performa algoritma untuk melakukan klasifikasi.

1.4 Batasan Masalah

Terdapat beberapa batasan masalah penelitian antara lain:

1. *Dataset* lagu daerah yang digunakan hanya 30 provinsi, diambil dari buku dan internet, dan dianggap sudah valid .
2. Tahap *preprocessing* pada tidak menggunakan *stemming* dan *stopwords removal* karena penelitian ini merupakan bagian dari penelitian besar untuk kamus bahasa sehingga setiap kata pada *dataset* dianggap penting.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah:

1. Sebagai referensi untuk penelitian terkait klasifikasi dokumen dengan metode *Multinomial Naïve Bayes*.
2. Menambah wawasan dan pengetahuan dalam klasifikasi teks lagu daerah dan lagu nasional.

1.6 Definisi Operasional

Untuk menghindari perbedaan penafsiran istilah yang digunakan, maka dituliskan beberapa definisi operasional sebagai berikut.

1. *Multinomial Naive Bayes* merupakan salah satu metode *text mining* yang digunakan untuk mengklasifikasikan teks lagu daerah dan lagu

nasional.

2. Klasifikasi adalah proses pengelompokan data berdasarkan ciri yang sama.

BAB II

KAJIAN PUSTAKA

Lagu daerah merupakan warisan budaya nusantara yang perlu dilestarikan. Lagu daerah adalah lagu yang ada pada daerah tertentu dengan bentuk yang sangat sederhana dan menggunakan bahasa dari daerah tersebut (Fadillah & Wahyuni, 2021). Lagu daerah biasanya diperdengarkan pada saat hiburan, pesta, dan acara adat (Fadillah & Wahyuni, 2021). Lagu daerah biasanya menceritakan kebiasaan masyarakat setempat, tradisi perjuangan suatu daerah, dan nilai-nilai budaya (Setiowati, 2020). Lagu daerah memiliki tujuan untuk memperkenalkan budaya dan adat istiadat dari suatu daerah tertentu (Saputro et al., 2019). Lagu daerah di Indonesia terdapat kurang lebih 439 lagu daerah yang mewakili masing-masing provinsi (Setiadi, 2019).

Sering kali orang-orang tidak mengingat bahkan mengetahui lagu daerah dan asalnya, karena begitu banyak lagu daerah di Indonesia. Tanpa adanya pengetahuan lagu daerah di Indonesia dapat menimbulkan berbagai masalah. Salah satunya kasus *overclaiming* yang dilakukan pada lagu daerah Rasa Sayang-Sayange yang diklaim budaya dari negara lain. Oleh karena itu, dengan kemajuan teknologi di era modern ini, diperlukan suatu cara untuk mengklasifikasi lagu daerah di Indonesia.

Penelitian terkait klasifikasi lagu daerah sudah pernah dilakukan dengan metode *tf-idf* dan metode *Multinomial Naïve Bayes* berdasarkan liriknya (Saputro et al., 2019). Penelitian lainnya dilakukan untuk tingkat kesedihan pada lagu Didi Kempot menggunakan metode *Multinomial Naïve Bayes* (Damayanti et al., 2021). Hasil dari kedua penelitian tersebut menyatakan bahwa semakin banyak kelas yang dilibatkan pada saat pengujian menghasilkan nilai akurasi yang menurun. Untuk mendapatkan hasil akurasi yang lebih baik dapat dilakukan penyeimbangan data dengan menambahkan jumlah data. Metode *Multinomial Naïve Bayes* dapat meningkatkan performa jika didukung dengan jumlah data yang lebih besar.

Metode klasifikasi teks *Multinomial Naïve Bayes* (MNB) merupakan salah satu metode klasifikasi dari *Naïve Bayes* (Abbas et al., 2019). Metode ini menggunakan proses perhitungan dengan jumlah frekuensi kemunculan kata pada sebuah dokumen. Proses perhitungan dilakukan dengan cepat, sederhana dan

menghasilkan akurasi yang optimal dibandingkan metode *Naïve Bayes* lainnya (Farisi et al., 2019). Proses klasifikasi metode MNB memiliki permasalahan data tidak seimbang yang dapat menyebabkan nilai performa yang menurun. Maka perlu dilakukan *upsampling* data dengan menggunakan *Synthetic Minority Over-Sampling Technique* (Ardhana et al., 2019).

Synthetic Minority Over-Sampling Technique atau SMOTE adalah teknik yang dapat digunakan untuk mengatasi ketidakseimbangan data. Teknik SMOTE tidak hanya menduplikasi data yang sama melainkan akan membuat sampel baru. Hasil sampel data yang baru menyerupai data asli dari kelas minoritas untuk menyeimbangkan *dataset*. Data dari hasil replikasi tersebut dikenal dengan data sintesis (Safitri & Muslim, 2020).

