Telematika: Jurnal Informatika dan Teknologi Informasi *ISSN: 1829-667X / E-ISSN: 2460-9021*
Vol. 18, No. 1, Februari 2021, pp.1-10 *DOI:10.31515/telematika.v18i1.xxxx*

Text Classification Of Traditional and National Songs Using Naïve Bayes Algorithm

Klasifikasi Teks Lagu Daerah Dan Lagu Nasional Menggunakan Algoritma Naïve Bayes

Triyanti Simbolon[1], Aji Prasetya Wibawa[2], Ilham Ari Elbaith Zaeni[3]

[1,2,3] Departemen Teknik Elektro, Teknik Informatika, Universitas Negeri Malang

[1]tryntsmbln@gmail.com, [2*]aji.prasetya.ft@um.ac.id, [3]ilham.ari.f t@um.ac.id
**: Penulis korespondensi* (*corresponding author*)

Informasi Artikel

Received: December 2020
Revised: January 2021
Accepted: January 2021
Published: February 2021
Menggunakan style info

Abstract

This research was made with the aim of proving the multinomial naive Bayes algorithm for text classification of folk songs and national songs. The Naïve Bayes method was chosen because this method determines document categories not only by the appearance of one word in one document but in all documents so that the resulting accuracy is greater and stable than other methods. The tested data consists of 480 folk songs and 90 national songs which will be divided into 6 scenarios, namely scenarios of 2 labels, 4 labels, and 31 labels using and without using SMOTE. The dataset will go through several stages, the first stage is the preprocessing process consisting of case folding, remove punctuation, tokenizing, TF-IDF, then classification will be carried out using naive Bayes multinomial, then tested with k-fold cross validation and SMOTE resampling. Based on the results of the study, the best scenario is to use SMOTE on 2 labels with an accuracy of 93.75%. These results prove that the MNB classification works well if the data label class is small.

Keywords:traditional songs; national songs; multinomial naïve bayes; SMOTE
Kata kunci: Lagu Daerah; Lagu Nasional; Naïve Bayes Multinomial; SMOTE

Abstrak

Penelitian ini dibuat dengan tujuan membuktikan bahwa algoritma naïve bayes multinomial untuk pengklafisikasian teks pada lagu daerah dan lagu nasional. Metode Naïve Bayes dipilih karena metode ini menentukan kategori dokumen tidak hanya dengan kemunculan satu kata pada satu dokumen tetapi pada seluruh dokumen sehingga akurasi yang dihasilkan lebih besar dan stabil dibandingkan metode lainnya. Data yang diuji terdiri dari 480 lagu daerah dan 90 lagu nasional yang akan dibagi menjadi 6 skenario

yaitu skenario 2 label, 4 label, dan 31 label dengan menggunakan dan tanpa menggunakan SMOTE. Dataset akan melalui beberapa tahapan, tahapan pertama adalah proses preprocessing terdiri dari case folding, remove punctuation, tokenazing, TF-IDF, selanjutnya akan dilakukan klasifikasi dengan naïve bayes multinomial, lalu diuji dengan k-fold cross validation dan resampling SMOTE. Berdasarkan hasil penelitian, skenario terbaik adalah dengan menggunakan SMOTE pada 2 label dengan hasil akurasi 93,75%. Hasil tersebut membuktikan bahwa klasifikasi MNB bekerja dengan baik jika kelas label data sedikit.

1. Pendahuluan

Indonesia merupakan negara yang memiliki keanekaragaman suku dan budaya. Salah satu kebudayaan yang ada di Indonesia adalah lagu daerah. Pada setiap daerah memiliki ciri khas lagu yang berbeda. Ciri khas tersebut terletak pada lirik lagu yang mewakilkan bahasa dari daerah tersebut. Selain lagu daerah, Indonesia juga memiliki karya seni 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 [1]. Namun pada saat ini lagu daerah dan lagu nasional sudah jarang diperdengarkan maupun dinyanyikan oleh masyarakat. Hal tersebut mengakibatkan menurunnya pengetahuan dan minat masyarakat pada lagu daerah dan lagu nasional.

