SKANIKA: Sistem Komputer dan Teknik Informatika
Volume xx, Nomor xx, Bulan Tahun, Halaman 193-203 E-ISSN: 2721-4788

Implementasi Algoritma Multinomial Naïve Bayes untuk Mendeteksi *Tweet* Ujaran Kebencian Bahasa Indonesia Terhadap PSSI

Riskiana Wulan[1*], Indra Hertanto[2]
[1,2]Fakultas Teknologi Informasi, Teknik Informatika, Universitas Budi Luhur, Jakarta, Indonesia
E-mail: [1*]riskiana.wulan@budiluhur.ac.id, [2]indra.hertanto@budiluhur.ac.id
(*: corresponding author)

Abstrak

Penelitian ini berfokus pada penerapan algoritma Multinomial Naive Bayes untuk mendeteksi ujaran kebencian pada *tweet* berbahasa Indonesia serta menguji tingkat akurasinya. Menurut *The 2022 World Football Report* sekitar 69% penduduk Indonesia menunjukkan minat yang tinggi terhadap sepak bola, menciptakan lingkungan digital yang positif. *Dataset* yang digunakan terdiri dari data *tweet* terkait PSSI dan politik yang diambil dari twitter sebanyak 2210 *tweet,* yang kemudian dilabel secara manual ke dalam tiga kelas yaitu non-HS (*Hate Speech*), penghinaan dan provokasi. Sebelum membagi *Dataset* menjadi *data train* dan *data testing*, diterapkan teknik *undersampling* untuk menangani ketidakseimbangan kelas, dengan tujuan memastikan distribusi yang seimbang antara ketiga kategori tersebut. Setelah dilakukan *undersampling*, *dataset* pelatihan terdiri dari *350 tweet* dan *dataset* pengujian terdiri dari 88 *tweet*. Evaluasi terhadap setiap metode dilakukan menggunakan *matrix precision*, *recall*, dan *F1-score*. Hasil penelitian mengindikasikan bahwa algoritma Multinomial Naïve Bayes memperoleh akurasi sebesar 62%. Hasil akurasi ini diharapkan dapat bermanfaat untuk mengembangkan model deteksi ujaran kebencian yang efektif dan akurat pada platform media sosial, khususnya Twitter, sehingga dapat membantu mengurangi kesadaran rakyat indonesia akan bahayanya penyebaran ujaran kebencian.

Kata kunci: Ujaran kebencian, *Text Mining*, PSSI, Multinomial Naïve Bayes, TF-IDF

Abstract

This study focuses on the application of the Multinomial Naive Bayes algorithm to detect hate speech in Indonesian tweets and test its accuracy level. According to The 2022 World Football Report, around 69% of Indonesia's population shows a high interest in football, creating a positive digital environment. The Dataset used consists of tweet data related to PSSI and politic taken from Twitter, which is then manually labeled into three classes, namely non-HS (Hate Speech), insults and provocations. The Dataset used consists of 2,210 tweets taken from Twitter, then manually labeled into three classes, namely non-HS (Hate Speech), insults, and provocations. Before dividing the Dataset into train and test data, an undersampling technique was applied to handle class imbalance, with the aim of ensuring a balanced distribution between the three categories. After undersampling, the training Dataset consisted of 350 tweets and the test Dataset consisted of 88 tweets. Evaluation of each method was carried out using matrix precision, recall, and F1-score. The results of the study indicate that the Multinomial Naïve Bayes algorithm obtained an accuracy of 62%. This accuracy result is expected to be useful for developing an effective and accurate hate speech detection model on social media platforms, especially Twitter, so that it can help reduce the awareness of the Indonesian people about the dangers of the spread of hate speech.

***Keywords**: Hate Speech, Text Mining, PSSI, Multinomial Naïve Bayes, TF-IDF*

1. PENDAHULUAN

Tweet pada twitter dapat memiliki sifat positif dan negatif. *Tweet* bersifat negatif perlu perhatian khusus karena dapat mengandung ujaran kebencian. Salah satu sektor dalam teknologi informasi yang telah mengalami kemajuan yang sangat cepat dalam beberapa dekade terakhir adalah teknologi informasi yang berkaitan dengan media sosial. Istilah milenial merujuk kepada generasi masyarakat yang mahir dalam menggunakan media sosial dan teknologi terbaru dalam kegiatan sehari-hari[1]. Menurut Statista Research Department, Indonesia merupakan negara dengan pengguna twitter peringkat 5 didunia dengan jumlah lebih dari 24,85 juta pengguna[2]. Kemudian disebutkan juga di dalam artikel yang berjudul *More than half of adults across 34 countries plan to watch the 2022 FIFA World Cup* bahwa 69% penduduk Indonesia tertarik dengan sepak bola[3]. Menurut Litbang Kompas, kegagalan Indonesia dalam menjadi penyelenggara Piala Dunia U-20 telah menimbulkan rasa kecewa di antara sebagian masyarakat. Banyak masyarakat menyalahkan narasi-narasi berbau politis sebagai penyebab utama kegagalan tersebut. Eskalasi

