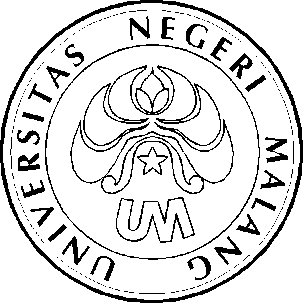
**IMPLEMENTASI SMOTE PADA KLASIFIKASI LAGU DAERAH DAN LAGU NASIONAL DENGAN METODE NAIVE BAYES**

**SKRIPSI**

OLEH  
TRIYANTI SIMBOLON  
NIM 180535632505

****

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

**IMPLEMENTASI SMOTE PADA KLASIFIKASI LAGU DAERAH DAN LAGU NASIONAL DENGAN 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 diuji.

Malang, (Tanggal Sidang)

Pembimbing I,

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

Pembimbing II,

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

LEMBAR PENGESAHAN

Skripsi oleh Triyanti Simbolon ini telah dipertahankan di depan dewan penguji

pada tanggal (Tanggal Sidang)

Dewan Penguji

……………………………. Ketua  
NIP. 0000000000

……………………………. Anggota  
NIP. 0000000000

……………………………. Anggota  
NIP. 0000000000

Mengesahkan, Mengetahui,  
Dekan Fakultas Teknik Ketua Jurusan Teknik Elektro

Prof. Dr. Marji, M.Kes. Aji Prasetya Wibawa, S.T., M.M.T., Ph.D.  
NIP. 195902031984031001 NIP.197912182005011001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Triyanti Simbolon

NIM : 180535632505

Jurusan/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, tgl

Yang membuat pernyataan

MATERAI

Triyanti Simbolon

RINGKASAN

S, Triyanti. 2022. *Implementasi SMOTE Pada Klasifikasi Lagu Daerah Dan Lagu Nasional Dengan Metode Naïve Bayes*. Skripsi, Program studi S1 Teknik Informatika, Jurusan 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 melestarikan kebudayaan Indonesia dengan memanfaatkan teknologi dan informasi dapat dilakukan dengan cara klasifikasi lagu daerah dan lagu nasional. *Dataset* klasifikasi yang akan diuji terdiri dari 480 lagu daerah dan 90 lagu nasional yang akan dibagi menjadi skenario 2 label, 4 label, dan 31 label dengan menggunakan dan tanpa menggunakan SMOTE. *Dataset* yang akan diklasifikasi, sebelumnya harus melalui beberapa tahapan, tahapan pertama adalah *preprocessing text* yang terdiri dari *case folding*, *remove punctuation*, *tokenizing*, dan *indexing with term frequency*. Teknikyang digunakan mengatasi *dataset* yang tidak seimbang adalah *upsampling* dengan menggunakan teknik *Synthetic Minority Over-Sampling Technique* (SMOTE). Kemudian *dataset* akan dilakukan klasifikasi dengan metode *Naïve Bayes Multinomial* yang akan menghitung jumlah frekuensi kemunculan kata pada suatu dokumen. *Dataset* akan dievaluasi dengan menggunakan metode 10-*fold cross validation* dan akan ditampilkan dalam *confusion matrix* yang terdiri dari akurasi, presisi, dan *recall*. Berdasarkan hasil penelitian didapatkan skenario terbaik dengan menggunakan SMOTE pada 31 label dengan hasil akurasi 98,1%.

SUMMARY

S, Triyanti. 2022. *SMOTE Implementation in Traditional and National Songs Classification Using Naïve Bayes Method.* Skripsi, Program studi S1 Teknik Informatika, Jurusan 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.

***Keywords: Traditional 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. In an effort to preserve Indonesian culture by utilizing technology and information, this can be done by classifying traditional songs and national songs. The classification dataset to be tested consists of 480 traditional songs and 90 national songs which will be divided into scenarios of 2 labels, 4 labels, and 31 labels using and without using SMOTE. The dataset to be classified, previously had to go through several stages, the first stage was preprocessing text which consisted of case folding, remove punctuation, tokenizing, and indexing with term frequency. The technique used to overcome unbalanced datasets is upsampling using the Synthetic Minority Over-Sampling Technique (SMOTE). Then the dataset will be classified using the Naïve Bayes Multinomial method which will calculate the number of occurrences of words in a document. The dataset will be evaluated using the 10-fold cross validation method and will be displayed in a confusion matrix consisting of accuracy, precision, and recall. Based on the results of the study, 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 “*Implementasi SMOTE Pada Klasifikasi Lagu Daerah Dan Lagu Nasional Dengan Metode Naïve Bayes*”. Tujuan penulisan ini disusun guna untuk melengkapi salah satu syarat dalam menyelesaikan jenjang Strata 1 jurusan 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 Jurusan Teknik Elektro sekaligus Pembimbing I 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 dan juga selaku 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 2