Dalam upaya mempermudah masyarakat untuk mengetahui tentang lagu daerah dan lagu nasional yang ada di indonesia maka dibuatlah sebuah klasifikasi teks pada lirik lagu. Klasifikasi teks bekerja dengan cara mengelompokkan objek berdasarkan ciri – ciri yang dimiliki oleh objek klasifikasi [2]. Metode-metode yang digunakan dalam klasifikasi teks antara lain *Naïve Bayes* yang berbasis pada probabilitas [3]*, Support Vector Machine* yang berbasis pada kernel [4]*, K-Nearest Neighbor* yang berbasis pada nilai/jarak tetangga terdekat [5]*,* dan *Decision Tree* yang berbasis pada jumlah pohon [6]. Pada setiap metode memiliki kelebihan dan kekurangan masing-masing.

Penelitian ini menggunakan metode *Multinomial Naïve Bayes* (MNB) yang merupakan salah satu model klasifikasi naïve bayes [7]. Metode ini bekerja dengan menerapkan teori peluang atau probabilitas, dimana algoritma akan menentukan kategori dokumen tidak hanya berdasarkan kemunculan kata pada satu dokumen melainkan dengan menghitung kemunculan kata pada seluruh dokumen. MNB dipilih karena metode ini memiliki pemodelan dengan pendekatan yang dominan dan lebih efisien dibandingkan dengan metode lainnya [8]. Tetapi dalam proses pengerjaan metode *naïve bayes multinomial* terdapat permasalahan pada ketidakseimbangan kelas data [9].

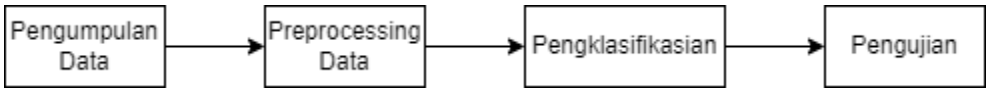
Permasalahan ketidakseimbangan kelas data dapat ditangani dengan dua pendekatan, yaitu pendekatan data dan pendekatan algoritma [10]. Pada penelitian ini dilakukan

pendekatan data dengan menggunakan *Synthetic Minority Over-Sampling Technique* (SMOTE). Penggunaan teknik SMOTE dapat menghasilkan nilai akurasi yang baik dan dengan cara kerja yang efektif. Teknik ini menambahkan kelas minoritas dengan membangkitkan data buatan atau sintesis berdasarkan k-tetangga terdekat (*k- nearest neighbor*) antar kelas minoritas [11].

Dari uraian diatas maka dapat disimpulkan bahwa metode MNB dan teknik SMOTE dapat diterapkan untuk mengklasifikasikan teks lagu daerah dan lagu nasional pada data yang tidak seimbang dan data yang seimbang. Dengan adanya klasifikasi ini, diharapkan dapat mempermudah pengelompokkan lagu daerah dan lagu nasional, serta meningkatkan efisiensi waktu ketika mencari lagu daerah dan lagu nasional berdasarkan asalnya.

2. Metode/Perancangan

Perancangan pada penelitian ini menggunakan beberapa tahapan yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, penelitian ini dimulai dengan pengumpulan data. Dataset yang digunakan pada penelitian ini berasal dari lirik lagu-lagu daerah dan lagu-lagu nasional Indonesia yang diambil dari beberapa sumber diantaranya yaitu buku lagu daerah, buku lagu nasional, artikel-artikel yang ada di internet. Dataset lirik lagu yang sudah digunakan pada penelitian ini terdiri dari 480 lagu daerah yang mewakili 30 provinsi dan 90 lagu nasional.

