Journal of Information Technology and Computer Science (INTECOMS)
Volume 5 Nomor 1, Juni 2022
e-ISSN : 2614-1574
p-ISSN : 2621-3249

KLASIFIKASI JENIS KEKERASAN PADA PEREMPUAN DAN ANAK DENGAN ALGORITMA MULTINOMIAL NAIVE BAYES

CLASSIFICATION OF TYPES OF VIOLENCE AGAINST WOMEN AND CHILDREN USING THE MULTINOMIAL NAIVE BAYES ALGORITHM

Giat Subroto[1], Nina Sulistiyowati[2], Azhari Ali Ridha[3]
Universitas Singaperbangsa Karawang[1,2,3]
giat.subroto17103@student.unsika.ac.id

ABSTRACT

Reports of cases of violence and sexual harassment against women and children received by Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A) in recapitulating and grouping case reports are still done manually. The study was conducted to create a classification model based on the chronology of events in case reports into several categories of types of violence by using Text Mining. The steps carried out are in accordance with the stages in the Knowledge Discovery in Database (KDD) method, namely data selection, preprocessing, transformation, modeling, and evaluation. Word weighting is done using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. To handle the imbalanced dataset, an oversampling process was carried out using the Random Oversampling algorithm. The algorithm used to perform the classification is the Multinomial Naïve Bayes algorithm.

***Keywords**: Multinomial Naïve Bayes, Random Oversampling, Text Mining, TF-IDF*

Laporan kasus tindak kekerasan dan pelecehan seksual pada perempuan dan anak yang diterima oleh Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A) dalam melakukan rekap dan pengelompokan laporan kasus masih dilakukan secara manual. Penelitian dilakukan untuk membuat model klasifikasi berdasarkan kronologi kejadian pada laporan kasus kedalam beberapa kategori jenis kekerasan dengan memanfaatkan *Text Mining*. Tahapan yang dilakukan sesuai dengan tahapan pada metode *Knowledge Discovery in Database* (KDD) yaitu *data selection*, *preprocessing*, *transformation*, *modeling*, dan *evaluation*. Pembobotan kata dilakukan menggunakan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF). Untuk menangani *imbalance dataset* dilakukan proses *oversampling* menggunakan algoritma *Random Oversampling*. Algoritma yang digunakan untuk melakukan klasifikasi yaitu algoritma *Multinomial Naïve Bayes*.

***Kata Kunci**: Multinomial Naïve Bayes, Random Oversampling, Text Mining, TF-IDF*

PENDAHULUAN

Berdasarkan data yang diperoleh melalui Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A), jumlah kasus tindak kekerasan dan pelecehan seksual dari tahun 2016 sampai tahun 2020 mengalami kenaikan setiap tahunnya. Pada tahun 2016 jumlah kasus yang tercatat sebanyak 49 kasus, pada tahun 2017 tercatat 56 kasus, pada tahun 2018 tercatat 71 kasus, pada tahun 2019 tercatat 88 kasus, dan pada tahun 2020 tercatat 98 kasus.

Laporan kasus tindak kekerasan dan pelecehan seksual yang diterima oleh Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A) berasal dari masyarakat yang membuat laporan di Pusat Pelayanan Terpadu Pemberdayaan Perempuan dan Anak (P2TP2A). Sebelum diserahkan ke Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A) laporan tersebut akan direkap dan dikelompokkan terlebih dahulu berdasarkan jenis kekerasannya. Proses rekap dan pengelompokkan laporan masih dilakukan secara manual.

Berkaitan dengan permasalahan tersebut dapat dilakukan pembuatan model klasifikasi untuk melakukan pengelompokkan laporan kasus dengan

memanfaatkan *text mining*. Laporan kasus dikelompokkan berdasarkan jenis kekerasan. Penelitian ini dilakukan dengan tahapan yang ada pada metode *Knowledge Discovery in Database* (KDD) dengan tahapan yang dilakukan yaitu *data selection*, *preprocessing*, *transformation*, *data mining*, *interpretation/evaluation* (Adi, 2018).