masalah di media sosial semakin bertambah pada 29 Maret 2023 setelah FIFA secara resmi menarik kembali status Indonesia sebagai penyelenggara Piala Dunia U-20[4]. Sejak saat itu, telah terjadi peningkatan kasus yang terkonfirmasi sebagai tindakan ujaran kebencian dari waktu ke waktu. Ujaran Kebencian atau *Hate Speech* merujuk pada ungkapan yang menyerang kelompok atau individu tertentu karena sifat yang dimiliki oleh mereka dan yang berpotensi mengganggu keharmonisan sosial[5]. Dengan menggunakan algoritma *Machine Learning*, data bisa diubah menjadi tindakan cerdas. Algoritma menerima data dan mengidentifikasi pola yang menarik dan mengubahnya menjadi tindakan. Proses pembelajaran mesin memiliki tiga komponen seperti *input* data, Abstraksi dan Generalisasi[6]. Algoritma Multinomial Naïve Bayes adalah salah satu teknik pembelajaran berbasis probabilitas yang berlandaskan pada teorema Bayes, serta sering digunakan dalam bidang *Natural Language Processing* (NLP). Metode ini bekerja dengan memanfaatkan konsep frekuensi istilah, yang menggambarkan seberapa sering kata tertentu muncul dalam suatu teks. Model ini mempertimbangkan dua aspek penting: Pertama, adanya kata di dalam dokumen, dan kedua, seberapa sering kata itu muncul dalam dokumen tersebut.

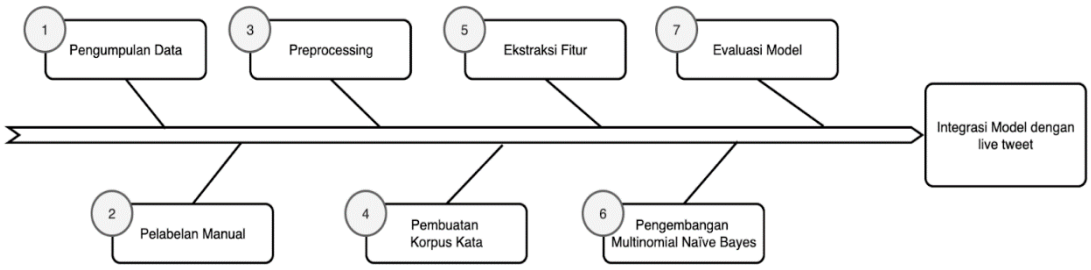
Beberapa penelitian sebelumnya telah dilakukan untuk menganalisis opini masyarakat menggunakan data dari media sosial dengan memanfaatkan algoritma tertentu. Salah satu penelitian menggunakan Algoritma Naïve Bayes Multinomial untuk mengklasifikasikan pandangan pemerintah terhadap penanganan Covid-19 dengan data yang berasal dari Twitter dan memperoleh tingkat akurasi sebesar 74% melalui analisis 2000 data[7].

Penelitian lainnya yang berjudul Penerapan Algoritma Multinomial Naïve Bayes, TF-IDF, serta Confusion Matrix untuk Klasifikasi Saran Monitoring dan Evaluasi Mahasiswa Terhadap Dosen di Program Studi Teknik Informatika Universitas Dayanu Ikhsanuddin mencapai akurasi 85% menggunakan 1037 data[8]. *Confusion matrix* merupakan tabel kontingensi yang digunakan untuk mengevaluasi kinerja model klasifikasi, di mana elemen diagonal melambangkan klasifikasi yang benar dan elemen di luar diagonal melambangkan kesalahan klasifikasi[9].

Penelitian yang lain juga dilakukan untuk meningkatkan kinerja klasifikasi algoritma Naive Bayes melalui penerapan pembobotan fitur dan kalibrasi Laplace dalam konteks manajemen risiko trafik. Hasil simulasi numerik menunjukkan bahwa dengan jumlah sampel besar, akurasi algoritma ini mencapai lebih dari 99% dan sangat stabil. Jika jumlah atribut sampel kurang dari 400 dan kategori kurang dari 24, akurasinya tetap di atas 95%[10].

Penelitian sebelumnya juga telah menerapkan algoritma Multinomial Naïve Bayes untuk melakukan klasifikasi spesialisasi peminatan siswa SMA berdasarkan nilai peminatan rata-rata siswa dengan dengan *confusion matrix*. Hasil menunjukkan akurasi yang cukup tinggi yaitu 96.19% dengan perbandingan *data training* 70% dan 30% *data testing*[11] . Dalam penelitian ini, perhatian kami difokuskan pada evaluasi akurasi algoritma dalam mengidentifikasi *tweet* yang mengandung ujaran kebencian dengan berbagai kategori, sekaligus menganalisis kinerja komputasi algoritma ketika dijalankan secara *real-time*. Penelitian ini juga diharapkan dapat memberikan kesadaran bagi masyarakat akan bahayanya ujaran kebencian yang disampaikan melalui media sosial serta manfaat bagi pemerintah untuk mendapatkan kesan dari masyarakat terhadap kepercayaan masyarakat kepada PSSI.