Tahap kedua adalah tahap preprocessing data. Tahapan ini digunakan untuk mempersiapkan teks yang terdapat dalam dataset supaya memperoleh hasil penelitian yang maksimal [12]. Umumnya terdapat beberapa tahapan preprocessing untuk dokumen teks, yaitu *remove punctuation, case folding, tokenizing, stopwords removal, stemming,* dan *indexing with term frequency (TF)*. Namun dalam penelitian ini *stopwords removal* dan *stemming* tidak digunakan karena semua informasi dan kata yang ada pada dataset dianggap penting untuk pengklasifikasian.

Remove punctuation merupakan proses menghilangkan karakter non-alphabet, seperti tanda tanya (?), tanda koma (,), tanda elipsis (…), dan lain sebagainya. Hal ini bertujuan untuk mengurangi *noise* yang dapat menimbulkan proses perhitungan dalam pengklasifikasian tidak optimal [13]. Hasil dari proses *remove punctuation* ditampilkan pada Tabel 1.

Pseudocode

```
Import library (pandas, numpy, matplotlib, pyplot, re)

    Membaca dataset .csv (dataset_final.csv) dengan library pd.read_csv
    sebagai variable df_FF
```

Mengubah isi Category menjadi Lagu Daerah dan Lagu Nasional menggunakan fungsi str.replace

Membuat variable cleaning = list
('0123456789!#$%&*()+,./;:<>+={}[]|\') dan petik_2 = list('"""""")

For cln in cleaning:
    hapus cln
    (df_FF .str.replace(cln,' '))
For ptk in petik_2:
    ubah ptk menjadi petik satu
    (df_FF.str.replace(ptk,"''))

Table 1. Hasil *Remove Punctuation*

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

Case folding merupakan proses mengubah huruf kapital menjadi huruf kecil pada lirik lagu. Hal ini dilakukan untuk mengurangi redudansi data yang akan digunakan dalam proses pengklasifikasian sehingga proses perhitungan menjadi optimal [14]. Hasil dari proses *case folding* ditampilkan pada Tabel 2.

Pseudocode

Mengubah format data pada variable df_FF menjadi lowercase menggunakan fungsi ft_FF.Text = df_FF.Text.str.lower()

Table 2. Hasil Case Folding

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

Tokenizing merupakan proses 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 [15]. Hasil dari proses *tokenizing* ditampilkan pada tabel 3.

Pseudocode

Import library (nltk, word tokenize)
    Melakukan tokenizing pada teks dengan fungsi tokenize = nltk.tokenize.word_tokenize(df_category)

Table 3. 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’]

Selanjutnya, proses pembobotan kata dengan menggunakan metode *term frequency* (TF). TF digunakan untuk mengukur berapa kali suatu kata atau frasa muncul dalam dokumen [16]. Nilai TF yang tinggi mengidentifikasikan bahwa kata tersebut penting pada proses penelitian ini. Selain itu, nilai TF juga dapat digunakan untuk menentukan letak kelas dari kata yang sama pada beberapa kelas dengan melihat nilai TF yang lebih tinggi paling mempengaruhi identifikasi kelas [17].

Tahap ketiga adalah proses klasifikasi yang bertujuan untuk mengetahui kategori kata pada suatu dokumen. Metode klasifikasi yang digunakan pada penelitian ini adalah MNB. Pada MNB, sistem perhitungan kategori dokumen tidak hanya ditentukan dari munculnya suatu kata dalam satu data tetapi juga jumlah frekuensi kemunculan kata dalam seluruh data [18]. MNB memulai prosesnya dengan menetapkan label data dan melakukan ekstraksi nilai dari label tersebut. Selanjutnya dilakukan proses perhitungan jumlah dokumen, jumlah kelas, dan jumlah kata dari seluruh data [19]. 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 [20].