Penelitian sebelumnya menyimpulkan bahwa semua *classifier* mendapatkan manfaat dari teknik *oversampling* dengan SMOTE sebagai teknik yang menghasilkan performa terbaik (Hairani et al., 2020). Penelitian dengan menggunakan teknik SMOTE untuk mengatasi data yang tidak seimbang juga pernah dilakukan oleh (Sulistiyowati & Jajuli, 2020). Metode SMOTE dapat menangani permasalahan *imbalanced data* dan SMOTE baik dikombinasikan dengan algoritma klasifikasi *Naïve Bayes*. Oleh karena itu, untuk mengatasi *dataset* yang tidak seimbang digunakan teknik SMOTE.

Setelah dilakukan optimasi dan penerapan model, langkah terakhir yang perlu dilakukan adalah validasi. Salah satu bentuk model validasi yang populer adalah *K-fold Cross-Validation*. Dalam pendekatan ini, pertama *file* data dikompilasi dari n kasus dan kemudian biasanya diacak dan dibagi menjadi k segmen yang sama. Segmen k pertama, yang terdiri dari n/k kasus, disisihkan dan model diparameterisasi dengan sisa kasus $(n-n/k)$, kemudian diuji terhadap segmen pertama 8 untuk tingkat kesalahan klasifikasi. Selanjutnya dari *full case file* k segmen kedua disisihkan dan model diparameterisasi dengan sisa kasus, kemudian diuji terhadap segmen kedua, begitu seterusnya untuk semua k segmen (Marcot & Hanea, 2020).

BAB III

METODE PENELITIAN

Metode yang digunakan pada penelitian ini terdiri dari beberapa tahapan. Desain dari alur penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1. Desain Penelitian

3.1 Pengumpulan Data

Pengumpulan data dilakukan dengan metode studi literatur yang bersumber dari buku dan dokumen tertulis yang tersedia di internet. Data yang sudah terkumpul akan digabungkan dan dimasukkan ke dalam Ms. Excel. Data selanjutnya akan dipisah berdasarkan asal daerah serta sumber pengambilan data. Berdasarkan Rancangan Undang-undang Daerah Otonomi Baru (RUU DOB) tahun 2022 yang disahkan oleh DPR, saat ini Indonesia memiliki 37 provinsi. Asal lagu daerah yang digunakan pada penelitian ini adalah 30 provinsi. Provinsi lainnya merupakan provinsi baru yang lagu daerahnya masih mengikuti provinsi lama. Contohnya ada pada provinsi baru Papua Pegunungan, Papua Selatan dan Papua tengah yang sebelumnya merupakan bagian dari Papua. Pada *dataset* yang digunakan data yang diambil hanya dari provinsi Papua.

Dataset yang sudah dikumpulkan sejumlah 480 lagu daerah. Pengujian yang dilakukan dengan *dataset* tersebut mendapatkan hasil akurasi yang rendah yaitu kurang lebih 25%. Hal tersebut dikarenakan perbedaan jumlah data antara label maksimum dan minimum yang sangat tinggi. Untuk mengatasi ketidakseimbangan data, maka dilakukan teknik SMOTE. Setelah dilakukan teknik SMOTE hasil akurasi masih rendah yaitu kurang lebih 40%. Untuk meningkatkan nilai akurasi, maka ditambahkan data baru dengan label lagu Nasional sejumlah 90 lagu. Pembagian label juga dilakukan untuk mengetahui perbandingan performa klasifikasi pada banyak label. Data total yang sudah

dikumpulkan sejumlah 570 lagu. Total data dari setiap label ditampilkan pada Tabel 3.1.

Tabel 3.1. Pembagian Label *Dataset*

2 Label	3 Label	31 Label	Total Data
Lagu Daerah	Barat	Aceh	10
		Bangka Belitung	10
		Banten	10
		Bengkulu	17
		Kalimantan	13
		Jambi	16
		Jawa Barat	42
		Jawa Tengah	15
		Jawa Timur	19
		Kalimantan Barat	10
		Kalimantan Tengah	12
		Lampung	13
		Riau	15
		Sumatera Barat	34
		Sumatera Selatan	41
		Sumatera Utara	28
		Kalimantan	11
	Tengah	Bali	14
		Gorontalo	11
		Kalimantan Selatan	10
		Kalimantan Timur	10
		Kalimantan Utara	4
		Nusa Tenggara Barat	10
		Nusa Tenggara Timur	10
		Sulawesi Selatan	11
		Sulawesi Tengah	14
		Sulawesi Tenggara	7
		Sulawesi Utara	20
	Timur	Maluku	26
		Papua	17
Lagu Nasional	Lagu Nasional	Lagu Nasional	90
Total Data			570

3.2 Preprocessing

Sebelum dilakukan klasifikasi, data perlu diproses terlebih dahulu menggunakan proses *text preprocessing*. *Text preprocessing* bertujuan untuk mempersiapkan teks dalam *dataset* agar memperoleh hasil yang optimal (Haddi et

al., 2013). *Text preprocessing* dilakukan dengan 4 tahapan yaitu *cleansing*, *casefolding*, *tokenizing*, dan *indexing with term frequency* (TF). *Cleansing* bertujuan untuk menghilangkan karakter *non-alphabet* (Squicciarini et al., 2017). *Case folding* bertujuan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil (Jumeilah, 2017). *Tokenizing* bertujuan untuk memotong sebuah teks menjadi kata oleh tanda baca atau spasi, sehingga didapatkan kata tunggal saja. Kata tunggal yang merupakan hasil dari *tokenizing* disebut *token*. *Token* akan digunakan sebagai *input* untuk proses selanjutnya (Sabrani et al., 2020). Hasil dari proses *cleansing*, *casefolding*, *tokenizing* ditampilkan pada Tabel 3.2.

