MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 8 | No. 2 | Halaman 254 - 266
DOI : https://doi.org/10.26760/mindjournal.v8i2.254-266 Desember 2023

Deteksi Seksisme Online menggunakan Support Vector Machine dan Naïve Bayes

DIYANK SHABIRA[1], SARIFUDDIN MADENDA[2], AL HAFIZ AKBAR MAULANA SIAGIAN[3], SLAMET RIYANTO[4]

[1,2]Program Studi Magister Manajemen Sistem Informasi, Universitas Gunadarma
[3,4]Badan Riset dan Inovasi Nasional
Email: diyank.shabira@gmail.com

Received 24 September 2023 | *Revised* 21 November2023 | *Accepted* 7 Desember 2023

ABSTRAK

Seksisme online menjadi topik penting di media sosial yang mempengaruhi perkembangan internet, menimbulkan efek negatif dan menjadi ancaman serius bagi wanita yang menjadi target. Penelitian ini menggunakan *machine learning* untuk mendeteksi seksisme pada kalimat bahasa Inggris. Algoritma yang digunakan adalah *Support Vector Machine* dan *Naive Bayes*. *Grid search* diterapkan pada model untuk mencari kombinasi *hyperparameter* terbaik sehingga menghasilkan skor terbaik. Pelatihan dibagi menjadi dua tugas, yaitu (1) pelatihan model menggunakan data tanpa penanganan *imbalanced* dan (2) pelatihan model menggunakan data yang telah dilakukan SMOTE. Hasil dari pelatihan model menunjukkan model SVM+SMOTE menghasilkan rata-rata skor F1 terbaik paling tinggi yaitu sebesar 0,96. Pengujian menggunakan data uji menunjukkan model SVM+SMOTE menghasilkan skor F1 tertinggi, yaitu sebesar 0,90 dengan 1467 kalimat diklasifikasikan benar 'not sexist', 47 kalimat 'not sexist' diklasifikasikan sebagai 'sexist', 189 kalimat 'sexist' diklasifikasikan benar dan 297 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'.

***Kata kunci**: Seksisme, Deteksi, SVM, Naive Bayes, SMOTE*

ABSTRACT

Online sexism has become a significant issue on social media, impacting internet progress and posing a serious threat to targeted women. This research uses machine learning to detect sexism in English sentences. The algorithms used are Support Vector Machine and Naive Bayes. Grid search is applied in the model to find the best combination of hyperparameters to produce the best score. The training is divided into two tasks: (1) training the model using unhandle the imbalanced data and (2) training the model using data with SMOTE. The training results show that the SVM+SMOTE model produces the highest average best F1 score is 0.96. The testing results show that the SVM+SMOTE model produces the highest F1 score is 0.90 with 1467 sentences correctly classified as 'not sexist', 47 'not sexist' sentences classified as 'sexist', 189 sentences classified as 'sexist' correctly and 297 'sexist' sentences were classified as 'not sexist'.

***Keywords**: Sexism, Detection, SVM, Naive Bayes, SMOTE*

1. PENDAHULUAN

Media sosial memudahkan setiap orang untuk menemukan, berkenalan, dan berinteraksi dengan orang lain sehingga membuat setiap orang bebas (hampir tanpa batas) berbagi informasi **(Appel et al., 2020)**. Kemudahan komunikasi ini dapat menjadi dampak positif yang memudahkan setiap orang untuk memperoleh informasi aktual tanpa harus menunggu lama, namun juga menjadi dampak negatif diantaranya menyebabkan terjadinya *cyber bullying*. Menurut Kamus Cambridge **(Cambridge, 2023)**, *cyber bullying* merupakan aktivitas penindasan di dunia maya berupa tindakan menyakiti seseorang secara langsung atau tidak langsung, sadar atau tidak sadar dengan menyampaikan konten, komentar atau pesan yang menyebabkan orang lain merasa tidak nyaman. *Cyber bullying* dapat terjadi pada siapapun, namun faktanya sebagian besar korban kasus *cyber bullying* adalah wanita sehingga mengarah pada tindakan seksisme yang menjurus pada isu ketimpangan gender **(Napier et al., 2020).**

Seksisme adalah paham atau bentuk prasangka negatif terhadap kelompok lain karena perbedaan gender yang umumnya menyerang wanita dan cenderung menjurus ke tindakan diskriminasi. Di dalam paham seksisme, wanita sering diabaikan hak dan kemampuannya, dinilai lemah, dan lebih rendah posisinya dibanding pria. Saat ini tidak dapat dipungkiri bahwa media massa adalah salah satu aspek yang berpengaruh terhadap isu seksisme, terutama media online seperti situs jejaring sosial, forum, dan video games **(Fox et al., 2015)**.

Seksisme online telah menjadi topik penting di media sosial yang mempengaruhi perkembangan internet dan dapat menimbulkan efek negatif. Pendeteksian seksisme online menjadi sangat penting karena seksisme online telah menjadi ancaman serius yang dapat menyebabkan kerugian bagi wanita yang menjadi target, merusak kesan ramah lingkungan di ruang online, serta menciptakan ketidakseimbangan sosial dan ketidakadila **(Kirk et al., 2023)**. Sebuah penelitian bersama Young Women's Trustand dari University College London, menemukan bahwa wanita muda yang menjadi sasaran seksisme kemungkinan besar mengalami depresi klinis **(Rezkisari, 2019)**.