2. METODE PENELITIAN



Gambar 1. Diagram Alir Implementasi Algoritma Multinomial Bayes

Berikut adalah penjelasan yang lebih jelas mengenai Gambar 1 diagram alir implementasi Algoritma Multinomial Bayes:

2.1 Pengumpulan Data

Sumber data yang digunakan dalam penelitian ini diambil dari situs media sosial X yang berfokus pada *tweet* berbahasa indonesia mulai dari 29 maret 2023 sampai dengan 27 Januari 2025 dengan jumlah *dataset* yang didapat sebanyak 2210 *tweet*. Kumpulan data yang dikumpulkan terdiri dari *tweet* yang relevan selama periode tersebut, dengan kata kunci seperti PSSI dan politik. Selanjutnya, informasi ini akan dipecah ke dalam dua bagian: data pelatihan dan data pengujian. Data pelatihan mencakup sekumpulan data yang telah dilabeli dengan klasifikasi dan digunakan untuk mengembangkan model kata, sementara data pengujian tidak memiliki label klasifikasi dan digunakan untuk menguji efektivitas model yang telah dikembangkan.

2.2 Pelabelan Manual

Dalam proses pelabelan data *tweet* terdapat tiga kelas yaitu non hs *(hate speech)*, penghinaan dan provokasi. *Tweet* akan dianalisa menggunakan makna linguistik dengan konsep konseptual. Konsep konseptual merupakan makna yang bebas konteks jadi kata atau kalimat diartikan sesuai arti gramatikal tanpa mempertimbangkan konteks.

2.3 *Preprocessing* Data

Data yang terkumpul selanjutnya akan melalui tahap *preprocessing*. Proses ini terdiri dari beberapa langkah, yaitu:

a. *Case Folding*: Pada fase ini, semua huruf besar diubah menjadi huruf kecil. Tujuannya adalah untuk menyeragamkan format teks agar lebih mudah diproses.
b. *Symbol Removal*: Langkah ini melibatkan penghapusan simbol dan karakter khusus, termasuk tanda *baca* seperti koma, titik, tanda tanya, tanda seru, serta elemen lainnya seperti mentions, tautan, angka, dan karakter khusus seperti simbol dolar ($), persen (%), atau tanda bintang (*).
c. *Slang Word:* Pada tahap ini, kata-kata tidak baku diubah menjadi bentuk resmi atau bakunya. Kata-kata tersebut biasanya berupa singkatan atau istilah dalam bahasa gaul.
d. *Stemming*: Proses ini bertujuan untuk mengembalikan kata-kata berimbuhan ke bentuk dasar sesuai *dengan* aturan tertentu.
e. *Stop Word Removal*: Langkah ini diterapkan untuk memilih kata-kata yang signifikan dan menghilangkan kata-kata yang tidak berperan dalam proses klasifikasi. Hal ini bertujuan untuk mengurangi jumlah kata yang akan digunakan dalam pelatihan model dan meningkatkan efisiensi penghitungan bobot istilah. Proses ini membutuhkan kamus sinonim untuk mengidentifikasi kata-kata dengan makna serupa [12].
f. Post *preprocessing* data *Cleaning* dan *Balancing*: Langkah *cleaning* untuk membersihkan data yang menjadi *kosong* dan duplikat karena proses *preprocessing*. *Balancing* dilakukan untuk memastikan sebaran data label yang dimiliki proporsional.

2.4 Pembuatan Korpus Data

Pembuatan korpus kata yaitu proses pengambilan seluruh *Dataset training* lalu diiterasi satu persatu untuk diambil kata-katanya, dan dimasukkan ke dalam korpus jika kata tersebut belum tersedia.

2.5 Ekstraksi Fitur

Setelah *preprocessing*, langkah berikutnya adalah ekstraksi fitur. Fitur-fitur ini merupakan aspek-aspek khusus dari data yang akan digunakan dalam model analisis. TF-IDF merupakan suatu metode yang dipakai untuk memberikan nilai atau bobot pada setiap kata dalam kumpulan teks dibandingkan dengan semua dokumen, dengan maksud untuk mengetahui seberapa pentingnya suatu kata dalam satu dokumen. Proses pembobotan TF-IDF melibatkan beberapa tahapan. Pertama, data dalam *Dataset* harus melalui tahapan *preprocessing*. Selanjutnya, *Dataset* tersebut membentuk korpus kata setelah *preprocessing*. Nilai *Term Frequency* (TF) dihitung untuk setiap kata dalam korpus terhadap seluruh *Dataset*. Selanjutnya, nilai *Document Frequency* (DF) dihitung untuk setiap kata dalam korpus kata. Setelah itu, *value* dari *Inverse Document Frequency* (IDF) dihitung untuk masing-masing kata di dalam korpus. Tahapan terakhir merupakan mengalikan nilai TF dan IDF.

2.6 Pengembangan model Multinomial Naïve Bayes

Setelah memperoleh nilai TF-IDF untuk setiap kata di dalam korpus, langkah berikutnya adalah menerapkan metode Multinomial Naïve Bayes. Nilai TF-IDF yang ada akan dipakai untuk membangun model. Model ini akan digunakan dalam klasifikasi data yang ingin diidentifikasi. Proses klasifikasi akan dilakukan dengan menggunakan algoritma Multinomial Naïve Bayes.