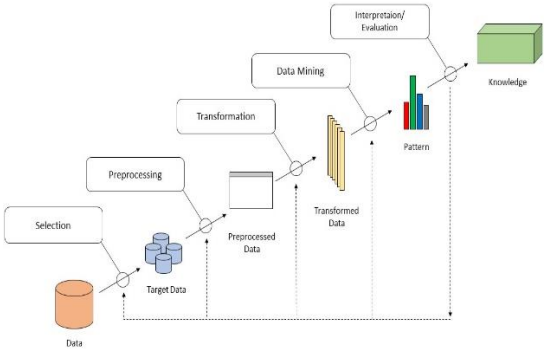
Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk memberikan suatu nilai bobot pada kata dalam dokumen. TF-IDF merupakan salah satu algoritma pencarian informasi yang menonjol karena sederhana dan efektif (Bahasyim, Widowati, & Husen, 2021). TD-IDF merupakan suatu fitur pembobotan kata yang paling popular dan paling sering digunakan serta memiliki nilai accuracy dan nilai recall yang cukup tinggi (Yutika, Adiwijaya, & Faraby, 2021).

Teknik *oversampling* digunakan pada penelitian ini karena terdapat data yang tidak seimbang antara satu kategori dengan kategori yang lainnya (*imbalance dataset*). Algoritma *oversampling* yang digunakan adalah *Random Oversampling* yaitu suatu metode yang digunakan untuk melakukan penambahan data kelas minoritas ke dalam data latih yang dilakukan secara acak (Fitriani, Yasin, & Tarno, 2021). *Random Oversampling* digunakan karena pada beberapa kategori pada data latih hanya memiliki 1 sampel data sehingga belum memenuhi syarat untuk metode lain seperti *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE mencari tetangga terdekat menggunakan algoritma *k-Nearest Neighbors* (kNN) sebagai dasar dalam melakukan pembuatan data sintetik menggunakan jarak *Euclidean* (Fahrudin, Buliali, & Fatichah, 2019).

Klasifikasi jenis kekerasan pada penelitian ini dilakukan menggunakan algoritma *Naïve Bayes*, karena algoritma *Naïve Bayes* merupakan algoritma yang dapat menentukan estimasi parameter untuk proses klasifikasi dengan jumlah data latih yang kecil (Manalu, Sianturi, & Manalu, 2017). Dalam beberapa penelitian sebelumnya juga algoritma *Naïve Bayes* memperoleh hasil lebih baik dibangdingkan dengan algoritma klasifikasi lainnya seperti pada penelitian yang melakukan komparasi algoritma *Naïve Bayes* dengan algoritma *Support Vector Machine* (SVM) dalam mengklasifikasikan teks yang divalidasi menggunakan *10-Fold Cross Validation* menghasilkan nilai rata-rata akurasi dengan AUC algoritma *Naïve Bayes* lebih tinggi dari algoritma *Support Vector Machine* (SVM) yaitu algoritma *Naïve Bayes* menghasilkan nilai akurasi 93,29% dengan AUC 0,525 sedangkan algoritma *Support Vector Machine* (SVM) menghasilkan nilai akurasi 92,61% dengan AUC 0,950 (Gunawan, Riana, Ardiansyah, Akbar, & Alfarizi, 2020). Pada penelitian lain yang juga melakukan komparasi algoritma *Naïve Bayes* dengan algoritma *Support Vector Machine* (SVM) dengan menggunakan *10-Fold Cross Validation*, hasil dari penelitian tersebut menyimpulkan bahwa algoritma *Naïve Bayes* lebih unggul dibandingkan dengan algoritma *Support Vector Machine* (SVM) karena dari hasil uji evaluasi algoritma *Naïve Bayes* memperoleh nilai akurasi sebesar 87% sedangkan algoritma *Support Vector Machine* (SVM) memperoleh nilai akurasi sebesar 56% (Muthia, 2018).

METODE

Penelitian ini menggunakan metode *Knowledge Discovery in Database* (KDD) dengan 5 tahapan yang dilakukan yaitu *data selection*, *preprocessing*, *transformation*, *data mining*, dan *interpretation/evaluation*.



Gambar 1. Tahapan KDD

Knowledge Discovery in Database (KDD) merupakan suatu proses ekstraksi non-trivial dari implisit suatu informasi yang tidak diketahui sebelumnya tetapi hasil yang diperoleh dari data tersebut terdapat potensi informasi (Ependi & Putra, 2019).

1. *Data Selection*

Pada tahap pertama dilakukan pengumpulan data dan pemilihan data yang digunakan untuk diproses pada tahap selanjutnya yaitu tahap *preprocessing*.