Pendeteksian seksisme online dilihat sebagai permasalahan klasifikasi dalam menentukan sebuah kalimat dikategorikan 'sexist' atau 'not sexist'. Permasalahan klasifikasi dapat diselesaikan dengan *machine learning*. *Machine learning* adalah bidang ilmu komputer yang memberikan pembelajaran kepada komputer untuk mencoba mengikuti bagaimana cara kerja manusia atau makhluk cerdas belajar dan menggeneralisasi **(Tanaka & Okutomi, 2014)**. Studi terkait pendeteksian seksisme telah dilakukan oleh peneliti terdahulu. Penelitian yang dilakukan oleh Kumar, et al. (2021) dalam penelitiannya melakukan deteksi seksisme pada tweet berbahasa Inggris dan Spanyol dengan membagi dua tugas dengan perbedaan kriteria. Tugas 1 merupakan klasifikasi biner membandingkan algoritma SVM, *Random Forest* dan LSTM. Tugas 1 memperoleh akurasi terbaik pada algoritma *Random Forest* sebesar 0,7115. Tugas 2 merupakan multiklasifikasi dengan membagi 6 kategori kelas (5 kelas kategori seksisme dan 1 non seksisme) dengan menggunakan dua algoritma yaitu SVM dan *Random Forest*. Tugas 2 memperoleh akurasi terbaik pada algoritma SVM sebesar 0,5923 **(Kumar et al., 2021)**.

Penelitian yang dilakukan Chiril, et al. (2020) tentang korpus beranotasi untuk deteksi seksisme pada tweet berbahasa Prancis menggunakan beberapa model klasifikasi yaitu SVM BoW, SVM BoW+URL/emoji, CNN, CNN+LSTM, BiLSTM *with attention* dan BERT. Penelitian tersebut memperoleh akurasi tertinggi dengan algoritma BERT sebesar 0,790 dengan akurasi per kelas yaitu non seksisme 0,843 dan seksisme 0,682 **(Chiril et al., 2020)**.

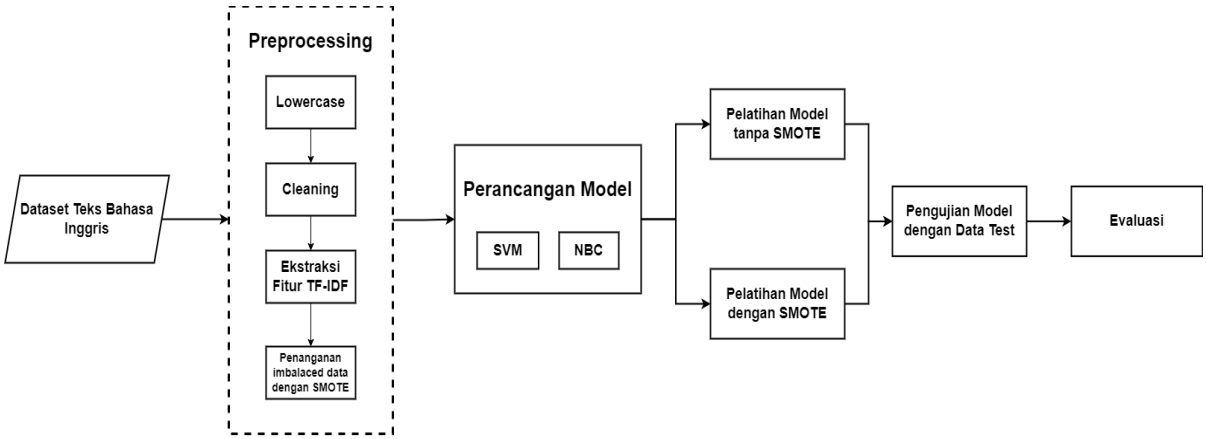
Penelitian Asogwa, et al. (2022) melakukan deteksi dan klasifikasi pada *hate messages* dengan menggunakan algoritma *Support Vector Machine* dan *Naive Bayes*. Dataset dikumpulkan dari *Unique Client Identifier* (UCI) secara online. Dataset diberi dua label yaitu 'non-offensive' dan 'offensive'. Hasil akurasi dari model SVM adalah 99,37% dan NB adalah 50,00%. Klasifikasi dengan SVM memberikan akurasi terbaik dibandingkan *Naive Bayes* **(Asogwa et al., 2022)**.

Berdasarkan permasalahan yang dijelaskan, penelitian ini mencoba merancang sistem yang mampu mendeteksi seksisme menggunakan metode *machine learning.* Algoritma *machine learning* yang digunakan adalah *Support Vector Machine* (SVM) dan *Naïve Bayes* dengan optimasi *hyperparameter*. Algoritma SVM dan *Naive Bayes* digunakan untuk menghasilkan model. Penelitian ini menerapkan metode *grid search* dalam melakukan optimasi *hyperparameter* untuk memperoleh optimasi terbaik dan *best score* dari model yang dibangun **(Fayed & Atiya, 2019)**.