2.7 Evaluasi Model.

Model yang telah dibuat dievaluasi menggunakan metrik *F1 Score* yang merupakan pengukuran yang menggabungkan nilai *Precision* dan *Recall* untuk menentukan seberapa baik model dalam mengenali teks. *Precision* mengukur seberapa banyak prediksi positif yang benar, sementara *Recall* mengukur seberapa banyak kasus positif yang berhasil dikenali oleh model. Dengan menggabungkan kedua metrik ini, *F1 Score* memberikan gambaran yang lebih seimbang tentang kinerja model, terutama pada data yang tidak seimbang, seperti dalam deteksi teks kebencian, di mana penting untuk menangani baik *false positive* maupun *false negative* secara adil. Semakin tinggi nilai *F1 Score*, semakin baik kemampuan model dalam mendeteksi teks yang relevan dengan target tanpa banyak kesalahan

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Crawling merupakan metode otomatis yang digunakan dalam pengumpulan dan menyusun data dari berbagai sumber, termasuk *website*, basis data, dan berkas [13]. Dalam penelitian ini, proses akuisisi data dilakukan dengan menjelajahi informasi yang terdapat dalam *tweet*. *Dataset* yang digunakan terdiri dari *tweet* yang diambil dari Twitter melalui proses *crawling*, menggunakan alat bernama *tweet* harvest (Twitter Crawler), yang dikembangkan dengan bahasa pemrograman Python. Proses *crawling Dataset* dilakukan pada 18 Juli 2024 – 27 Januari 2025 dengan kata kunci: PSSI, politik dan konfigurasi data *tweet* berbahasa indonesia. *Dataset* yang berhasil terkumpul kemudian disimpan dalam bentuk file csv. Total data yang terkumpul berjumlah 2210 *tweet*. *Sample Dataset* yang berhasil dikumpulkan terdapat di dalam Tabel 1 berikut:

Tabel 1. Tabel Sample *Dataset* dari Twitter

No.	Source	Text
1.	Twitter	@tempodotco BUMN tidak kau urus Erick? Kenapa gak lepas jabatan Menteri kalo gak mau urus BUMN? Apa peranmu atasi korupsi salah kelola dan hutang2 BUMN yg jumbo? Apa tanggung jawabmu angkat badut2 jd Komisaris2 BUMN? Kau gunakan BUMN & PSSI sbg batu loncatan utk karir politik & bisnis2mu ?
2.	Twitter	@SiaranBolaLive Ketum PSSI makin hari makin Tolol sepakbola di bawa jadi kendaraan politik. Udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini.

3.2 Pelabelan Manual

Setelah data dikumpulkan, kemudian akan diberi label secara manual. Pelabelan manual dilakukan dalam penelitian ini karena berbagai alasan penting, khususnya untuk konteks Indonesia. Bahasa Indonesia memiliki nuansa budaya, slang, dan variasi regional yang kompleks, sehingga pelabelan manual diperlukan untuk memastikan interpretasi yang akurat terhadap sarkasme, makna ganda, atau referensi budaya. Ujaran kebencian, provokasi, dan penghinaan juga sering kali bersifat kontekstual, sehingga membutuhkan pemahaman manusia terhadap niat dan konteks, yang sulit dicapai oleh model *machine learning* saat ini. Selain itu, model otomatis yang ada umumnya tidak dirancang untuk menangani *Dataset* khusus Indonesia, sehingga berpotensi menghasilkan bias atau ketidakakuratan. Dengan pelabelan manual, penelitian ini dapat menjamin kualitas *data ground truth* yang tinggi dan bebas bias, yang sangat penting untuk melatih dan mengevaluasi model secara efektif. Ukuran *Dataset* yang digunakan (sekitar 2000 sampel) juga memungkinkan pelabelan manual dilakukan secara praktis, memberikan keakuratan yang lebih tinggi dibandingkan pendekatan otomatis.

Dalam proses pelabelan data *tweet* terdapat tiga kelas yaitu non hs (*hate speech*), penghinaan dan provokasi. *Tweet* akan dianalisa menggunakan makna linguistik dengan konsep konseptual.

Konsep konseptual merupakan makna yang bebas konteks jadi kata atau kalimat diartikan sesuai arti gramatikal tanpa mempertimbangkan konteks. Berikut ini adalah contoh data pelabelan manual yang terlihat pada contoh di dalam Tabel 2:

Tabel 2. Sample *Dataset* Hasil *Labeling* Manual

No.	Sumber	*Tweet*	Label
1.	Twitter	@tempodotco BUMN tidak kau urus Erick? Kenapa gak lepas jabatan Menteri kalo gak mau urus BUMN? Apa peranmu atasi korupsi salah kelola dan hutang2 BUMN yg jumbo? Apa tanggung jawabmu angkat badut2 jd Komisaris2 BUMN? Kau gunakan BUMN & PSSI sbg batu loncatan utk karir politik & bisnis2mu ?	provokasi
2.	Twitter	@SiaranBolaLive Ketum PSSI makin hari makin Tolol sepakbola di bawa jadi kendaraan politik. Udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini.	penghinaan