Tabel 3.2. Hasil Proses *Cleansing*, *Casefolding*, *Tokenizing*

<i>Input</i>	<i>Output Cleansing</i>	<i>Output Casefolding</i>	<i>Output Tokenizing</i>
Bungong jeumpa, Bungong jeumpa, Meugah di Aceh... Bungong teuleubeh, teuleubeh, Indah lagoina...	Bungong jeumpa Bungong jeumpa Meugah di Aceh Bungong teuleubeh teuleubeh Indah lagoina	bungong jeumpa bungong jeumpa meugah di aceh bungong teuleubeh teuleubeh indah lagoina	['bungong', 'jeumpa', 'bungong', 'jeumpa', 'meugah', 'di', 'aceh', 'bungong', 'teuleubeh', 'teuleubeh', 'indah', 'lagoina']

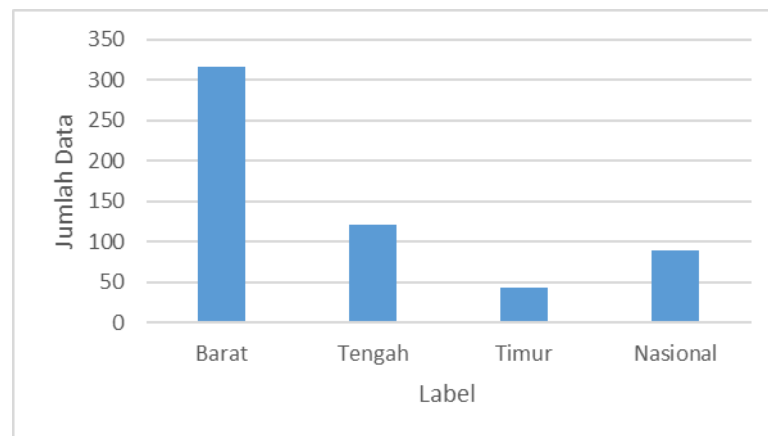
Dari Tabel 3.2, hasil token akan digunakan untuk proses pembobotan kata dengan menggunakan metode *term frequency* (TF). TF digunakan untuk mengukur berapa kali suatu kata atau frasa muncul dalam dokumen (Harahap et al., 2021). Nilai TF yang tinggi mengidentifikasikan bahwa kata tersebut penting dalam dokumen. Nilai TF dapat digunakan untuk menentukan letak kelas dari kata yang sama pada beberapa kelas (Weggenmann & Kerschbaum, 2018). Jumlah *term* yang ada pada *dataset* yaitu sebanyak 11.698 *term*. Contoh frekuensi kemunculan salah satu kata di setiap kelas ditampilkan pada Tabel 3.3.

Tabel 3.3. *Term Frequency*

	sengko	rindang	beta	bangsa
Total Frekuensi	134	16	52	70
Frekuensi (Barat)	134	0	0	11
Frekuensi (Tengah)	0	16	10	1

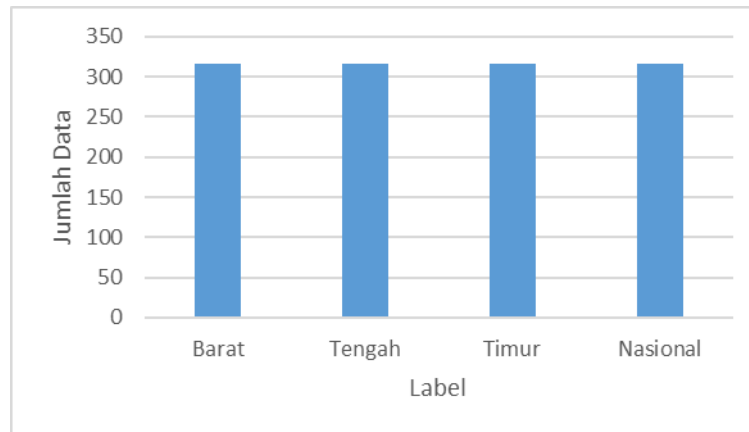
Frekuensi (Timur)	0	0	37	1
Frekuensi (Nasional)	0	0	5	57

Dataset yang tidak seimbang dapat diatasi dengan menggunakan metode *undersampling* dan *oversampling*. Metode yang digunakan adalah metode *oversampling* menggunakan teknik *Synthetic Over-Sampling Technique* (SMOTE). SMOTE dipilih karena tidak menghapus *instance* dari *dataset* yang mungkin membawa beberapa informasi penting. SMOTE dilakukan dengan tujuan untuk mengetahui pengaruhnya terhadap performa dari metode *Multinomial Naïve Bayes*. Pada Gambar 3.2 ditampilkan jumlah data di setiap kelas label yang berbeda dan jarak nilai antara kelas yang sangat jauh.