Penelitian ini bertujuan untuk menghasilkan model yang digunakan untuk mendeteksi seksisme online. Penelitian ini melakukan dua tugas dengan kriteria berbeda. Tugas 1 melakukan pelatihan dengan dua model berbeda menggunakan algoritma SVM dan *Naive Bayes* dengan data tanpa penanganan *imbalanced* pada kelas sedangkan tugas 2 merupakan pelatihan dengan dua model berbeda menggunakan algoritma SVM dan *Naive Bayes* dengan menerapkan SMOTE dalam mengatasi *imbalanced* data pada kelas **(Xu et al., 2020)**. Kemudian model yang dibangun akan dievaluasi dengan *confusion matrix* dengan mengetahui skor F1 sebagai parameter evaluasi.

2. METODE PENELITIAN

Penelitian ini mampu mendeteksi kalimat yang mengandung unsur seksisme dalam kalimat berbahasa Inggris. Sistem ini menggunakan metode *Support Vector Machine* dan *Naïve Bayes* untuk mengklasifikasikan kalimat menjadi kategori "Sexist" atau "Not Sexist".



Gambar 1. Metode Penelitian

Gambar 1 menampilkan alur dari metode penelitian. Penelitian ini menggunakan data sekunder publik berupa kalimat bahasa Inggris sebanyak 16000 data yang diambil dari SemEval-2023. Dataset terdiri dari 14000 data latih dan 2000 data uji. Data sudah dilakukan pelabelan oleh tiga orang annotator terlatih dan divalidasi oleh dua orang pakar. Data dikategorikan dalam dua kelas yaitu "sexist" dan "not sexist". Kelas 'not sexist' diberi label 0 dan kelas 'sexist' diberi label dengan 1. Proporsi pembagian data pada masing-masing dataset dapat dilihat pada Tabel 1 untuk proporsi data latih dan Tabel 2 untuk proporsi data uji.

Tabel 1. Proporsi Data Latih

Label	Total
Not Sexist	10602
Sexist	3398

Tabel 2. Proporsi Data Uji

Label	Total
Not Sexist	1514
Sexist	486

Tabel 1 menampilkan pembagian data yang digunakan untuk pelatihan yaitu sebanyak 10602 data kategori 'not sexist' dan 3398 data 'kategori sexist'. Tabel 2 menunjukan pembagian data untuk pengujian sebesar 1514 data kategori 'not sexist' dan 486 data dengan kategori 'sexist'.

Dataset berupa teks berbahasa Inggris akan dilakukan *preprocessing* untuk menghasilkan data yang terstruktur. *Preprocessing* dimulai dengan proses *lowercase* untuk mengubah semua huruf menjadi bentuk huruf kecil sehingga kata yang sama akan memiliki dimensi yang sama pula **(Duong & Nguyen-Thi, 2021).** Selanjutnya dilakukan *cleaning*, yang merupakan proses pembersihan persiapan data. Proses *cleaning* melakukan penghapusan URL, *hashtag*, tanda baca, *mention*, dan *whitespace* agar memperoleh data yang terstruktur sesuai dengan standar yang sama **(Brownlee, 2020)**. Selain itu, pada proses *cleaning* juga melakukan penghapusan data duplikat dan data null. Setelah data dibersihkan pada proses *cleaning*, data akan diekstraksi fitur dengan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) untuk setiap token (kata) pada setiap dokumen di dalam korpus **(Chen et al., 2020)**. Nilai TF dihitung berdasarkan jumlah kemunculan kata dalam tiap dokumen, sedangkan IDF dihitung berdasarkan jumlah kemunculan kata pada keseluruhan dokumen. Setelah melalui proses normalisasi, nilai TF akan dibandingkan terhadap nilai IDF. Perhitungan bobot TF-IDF dapat dilihat pada Persamaan (1).

$$w(t_k) = tf_k \cdot log(\frac{N}{df_k}) \quad (1)$$

tf_k = Frekuensi kemunculan kata (t_k) dalam dokumen

N = Jumlah total dokumen

df_k = Frekuensi kemunculan kata dalam dokumen

$w(t_k)$ = Bobot TF-IDF

Tahap akhir dari *preprocessing* adalah mengatasi *imbalanced* kelas pada data. Kelas yang *imbalanced* adalah masalah yang biasa ditemukan pada klasifikasi *machine learning* yang

terdapat ketidakseimbangan pada perbandingan data tiap kelas. Penelitian ini menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi data yang *imbalanced*.

Proporsi data latih yang ditunjukkan pada Tabel 1 memperlihatkan ketidakseimbangan jumlah data pada kedua kelas. Kelas 'not sexist' jauh lebih banyak daripada kelas 'sexist'. Data yang tidak seimbang tersebut dapat menyebabkan hasil pengklasifikasian data menjadi tidak optimal. SMOTE memanfaatkan jarak Euclidean untuk menghasilkan data sintetik untuk kelas minoritas berdasarkan tetangga terdekatnya. Data baru dibuat sebagai duplikat dari data asli berdasarkan fiturnya **(Umer et al., 2021)**.