2. *Preprocessing*

Pada tahap *preprocessing* dilakukan beberapa proses yaitu:

a. *Case Folding*, proses ini merupakan proses mengubah huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*) di dalam suatu dokumen.
b. *Cleaning*, proses yang digunakan untuk menghilangkan simbol atau karakter khusus serta menghilangkan angka yang ada di dalam suatu dokumen.
c. *Tokenizing*, proses pemecahan atau penguraian kalimat di dalam dokumen menjadi kata-kata.
d. *Spell Normalization*, proses untuk melakukan normalisasi pada kata singkatan atau kata gaul di dalam dokumen.
e. *Filtering*, proses untuk mengambil kata yang memiliki makna dengan cara menghilangkan atau menghapus kata yang tidak memiliki makna di dalam dokumen.
f. *Stemming*, proses untuk menemukan kata dasar dari kata yang telah diproses pada tahap *filtering*.

3. *Tranformation*

Transformation merupakan tahap yang digunakan untuk melakukan transformasi pada data yang digunakan sehingga data tersebut sesuai dengan yang diharapkan dan siap untuk diolah pada tahap selanjutnya yaitu tahap *data mining* atau *modeling*.

Terdapat 2 proses yang dilakukan pada tahap *transformation* yaitu:

a. *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan teknik atau cara untuk memberikan bobot nilai hubungan pada suatu kata terhadap dokumen (Rokhim & Yaqin, 2017).

TF-IDF adalah metode yang menggabungkan dua jenis konsep dalam melakukan perhitungan suatu bobot yaitu menghitung nilai *term frequency* pada suatu dokumen dan menghitung nilai *inverse document frequency* yang mengandung kata tersebut. Frekuensi kemunculan setiap kata pada dokumen menandakan seberapa penting kata tersebut di dalam dokumen (Riyani, Naf'an, & Burhanuddin, 2019).

Untuk memperoleh nilai TF-IDF perlu dapat dilakukan dengan beberapa rumus perhitungan(Bahasyim, Widowati, & Husen, 2021)

Langkah pertama yaitu mencari nilai *Term Frequency* (TF) dengan rumus berikut:

$$\text{TF} = \frac{\text{Total appearance of a word document}}{\text{Total words in a document}}$$

Langkah kedua mencari nilai *Inverse Document Frequency* (IDF) dengan rumus berikut:

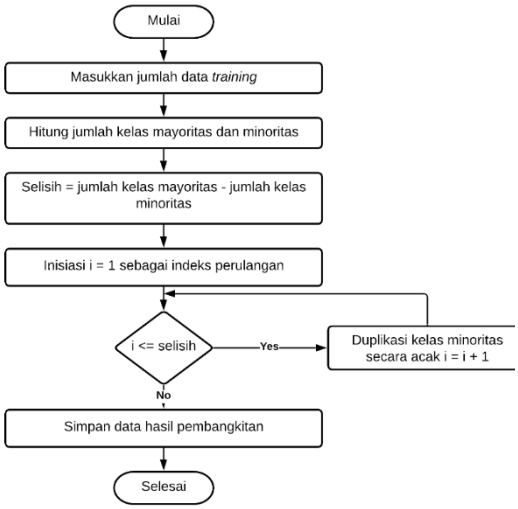
$$\text{IDF} = \log \frac{\text{All document number}}{\text{Document Frequency}}$$

Terakhir untuk memperoleh nilai TF-IDF dapat menggunakan rumus berikut:

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

b. *Random Oversampling*

Random Oversampling merupakan metode untuk menambah jumlah data dari kelas minoritas ke data latih yang dilakukan secara acak. Proses dalam menambahkan data dilakukan secara berulang sampai data kelas minoritas sama dengan jumlah data kelas mayoritas. (Fitriani, Yasin, & Tarno, 2021).



Gambar 2. Flowchart Random Oversampling

4. *Data Mining*

Tahap ini merupakan tahap yang dilakukan untuk mencari suatu pola atau informasi pada suatu data yang telah dipilih melalu proses sebelumnya. Pencarian tersebut dilakukan menggunakan algoritma atau teknik tertentu.