Gambar 3.2. Jumlah Data Setiap Kelas Sebelum SMOTE

Berdasarkan Gambar 3.2 didapatkan data yang tidak seimbang sehingga perlu dilakukan tahapan SMOTE. SMOTE diawali dengan menghitung perbedaan jumlah data antara kelas mayoritas dan kelas minoritas. Selanjutnya dilakukan perhitungan persentase duplikasi yang diinginkan pada kelas minoritas dan dilakukan pemilihan jumlah k. Nilai k dilakukan dengan beberapa kali percobaan, dimana batas akhir nilai k merupakan nilai maksimal pada kelas minoritas. Tahapan terakhir data sintetis akan dibuat sebanyak persentase duplikasi yang diinginkan antara data dengan nilai k yang telah dipilih. Sehingga dihasilkan data yang sudah seimbang yang dapat dilihat pada Gambar 3.3.



Gambar 3.3. Jumlah Data Setiap Kelas Sesudah SMOTE

3.3 Klasifikasi

Metode klasifikasi yang digunakan adalah *Multinomial Naïve Bayes* (MNB). MNB diterapkan dengan tujuan untuk membentuk model yang sesuai. Pada metode MNB, sistem perhitungan kategori dokumen ditentukan dari frekuensi masing-masing kemunculan kata dalam sebuah dokumen (Setianingrum et al., 2018). Frekuensi kemunculan kata digunakan untuk menghitung probabilitas kemunculan kata pada suatu label.

Metode ini memulai prosesnya dengan menentukan label data dan melakukan ekstraksi nilai dari label tersebut. Selanjutnya dilakukan proses perhitungan jumlah dokumen, jumlah kelas, dan jumlah kata dari seluruh data. Kemudian dilakukan perhitungan *prior probability* untuk menghitung probabilitas kelas asal lagu. Lalu *post probability* yang bertujuan untuk menghitung probabilitas suatu kata yang masuk ke dalam kelas asal lagu (Mayasari & Indarti, 2022). MNB tidak bisa mengukur dengan tepat jumlah data prediksi sehingga mengurangi nilai akurasi pada seleksi atribut. Untuk mengatasi hal tersebut diperlukan pengoptimalan pada atributnya.

3.4 Evaluasi

Pengujian *dataset* akan dievaluasi dengan menggunakan *k-fold cross validation*. *K-fold cross validation* digunakan untuk menghilangkan bias pada *dataset*. *K-fold cross validation* dilakukan dengan memisahkan data menjadi dua bagian, yaitu data uji dan data latih. *K-fold cross validation* bekerja dengan

membagi data yang akan diuji dan dilatih menjadi himpunan bagian k dengan ukuran yang hampir sama (Mardiana et al., 2022).

Metode *fold cross validation* merupakan metode validasi yang paling umum digunakan dan diketahui banyak memberikan hasil yang baik (Marcot & Hanea, 2020). Nilai k *fold* yang digunakan sebesar 10. Lalu perhitungan performa evaluasi model klasifikasi yang digunakan adalah *confusion matrix*. *Confusion Matrix* melakukan perhitungan nilai akurasi, presisi, dan *recall* dari setiap label.

BAB IV

HASIL DAN PEMBAHASAN

Pengujian dilakukan menggunakan metode *Multinomial Naïve Bayes* (MNB). *Dataset* dibagi ke dalam enam skenario dimana masing-masing skenario memiliki perbedaan pada jumlah label dan tahapan SMOTE yang ditampilkan pada Tabel 4.1.

Tabel 4.1. Skenario Pengujian

SKENARIO	LABEL		SMOTE
	LAGU NASIONAL	LAGU DAERAH	
1	90	480	-
2	480	480	K=2, 3, 13, 22, 33, 42, 53, 62, 73, 82, 89.
3	90	Barat: 316	-
		Tengah: 121	
		Timur: 43	
4	316	Barat: 316	K=2, 3, 13, 22, 33, 42.
		Tengah: 316	
		Timur: 316	
5	90	Maksimal: 42 (Jawa Barat)	-
		Minimal: 4 (Kalimantan Utara)	
6	90	30 Provinsi: 90	K=2, 3.

Berdasarkan Tabel 4.1, skenario 1 memiliki 2 label yaitu label Lagu Nasional sebanyak 90 lagu dan label Lagu Daerah sebanyak 480 lagu. Skenario 1 diuji tanpa menggunakan teknik SMOTE. Skenario 2 memiliki 2 label dengan jumlah *dataset* pada label Lagu Nasional sama dengan label Lagu Daerah sebesar 480 lagu. Jumlah data pada label yang sama pada skenario 2 terjadi karena pengujian dilakukan dengan menggunakan teknik SMOTE. Nilai k pada skenario

2 sebesar 2, 3, 13, 22, 33, 42, 53, 62, 73, 89. Nilai k ditentukan secara acak hingga 89 yang diambil dari jumlah data pada label minimum sebelum SMOTE yaitu 90 dikurangi 1.