Penelitian ini merancang model dengan dua algoritma yaitu *Support Vector Machine* (SVM) dan *Naive Bayes*. Kedua model melakukan optimasi parameter untuk mendapatkan nilai *hyperparameter* terbaik dari setiap model. Optimasi parameter menggunakan metode *grid search*. Metode *grid search* akan mencoba seluruh kombinasi dari semua nilai *hyperparameter* yang ditentukan dengan teknik *brute force* untuk mengetahui nilai *hyperparameter* terbaik dari setiap pengujian sehingga mendapatkan model terbaik. Model *grid search* yang digunakan adalah *Grid Search CV* yang mengevaluasi model menggunakan *K-fold Cross Validation*. Skor terbaik diperoleh dari nilai rata-rata terbesar yang dihasilkan dari *fold* pada setiap kombinasi parameter. Nilai *cross validation* yang digunakan adalah 5. Nilai 5 *fold cross validation* berarti data akan dibagi menjadi 5 *fold*, sehingga menghasilkan 5 subset data. *Cross validation* akan menggunakan 4 *fold* untuk pelatihan dan 1 *fold* untuk validasi.

Algoritma *Support Vector Machine* (SVM) digunakan untuk dapat mengenali pola kalimat dari data latih dan data uji sehingga dapat mengidentifikasi kalimat. Model SVM dibuat menggunakan library SVM pada scikit-learn yaitu SVC. Penelitian ini melakukan percobaan dengan kernel *Radial Basis Function* (RBF). Kernel berjalan menggunakan fungsi SVC dengan parameter *default*. Parameter yang akan dicari nilai optimalnya adalah parameter C dan gamma. Parameter C (Cost) merupakan parameter regularisasi yang menentukan penalti akibat kesalahan klasifikasi **(Yu et al., 2019)**. Jika nilai C besar maka optimasi akan memilih *hyperplane* dengan margin kecil agar pelatihan lebih baik karena kesalahan klasifikasi atau eror lebih kecil. Sebaliknya, nilai C yang kecil menyebabkan margin pada *hyperplane* akan besar sehingga jumlah kesalahan klasifikasi meningkat. Nilai C yang digunakan adalah 0,01, 0,1, 1, 10, 100. Parameter gamma digunakan sebagai ukuran kesamaan antara dua titik. Parameter gamma yang digunakan adalah parameter *default*.

Algoritma *Naive Bayes* merupakan salah satu algoritma yang digunakan pada teknik klasifikasi. *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya **(Bustami, 2013)**. Model *Naive Bayes* menggunakan pustaka *MultinomialNB*. Parameter yang akan dicari nilai optimalnya adalah parameter alpa. Skor terbaik diambil dari nilai rata-rata terbesar dari kombinasi parameter. Nilai alpha yang kecil menghasilkan variansi tinggi sehingga dapat menyebabkan *overfitting* sedangkan nilai alpha yang besar menghasilkan bias yang tinggi sehingga dapat menyebabkan *underfitting*. Nilai alpa yang digunakan adalah 0,01, 0,1, 0,5, 1, 10.

Setelah perancangan model dilakukan pelatihan model. Penelitian ini membagi dua tugas pada pelatihan model. Tugas 1 melakukan pelatihan model dengan data yang tidak dilakukan penanganan *imbalanced* sedangkan tugas 2 merupakan pelatihan model dengan data yang telah dilakukan penanganan *imbalanced* menggunakan SMOTE. Pembagian dua tugas ditujukan untuk mengetahui pengaruh penanganan *imbalanced* data terhadap masing-masing algoritma yang digunakan dalam pengklasifikasian.

Pengujian model dilakukan dengan menggunakan data uji. Data uji juga telah di *preprocessing* untuk mendapatkan data yang bersih dan terstruktur, namun tidak dilakukan penanganan *imbalanced*. Hasil klasifikasi data uji dengan model yang dibangun akan dibandingkan dengan label aktual yang telah diberikan oleh annotator.

Tahap terakhir model akan dievaluasi. Evaluasi model bertujuan untuk menilai kinerja model. Evaluasi pada penelitian ini menggunakan *confusion matrix* sebagai matriks peforma. Matriks performa memberikan informasi tentang kemampuan model dalam mengklasifikasi data dengan baik. Informasi tersebut dapat digunakan dalam pengambilan keputusan lanjutan terkait perbaikan model untuk meningkatkan hasil prediksi **(Powers, 2020).** *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan **(Rahman et al., 2017).**

Tabel 3. Confusion Matrix

Nilai Prediksi	Nilai Aktual Positif	Nilai Aktual Negatif
Positif	TP	FP
Negatif	FN	TN

Berdasarkan Tabel 3 *confusion matrix* diatas menjelaskan:

a. *True Positives* (TP) adalah jumlah record data positif yang diklasifikasikan benar sebagai nilai positif.
b. *False Positives* (FP) adalah jumlah record data negatif yang diklasifikasikan sebagai nilai positif.
c. *False Negatives* (FN) adalah jumlah record data positif yang diklasifikasikan sebagai nilai negatif.
d. *True Negatives* (TN) adalah jumlah record data negatif yang diklasifikasikan benar sebagai nilai negatives.