Penelitian ini menggunakan algoritma *Multinomial Naïve Bayes* untuk melakukan proses pencarian pola atau pembuatan model klasifikasi. Pada formula Multinomial Naïve Bayes Classifier, kelas suatu dokumen tidak hanya ditentukan berdasarkan kata yang muncul tetapi berdasarkan jumlah kemunculannya juga (Kalokasari, Shofi, & Setyaningrum, 2017). Terdapat beberapa rumus perhitungan dalam pada algoritma *Multinomial Naïve Bayes* (Rahman, Wiranto, & Doewes, 2017).

Untuk memperhitungkan kelas dari suatu dokumen, rumus yang digunakan yaitu:

$$P(c|\text{term dokumen d}) = P(c) \times P(t_1|c) \times P(t_2|c) \times \ldots \times P(t_n|c)$$

Probabilitas *prior* kelas c dapat ditentukan menggunakan rumus berikut:

$$P(c) = \frac{N_c}{N}$$

Probabilitas dari kata ke-n ditentukan menggunakan metode *laplacian smoothing* dengan rumus sebagai berikut:

$$P(t_n|c) = \frac{\text{count}(t_n,c)+1}{\text{count}(c)+|V|}$$

5. *Interpretaion/Evaluation*

Tahap *evaluation* merupakan tahapan yang mencakup beberapa pemeriksaan untuk membuktikan apakah pola yang diperoleh sudah sesuai dengan fakta atau hipotesa yang ada sebelumnya.

Pada penelitian ini hasil evaluasi berupa nilai *accuracy*, *precision*, *recall*, dan *f-measure*.

Accuracy merupakan jumlah data yang diklasifikasikan dengan benar. *Accuracy* dapat dihitung menggunakan rumus: (Ruhyana, 2019)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}}$$

Precision dan *recall* digunakan untuk mengetahui relevansi dan ketepatan sistem dalam melakukan

pencarian informasi yang diminta oleh pengguna (Ruhyana, 2019).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

F-measure merupakan nilai yang diperoleh dari pengukuran *precision* dan *recall* antara kelas hasil *cluster* dan kelas sebenarnya (Ruhyana, 2019). *F-measure* dapat dicari menggunakan rumus berikut:

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

HASIL DAN PEMBAHASAN

1. *Data Selection*

Pada tahap data selection dilakukan pengumpulan data dan melakukan pemilihan data yang digunakan. Data tersebut berisi data kronologi kejadian tindak kekerasan dan pelecehan seksual dan jenis kekerasan dari setiap kronologi kejadiannya. Data tersebut diperoleh dari Dinas Pemberdayaan Perempuan dan Perlindungan Anak (DP3A). Pada penelitian ini atribut jenis kekerasan digunakan sebagai label. Setelah dilakukan proses pengumpulan dan pemilihan data yang dapat digunakan pada penelitian ini, diperoleh jumlah data pada masing-masing label dapat dilihat pada tabel 1.

Tabel 1. Jumlah Data Pada Tiap Kategori

Kategori	Jumlah
KDRT	61
PELECEHAN	49
ANAK	40
PENELANTARAN	29
PEMERKOSAAN	27
KEKERASAN	15
LAINNYA	10
PERSELINGKUHAN	9
PERCERAIAN	5
TRAFFICKING	4
TKI	4
ORANG HILANG	4
PENCULIKAN	3
PEMBUNUHAN	1

2. *Preprocessing*

Dataset yang telah dikumpulkan dapat dikatakan belum siap untuk dianalisis, sehingga *preprocessing* merupakan tahap yang sangat penting sebelum dataset siap untuk diolah menggunakan algoritma ditahap selanjutnya. Tahap *preprocessing* digunakan untuk melakukan pembersihan data seperti mengubah huruf besar menjadi huruf kecil, menghilangkan angka dan simbol, memperbaiki kata yang tidak normal dan lain sebagainya agar data siap untuk digunakan pada tahap selanjutnya. Tahapan yang dilakukan pada *preprocessing* yaitu sebagai berikut:

a. *Case Folding*

Pada tahap *case folding* huruf-huruf besar akan diubah menjadi huruf kecil. Gambar 3 menunjukkan masih ada huruf yang berbentuk huruf besar seperti kata “Awal” dan “Sampai”.