1.3 Tujuan Penelitian 2

1.4 Batasan Masalah 2

1.5 Manfaat Penelitian 2

1.6 Definisi Operasional 3

**BAB II LANDASAN TEORI 4**

**BAB III METODE PENELITIAN 6**

3.1 *Business Understanding* 6

3.2 *Data Understanding* 6

3.3 *Data Preparation* 7

3.3.1 *Cleansing* 8

3.3.2 *Casefolding* 8

3.3.3 *Tokenizing* 8

3.3.4 *Term Frequency* 9

3.3.5 SMOTE 9

3.4 *Modeling* 11

3.5 *Evaluation* 11

**BAB IV HASIL DAN PEMBAHASAN 12**

**BAB V KESIMPULAN DAN SARAN 17**

5.1 Kesimpulan 17

5.2 Saran 17

**DAFTAR PUSTAKA 18**

**LAMPIRAN 23**

DAFTAR TABEL

**Tabel 3.1** Pembagian Label *Dataset* 7

**Tabel 3.2** Hasil *Data* *Cleansing* 8

**Tabel 3.3** Hasil *Data Case Folding* 8

**Tabel 3.4** Hasil *Data Tokenizing* 9

**Tabel 3.5** Term Frequency 9

**Tabel 4.1** Skenario pengujian 12

**Tabel 4.2** Perbandingan Performa Klasifikasi Pada Skenario 1 dan 2 13

**Tabel 4.3** Perbandingan Performa Klasifikasi Pada Skenario 3 dan 4 14

**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** Jumlah Data Setiap Kelas Sebelum SMOTE 10

**Gambar 3.2** Jumlah Data Setiap Kelas Setelah SMOTE 10

**DAFTAR LAMPIRAN**

**Lampiran 23**

BAB I

PENDAHULUAN

* 1. Latar Belakang Masalah

Indonesia merupakan negara yang memiliki keanekaragaman bahasa dan budaya. Salah satu kebudayaan yang ada pada setiap provinsi di Indonesia adalah lagu daerah (Setiowati, 2020). Lagu daerah memiliki ciri khas yang berbeda-beda yaitu terletak pada lirik lagu yang mewakilkan bahasa dari daerah tersebut. Selain lagu daerah, Indonesia juga memiliki budaya lainnya yaitu lagu nasional. Lagu nasional atau yang bisa disebut juga sebagai lagu kebangsaan merupakan lagu yang liriknya mewakilkan satu bahasa dari sebuah negara yang mengekspresikan rasa nasionalisme dan patriotisme (Yati et al., 2020).

Seiring berkembangnya zaman, semakin banyak lagu-lagu modern dan lagu-lagu asing yang masuk ke Indonesia, serta kurangnya media atau alat bantu dan kesadaran masyarakat akan pentingnya mengenal budaya nusantara dapat menyebabkan minat masyarakat khususnya anak-anak dan remaja menjadi menurun. Sebagaimana dapat diketahui bahwa pelestarian serta pemahaman budaya nusantara terlebih pada lagu daerah dan lagu nasional sangat penting dalam pembentukan jati diri dan watak dari sebuah bangsa (Imam & Sismoro, 2015). Dalam upaya melestarikan kebudayaan Indonesia dengan memanfaatkan teknologi dan informasi dapat dilakukan dengan cara klasifikasi teks. Klasifikasi teks bekerja dengan cara mengelompokkan objek berdasarkan ciri – ciri yang dimiliki oleh objek klasifikasi (Aji, Guntur, & Fathony, 2018). Klasifikasi yang dilakukan pada penelitian ini berdasarkan dengan lirik dan asal daerahnya yang diharapkan dapat membantu masyarakat dalam mempermudah pengetahuan terhadap asal dari lagu daerah tersebut.

Metode klasifikasi teks yang digunakan pada penelitian ini adalah *Multinomial Naïve Bayes* (MNB) yang merupakan salah satu metode klasifikasi dari *Naïve Bayes* (Abbas et al., 2019). Metode ini dipilih karena proses perhitungan dilakukan dengan jumlah frekuensi kemunculan kata pada sebuah dokumen dan menghasilkan proses perhitungan yang cepat, sederhana dan menghasilkan akurasi yang optimal dibandingkan metode *Naïve Bayes* lainnya (Farisi et al., 2019). Pada proses klasifikasi metode MNB terdapat permasalahan data yang tidak seimbang sehingga perlu dilakukan pendekatan data *upsampling* dengan menggunakan *Synthetic Minority Over-Sampling Technique* (SMOTE) (Ardhana et al., 2019). *Dataset* yang digunakan dalam penelitian ini merupakan *dataset* Lagu Daerah yang mewakilkan 30 Provinsi di Indonesia serta Lagu Nasional.