Skenario lainnya memiliki penjelasan yang sama seperti skenario 1 dan 2. Perbedaan pembagian label terletak pada label Lagu Daerah dan nilai k. Masing-masing skenario dengan teknik SMOTE memiliki batas akhir dari nilai k yang berbeda-beda. Hal ini berdasarkan jumlah data dari masing-masing label minimal skenario yang belum dilakukan pengujian dengan teknik SMOTE.

Selanjutnya akan dilakukan perbandingan klasifikasi skenario berdasarkan jumlah label yang sama. Hasil perbandingan klasifikasi skenario 1 dan 2 dengan jumlah label 2 dapat dilihat pada Tabel 4.2. Tabel 4.2 menampilkan hasil pengujian dengan dan tanpa menggunakan teknik SMOTE. Pada hasil pengujian didapatkan performa klasifikasi terbaik dengan nilai $k=2$ yang memiliki akurasi sebesar 93,854%, presisi sebesar 94,479%, dan *recall* sebesar 93,929%. Perbedaan hasil akurasi pada pengujian tanpa dan dengan teknik SMOTE kurang lebih sebesar 3%, dimana saat teknik SMOTE dilakukan terdapat kenaikan performa.

Tabel 4.2. Perbandingan Performa Klasifikasi Pada Skenario 1 dan 2

Nilai k	Akurasi (%)	Presisi (%)	<i>Recall</i> (%)
-	90,350	94,856	69,333
2	93,854	94,479	93,929
3	93,854	94,467	93,922
13	93,125	93,897	93,193
22	93,541	94,243	93,580
33	93,333	94,058	93,399
42	93,854	94,479	93,923
53	93,437	94,145	93,473
62	93,437	94,145	93,515
73	92,916	93,763	92,957
89	93,125	93,918	93,168

Hasil perbandingan klasifikasi skenario 3 dan 4 dengan jumlah label 4 dapat dilihat pada Tabel 4.3. Tabel 4.3 menampilkan hasil pengujian dengan dan tanpa menggunakan teknik SMOTE. Pada hasil pengujian didapatkan performa klasifikasi terbaik dengan nilai $k=33$ yang memiliki akurasi sebesar 91,768%, presisi sebesar 92,378%, dan *recall* sebesar 91,765%. Perbedaan hasil akurasi pada pengujian tanpa dan dengan teknik SMOTE kurang lebih sebesar 26%, dimana saat teknik SMOTE dilakukan terdapat kenaikan performa.

Tabel 4.3. Perbandingan Performa Klasifikasi Pada Skenario 3 dan 4

Nilai k	Akurasi (%)	Presisi (%)	<i>Recall</i> (%)
-	65,789	42,356	41,225
2	90,664	91,119	90,853
3	90,505	91,131	90,517
13	90,583	91,213	90,687
22	91,612	92,234	91,639
33	91,768	92,378	91,765
42	91,135	91,826	91,224

Hasil perbandingan nilai klasifikasi pada skenario 5 dan 6 dengan jumlah label 31 dapat dilihat pada Tabel 4.4. Tabel 4.4 menampilkan hasil pengujian dengan dan tanpa menggunakan teknik SMOTE. Pada hasil pengujian didapatkan performa klasifikasi terbaik dengan nilai $k=3$ yang memiliki akurasi sebesar 98,1%, presisi sebesar 97,947%, dan *recall* sebesar 98,271%. Perbedaan hasil akurasi pada pengujian tanpa dan dengan teknik SMOTE kurang lebih sebesar 63%, dimana saat teknik SMOTE dilakukan terdapat kenaikan performa.

Pengujian skenario tanpa dan dengan SMOTE pada 31 label mengalami kenaikan yang signifikan dibandingkan kenaikan pada skenario lainnya. Pada 31 label, data terkecil ada pada provinsi Kalimantan Utara yaitu 4 lagu sedangkan data terbesar ada pada provinsi Jawa Barat yaitu 90 lagu. Skenario 5 memiliki perbedaan jumlah data besar yang dapat meningkatkan *bias* sehingga menurunkan performa algoritma MNB. Sedangkan pada skenario 6, performa yang dihasilkan meningkat dengan signifikan karena jumlah data sudah seimbang. SMOTE

menduplikasi data sintetis pada label terkecil dalam jumlah banyak, yaitu sekitar 86 data sintetis baru.

Tabel 4.4. Perbandingan Performa Klasifikasi Pada Skenario 5 dan 6

Nilai k	Akurasi (%)	Presisi (%)	Recall (%)
-	25,964	13,983	11,546
2	97,885	97,799	98,101
3	98,100	97,947	98,271

Hasil perbandingan nilai performa terbaik dari pengujian seluruh skenario pada label yang berbeda dapat dilihat pada Tabel 4.5. Pada Tabel 4.5 dapat disimpulkan bahwa nilai k yang kecil cenderung menghasilkan hasil performa yang tinggi. Hal ini dapat terjadi karena kinerja SMOTE menghasilkan data sintetis yang memiliki kemiripan tinggi dengan data asli. Kinerja SMOTE dapat mempengaruhi performa algoritma MNB karena persebaran data antar label akan lebih *distinct* sehingga pengenalan data dalam proses klasifikasi lebih mudah.