Awal stelah 1 hari lebaran pergi ke surabaya untuk silahturahmi di jemput di stasiun pasar turi tgl 17-Juni-2019 diambil tetapi mereka tidak mengizinkannya bahkan saya tidak sampai bertemu dngn anak sy. Sampai sekarang masuk minggu ke-3 anak sy tdk masuk sekolah dan disana pun anak sy tdk sekolah formal hanya sekolah agama dan itu sekolah pun tdk membutuhkan surat-surat dan disana bukan ayahnya yg ngasuh melainkan bibi dari ayahnya. harapan sy: semoga bapak/ibu bisa membantu mengambil anak sy kembali.

Gambar 3. Sebelum Proses Case Folding

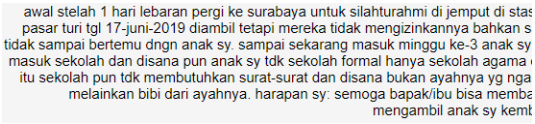
awal stelah 1 hari lebaran pergi ke surabaya untuk silahturahmi di jemput di stasiun pasar turi tgl 17-juni-2019 diambil tetapi mereka tidak mengizinkannya bahkan saya tidak sampai bertemu dngn anak sy. sampai sekarang masuk minggu ke-3 anak sy tdk masuk sekolah dan disana pun anak sy tdk sekolah formal hanya sekolah agama dan itu sekolah pun tdk membutuhkan surat-surat dan disana bukan ayahnya yg ngasuh melainkan bibi dari ayahnya. harapan sy: semoga bapak/ibu bisa membantu mengambil anak sy kembali.

Gambar 4. Sesudah Proses Case Folding

Setelah proses *case folding* dilakukan huruf yang tadinya masih menggunakan huruf besar sudah diubah menjadi huruf kecil. Seperti ditunjukkan pada gambar 4, huruf yang ada pada kata “awal” dan “sampai” sudah tidak menggunakan huruf besar.

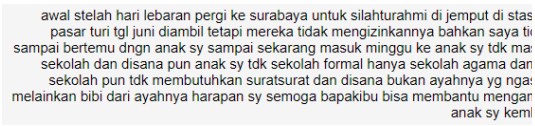
b. *Cleaning*

Tahap *cleaning* digunakan untuk menghilangkan simbol atau karakter khusus dan angka di dalam dokumen.



Gambar 5. Sebelum Proses Cleaning

Gambar 5 menunjukkan dokumen sebelum proses *cleaning* dilakukan, dalam dokumen tersebut masih terdapat tanda strip (-), titik (.) dan angka.

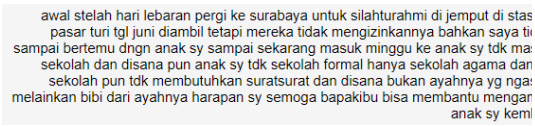


Gambar 6. Setelah Proses Cleaning

Setelah proses *cleaning* dilakukan, tanda strip (-), titik (.) maupun angka sudah dihilangkan dari dokumen seperti ditunjukkan pada gambar 6.

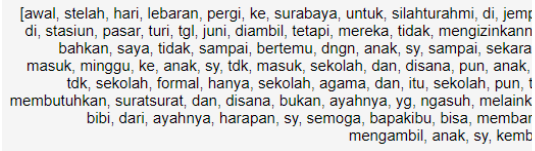
c. *Tokenizing*

Pada tahap *tokenizing* kalimat di dalam dokumen akan dipecah menjadi kata-kata yang dipisahkan dengan tanda koma (,).



Gambar 7. Sebelum Proses Tokenizing

Gambar 7 menunjukkan dokumen yang belum diproses pada tahap *tokenizing* data pada dokumen tersebut masih berupa kalimat.

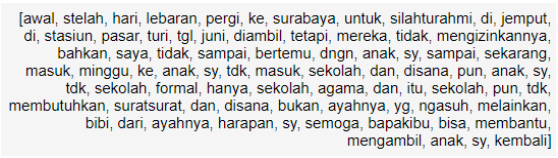


Gambar 8. Setelah Proses Tokenizing

Setelah proses *tokenizing* data berupa kalimat telah dipecah menjadi kata-kata seperti ditunjukkan pada gambar 8. Setiap kata dipisahkan dengan tanda koma (,).