3. HASIL DAN PEMBAHASAN

3.1. Preprocessing

Preprocessing dilakukan pada data latih dan data uji. *Preprocessing* terdiri dari empat proses, yaitu *lowercase, cleaning*, ektraksi fitur, dan penanganan *imbalanced* dengan SMOTE. Hasil dari tahap *preprocessing* adalah data yang bersih dan terstruktur.

a. Lowercase
   Lowercase adalah proses untuk mengubah huruf menjadi bentuk huruf kecil agar kata yang sama memiliki dimensi yang sama.

Tabel 4. Hasil Proses Lowercase

Sebelum Lowercase	Sesudah Lowercase
How the FUCK is this woman still an MP!!!??? @MPP__	how the fuck is this woman still an mp!!!???@mp__
ugly chicks A pussy is a pussy.#MeooWW	ugly chicks a pussy is a pussy.#meooww
Every woman wants to be a Model. it's codeword for "I get Everything for Free and people want me";))	every woman wants to be a model. it's codeword for "i get everything for free and people want me";))

Dilihat pada Tabel 4, hasil dari proses *lowercase* semua huruf menjadi bentuk huruf kecil. Hasil tersebut juga menunjukkan kalimat masih mengandung tanda baca, *mention*, *hashtags*.

b. Cleaning
Data yang telah di *lowercase* selanjutnya akan dibersihkan pada proses cleaning. *Cleaning* merupakan proses menghapus tanda baca, *mentions, hashtag*, URL dan *whitespace* sehingga menghasilkan data yang terstruktur dan memiliki standar yang sama.

Tabel 5. Hasil Proses *Cleaning*

Sebelum Cleaning	Sesudah Cleaning
how the fuck is this woman still an mp!!!???@mp__	how the fuck is this woman still an mp
ugly chicks a pussy is a pussy.#meooww	ugly chicks a pussy is a pussy
every woman wants to be a model. it's codeword for "i get everything for free and people want me";))	every woman wants to be a model it s codeword for i get everything for free and people want me

Contoh dari hasil proses *cleaning* pada Tabel 5 menunjukkan bahwa tanda baca, *mention*, dan *hashtag* telah dihapus. URL dan *whitespace* pada kalimat juga akan dihapus pada proses ini. Kemudian dilakukan penghapusan data duplikat dan data null untuk menghilangkan *noise* pada data.

c. Ekstraksi Fitur
Ekstraksi fitur dilakukan dengan memberikan bobot pada setiap kata. Proses pembobotan dilakukan dengan merepresentasikan kata dalam bentuk numerik. Pembobotan menggunakan pustaka *sklearn* yaitu modul feature_extraction.text.

```
(0, 2701)       0.6545197415912966
(0, 13359)      0.35414709314197396
(0, 19311)      0.5412770079518656
(0, 4101)       0.3914114777941859
(1, 10136)      0.2231686448116929
(1, 16621)      0.3114290009699667
(1, 19201)      0.16861016485721642
(1, 6994)       0.2776518708325439
(1, 17997)      0.29676968578880386
(1, 12107)      0.29923914806170204
(1, 18633)      0.3773615389303172
(1, 11040)      0.4019547069258976
(1, 2239)       0.31394142974211386
(1, 729)        0.33260999765123384
(1, 19385)      0.23856554820861067
(2, 11319)      0.8556235098380637
(2, 6802)       0.38826016582102224
(2, 19201)      0.34229001307228507
(3, 5643)       0.37334488264087895
(3, 6247)       0.19301405236410202
(3, 1418)       0.46522280339988725
(3, 8672)       0.46522280339988725
(3, 19249)      0.24459093219991018
(3, 11172)      0.24570719958451298
(3, 9059)       0.24638985115552506
:       :
(13988, 15731)          0.4301393456403773
(13988, 4485) 0.3766482105950002
(13988, 9228) 0.43262388684892866
(13988, 8465) 0.35363028016186554
(13988, 4942) 0.27487299170023666
(13988, 7553) 0.2753858316622378
(13988, 18084)          0.31794794873161986
(13988, 9983) 0.18758094141184864
(13988, 10136)          0.2706075400610211
(13989, 17328)          0.44311606349561816
(13989, 9520) 0.4917686976052374
(13989, 4154) 0.5283541626963819
(13989, 4152) 0.38900006331467823
```

Gambar 3. Hasil Pembobotan TF-IDF

Berdasarkan Gambar 3 hasil pembobotan TF-IDF, angka yang ditandai dengan kotak merah menunjukkan nomor baris dari masing-masing data. Angka pada kotak kuning merupakan penomoran integer unik untuk setiap kata dan angka dan kotak hijau adalah hasil pembobotan dengan TF-IDF. Setelah didapatkan bobot, data dapat diproses pada model klasifikasi.

d. Penanganan *Imbalanced data* dengan SMOTE
   Tahap akhir dari *preprocessing* adalah penanganan *imbalanced* pada data latih. Data latih dengan kategori yang memiliki jumlah data lebih kecil akan di *oversampling* menggunakan SMOTE. Data latih dengan kategori 'sexist' telah dilakukan *oversampling* menjadi 10593, jumlah data kategori 'sexist' tersebut menjadi sama banyaknya dengan data kategori 'not sexist'. Data tersebut akan digunakan dalam pelatihan model.