Tabel 4.5. Perbandingan Skenario Dengan Nilai Performa Terbaik

Label	Nilai k	Akurasi (%)	Presisi (%)	Recall (%)
2	2	93,854	94,479	93,929
4	33	91,768	92,378	91,765
31	3	98,100	97,947	98,271

Berdasarkan tabel-tabel hasil pengujian ini, diketahui bahwa terdapat beberapa faktor yang mempengaruhi kinerja klasifikasi metode MNB. Faktor pertama adalah jumlah label, dapat dilihat pada hasil akurasi skenario yang tidak menggunakan SMOTE. Skenario 1 dengan jumlah label 2 menghasilkan performa yang lebih baik dibandingkan skenario 3 dan 5. Kinerja metode MNB dapat lebih mudah melakukan klasifikasi pada jumlah label yang sedikit.

Faktor kedua adalah penggunaan SMOTE. Pada *dataset* yang dilakukan pengujian dengan menggunakan teknik SMOTE mendapatkan hasil performa yang lebih baik dibandingkan tanpa menggunakan teknik SMOTE. Hal ini

disebabkan oleh meratanya persebaran data pada setiap label saat teknik SMOTE diterapkan. Jika teknik SMOTE tidak diterapkan terdapat perbedaan jumlah data pada label yang tidak seimbang sehingga meningkatkan bias data.

Faktor lainnya adalah nilai k SMOTE pada skenario pengujian yang sebenarnya tidak terlalu mempengaruhi performa MNB dalam melakukan klasifikasi. Perbedaan hasil yang didapatkan dari beberapa pengujian dengan nilai k yang berbeda-beda pada setiap skenario hanya sebesar 1-2%.

Tabel 4.6. Perhitungan Probabilitas Kalimat

Keterangan	Sebelum SMOTE		Setelah SMOTE	
	Tengah	Timur	Tengah	Timur
Frekuensi "jangan"	24	8	34	198
Frekuensi "mama"	13	13	36	72
Frekuensi "marah"	7	3	7	16
Frekuensi "beta"	10	37	15	205
<i>Prior</i>	0,214	0,073	0,25	0,25
<i>Post 1</i> "jangan"	$5,51 \times 10^{-4}$	$1,46 \times 10^{-4}$	$3,59 \times 10^{-4}$	$2,01 \times 10^{-3}$
<i>Post 1</i> "mama"	$3,08 \times 10^{-4}$	$3,4 \times 10^{-4}$	$3,79 \times 10^{-4}$	$7,39 \times 10^{-4}$
<i>Post 1</i> "marah"	$1,76 \times 10^{-4}$	$9,74 \times 10^{-5}$	$8,21 \times 10^{-5}$	$1,72 \times 10^{-4}$
<i>Post 1</i> "beta"	$2,42 \times 10^{-4}$	$9,25 \times 10^{-4}$	$1,64 \times 10^{-4}$	$2,08 \times 10^{-3}$
<i>Post 2</i>	<u>$1,55 \times 10^{-14}$</u>	$3,26 \times 10^{-16}$	$4,57 \times 10^{-16}$	<u>$1,32 \times 10^{-13}$</u>

Pada Tabel 4.6 dilakukan perhitungan dengan menggunakan contoh kalimat “jangan mama marah beta” dari lagu “Ayo Mama” dengan label Indonesia bagian Timur. Pada Tabel 4.6 dapat dilihat bahwa sebelum dilakukan SMOTE, frekuensi dan *post probability* 1 kata “jangan”, “mama”, “marah” lebih besar di label Tengah. Kemudian dilakukan perhitungan *post probability* 2, yaitu perhitungan akhir untuk menentukan label asal daerah kalimat “jangan mama marah beta”. Kalimat tersebut mendapatkan hasil klasifikasi label Tengah. Hal ini dikarenakan *prior probability* pada label Tengah lebih besar dari label Timur. Pada kasus lain, jika frekuensi dan *post probability* 1 mengarah pada label yang benar, kalimat akan diklasifikasikan ke label yang memiliki hasil *prior probability* lebih besar.

Setelah dilakukan SMOTE kalimat “jangan mama marah beta” dapat diklasifikasikan dengan benar ke label Timur. *Prior probability* setelah dilakukan

SMOTE pada masing-masing label seimbang yaitu 0,25. Seperti yang sudah dijelaskan sebelumnya bahwa *prior probability* mempengaruhi perhitungan akhir probabilitas kalimat “jangan mama marah beta”. *Post probability* 2 merupakan perhitungan akumulasi dari *prior probability* dan *post probability* 1.

Berdasarkan penjelasan tersebut, teknik SMOTE diketahui dapat meningkatkan performa metode MNB karena dapat mempengaruhi *prior probability* pada masing-masing label. Label dengan *prior probability* yang kecil akan cenderung diabaikan dan kata/kalimat akan diklasifikasikan ke label dengan nilai yang lebih besar. Oleh karena itu, dengan menggunakan SMOTE dihasilkan nilai *prior probability* masing-masing label yang seimbang sehingga metode MNB dapat melakukan klasifikasi dengan lebih tepat.