* 1. Rumusan Masalah

Berdasarkan latar belakang yang diuraikan, didapatkan rumusan masalah penelitian antara lain:

1. Bagaimana kinerja algoritma *Multinomial Naives Bayes* dalam melakukan klasifikasi lagu daerah dan lagu nasional?
2. Bagaimana pengaruh metode *Synthetic Minority Over-Sampling Technique* (SMOTE)terhadap performa algoritma dalam melakukan klasifikasi?
   1. Tujuan Penelitian

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

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

Dalam penelitian ini didasarkan terhadap beberapa asumsi dan keterbatasan yang membatasi masalah penelitian. Keterbatasan pada penelitian ini antara lain:

1. *Dataset* yang digunakan pada penelitian ini adalah 570 lagu yang terdiri dari 480 lagu daerah yang mewakilkan 30 provinsi dan 90 lagu nasional.
2. Metode yang digunakan pada penelitian ini adalah algoritma *Multinomial Naive Bayes.*
3. Tahap *preprocessing* pada penelitian ini tidak menggunakan *stamming* dan *stopwords* *removal* karena penelitian ini merupakan bagian dari penelitian besar untuk kamus bahasa sehingga setiap kata pada *dataset* dianggap penting.
   1. Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini dapat dilihat dari sudut pandang teoritis dan sudut pandang praktis.

1. Manfaat Teoritis

Hasil dari penelitian ini diharapkan dapat digunakan sebagai referensi untuk penelitian-penelitian terkait klasifikasi dokumen dengan metode *text* *mining*.

1. Manfaat Praktis

Hasil dari penelitian ini diharapkan dapat memberi kesempatan untuk menambah wawasan dan pengetahuan masyarakat dalam klasifikasi lagu daerah dan lagu nasional.

* 1. Definisi Operasional

Untuk menghindari perbedaan penafsiran istilah yang digunakan, maka penelitian ini menuliskan 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

LANDASAN TEORI

Lagu daerah adalah lagu yang ada pada daerah tertentu dengan bentuk yang sangat sederhana dan menggunakan bahasa daerah tersebut. Lagu daerah memiliki banyak makna, seperti makna dari kebiasaan kehidupan sehari-hari suatu kelompok masyarakat yang tinggal di daerah tersebut. Lagu daerah sangat mudah untuk dipahami dan diterima dalam kalangan masyarakat. Lagu daerah biasanya diperdengarkan pada tradisi-tradisi tertentu, seperti pada saat hiburan rakyat, pesta rakyat, acara adat dan lain sebagainya (Fadillah & Wahyuni, 2021).

Lagu nasional adalah lagu-lagu berbahasa Indonesia yang berisi tentang aspek kehidupan bangsa Indonesia (Witantina et al., 2020). Lagu nasional diciptakan untuk menumbuhkan nasionalisme, kepahlawanan dan mengorbankan semangat juang bangsa (Setiadi, 2019). Lagu nasional sudah seharusnya dapat dibawakan dengan benar, terutama lagu Indonesia Raya (Sekawael et al., 2022). Menurut kamus bahasa Indonesia, lagu adalah ragam nada atau suara yang berirama. Sedangkan nasional artinya bersifat kebangsaan yang berasal dari bangsa sendiri. Jadi, lagu nasional dapat diartikan sebagai ragam nada atau suara yang berirama, bersifat kebangsaan dan berasal dari bangsa sendiri.

*Text mining* adalah salah satu variasi dari *data mining* dan merupakan teknik yang dapat digunakan untuk melakukan klasifikasi dengan cara menemukan pola-pola menarik dari sekumpulan besar data tekstual (Foristek et al., 2020). Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan data (Ernawati, 2019). Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur . Berdasarkan struktur data teks yang tidak teratur, maka proses *text mining* memerlukan beberapa tahapan awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur (Joergensen Munthe et al., 2022).

Metode yang digunakan dalam *text mining* banyak jenisnya yaitu *Support Vector Machines* (SVM), *K-Nearest Neighbours* (KNN), *Naive Bayes*, dan *Decision Trees* (Pohon Keputusan). *Text mining* berbasis probabilitas adalah *Naive Bayes* (Catur & Ika, 2013), berbasis nilai/jarak tetangga terdekat adalah KNN (Kasanah et al., 2019), berbasis *kernel* adalah SVM (Mutawalli et al., 2019) dan berbasis jumlah pohon adalah *Decision Tree* (Hafizan & Putri, 2020)*.* Setiap metode ini memiliki karakteristik berbeda-beda.