3.3 *Pre-processing*

Pre-processing merupakan fase pertama dimana tujuannya adalah untuk menyiapkan dan membersihkan teks dari istilah yang tidak relevan. Proses ini dilakukan untuk mengorganisir data agar teks menjadi lebih teratur dengan cara menghilangkan elemen yang mengganggu. Tujuannya adalah untuk memudahkan proses klasifikasi. Dalam penelitian ini, tahap *pre-processing* meliputi beberapa langkah, yaitu *case folding*, penghapusan simbol, pengolahan kata *slang*, *stemming*, dan penghilangan kata henti (*stopword*).

a. *Case Folding*

Case Folding merupakan tahapan untuk mentransformasi semua huruf besar menjadi huruf kecil. Langkah ini bertujuan untuk memastikan format teks tetap konsisten, sehingga dapat lebih mudah dibaca dan diproses, seperti contoh di dalam Tabel 3.

Tabel 3. Perbandingan Hasil *Case Folding*

Sebelum	Setelah *Case Folding*
@SiaranBolaLive Ketum PSSI makin hari makin Tolol sepakbola di bawa jadi kendaraan politik. Udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini.	@siaranbolalive ketum PSSI makin hari makin tolol sepakbola di bawa jadi kendaraan politik. udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini.

b. *Symbol Removal*

Symbol Removal adalah suatu langkah untuk mengeliminasi simbol dan karakter khusus, termasuk tanda baca (seperti koma, titik, tanda tanya, dan tanda seru), nama atau tautan, angka (0-9), serta karakter tambahan seperti $, %, dan lainnya. Proses ini terdapat pada contoh yang ada pada Tabel 4.

Tabel 4. Perbandingan Hasil *Symbol Removal*

Sebelum	Sesudah Simbol Removal
@siaranbolalive ketum PSSI makin hari makin tolol sepakbola di bawa jadi kendaraan politik. udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini.	siaranbolalive ketum PSSI makin hari makin tolol sepakbola di bawa jadi kendaraan politik udah la gak usah berharap lagi untuk sepakbola kedepannya kalau ketemu federasi nya modelan macem gini

c. *Slangword*

Slangword bertujuan untuk mentransformasi kata-kata yang tidak resmi ke dalam istilah yang memenuhi kaidah bahasa yang benar dan siap untuk diolah. Istilah-istilah yang tidak resmi ini dapat mencakup akronim atau bahasa yang biasa digunakan sehari-hari. Contoh istilah non-baku yang kerap muncul di platform media sosial antara lain "begajulan" yang berarti nakal, "bumil" yang merujuk pada ibu hamil, dan "mager" yang berarti malas bergerak. Pada tahap ini, istilah-

istilah tersebut akan disesuaikan ke bentuk standar agar dapat diproses dengan lebih baik, seperti contoh yang terdapat di dalam Tabel 5.

Tabel 5. Perbandingan Hasil *Slangword*

Sebelum	Sesudah Slangword
apa **gak** takut disanksi PSSI ini	apa **tidak** takut disanksi PSSI in

d. *Stemming*

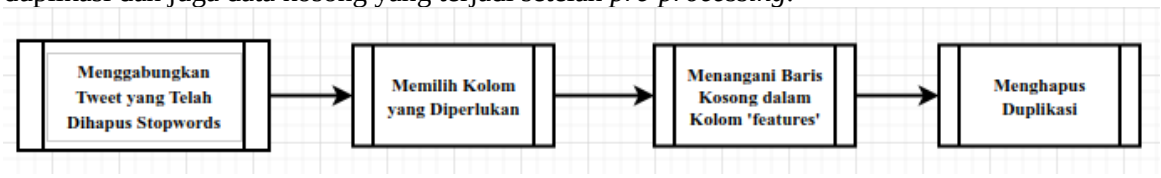
Stemming adalah suatu metode pemrosesan kata yang bertujuan untuk memperoleh bentuk dasar suatu kata setelah menghapus imbuhan berdasarkan ketentuan tertentu. Metode ini diperlukan karena kata-kata yang memiliki awalan, akhiran, atau sisipan dapat mempersulit pencarian kata-kata yang sejenis, yang terlihat di dalam Tabel 6 berikut:

Tabel 6. Perbandingan Hasil *Stemming*

Sebelum	Sesudah *Stemming*
apa **gak** takut disanksi PSSI ini	apa tidak takut sanksi PSSI ini

3.4 *Post Pre-Processing Cleaning* dan *Balancing*

Setelah data *tweet* dilakukan *pre-processing*, langkah selanjutnya adalah menghapus data duplikasi dan juga data kosong yang terjadi setelah *pre-processing*.