Penelitian Klasifikasi Teks Lagu Daerah dilakukan secara berkelompok dengan penggunaan metode yang berbeda-beda. Penelitian dengan menggunakan *Support Vector Machine* (Erlangga, 2022) yang bekerja berdasarkan beberapa jenis *kernel*. Metode ini menghasilkan performa terbaik dibandingkan metode lainnya karena data yang digunakan pada merupakan data berdimensi tinggi. SVM adalah metode klasifikasi yang optimal untuk data berdimensi tinggi, yang telah dibuktikan oleh (Chauhan et al., 2018).

Penelitian lainnya dilakukan dengan menggunakan *K-Nearest Neighbors* (KNN) berdasarkan nilai tetangga terdekat (Abdillah, 2022). KNN diuji dengan menggunakan SMOTE untuk melihat pengaruh terhadap algoritma tersebut. Pengaruh SMOTE pada algoritma KNN mendapatkan hasil yang tidak dapat diprediksi performanya. Pada pengujian dengan 2 dan 4 label mengalami penurunan jika menggunakan SMOTE, sedangkan dengan 31 label mengalami kenaikan performa.

Penelitian dengan menggunakan *Multinomial Naïve Bayes* memiliki kelebihan dimana kata yang muncul pada dua atau lebih label, masih dapat diklasifikasikan. Kekurangan pada metode ini terletak pada jumlah data yang tidak seimbang sehingga menghasilkan *prior probability* yang tidak seimbang juga. Hal tersebut dapat menyebabkan performa dari metode ini menurun karena pada proses perhitungannya dapat meningkatkan *bias*. Kekurangan lain dalam metode ini ada pada proses seleksi atribut. Optimasi kekurangan tersebut dapat

dilakukan pemberian bobot pada atribut dengan menggunakan metode *Particle Swarm Optimization* (PSO) (Aulianita & Rifai, 2018).

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pengujian yang telah dilakukan dapat disimpulkan bahwa algoritma *Naïve Bayes Multinomial* dapat digunakan untuk klasifikasi lagu daerah dan lagu nasional dan menghasilkan performa yang baik. Skenario pengujian yang tidak menggunakan SMOTE didapatkan hasil dengan akurasi yang lebih baik pada skenario yang memiliki 2 label.

Penggunaan teknik SMOTE berpengaruh untuk meningkatkan performa klasifikasi pada algoritma *Naïve Bayes Multinomial*. Pengaruh penggunaan SMOTE dapat dilihat pada skenario 2, 4, 6 yang memiliki performa yang lebih baik dibandingkan dengan skenario lainnya yang tidak menggunakan teknik SMOTE.

5.2 Saran

1. Penelitian selanjutnya dapat menambahkan *dataset* lagu daerah dari provinsi lain yang belum digunakan pada penelitian ini.
2. Penelitian selanjutnya dapat menggunakan metode *Bernoulli Naïve Bayes* maupun metode klasifikasi lainnya.
3. Penelitian selanjutnya dapat menggunakan *data preprocessing* seperti menambahkan proses *stemming* dan *stopword removal* untuk meningkatkan akurasi dan mempercepat proses klasifikasi.

DAFTAR PUSTAKA

- Abbas, M., Ali Memon, K., & Aleem Jamali, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *International Journal of Computer Science and Network Security (IJCSNS)*, 19(3), 62. <https://doi.org/10.13140/RG.2.2.30021.40169>
- Ardhana, A. P., Cahyani, D. E., & Winarno. (2019). Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods. *Journal of Physics: Conference Series*, 1306(1). <https://doi.org/10.1088/1742-6596/1306/1/012049>
- Aulianita, R., & Rifai, A. (2018). Optimasi Particle Swarm Optimization pada Naive Bayes untuk Sentiment Analysis Furniture. *Information Management for Educators and Professionals: Journal of Information Management*, 3(1), 31–40. <https://ejournal-binainsani.ac.id/index.php/IMBI/article/view/1043>
- Chauhan, V. K., Dahiya, K., & Sharma, A. (2018). Problem Formulations and Solvers in Linear SVM: a Review. *Artificial Intelligence Review*, 52(2), 803–855. <https://doi.org/10.1007/S10462-018-9614-6>
- Damayanti, S. L., Wibawa, A. P., & Pujiyanto, U. (2021). Klasifikasi Tingkat Kesedihan pada Lagu Didi Kempot dengan Menggunakan Metode Multinomial Naive Bayes. *Jurnal Informatika*, 1–12.
- Fadillah, S., & Wahyuni, S. (2021). Peningkatan Self-Awareness Anak Usia 5-6 Tahun Melalui Pembelajaran Lagu Daerah Riau. *Jurnal Pendidikan Anak Usia Dini (PERNIK)*, 4(1), 100–104. <https://doi.org/10.31851/PERNIK.V4I1.6801>
- Farisi, A. A., Sibaroni, Y., & Faraby, S. Al. (2019). Sentiment Analysis on Hotel Reviews Using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series*,. <https://doi.org/10.1088/1742-6596/1192/1/012024>
- Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Preprocessing in Sentiment Analysis. *Procedia Computer Science*, 17, 26–32.