Algoritma *Naive Bayes* adalah algoritma klasifikasi menggunakan probabilitas dan statistik berdasarkan teorema *Bayes* (Imandasari et al., 2019). Salah satu kelebihan *Naive Bayes* adalah tingkat akurasi yang tinggi dengan perhitungan sederhana. Asumsi bahwa semua fitur bersifat independen membuat algoritma *Naive Bayes* sangat cepat dibandingkan dengan algoritma yang rumit . Di sisi lain, asumsi bahwa semua fitur independen biasanya tidak terjadi di kehidupan nyata sehingga membuat algoritma *Naive Bayes* kurang akurat dibandingkan algoritma yang rumit. *Naive Bayes* memiliki beberapa jenis yaitu *Gaussian Naive Bayes, Multinomial Naive Bayes* dan *Bernoulli Naive Bayes* (Saputro et al., 2019). *Mulinomial Naive Bayes* merupakan model pengembangan dari algoritma bayes yang cocok dalam pengklasifikasian teks atau dokumen. *Multinomial Naive Bayes* adalah salah satu metode yang dipakai dengan memperhitungkan frekuensi masing-masing kemunculan *token* atau kata dalam sebuah dokumen dan probabilitas (Rahman & Doewes, 2017).

SMOTE atau *Synthetic Minority Over-Sampling Technique* adalah teknik yang dapat digunakan untuk mengatasi ketidakseimbangan data. Teknik SMOTE yang diusulkan oleh Chawla (Safitri & Muslim, 2020) tidak hanya menduplikasi data yang sama melainkan akan membuat sampel baru yang menyerupai data asli dari kelas minoritas (replikasi) untuk menyeimbangkan *dataset*. Data dari hasil replikasi tersebut dikenal dengan data sintetis. Penelitian yang telah dilakukan sebelumnya (Hairani et al., 2020) terkait menangani keseimbangan data dalam klasifikasi teks menyimpulkan bahwa semua *classifier* mendapatkan manfaat dari teknik *oversampling* dengan SMOTE sebagai salah satu teknik yang menghasilkan performa terbaik. Oleh karena itu, pada penelitian ini digunakan teknik SMOTE untuk menyeimbangkan *dataset*.

BAB III

METODELOGI PENELITIAN

3.1 Business Understanding

Tahapan *business understanding* dilakukan dengan memahami tujuan dari penelitian ini. Klasifikasi lagu daerah dan lagu nasional dilakukan berdasarkan asal daerah yaitu provinsi di Indonesia. Berdasarkan Rancangan Undang-undang Daerah Otonomi Baru (RUU DOB) tahun 2022 yang diusulkan oleh DPR, saat ini Indonesia memiliki 37 Provinsi. Asal dari lagu daerah yang dipilih untuk penelitian ini hanya 30 Provinsi saja, dikarenakan Provinsi Kepulauan Riau, Sulawesi Barat, Maluku Utara, Papua Barat, Papua Tengah, Papua Selatan, dan Papua Pegunungan merupakan provinsi baru yang belum memiliki lagu daerah.

3.2 Data Understanding

Tahapan *data understanding* dilakukan dengan mengumpulkan data lagu daerah dan lagu nasional serta penggabungan data. Pengumpulan data dilakukan dengan metode studi literatur yang bersumber dari buku lagu daerah, buku lagu nasional, berita dan artikel yang ada di internet. Data yang sudah dikumpulkan akan digabungkan dalam satu dokumen *Ms.Excel* dan dibagi berdasarkan asal daerah dari lagu tersebut serta sumber pengambilan data.

Data lirik lagu yang sudah dikumpulkan sejumlah 570 lagu yang terdiri dari 480 lagu daerah yang mewakili 30 provinsi dan 90 lagu nasional. Kemudian label pada data dibagi menjadi 3 yaitu dengan 2 label, dengan 4 label, dan dengan 31 label. Pembagian 2 label dibagi berdasarkan jenis lagu yang ada pada data yaitu Lagu Daerah dan Lagu Nasional. Pembagian 4 label dibagi berdasarkan pembagian waktu di Indonesia yaitu Indonesia bagian Barat, Indonesia bagian Tengah, dan Indonesia bagian Timur, serta ditambahkan Lagu Nasional. Pembagian 31 label dibagi berdasarkan 30 serta ditambahkan satu label Lagu Nasional. Skenario pembagian label dapat dilihat pada Tabel 3.1.