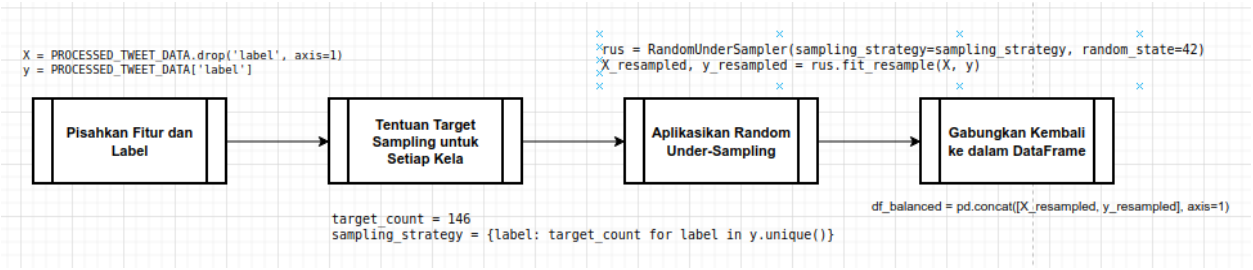
Gambar 2. Tahapan *Post Pre-Processing Cleaning*

Gambar 2 menjelaskan tahapan yang dilakukan pada *Post Pre-Processing* antara lain yaitu dimulai dengan menggabungkan kembali kata-kata yang telah dibersihkan dari *stopwords* menjadi sebuah string untuk setiap *tweet*, dan menyimpannya dalam kolom baru *features*. Kemudian dilanjutkan dengan memilih hanya kolom *features* dan label dari *dataset* yang telah dibersihkan untuk analisis lebih lanjut. Berikutnya menghapus baris di mana kolom *features* memiliki string kosong, karena data tersebut tidak memberikan kontribusi pada analisis model. Dan yang terakhir menghapus duplikat berdasarkan kolom *features* untuk menghindari data yang sama diproses berulang kali, dan agar *Dataset* lebih representatif. Setelah dilakukan tahap *post preprocessing* dari total data awal 2211 terdapat 3 data yang kosong dan 34 duplikasi data, sehingga menjadikan total data *tweet* berjumlah 2042 dengan distribusi label yang terlihat di dalam Tabel 7.

Tabel 7. Distribusi Data *Tweet* Setelah Proses *Cleaning*

Kelas	Jumlah
non hs	1.612
provokasi	146
penghinaan	284
Total	2.042

Data *training* untuk setiap kategori non hs, provokasi, dan penghinaan disarankan memiliki jumlah yang sama dan setelah dianalisa kategori terendah berjumlah 146 data pada kategori provokasi, sehingga dilakukan proses *balancing* agar data kategori lainnya juga sama dengan proses yang ditunjukkan pada Gambar 3.



Gambar 3. Proses *Balancing*

Pada Gambar 3 merupakan proses *balancing* menggunakan *RandomUnderSampler* dari pustaka imblearn. Pertama, *dataset* PROCESSED_*TWEET*_DATA dipisahkan menjadi dua bagian: X yang berisi fitur dan y yang berisi label (target). Kemudian, strategi pemilihan sampel ditentukan dengan menetapkan jumlah target untuk setiap kelas, di sini ditentukan sebanyak 146 sampel untuk setiap kelas. Proses *undersampling* dilakukan dengan menerapkan *RandomUnderSampler* yang bertujuan mengurangi jumlah sampel dari kelas mayoritas sehingga setiap kelas dalam *dataset* memiliki jumlah sampel yang seimbang, yaitu 146 setiap kelas dengan total data menjadi 350 data. Dari 2.042 data, sebanyak 1.692 data tidak digunakan dalam proses *training* dengan rincian 1.466 kelas non hs dan 138 kelas penghinaan. Hasil dari proses ini adalah dua variabel X_resampled dan y_resampled, yang kemudian digabungkan kembali menjadi sebuah DataFrame df_*balanced*. DataFrame ini berisi data yang telah di-sampling ulang dengan jumlah sampel yang seimbang, yang bertujuan untuk mencegah bias pada model akibat ketidakseimbangan kelas dalam *Dataset* asli. Klasifikasi merupakan metode dalam *data mining* yang digunakan untuk memprediksi keterkaitan antara data dalam suatu himpunan data. Prediksi ini dilakukan dengan membagi data ke dalam sejumlah kategori yang berbeda dengan memperhatikan aspek-aspek tertentu [14].

3.5 Pemisahan Data *Training* dan *Testing*

Tabel 8 memperlihatkan proses pemisahan antara data pelatihan dan pengujian. Dari total data *tweet* sebanyak 2210, *dataset* ini telah dibagi menjadi *data training* sebesar 350 data (80%) dan *data testing* sebanyak 88 data (20%) dengan mempertahankan distribusi yang seimbang untuk masing-masing kategori label: *non-hs* (non-hate speech), penghinaan, dan provokasi.