Pseudocode

```
Import library (numpy, pandas, matplotlib, operator)

Melakukan training label dengan fungsi train_label = open(‘dataset’)

Ekstraksi values/nilai dari label dengan fungsi lines = train_label .readlines()

Mengambil jumlah dokumen dengan fungsi total = len(lines)

Menghitung frekuensi kemunculan tiap kelas lagu
for line in lines:
  val = int(line.plot()[0])
  pi[val] += 1

Probabilitas kelas lagu dengan total dokumen
for key in pi:
  pi[key] /= total

Menghitung probabilitas tiap kata sesuai kelas dialek
pb_ij = df.groupby([‘classIdx’,’wordIdx’])
```

```
pb_j = df.groupby(['classIdx'])
Pr = (pb_ij['count'].sum() + a) / (pb_j['count'].sum())

Melakukan smoothing
if smooth:
  probability=Pr_dict[wordIdx][classIdx]
  power = np.log(1+new_dict[docIdx][wordIdx])

Mengambil kelas dengan probabilitas tertinggi
max_score = max(score_dict, key=score_dict.get)
prediction.append(max_score)
```

Salah satu kelebihan metode ini adalah MNB dianggap sebagai pemodelan dengan pendekatan yang dominan dan lebih efisien daripada model Naïve Bayes lainnya [21]. Selain kelebihan tersebut, metode ini memiliki beberapa kelemahan salah satunya MNB sangat sensitif dalam pemilihan fitur [22]. Jika jumlah fitur yang dimiliki terlalu banyak dalam proses klasifikasi, MNB tidak hanya bekerja meningkatkan waktu penghitungan tetapi juga dapat menurunkan hasil akurasi [23].

Tahap keempat adalah pengujian dataset. Pada penelitian ini pengujian dataset dibagi menjadi enam skenario, dimana 3 skenario dilakukan dengan menggunakan teknik SMOTE dan 3 skenario lainnya tanpa menggunakan teknik SMOTE. Teknik SMOTE bekerja menyetarakan kelas minoritas dan kelas mayoritas dengan cara menduplikasikan sampel kelas minoritas secara acak. Penggunaan teknik SMOTE menghasilkan hasil yang baik dan cara yang efektif untuk menangani ketidakseimbangan kelas yang mengalami overfitting pada teknik oversampling untuk memproses kelas minoritas [24]. Pada penelitian ini 3 skenario dilakukan pembagian label yang berbeda-beda, dan jumlah dataset pada label berbeda antara kelas mayoritas dan kelas minoritas. Pada penelitian ini, skenario 1 dan 2 memiliki perbedaan 480 label pada lagu daerah dan 90 label pada lagu nasional. Nilai k pada skenario 1 dan 2 adalah 9, 19, 29, 39, 49, 59, 69, 79, 89, dimana nilai k=89 adalah nilai maksimal. Nilai k yang paling maksimal ditentukan oleh jumlah label pada kelas minoritas yaitu 90 -1 (Chawla et al., 2002) [25]. Berlaku juga dalam penjelasan nilai k teknik SMOTE pada skenario pengujian lainnya.

Pseudocode

```
Input: Jumlah minoritas kelas, Jumlah SMOTE N%, nilai dari k
Output: (N/100) * T

If N < 100
        Then random dari jumlah T minoritas
        T = (N/100)*T
        N = 100
N = (int(N/100))
        For i <- 1 to T
        Hitung k tetangga terdekat untuk i dan simpan N-array

Mengidentifikasi populasi (N, i, N-array)
```

```
Memilih jumlah random antara 1 dan k yang disebut nn.
Tahap ini memilih 1 k dari i
For attr <- 1 to jumlah atribut
        Hitung
        dif = Sample[N-array[nn][attr]] –
        Sample[i][attr]]
        gap = rand(0,1)
        Sintetis[new][attr] = Sample[i][attr] + gap *dif
new++
N = N – 1

return(populasi akhir)
```

Lalu pengujian dataset divalidasi dengan menggunakan *k-fold cross validation* dimana nilai k pada penelitian ini adalah 10. Cross validation bekerja dengan cara membagi data menjadi himpunan bagian k dengan ukuran yang hampir sama, yang dimana data pada klasifikasi yang diuji dan dilatih sebanyak k=10. Pada setiap pengulangan, data latih berasal dari salah satu himpunan dan data penguji dari sub kelompok data k lainnya [26].