**Tabel 3.1.** Pembagian Label *Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| **Skenario 1** | **Skenario 2** | **Skenario**  **3** | **Total Data** |
| Lagu Daerah | Barat | Aceh | 10 |
| Bangka Belitung | 10 |
| Banten | 10 |
| Bengkulu | 17 |
| Jakarta | 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 |
| Yogyakarta | 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 |

3.3 Data Preparation

Sebelum dilakukan klasifikasi, data perlu diproses terlebih dahulu. Teknik yang digunakan adalah *text* *preprocessing* yang bertujuan untuk mempersiapkan teks yang terdapat dalam *dataset* agar memperoleh hasil yang optimal (Haddi et al., 2013). *Text* *preprocessing* untuk penelitian ini hanya dilakukan 4 tahapan yaitu *cleansing, casefolding, tokenizing*, dan *indexing with term frequency* (TF). Tahapan *stopwords removal* dan *stemming* tidak digunakan pada penelitian ini, dikarenakan semua informasi dan kata yang ada pada *dataset* dianggap penting untuk pengklasifikasian pada penelitian ini.

**3.3.1 Cleansing**

*Data cleansing* bertujuan untuk menghilangkan karakter *non-alphabet*, seperti tanda tanya (?), tanda koma (,), tanda elipsis (…), dan lain sebagainya.Dalam tahap kerjanya, *dataset* dibersihkan melalui beberapa proses seperti mengisi nilai yang hilang, menghaluskan *noisy data*, dan menyelesaikan inkonsistensi yang ditemukan (Squicciarini et al., 2017). Hasil dari proses *data cleansing* ditampilkan pada Tabel 3.2.

**Tabel 3.2**. Hasil *Data Cleansing*

|  |  |
| --- | --- |
| **INPUT** | **OUTPUT** |
| Bungong jeumpa, Bungong jeumpa, Meugah di Aceh…  Bungong teuleubeh, teuleubeh, Indah lagoina… | Bungong jeumpa Bungong jeumpa Meugah di Aceh  Bungong teuleubeh teuleubeh Indah lagoina |

**3.3.2 Case Folding**

*Case folding* bertujuan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter* (Jumeilah, 2017). Hasil dari proses *case folding* ditampilkan pada Tabel 3.3.

**Tabel 3.3.** Hasil *Case Folding*

|  |  |
| --- | --- |
| **INPUT** | **OUTPUT** |
| Bungong jeumpa Bungong jeumpa Meugah di Aceh  Bungong teuleubeh teuleubeh Indah lagoina | bungong jeumpa bungong jeumpa meugah di aceh  bungong teuleubeh teuleubeh indah lagoina |

**3.3.3 Tokenizing**

*Tokenizing* bertujuan untuk memotong sebuah teks menjadi kata oleh tanda baca atau spasi, sehingga yang didapatkan hanya kata tunggal saja. Kumpulan dari beberapa kata tunggal disebut dengan *token*. *Token* digunakan sebagai *input* untuk proses selanjutnya (Sabrani et al., 2020). Hasil dari proses *tokenizing* ditampilkan pada Tabel 3.4.

**Tabel 3.4**. Hasil *Tokenizing*

|  |  |
| --- | --- |
| **INPUT** | **OUTPUT** |
| bungong jeumpa bungong jeumpa meugah di aceh  bungong teuleubeh teuleubeh indah lagoina | [‘bungong’,’jeumpa’,’bungong’, ’jeumpa’,’meugah’,’di’,’aceh’, ‘bungong’,’teuleubeh’,’teuleubeh’, ‘indah’,’lagoina’] |

**3.3.4 Term Frequency (TF)**

Selanjutnya, 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 pada proses penelitian ini. Nilai TF dapat digunakan untuk menentukan letak kelas dari kata yang sama pada beberapa kelas dengan melihat nilai TF yang lebih tinggi paling mempengaruhi identifikasi kelas (Weggenmann & Kerschbaum, 2018). Masing-masing frekuensi kemunculan salah satu kata di setiap kelas dapat dilihat pada Tabel 3.5.

**Tabel 3.5.** *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 |

**3.3.5 SMOTE**

Jumlah data yang telah dikumpulkan didapatkan hasil yang tidak seimbang. Metode yang digunakan untuk menangani *dataset* yang tidak seimbang yaitu dengan *undersampling* dan *oversampling*. Metode yang digunakan adalah metode *oversampling* karena metode ini tidak menghapus *instance* dari *dataset* yang mungkin membawa beberapa informasi penting. Metode *oversampling* yang digunakan adalah *Synthetic Over-Sampling Technique* (SMOTE). SMOTE dilakukan dengan tujuan untuk mengetahui pengaruhnya terhadap performa dari metode *Multinomial Naïve Bayes*. Dapat dilihat pada Gambar 3.1 bahwa jumlah data di setiap kelas label berbeda dan jarak nilai antara kelas minimum dan maksimum sangat jauh sehingga perlu digunakan metode tersebut.