<https://doi.org/10.1016/J.PROCS.2013.05.005>

- Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE for Handling Class Imbalance in The Classification of Diabetes with C4.5, SVM, and Naive Bayes. *Jurnal Teknologi Dan Sistem Komputer*, 8(2), 89–93. <https://doi.org/10.14710/jtsiskom.8.2.2020.89-93>
- Harahap, C. N., Marthasari, G. I., & Hayatin, N. (2021). Perbandingan Klasifikasi Berita Hoax Kategori Kesehatan Menggunakan Naïve Bayes dan Multinomial Naïve Bayes. *Jurnal Repositor*, 3(4), 419–424. <http://repositor.umm.ac.id/index.php/repositor/article/view/1380>
- Imam, D. S., & Sismoro, H. (2015). Rancang Bangun Aplikasi Mobile Sebagai Media Pelestarian Lagu Tradisional dan Nasioal Indonesia Berbasis Android. *Jurnal Ilmiah DASI*, 16(1), 40–42.
- Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal Rekayasa Sistem Dan Teknologi Informasi (RESTI)*, 1(1), 19–25. <https://doi.org/10.29207/resti.v1i1.11>
- Marcot, B. G., & Hanea, A. M. (2020). What is an Optimal Value of K in K-Fold Cross-Validation in Discrete Bayesian Network Analysis? *Computational Statistics*, 36(3), 2009–2031. <https://doi.org/10.1007/S00180-020-00999-9>
- Mardiana, L., Kusnandar, D., & Satyahadewi, N. (2022). Analisis Diskriminan dengan K Fold Cross Validation untuk Klasifikasi Kualitas Air di Kota Pontianak. *Bimaster*, 11(1), 97–102.
- Mayasari, L., & Indarti, D. (2022). Klasifikasi Topik Tweet Mengenai Covid Menggunakan Metode Multinomial Naïve Bayes dengan Pembobotan Tf-Idf. *Jurnal Ilmiah Informatika Komputer*, 27(1), 43–53. <https://doi.org/10.35760/ik.2022.v27i1.6184>
- Sabrani, A., Wedashwara W., I. G. W., & Bimantoro, F. (2020). Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA)*, 2(1), 89–100. <https://doi.org/10.29303/jtika.v2i1.87>

- Safitri, A. R., & Muslim, M. A. (2020). Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms. *Journal of Soft Computing Exploration*, 1(1), 70–75.
- Saputro, P. H., Aristian, M., & Listianing, T. D. (2019). Klasifikasi Lagu Daerah Indonesia Berdasarkan Lirik Menggunakan Metode Tf-Idf dan Naïve Bayes. *Jurnal Teknologi Informasi Dan Terapan*, 4(1), 47–52. <https://doi.org/10.25047/jtit.v4i1.20>
- Setiadi, G. (2019). Eksegesis Syair Lagu Wajib Nasional Berdasarkan Kajian Hermeneutik Guna Memahami Makna dan Pesan Kepahlawanan untuk Penanaman Karakter pada Anak. *Jurnal Heritage*, 7(1), 10–22. <https://doi.org/10.35891/HERITAGE.V7I1.1568>
- Setianingrum, A. H., Kalokasari, D. H., & Shofi, I. M. (2018). Implementasi Algoritma Multinomial Naive Bayes Classifier. *Jurnal Teknik Informatika*, 10(2), 109–118. <https://doi.org/10.15408/jti.v10i2.6822>
- Setiowati, S. P. (2020). Pembentukan Karakter Anak pada Lagu Tokecang, Jawa Barat. *Jurnal Ilmu Budaya*, 8(1), 172. <https://doi.org/10.34050/jib.v8i1.9980>
- Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment Analysis During Hurricane Sandy in emergency Response. *International Journal of Disaster Risk Reduction*, 21, 213–222. <https://doi.org/10.1016/j.ijdrr.2016.12.011>
- Sulistiyowati, N., & Jajuli, M. (2020). Integrasi Naive Bayes dengan Teknik Sampling Smote untuk Menangani Data Tidak Seimbang. *Nuansa Informatika*, 14(1), 34. <https://doi.org/10.25134/nuansa.v14i1.2411>
- Weggenmann, B., & Kerschbaum, F. (2018). Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining. *SynTF*.
- Wibawa, A. P., Purnama, M. G., Akbar, M. F., & Dwiyanto, F. A. (2018). Metode-Metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, 3(1), 134.
- Yati, N., Sofyan, F. S., & Saylendra, N. P. (2020). Peran Guru Membiasakan Menyanyikan Lagu Nasional sebagai Upaya Pembentukan Nasionalisme

Siswa. *Jurnal Pendidikan Pancasila Dan Kewarganegaraan (CIVICS)*, 5(2), 117–121. <https://doi.org/10.36805/civics.v5i2.1338>

LAMPIRAN

Beberapa lampiran *pseudocode*, dan data lainnya disimpan dalam *url* berikut

https://drive.google.com/drive/folders/1UvzOsDL2L8tFeVeEf0UW_fyb_gPy782i?usp=sharing