3. Hasil dan Pembahasan

Pada penelitian ini dataset akan diuji dalam enam skenario dimana masing-masing skenario memiliki perbedaan pada label dan tahap SMOTE yang ditampilkan pada **Tabel 4**.

SKENARIO	LABEL		SMOTE
	LAGU NASIONAL	**LAGU DAERAH**	
1	90	480	-
2	90	480	K = 9, 19, 29, 39, 49, 59, 69, 79, 89.
3	90	Barat: 312 Tengah: 121 Timur: 42	-
4	90	Barat: 312 Tengah: 121 Timur: 42	K = 11, 21, 31, 41.
5	90	30 Provinsi Maksimal: 42 (Jawa Barat) Minimal: 4 (Kalimantan Utara)	-
6	90	30 Provinsi Maksimal: 42 (Jawa Barat) Minimal: 4 (Kalimantan Utara)	K = 2, 3.

Pada skenario 1 terdapat 2 label yaitu label lagu daerah yang memiliki dataset sebanyak 480 lagu dan label lagu nasional yang memiliki dataset sebanyak 90 lagu.

Skenario 1 diuji tanpa menggunakan teknik SMOTE. Pada skenario 2 jumlah label sama dengan skenario 1, tetapi skenario 2 diuji menggunakan teknik SMOTE dengan nilai k = 9, 19, 29, 39, 49, 59, 69, 79, 89. Pada skenario lainnya penjelasan tabel sama seperti skenario 1 dan 2, dimana pembagian label ada pada label lagu daerah dan nilai k pada teknik SMOTE yang berbeda-beda, dimana nilai k maksimal adalah jumlah data dari nilai kelas dataset minimal.

Hasil perbandingan klasifikasi pada skenario 1 dan 2 yang menguji dataset dibagi menjadi 2 label dapat dilihat pada Tabel 5. Tabel 5 menunjukkan hasil pengujian dengan dan tanpa menggunakan teknik SMOTE, dimana pengujian dengan teknik SMOTE menggunakan nilai k= 9, 19, 29, 39, 49, 59, 69, 79, 89 dengan evaluasi *10-fold Cross Validation* dan metode MNB. Dari hasil pengujian tersebut disimpulkan bahwa nilai akurasi, presisi, dan recall yang terbaik adalah dengan nilai k=9.

Tabel 5. Perbandingan Performa Klasifikasi Pada Skenario 1 dan 2

Nilai k	Akurasi (%)	Presisi (%)	Recall (%)
9	92,92	93,86	92,93
19	92,81	93,75	92,82
29	93,12	94,07	93,13
39	93,54	94,35	93,53
49	93,23	94,12	93,24
59	93,33	94,12	93,34
69	93,44	94,20	93,45
79	93,65	94,41	93,66
89	93,75	94,48	93,76
-	89,12	94,31	65,56

Selanjutnya hasil perbandingan klasifikasi pada skenario 3 dan 4 yang membagi dataset menjadi 4 label dapat dilihat pada Tabel 6. Tabel 6 menunjukan bahwa pengujian menggunakan teknik SMOTE dengan nilai k = 11, 21, 31, 41 dan tanpa menggunakan teknik SMOTE didapatkan hasil nilai akurasi, presisi, dan recall yang terbaik adalah dengan nilai k = 41.