3.2. Pelatihan Model

Pelatihan model dibagi menjadi dua tugas, yaitu tugas 1 dan tugas 2. Tugas 1 merupakan pelatihan model dengan data tanpa penanganan *imbalanced* dan tugas 2 merupakan model dengan data yang telah dilakukan penanganan *imbalanced* menggunakan SMOTE. Kedua tugas menggunakan data yang telah di *preprocessing.* Model dibangun dengan menggunakan *grid search* sebagai optimasi parameter untuk menemukan parameter terbaik sehingga memperoleh hasil terbaik. Nilai cv yang digunakan adalah 5, sehingga akan dilakukan 5 iterasi untuk setiap kombinasi nilai *hyperparameter.*

Tabel 7. Rata-rata Skor F1 terbaik Model

Model	Rata-rata Skor F1 terbaik	Parameter
Naive Bayes	0,74	Alpha = 0,1
Naive Bayes+SMOTE	0,85	Alpha = 0,01
SVM	0,80	C = 10 dan C = 100
SVM+SMOTE	0,96	C = 100

Hasil pelatihan model yang pada Tabel 7 menunjukkan model *Naive Bayes, Naive Bayes*+SMOTE, SVM, dan SVM+SMOTE memperoleh rata-rata skor F1 terbaik sebesar 0,74, 0,85, 0,80, dan 0,96. Berdasarkan hasil pelatihan model tersebut, model dengan data yang telah dilakukan penanganan *imbalanced* menggunakan SMOTE memiliki nilai rata-rata skor F1 terbaik lebih tinggi daripada model dengan data tanpa penanganan *imbalanced*. Hal ini menunjukan bahwa SMOTE dapat meningkatkan kemampuan model dalam mengklasifikasi data. *Grid search* juga berpengaruh dalam perolehan skor yang tinggi karena *grid search* mencoba semua kombinasi dari seluruh *hyperparameter* sehingga memperoleh model terbaik dan menghasilkan skor terbaik.

3.3. Pengujian dengan Data Uji

Pengujian dilakukan dengan data uji terhadap model yang telah dibangun untuk mendeteksi kalimat sebagai 'sexist' atau 'not sexist'. Hasil pendeteksian akan dibandingkan dengan label aktual yang telah dilabeli oleh annotator.

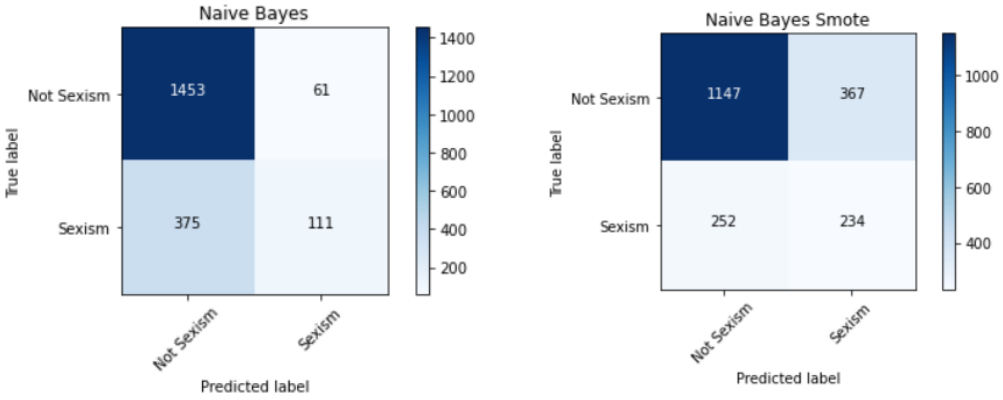
text	label	clean_test	true label	nb	nbs	svm	svms
I do not try to hard. I do what guys who get w...	0	i do not try to hard i do what guys who get wi...	not sexim	0	1	0	0
That sucks. We elected a Navy Seal in Missouri...	0	that sucks we elected a navy seal in missouri ...	not sexim	0	0	0	0
Walk away and never look back. If she reaches ...	0	walk away and never look back if she reaches o...	not sexim	0	0	0	0
basically I placed more value on her than I sh...	0	basically i placed more value on her than i sh...	not sexim	0	0	0	0
NYT and Asian "part ways." But why on earth wa...	0	nyt and asian part ways but why on earth was s...	not sexim	0	0	0	0
...	...	...	...	...	...	...	...
None ,but some of the girls I bang are on ecst...	0	none but some of the girls i bang are on ecsta...	not sexim	0	0	0	0
It's a constant isnt it, whenever women fall b...	1	it s a constant isnt it whenever women fall be...	sexism	0	1	0	0
yep. No fapping means you have to chase the va...	1	yep no fapping means you have to chase the vag...	sexism	0	1	0	0
I didn't say you had to stay with a crazy bitc...	1	i didn t say you had to stay with a crazy bitc...	sexism	1	1	1	1
Just get a good looking hooker you weak minded...	0	just get a good looking hooker you weak minded...	not sexim	0	1	0	0