Tabel 8. Pemisahan Data *Training* dan *Testing*

Dataset	**Label**	**Proportion (%)**	**Jumlah Data**
Training Set	non hs	33.14%	116
	penghinaan	33.43%	117
	provokasi	33.43%	117
Testing Set	non hs	34.09%	30
	penghinaan	32.95%	29
	provokasi	32.95%	29

3.6 Pembuatan Korpus data dan Ekstraksi Fitur

Berdasarkan setiap kolom pada Gambar 4 merepresentasikan kata-kata unik yang terdapat dalam *dataset* teks (sebagai fitur), sementara setiap baris merepresentasikan dokumen yang telah diubah menjadi vektor numerik. Nilai dalam tabel menunjukkan skor TF-IDF untuk setiap kata dalam dokumen tertentu, di mana nilai 0.0 berarti kata tersebut tidak muncul dalam dokumen tersebut. Jika sebuah kata memiliki nilai yang lebih tinggi, maka kata tersebut dianggap lebih penting atau lebih unik dalam dokumen yang bersangkutan. Berdasarkan informasi yang tersedia, tabel ini memiliki 350 baris, yang berarti terdapat 350 dokumen dalam data pelatihan, serta 2097 kolom, yang menunjukkan adanya 2097 kata unik yang dipilih oleh TF-IDF Vectorizer. Dengan

tabel ini, kita dapat melakukan analisis terhadap kata-kata yang memiliki bobot lebih tinggi dan mengidentifikasi fitur penting dalam setiap dokumen.

```
# Initialize Count Vectorizer
wc_vectorizer = CountVectorizer()

# Fit and transform training data, transform test data
X_train_wc = wc_vectorizer.fit_transform(X_train)
X_test_wc = wc_vectorizer.transform(X_test)

pd.DataFrame(X_train_wc.toarray(), columns=wc_vectorizer.get_feature_names_out())
```

	abaikan	abe	acak	acara	aceh	acungi	ada	adem	adik	adil	...	yes	ygg	you	your	youtube	zakat	zionis	zirkzee	zobirin	zoom
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
345	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
346	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
347	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
348	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
349	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

350 rows × 2097 columns

Gambar 4. Program Ekstraksi Fitur TF-IDF

3.7 Pengembangan Multinomial Naïve Bayes dan Evaluasi Model

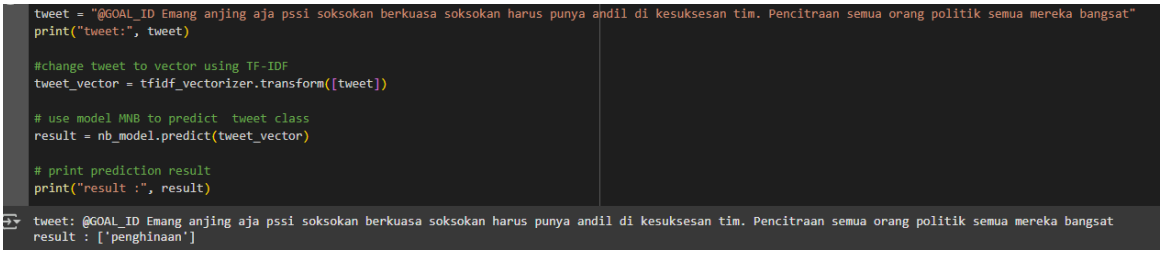
Dalam model multinomial, sebuah dokumen (pada penelitian kali ini adalah *tweet*) dianggap sebagai '*bags of words*'. Tidak ada urutan kata-kata yang dipertimbangkan, tetapi frekuensi setiap kata dalam *tweet* yang ditangkap[15]. Model multinomial Naive Bayes menggunakan *library* sklearn untuk model dan metrik evaluasi. Akurasi menyediakan gambaran sejauh mana model dapat melakukannya memprediksi dengan benar semua data yang diamati[16]. Tingkat ketelitian pada tahap ini diukur dengan nilai kinerja klasifikasi dengan melakukan pengujian model dan penilaian model[17].

Tabel 9. Hasil Akurasi Metode Multinomial Naïve Bayes

Kategori	**Presisi**	***Recall***	**F1-score**	**Jumlah Data (*Support*)**
Non-HS	66%	63%	64%	30
Penghinaan	60%	62%	61%	29
Provokasi	62%	62%	62%	29
Akurasi Keseluruhan	62%			88

Tabel 9 berisi hasil akurasi metode multinomial naive bayes yang telah dijalankan dari program untuk pengujian kedua yaitu dengan menggunakan 438 *Dataset* di mana *data training* sebesar 350 data (80%) dan *data testing* sebanyak 88 data (20%). Akurasi model adalah 62%. Confusion Matrix yang digunakan adalah 3x3 dari model klasifikasi sebanyak 3 kelas yaitu non hs, penghinaan, dan provokasi.

Gambar 5 merupakan hasil pengujian keakuratan dengan sampel *tweet* data dalam pengujian keberhasilan prediksi yang dilakukan menggunakan multinomial naive bayes. Terlihat bahwa model ini memperoleh hasil yang akurat untuk contoh *tweet* yang berisi kategori penghinaan. Hasil akurasi pada evaluasi model Multinomial Naïve Bayes tersebut didapatkan dari rumus yang dijelaskan pada Tabel 10.



Gambar 5. Hasil Pengujian Sampel *Tweet* Data

Tabel 10. Rumus Evaluasi Model Multinomial Naive Bayes

Metrik	Rumus	Deskripsi
Akurasi	$\frac{TP+TN}{TP+TN+FP+FN}$	Tingkat keseluruhan prediksi yang benar oleh model.
Presisi	$\frac{TP}{TP+FP}$	Seberapa banyak prediksi positif yang benar.
Recall (Sensitivitas)	$\frac{TP}{TP+FN}$	Seberapa banyak data positif yang berhasil terdeteksi.
F1-score	$\frac{2\times \text{Presisi}\times \text{Recall}}{\text{Presisi}+\text{Recall}}$	Nilai gabungan antara Precision dan *Recall* untuk keseimbangan keduanya.