Tabel 6. Perbandingan Performa Klasifikasi Pada Skenario 3 dan 4

Nilai k	Akurasi (%)	Presisi (%)	Recall (%)
11	91,38	92,25	91,25
21	91,46	92,15	91,41
31	91,61	92,47	91,31
41	91,77	92,57	91,79
-	65,26	50,38	40,07

Lalu hasil perbandingan nilai klasifikasi pada skenario 5 dan 6 yang membagi dataset menjadi 31 label dapat dilihat pada Tabel 7. Tabel 7 menunjukan bahwa pengujian SMOTE dengan nilai k = 2,3 dan tanpa menggunakan SMOTE didapatkan hasil nilai akurasi, presisi, dan recall yang terbaik adalah dengan nilai k = 2.

Tabel 7. Perbandingan Performa Klasifikasi Pada Skenario 5 dan 6

Nilai k	Akurasi (%)	Presisi (%)	Recall (%)
2	90,65	91,41	90,66
3	90,75	91,48	90,76
-	21,58	8,45	6,03

Hasil perbandingan dari nilai akurasi, presisi, dan recall terbaik dari pengujian seluruh skenario pada label yang berbeda dapat dilihat pada **Tabel 8**.

Table 8. Perbandingan Skenario Dengan Akurasi, Presisi, Dan Recall Terbaik

Skenario	**SMOTE**	**Akurasi (%)**	**Presisi (%)**	**Recall (%)**
2	k = 89	93,75	94,48	93,76
4	k = 41	91,77	92,57	91,79
6	k = 3	90,75	91,48	90,76

Berdasarkan hasil ini, diketahui bahwa klasifikasi dengan menggunakan metode MNB pada dataset lagu daerah dan lagu nasional dengan menggunakan teknik SMOTE menghasilkan akurasi, presisi dan recall lebih baik dibandingkan tanpa menggunakan teknik SMOTE. Hal ini mungkin disebabkan oleh meratanya persebaran kelas data yang diatasi oleh teknik SMOTE. Dimana jika dataset tidak menggunakan SMOTE terdapat perbedaan yang sangat jauh antara jumlah dataset maksimal dengan dataset minimal, contoh pada kasus skenario 31 label dimana kelas minimal ada pada Kalimantan Utara yang memiliki 4 data saja dan kelas maksimal pada Lagu Nasional yang memiliki 90 data.

Sedangkan nilai akurasi, presisi dan recall yang terbaik ada pada skenario 2 label. Hal tersebut disebabkan oleh bahasa daerah dan bahasa indonesia memiliki kosakata yang berbeda. Hal ini menyebabkan proses pengklasifikasian lebih mudah, karena setiap kelas terdiri dari data lagu yang lebih banyak. Selain itu ada faktor lainnya juga yaitu nilai k pada SMOTE, dimana dapat dilihat jika nilai k pada SMOTE semakin tinggi maka nilai akurasi akan semakin baik. Secara umum nilai k yang tinggi akan menghasilkan akurasi yang lebih baik, karena pengambilan kata-kata untuk pembentukan data sintetis yang lebih beragam pada teknik SMOTE.

4. Kesimpulan dan Saran

Berdasarkan pengujian yang telah dilakukan dapat disimpulkan bahwa algoritma Naïve Bayes Multinomial dapat digunakan untuk mengklasifikasi lagu daerah dan lagu nasional berdasarkan asal daerahnya. Penggunaan SMOTE berpengaruh untuk meningkatkan performa klasifikasi. Pengujian yang telah dilakukan saat ini hanya berdasarkan lirik lagu. Untuk kedepannya dalam upaya meningkatkan hasil dari akurasi dapat dipertimbangkan aspek lainnya seperti ditambahkannya audio dari lagu tersebut. Selanjutnya untuk penelitian selanjutnya dapat menggunakan data preprocessing seperti menambahkan proses stemming dan stopword removal. Dapat juga menggunakan model algoritma naïve bayes yang berbeda dan dataset yang berbeda juga.

5. Daftar Pustaka

[1] N. Yati, Fitri Silvia Sofyan, and Nadya Putri Syailendra, “Peran guru membiasakan

menyanyikan lagu nasional sebagai upaya pembentukan nasionalisme siswa," *Civ. J. Pendidik. Pancasila dan Kewarganegaraan*, vol. 5, no. 2, pp. 117–121, 2020, doi: 10.36805/civics.v5i2.1338.