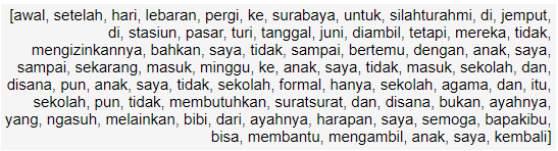
d. *Spell Normalization*

Tahap ini merupakan tahap yang digunakan untuk melakukan normalisasi setiap kata di dalam dokumen.



Gambar 9. Sebelum Proses Spell Normalization

Gambar 9 menunjukkan kata-kata yang belum dinormalisasi sehingga terdapat kata yang masih disingkat seperti kata “sy”, “tgl”, “stelah”, “tdk” dan lain sebagainya.

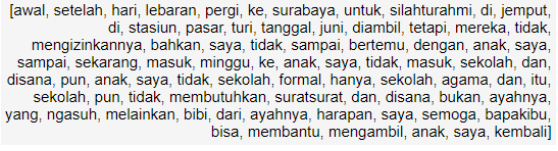


Gambar 10. Setelah Proses Spell Normalization

Setelah proses normalisasi dilakukan, kata-kata yang disingkat sebelumnya sudah dikembalikan ke bentuk normalnya. Gambar 10 menunjukkan kata-kata yang telah dinormalkan seperti kata “saya”, “tanggal”, “setelah”, “tidak” dan lain sebagainya.

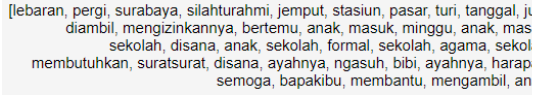
e. *Filtering*

Tahap selanjutnya adalah tahap *filtering* yang merupakan tahap untuk mengambil kata-kata bermakna dengan cara menghilangkan kata yang tidak memiliki makna di dalam dokumen.



Gambar 11. Sebelum Proses Filtering

Gambar 11 menunjukkan dokumen yang masih menyimpan kata-kata yang tidak memiliki makna di dalam dokumen seperti kata “awal”, “setelah”, “hari”, “ke” dan lain sebagainya.

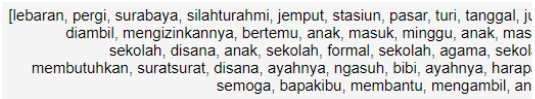


Gambar 12. Setelah Proses Filtering

Setelah proses *filtering* dilakukan, kata-kata tidak bermakna seperti kata “awal”, “setelah”, “hari”, “ke” dan lain sebagainya telah dihilangkan dari dokumen seperti ditunjukkan pada gambar 12.

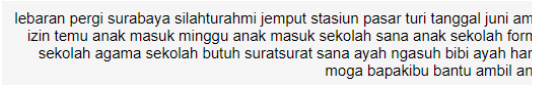
f. *Stemming*

Stemming merupakan tahap terakhir pada tahap *preprocessing*. *Stemming* dilakukan untuk mencari kata dasar dari setiap kata yang ada di dalam dokumen.



Gambar 13. Sebelum Proses Stemming

Gambar 13 menunjukkan kata-kata yang belum diproses pada tahap *stemming* terdapat kata yang belum menggunakan kata dasarnya seperti kata “mengizinkannya”, “bertemu”, “disana”, “mengambil” dan lain sebagainya.



Gambar 14. Setelah Proses Stemming

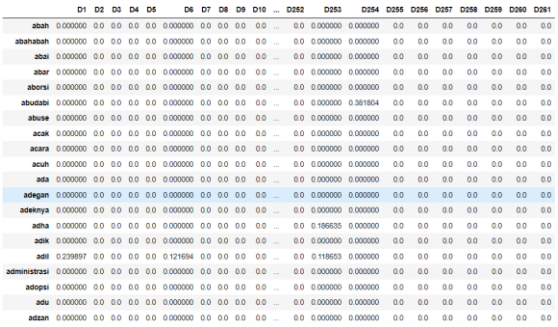
Setelah proses *stemming* dilakukan, setiap kata yang telah diproses sudah diubah menjadi kata dasarnya seperti ditunjukkan pada gambar 14, kata-kata yang sudah diproses yaitu “izin”, “temu”, “sana”, “ambil” dan lain sebagainya.