**Gambar 3.1**. Jumlah Data Setiap Kelas Sebelum SMOTE

Tahapan SMOTE diawali dengan menghitung perbedaan jumlah data antara kelas mayoritas dan kelas minoritas. Lalu dilakukan perhitungan persentase duplikasi yang diinginkan pada kelas minoritas. Kemudian dilakukan pemilihan jumlah k. Pada penelitian ini dilakukan beberapa kali percobaan dengan nilai k, di mana batas akhir pada nilai k yang digunakan merupakan nilai maksimal jumlah data pada kelas yang paling minoritas dalam *dataset*. 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 3.2.

**Gambar 3.2.** Jumlah Data Setiap Kelas Sesudah SMOTE

3.4 Modeling

Metode klasifikasi yang digunakan pada penelitian ini adalah *Multinomial Naïve Bayes* (MNB). Pada metode MNB, sistem perhitungan kategori dokumen tidak hanya ditentukan dari munculnya suatu kata dalam satu data tetapi juga jumlah frekuensi kemunculan kata dalam seluruh data (Setianingrum et al., 2018). Frekuensi kemunculan kata pada perhitungan *Multinomial Naïve Bayes* 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 proses perhitungan *prior probability* yang bertujuan untuk menghitung probabilitas kelas asal lagu dan *post probability* yang bertujuan untuk menghitung probabilitas suatu kata yang masuk ke dalam kelas asal lagu (Mayasari & Indarti, 2022). Kelemahan pada metode ini adalah MNB tidak bisa mengukur dengan tepat jumlah data pada prediksi sehingga mengurangi nilai akurasi pada seleksi atribut, maka diperlukan pengoptimalan pada atributnya.

3.5 Evaluation

Pengujian *dataset* akan dievaluasi dengan menggunakan *k-fold cross validation*. *K-fold cross validation* digunakan untuk menghilangkan bias pada *dataset*. *K-fold cross validation* memisahkan data menjadi dua bagian, yaitu data uji dan data latih. *K-fold cross validation* bekerja dengan cara membagi data yang akan diuji dan dilatih menjadi himpunan bagian k dengan ukuran yang hampir sama (Mardiana et al., 2022). Pada penelitian ini nilai k sebesar 10, karena *10-fold cross validation* merupakan metode validasi yang paling umum digunakan dan penggunaan *fold*=10 diketahui banyak memberikan hasil yang baik (Marcot & Hanea, 2020).

Lalu perhitungan performa evaluasi kinerja model klasifikasi yang dipakai dalam penelitian ini adalah *confusion* *matrix. Confusion matrix* merepresentasikan prediksi dan kondisi sebenarnya dari data yang dihasilkan oleh metode. *Confusion Matrix* dapat dilakukan perhitungan nilai akurasi, presisi, dan *recall* dari setiap label.

BAB IV

HASIL DAN PEMBAHASAN

Pada penelitian ini *dataset* akan diuji ke dalam enam skenario di mana masing-masing skenario memiliki perbedaan pada 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. |

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 jumlah label Lagu Daerah sebesar 480 karena dilakukan pengujian dengan menggunakan teknik SMOTE dan nilai k SMOTE sebesar 2, 3, 13, 22, 33, 42, 53, 62, 73, 89. Nilai k ditentukan secara acak hingga batas maksimum yang merupakan jumlah data pada label minimum sebelum SMOTE yaitu 90 dikurangi 1 menurut (Chawal et al., 2002).

Skenario lainnya memiliki penjelasan tabel yang sama seperti skenario 1 dan 2, perbedaan pembagian label yang terletak pada label Lagu Daerah dengan nilai k pada teknik SMOTE secara acak dan memiliki batas akhir berbeda-beda tergantung pada jumlah data dari nilai minimal label masing-masing skenario pembagian label.

Hasil perbandingan klasifikasi pada skenario 1 dan 2 dapat dilihat pada Tabel 4.2. Tabel 4.2 menunjukkan hasil pengujian dengan dan tanpa menggunakan teknik SMOTE. Pengujian dengan menggunakan teknik SMOTE diuji dengan nilai k= 2, 3, 13, 22, 33, 42, 53, 62, 73, 89 dan ditampilkan dengan *confusion matrix* yang terdiri dari akurasi, presisi, dan *recall*. Dari hasil pengujian tersebut disimpulkan bahwa nilai performa klasifikasi yang terbaik dengan teknik SMOTE nilai k=2.