[2] F. A. D. Aji Prasetya Wibawa, Muhammad Guntur Aji Purnama, Muhammad Fathony Akbar, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, p. 134, 2018.

[3] I. N. D. Catur Supriyanto, "Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naïve Bayes," *Simantik 2013*, vol. 2013, no. November, pp. 156–160, 2013.

[4] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto)," *J. Inform. dan Rekayasa Elektron.*, vol. 2, no. 2, p. 43, 2019, doi: 10.36595/jire.v2i2.117.

[5] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.

[6] H. Hafizan and A. N. Putri, "Penerapan Metode Klasifikasi Decision Tree Pada Status Gizi Balita Di Kabupaten Simalungun," *KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 1, no. 2, pp. 68–72, 2020, doi: 10.30645/kesatria.v1i2.23.

[7] M. Abbas, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 19, no. 3, p. 62, 2019, doi: 10.13140/RG.2.2.30021.40169.

[8] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012024.

[9] A. P. Ardhana, D. E. Cahyani, and Winarno, "Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods," *J. Phys. Conf. Ser.*, vol. 1306, no. 1, 2019, doi: 10.1088/1742-6596/1306/1/012049.

[10] B. Santoso, H. Wijayanto, K. A. Notodiputro, and B. Sartono, "Synthetic over Sampling Methods for Handling Class Imbalanced Problems : A Review," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 58, no. 1, Apr. 2017, doi: 10.1088/1755-1315/58/1/012031.

[11] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020.

[12] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, no. December, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[13] A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment analysis during Hurricane Sandy in emergency response," *Int. J. Disaster Risk Reduct.*, vol. 21, pp. 213–222, 2017, doi: 10.1016/j.ijdrr.2016.12.011.

[14] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[15] A. Sabrani, I. G. W. Wedashwara W., and F. Bimantoro, "Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 2, no. 1, pp. 89–100, 2020, doi: 10.29303/jtika.v2i1.87.

[16] C. N. Harahap, G. I. Marthasari, and N. Hayatin, "Perbandingan Klasifikasi Berita Hoax Kategori Kesehatan Menggunakan Naïve Bayes dan Multinomial Naïve Bayes," *J. Repos.*, vol. 3, no. 4, pp. 419–424, 2021, [Online]. Available: http://repositor.umm.ac.id/index.php/repositor/article/view/1380

[17] B. Weggenmann and F. Kerschbaum, "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," pp. 305–314, 2018.

[18] A. H. Setianingrum, D. H. Kalokasari, and I. M. Shofi, "Implementasi Algoritma Multinomial Naive Bayes Classifier," *J. Tek. Inform.*, vol. 10, no. 2, pp. 109–118, 2018, doi: 10.15408/jti.v10i2.6822.

[19] L. Mayasari and D. Indarti, "Klasifikasi Topik Tweet Mengenai Covid Menggunakan Metode Multinomial Naïve Bayes Dengan Pembobotan Tf-Idf," *J. Ilm. Inform. Komput.*, vol. 27, no. 1, pp. 43–53, 2022, doi: 10.35760/ik.2022.v27i1.6184.

[20] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.

[21] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, pp. 257–261, Mar. 2017, doi: 10.1109/NGCT.2016.7877424.

[22] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009, doi: 10.1016/J.ESWA.2008.06.054.

[23] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Syst.*, vol. 36, pp. 226–235, Dec. 2012, doi: 10.1016/J.KNOSYS.2012.06.005.

[24] S. A. Putri, "Integrasi Teknik Smote Bagging Dengan Information Gain Pada Naive Bayes Untuk Prediksi Cacat Software," *J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 2, no. 2, pp. 22–31, 2017.

[25] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[26] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak," *Bimaster*, vol. 11, no. 1, pp. 97–102, 2022.