Gambar 4. Hasil Klasifikasi dengan Data uji

Gambar 4 menampilkan hasil deteksi kalimat data uji dari beberapa model. Hasil tersebut menunjukkan masih terdapat beberapa model yang mendeteksi kalimat tidak sesuai dengan label aktualnya. Kalimat yang mengandung kata 'fuuuck', yang merupakan kata memanjang dari kata 'fuck' dilabeli sebagai 'sexist' pada label aktual, namun dideteksi 'not sexist' pada model *Naive Bayes*, SVM, dan SVM+SMOTE.

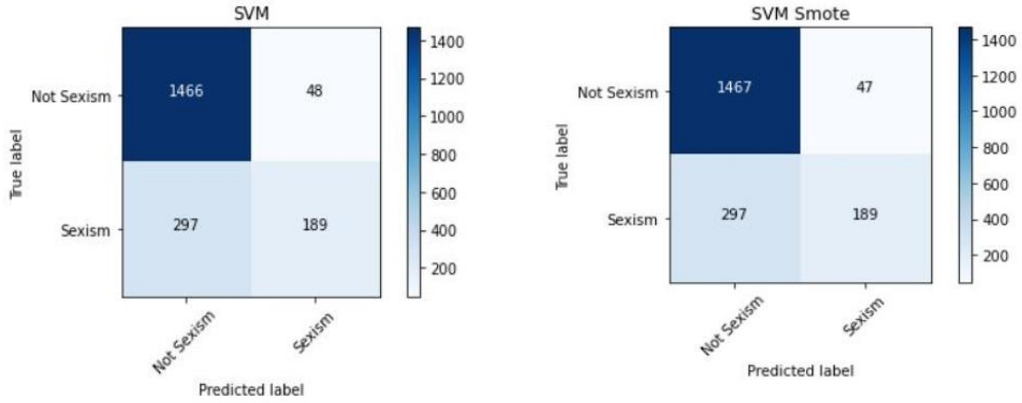
3.4. Evaluasi

Evaluasi model menggunakan *confusion matrix* untuk menghitung kinerja model klasifikasi. Penelitian ini mengukur kinerja model berdasarkan nilai skor F1.



Gambar 5. *Confusion Matrix* Data Uji Model *Naive Bayes* dan Model *Naive Bayes*+SMOTE

Confusion matrix pada Gambar 5 menunjukkan model *Naive Bayes* mengklasifikasikan 1453 kalimat 'not sexist' dengan benar dan mengklasifikasikan dengan salah 61 kalimat 'not sexist' sebagai 'sexist', sedangkan 111 kalimat 'sexist' terklasifikasikan benar dan 375 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'. Model *Naive Bayes*+SMOTE mengklasifikasikan 1147 kalimat 'not sexist' dengan benar dan melakukan kesalahan pada 367 kalimat 'not sexist' diklasifikasikan sebagai 'sexist', sedangkan sebanyak 234 kalimat 'sexist' diklasifikasikan dengan benar dan 252 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'. Skor F1 yang diperoleh model *Naive Bayes* dan *Naive Bayes*+SMOTE adalah 0,87 dan 0,79.



Gambar 6. *Confusion Matrix* Data Uji Model SVM dan SVM+SMOTE

Gambar 6 menunjukkan *confusion matrix* model SVM mengklasifikasikan 1466 kalimat 'not sexist' dengan benar dan mengklasifikasikan dengan salah 48 kalimat 'not sexist' sebagai 'sexist', sedangkan 189 kalimat 'sexist' terklasifikasikan benar dan 297 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'. Model SVM+SMOTE mengklasifikasikan 1467 kalimat 'not sexist' dengan benar dan melakukan kesalahan pada 47 kalimat 'not sexist' diklasifikasikan sebagai 'sexist', sedangkan sebanyak 189 kalimat 'sexist' diklasifikasikan dengan benar dan 297 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'. Skor F1 yang diperoleh model SVM dan SVM+SMOTE adalah 0,89 dan 0,90.