**Tabel 4.2.** Perbandingan Performa Klasifikasi Pada Skenario 1 dan 2

|  |  |  |  |
| --- | --- | --- | --- |
| Nilai k | Akurasi (%) | Presisi (%) | *Recall* (%) |
| - | 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 |

Selanjutnya hasil perbandingan klasifikasi pada skenario 3 dan 4 dengan *dataset* sejumlah 4 label dapat dilihat pada Tabel 4.3. Tabel 4.3 menunjukkan pengujian menggunakan teknik SMOTE dengan nilai k = 2, 3, 13, 22, 33, 42 dan tanpa menggunakan teknik SMOTE yang didapatkan hasil nilai akurasi, presisi, dan *recall* terbaik adalah nilai k = 33.

**Tabel 4.3.** Perbandingan Performa Klasifikasi Pada Skenario 3 dan 4

|  |  |  |  |
| --- | --- | --- | --- |
| Nilai k | Akurasi (%) | Presisi (%) | *Recall* (%) |
| - | 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 |

Lalu hasil perbandingan nilai klasifikasi pada skenario 5 dan 6 yang membagi *dataset* menjadi 31 label dapat dilihat pada Tabel 4.4. Tabel 4.4 menunjukkan bahwa pengujian SMOTE dengan nilai k = 2,3 dan tanpa menggunakan SMOTE didapatkan hasil nilai akurasi, presisi, dan *recall* yang terbaik adalah nilai k = 3.

**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 dari nilai akurasi, presisi, dan *recall* terbaik dari pengujian seluruh skenario pada label yang berbeda dapat dilihat pada Tabel 4.5.

**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 hasil pengujian ini, diketahui bahwa terdapat beberapa faktor yang mempengaruhi kinerja klasifikasi metode MNB. Faktor pertama adalah jumlah label, di mana jika tanpa menggunakan SMOTE semakin sedikit jumlah label maka semakin tinggi performa akurasi. Hal tersebut disebabkan dengan perhitungan metode klasifikasi MNB yang menghitung jumlah frekuensi kemunculan kata pada label Lagu Daerah dan Lagu nasional dengan mudah karena kata-kata pada lagu daerah yaitu bahasa daerah yang sangat berbeda dengan kata-kata pada lagu nasional yaitu bahasa Indonesia.

Faktor kedua adalah penggunaan SMOTE, di mana pada *dataset* lagu daerah dan lagu nasional yang dilakukan pengujian dengan menggunakan teknik SMOTE mendapatkan hasil akurasi, presisi dan *recall* 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 biasdata.

Faktor lainnya adalah nilai k SMOTE pada skenario pengujian yang sebenarnya tidak terlalu mempengaruhi performa MNBdalam 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 |
| *Prior* | 0,214 | 0,073 | 0,25 | 0,25 |
| *Post* 1 "jangan" | 5,51 x 10-4 | 1,46 x 10-4 | 3,59 x 10-4 | 2,01 x 10-3 |
| *Post* 1 "mama" | 3,08 x 10-4 | 3,4 x 10-4 | 3,79 x 10-4 | 7,39 x 10-4 |
| *Post* 1 "marah" | 1,76 x 10-4 | 9,74 x 10-5 | 8,21 x 10-5 | 1,72 x 10-4 |
| *Post* 1 "beta" | 2,42 x 10-4 | 9,25 x 10-4 | 1,64 x 10-4 | 2,08 x 10-3 |
| *Post* 2 | 1,55 x 10-14 | 3,26 x 10-16 | 4,57 x 10-16 | 1,32 x 10-13 |

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 tersebut dapat dilihat bahwa sebelum dilakukan SMOTE, frekuensi dan *post probability* 1 kata “jangan”, “mama”, “marah” lebih besar di label Tengah kecuali kata “beta” yang ditempatkan benar di label Timur. Kemudian saat 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 probabilit*y 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” dikarenakan *post probability* 2 merupakan perhitungan akumulasi dari *prior probability* dan *post probability* 1.

Berdasarkan penjelasan tersebut, teknik SMOTE diketahui dapat meningkatkan performa metodeMNB 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.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pengujian yang telah dilakukan dapat disimpulkan bahwa algoritma *Naïve Bayes Multinomial* dapat digunakan untuk mengklasifikasi lagu daerah dan lagu nasional berdasarkan asal daerahnya. Skenario yang tidak menggunakan SMOTE didapatkan hasil dengan akurasi yang lebih baik pada skenario 1 yang memiliki 2 label. Penggunaan metode 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.