3. *Transformation*

Pada tahap data transformation terdapat 2 proses yang dilakukan yaitu pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan melakukan *oversampling* menggunakan metode *Random Oversampling*.

a. *Term Frequency-Inverse Document Frequency* (TF-IDF)

Seperti terlihat pada gambar 15 dapat kita diketahui bahwa terdapat beberapa kata yang telah diberi bobot nilai dari hasil pembobotan seperti kata-kata seperti kata “abudabi” memiliki bobot 0,381804 pada dokumen ke 254 (D254), kata “adha” memiliki bobot 0,186635 pada dokumen ke 253 (D253), dan kata “adil” memiliki bobot 0,121694 pada dokumen ke 6 (D6), dan 0,118653 pada dokumen ke 253 (D253).



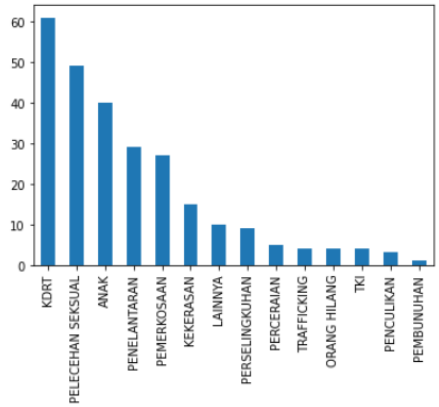
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	...	D252	D253	D254	D255	D256	D257	D258	D259	D260	D261
abah	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
abahabah	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
abai	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
abar	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aborsi	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
abudabi	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.381804	0.0	0.0	0.0	0.0	0.0	0.0	0.0
abuse	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
acak	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
acara	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
acuh	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ada	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adegan	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adeknya	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adha	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.186635	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adik	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adil	0.239897	0.0	0.0	0.0	0.0	0.121694	0.0	0.0	0.0	0.0	...	0.0	0.118653	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
administrasi	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adopsi	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adu	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
adzan	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 15. Hasil Proses TF-IDF

b. *Random Oversampling*

Seperti terlihat pada gambar 16, terdapat ketidakseimbangan jumlah data dari beberapa kategori. Perbedaan jumlah data yang cukup jauh seperti antara kategori KDRT, pelecehan seksual, dan anak jika dibandingkan kategori pembunuhan, penculikan, dan orang hilang. Dari ketidakseimbangan data tersebut maka dapat dilakukan proses oversampling pada data latih menggunakan metode Random Oversampling sehingga data yang digunakan seimbang antara satu kategori dengan kategori lainnya. *Oversampling* dilakukan pada data

latih saja untuk menghindari *overfitting* pada data yang akan diolah.



Gambar 16. Data Sebelum Proses Oversampling

Setelah proses *oversampling* dilakukan maka data pada masing-masing kategori yang tadinya tidak seimbang menjadi seimbang. Hasil *oversampling* pada 4 skenario yang telah dibuat dapat dilihat pada tabel 2.

Tabel 2. Hasil Oversampling

Kategori	Skenario/Perbandingan			
	1/60 :40	2/70 :30	3/80 :20	4/90 :10
KDRT	39	44	52	56
PELECEHAN SEKSUAL	39	44	52	56
ANAK	39	44	52	56
PENELANTARAN	39	44	52	56
PEMERKOSAAN	39	44	52	56
KEKERASA	39	44	52	56
LAINNYA	39	44	52	56
PERSELINGKUHAN	39	44	52	56
PERCERAIA	39	44	52	56
TRAFFICKING	39	44	52	56
TKI	39	44	52	56
ORANG HILANG	39	44	52	56
PENCULIKA	39	44	52	56
PEMBUNUHAN	39	44	52	56

4. *Data Mining/Modeling*

Tahap ini adalah tahap untuk menerapkan teknik klasifikasi menggunakan algoritma *Multinomial Naïve Bayes*. Gambar 17 adalah proses yang digunakan pada tahap *data mining/modeling*.

```
from sklearn.naive_bayes import MultinomialNB

NB = MultinomialNB()
model = NB.fit(X_train_resampled, y_train_resampled)
y_pred = model.predict(X_test)

```

Gambar 17. Proses Data Mining/Modeling

5. *Evaluation*

Setelah proses pembuatan model selesai, model yang telah dibuat dari masing-masing skenario akan diuji. Nilai yang diperoleh dari hasil uji model yaitu nilai *accuracy*, *precision*, *recall*, dan *f-measure* (*f1-score*). Hasil yang diperoleh dari masing-masing skenario dapat dilihat pada tabel 3.