4. KESIMPULAN

Penelitian ini menggunakan *machine learning* untuk mendeteksi seksisme pada kalimat berbahasa Inggris. Kalimat bahasa Inggris di *preprocessing* untuk menghasilkan data yang

bersih dan terstruktur sehingga dapat digunakan pada proses klasifikasi. Algoritma yang digunakan untuk membangun model klasifikasi adalah *Support Vector Machine* dan *Naive Bayes*. Optimasi parameter menggunakan Grid Search CV diterapkan pada kedua model untuk mendapatkan nilai *hyperparameter* terbaik sehingga menghasilkan skor terbaik dari model terbaik. Pelatihan model dibagi menjadi dua tugas, yaitu tugas 1 adalah pelatihan model dengan menggunakan data tanpa penanganan *imbalanced* dan tugas 2 adalah pelatihan model dengan data yang telah dilakukan penanganan *imbalanced* menggunakan SMOTE. Pelatihan model menghasilkan nilai rata-rata skor F1 untuk model *Naive Bayes, Naive Bayes*+SMOTE, SVM, dan SVM+SMOTE yaitu sebesar 0,74, 0,85, 0,80, dan 0,96. Model diuji menggunakan data uji yang telah di *preprocessing* selanjutnya hasil deteksi model akan dibandingkan dengan label aktualnya. Pengujian menghasilkan skor F1 tertinggi pada model SVM+SMOTE sebesar 0,90 dengan 1467 kalimat diklasifikasikan 'not sexist' dengan benar dan melakukan kesalahan pada 47 kalimat 'not sexist' diklasifikasikan sebagai 'sexist', sedangkan sebanyak 189 kalimat 'sexist' diklasifikasikan dengan benar dan 297 kalimat 'sexist' diklasifikasikan sebagai 'not sexist'. Hasil pengujian pada Gambar 4 menunjukkan beberapa model mendeteksi kalimat berbeda dengan label aktualnya. Model SVM dengan data yang telah dilakukan penanganan *imbalanced* menggunakan SMOTE memiliki performa model lebih baik dari pada model lain. Performa model dalam melakukan deteksi seksisme dipengaruhi oleh *preprocessing* data, algoritma yang digunakan, dan parameter yang dilatih.

DAFTAR RUJUKAN

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The Future of Social media in Marketing. Journal of the Academy of Marketing Science, 79-95.

Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2022, March 21). Hate Speech Classification Using SVM and Naive Bayes. Diambil kembali dari ArXiv: arxiv.org

Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Melbourne: Machine Learning Mastery. Diambil kembali dari Machine Learning Mastery: machinelearningmastery.com

Bustami, B. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. TECHSI-Jurnal Teknik Informatika, 127-146.

Cambridge. (2023, June 14). Sexism. Diambil kembali dari Cambridge Dictionary: https://dictionary.cambridge.org/dictionary/english/sexism

Chen, J., Kudjo, P. K., Mensah, S., Brown, S. A., & Akorfu, G. (2020). An Automatic Software Vulnerability Classification Framework Using Term Frequency-Inverse Gravity Moment and Feature Selection. Journal of Systems and Software, 1-20.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., & Coulomb-Gully, M. (2020). An Annotated Corpus for Sexism Detection in French Tweets. Conference on Language Resources and Evaluation, (hal. 1397-1403).

Duong, H. T., & Nguyen-Thi, T. A. (2021). A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis. Computational Social Networks, 1-16.

Fayed, H. A., & Atiya, A. F. (2019). Speed Up Grid-Search for Parameter Selection of Support Vector Machines. Applied Soft Computing, 1-16.

Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating Online Sexism Offline: Anonymity, Interactiviy, and the Effect of Sexist Hashtags on Social Media. Computers in Human Behavior, 436-442.

Kirk, H. R., Vidgen, B., Röttger, P., & Wenjie, Y. (2023, March 7). SemEval-2023 Task 10: Explainable Detection of Online Sexism. Diambil kembali dari ArXiv: arxiv.org

Kumar, R., Pal, S., & Pamula, R. (2021). Sexism Detection in English and Spanish Tweets. Conference of the Spanish Society for Natural Language Processing, (hal. 500-505).

Napier, J. L., Suppes, A., & Bettinsoli, M. L. (2020). Denial of Gender Discrimination is Associated with Better Subjective Well-Being Among Women: A System Justification Account. European Journal of Social Psychology, 1191-1209.

Powers, D. M. (2020, October 10). Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness and Correlation. Diambil kembali dari ArXiv: arxif.org

Rahman, M. F., Alamsah, D., Darmawidjadja, M. I., & Nurma, I. (2017). Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). Jurnal Informatika, 36-45.

Rezkisari, I. (2019, December 2). Studi: Wanita Sasaran Seksisme Rentan Alami Depresi. Diambil kembali dari Republika: sindikasi.republika.co.id

Tanaka, M., & Okutomi, M. (2014). A Novel Inference of a Restricted Boltzmann Machine. International Conference on Pattern Recognition, (hal. 1526-1531).

Umer, M., Sadiq, S., Missen, M. M., Hameed, Z., Aslam, Z., Siddique, M. A., & Nappi, M. (2021). Scientific Papers Citation Analysis Using Textual Features and SMOTE Resampling Techniques. Pattern Recognition Letters, 250-257.

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A Hybrid Sampling Algorithm Combining M-Smote and Enn Based on Random Forest for Medical Imbalanced Data. ournal of Biomedical Informatics, 1-11.

Yu, S., Li, X., Zhang, X., & Wang, H. (2019). The OCS-SVM: An Objective-Cost-Sensitive Svm With Sample-Based Misclassification Cost Invariance. IEEE Access, 118931-118942.