5.2 Saran

Pengujian yang telah dilakukan saat ini hanya berdasarkan lirik lagu. Penelitian selanjutnya dalam upaya meningkatkan hasil dari akurasi dapat dipertimbangkan aspek lainnya seperti ditambahkannya audio dari lagu tersebut. Pada penelitian selanjutnya dapat menggunakan *data preprocessing* seperti menambahkan proses *stemming* dan *stopword removal*. Selain itu, dapat juga menggunakan metode *naïve bayes* yang berbeda dan *dataset* yang berbeda juga.

**DAFTAR PUSTAKA**

Abbas, M., Ali Memon, K., & Aleem Jamali, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network Security*, *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

Catur, S., & Ika, N. D. (2013). Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naïve Bayes. *Simantik 2013*, *2013*(November), 156–160.

Ernawati, I. (2019). Naive Bayes Classifier dan Support Vector Machine Sebagai Alternatif Solusi Untuk Text Mining. *Jurnal Teknologi Informasi Dan Pendidikan*, *12*(2), 33–39. https://doi.org/10.24036/TIP.V12I2.219

Fadillah, S., & Wahyuni, S. (2021). Peningkatan Self-Awareness Anak Usia 5-6 Tahun Melalui Pembelajaran Lagu Daerah Riau. *PERNIK : Jurnal Pendidikan Anak Usia Dini*, *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*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012024

Foristek, J. I., Santi, D., Nangi, J., & Ransi, N. (2020). Implementasi Naïve bayes Clasifier dalam Klasifikasi Jenis Berita. *Foristek*, *10*(1), 20–25. https://doi.org/10.54757/FS.V10I1.52

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-Processing in Sentiment Analysis. *Procedia Computer Science*, *17*, 26–32. https://doi.org/10.1016/J.PROCS.2013.05.005

Hafizan, H., & Putri, A. N. (2020). Penerapan Metode Klasifikasi Decision Tree pada Status Gizi Balita di Kabupaten Simalungun. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *1*(2), 68–72. https://doi.org/10.30645/kesatria.v1i2.23

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.

Imandasari, T., Irawan, E., Perdana Windarto, A., & Wanto, A. (2019). Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, *1*(0), 750–761. https://doi.org/10.30645/SENARIS.V1I0.81

Joergensen Munthe, C. E., Astuti Hasibuan, N., & Hutabarat, H. (2022). RESOLUSI: Rekayasa Teknik Informatika dan Informasi Penerapan Algoritma Text Mining Dan TF-RF Dalam Menentukan Promo Produk Pada Marketplace. *Media Online*, *2*(3), 110–115. https://djournals.com/resolusi

Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(1), 19–25. https://doi.org/10.29207/resti.v1i1.11

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

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 2020 36:3*, *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

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto). *Jurnal Informatika Dan Rekayasa Elektronik*, *2*(2), 43. https://doi.org/10.36595/jire.v2i2.117

Rahman, A., & Doewes, A. (2017). Online News Classification Using Multinomial Naive Bayes. *ITSMART: Jurnal Ilmiah Teknologi Dan Informasi*, *6*(1), 32–38.

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

Sekawael, L., Salamor, L., Bakker, R., Studi PPKn, P., & Pattimura, U. (2022). Analisis Pemahaman Siswa Terhadap Nilai Nasionalisme dalam Lagu Indonesia Raya. *PEDAGOGIKA: Jurnal Pedagogik Dan Dinamika Pendidikan*, *10*(1), 1–11. https://doi.org/10.30598/PEDAGOGIKAVOL10ISSUE1PAGE1-11

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

Weggenmann, B., & Kerschbaum, F. (2018). *SynTF: Synthetic and Differentially Private Term Frequency Vectors for PriWeggenmann, B., & Kerschbaum, F. (2018). SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining, 305–314. Retrieved from https://arxiv*. 305–314.

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.

Witantina, A., Budyartati, S., & Tryanasari, D. (2020). Implementasi Pembelajaran Lagu Nasional pada Pembelajaran SBDP di Sekolah Dasar. *Prosiding Konferensi Ilmiah Dasar*, *2*(Merdeka Belajar di Era Normal), 117–121.

Yati, N., Sofyan, F. S., & Saylendra, N. P. (2020). Peran Guru Membiasakan Menyanyikan Lagu Nasional Sebagai Upaya Pembentukan Nasionalisme Siswa. *CIVICS: Jurnal Pendidikan Pancasila Dan Kewarganegaraan*, *5*(2), 117–121. https://doi.org/10.36805/civics.v5i2.1338

**LAMPIRAN**

Untuk beberapa lampiran *pseudocode,* dan data lainnya disimpan dalam *url* berikut

<https://drive.google.com/drive/folders/1bpECEYCykm8-8BcitsMVDm_v25_D9qab?usp=sharing>