Tabel 3. Hasil Evaluasi Model

Skenario	*Accuracy*	*Precision*	*Recall*	*F-Measure*
1/60:40	39%	28%	28%	26%
2/70:30	47%	45%	42%	41%
3/80:20	45%	23%	20%	21%
4/90:20	41%	24%	26%	23%

SIMPULAN

Penelitian ini telah berhasil membangun model untuk melakukan klasifikasi jenis kekerasan pada perempuan dan anak menggunakan algoritma *Multinomial Naïve Bayes*. Berdasarkan hasil pengujian model yang dibuat menggunakan algoritma *Multinomial Naïve Bayes* dengan melakukan *oversampling* menggunakan *Random Oversampling*, diperoleh hasil terbaik pada model skenario 2 dengan nilai *accuracy* yang diperoleh yaitu 47%, nilai *precision* sebesar 45%, nilai *recall* sebesar 42%, dan nilai *f-measure* (*f1-score*) sebesar 41%.

DAFTAR PUSTAKA

Adi, S. (2018). Implementasi Algoritma Naive Bayes Classifier untuk Klasifikasi Penerima Beasiswa PPA di Universtias Amikom Yogyakarta. *Jurnal Mantik Penusa*, 11-16.

Bahasyim, S. R., Widowati, S., & Husen, J. H. (2021). Otomasi Penelusuran Kebutuhan ke Kode Program menggunakan TF-IDF. *Jurnal Tugas Akhir Fakultas Informatika*, 3272-3281.

Ependi, U., & Putra, A. (2019). Solusi Prediksi Persediaan Barang dengan Menggunakan Algoritma Apriori (Studi Kasus: Regional Part Depo Auto 2000 Palembang). *Jurnal Edukasi dan Penelitian Informatika*, 139-145.

Fahrudin, T., Buliali, J. L., & Fatichah, C. (2019). Enhancing The Performance of SMOTE Algorithm By Using Attribute Weighting Scheme and New Selective Sampling Method For Imbalance Data Set. *International Journal of Innovative Computing, Information and Control*, 423-444.

Fitriani, R. D., Yasin, H., & Tarno. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes. *Jurnal Gaussian*, 11-20.

Gunawan, D., Riana, D., Ardiansyah, D., Akbar, F., & Alfarizi, S. (2020). Komparasi Algoritma Support Vector Machine dan Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023. *Jurnal Teknik Komputer AMIK BSI*, 121-129.

Kalokasari, D. H., Shofi, I. M., & Setyaningrum, A. H. (2017). Implementasi Algoritma Multinomial Naive Bayes Classifier Pada Sistem Klasifikasi Surat Keluar (Studi Kasus: DISKOMINFO Kabupaten Tangerang). *Jurnal Teknik Informatika*, 109-118.

Manalu, E., Sianturi, F. A., & Manalu, M. R. (2017). Penerapan Algoritma Naive Bayes Untuk Memprediksi Jumlah Produksi Barang Berdasarkan Data Persediaan Dan Jumlah Pemesanan Pada Cv. Papadan Mama Pastries. *Jurnal Mantik Penusa*, 16-21.

Muthia, D. A. (2018). Komparasi Algoritma Klasifikasi Text Mining untuk Analisis Sentimen pada Review Restoran. *Jurnal PILAR Nusa Mandiri*, 69-74.

Rahman, A., Wiranto, & Doewes, A. (2017). Online News Classification Using Multinomial Naive Bayes. *Jurnal Ilmiah Teknologi dan Informasi*, 32-38.

Riyani, A., Naf'an, M. Z., & Burhanuddin, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Jurnal Linguistik Komputasional*, 23-27.

Rokhim, A., & Yaqin, A. A. (2017). Implementasi Metode Term Frequency Inversed Document Frequence (TF-IDF) dan Vector Space Model Pada Aplikasi Pemberkasan Skripsi Berbasis Web. *Jurnal SPIRIT*, 34-48.

Ruhyana, N. (2019). Klasifikasi Komentar Instagram Untuk Identifikasi Keluhan Pelanggan Jasa Pengiriman Barang Dengan Teknik SMOTE. *Faktor Exacta*, 280-290.

Yutika, C. H., Adiwijaya, & Faraby, S. A. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-

IDF dan Naive Bayes. *Jurnal Media Informatika Budidarma*, 422-430.