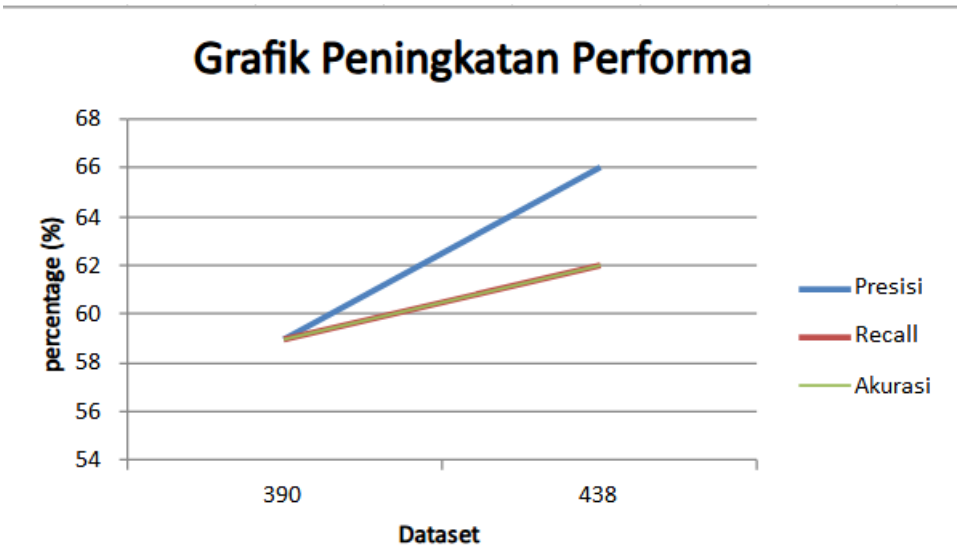
Keterangan:
TP (*True Positive*): Kasus yang benar-benar positif dan diklasifikasikan dengan benar.
TN (*True Negative*): Kasus yang benar-benar negatif dan diklasifikasikan dengan benar.
FP (*False Positive*): Kasus negatif yang salah diklasifikasikan sebagai positif.
FN (*False Negative*): Kasus positif yang salah diklasifikasikan sebagai negatif.

Dengan menggunakan metrik-metrik ini, kita dapat mengevaluasi kinerja model Multinomial Naïve Bayes dengan lebih menyeluruh dan merencanakan perbaikan jika diperlukan. Pengujian dilakukan sebanyak 2x terlihat dari perbandingan hasil yang didapat pada Tabel 11. *Dataset* pertama memiliki 390 data, dengan 312 data untuk *training* dan 78 data untuk *testing*. Model menghasilkan presisi sebesar 59%, *recall* 59%, dan akurasi 59%. Sementara itu, *dataset* kedua memiliki 438 data, dengan 350 data untuk *training* dan 88 data untuk *testing*. Hasil pengujian menunjukkan peningkatan performa, yaitu presisi 66%, *recall* 62%, dan akurasi 62% terlihat pada Gambar 6.

Tabel 11. Perbandingan Hasil Data *Testing* Pertama dan Kedua

Dataset	Jumlah *Record*		Proporsi *Testing*	Proporsi *Training*	Presisi	*Recall*	Akurasi
	Training	*Testing*					
390	312	78	20%	80%	59%	59%	59%
438	350	88	20%	80%	66%	62%	62%

Perbedaan akurasi antara kedua *dataset* disebabkan oleh beberapa faktor utama. Jumlah data training yang lebih besar pada *dataset* kedua memungkinkan model untuk mengenali pola dengan lebih baik, sehingga meningkatkan akurasi. Selain itu, distribusi data yang tidak seimbang pada *dataset* pertama bisa menyebabkan bias dalam model, yang berkontribusi terhadap penurunan akurasi. Keanekaragaman fitur juga berpengaruh, di mana *dataset* kedua mungkin memiliki fitur yang lebih representatif terhadap kategori yang dipelajari, sehingga model dapat melakukan generalisasi dengan lebih baik. Faktor lain yang perlu diperhatikan adalah *overfitting* atau *underfitting*, dimana *dataset* pertama kemungkinan mengalami *underfitting* karena jumlah data pelatihan yang lebih sedikit, sehingga model kesulitan menangkap pola yang kompleks. Dengan demikian, peningkatan jumlah data pelatihan dan keseimbangan distribusi data berperan penting dalam meningkatkan akurasi model Multinomial Naïve Bayes.



Gambar 6. Grafik Hasil Perbandingan Pengujian Berdasarkan Jumlah *Dataset*

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan hasil analisis dan evaluasi penelitian ini, penerapan metode Multinomial Naïve Bayes dengan *Dataset* berjumlah 438 *tweet*, terdiri dari 350 data pelatihan dan 88 data pengujian, menghasilkan akurasi sebesar 62%. Sebagai perbandingan, pada pengujian sebelumnya dengan *Dataset* yang lebih kecil, yaitu 390 *tweet* dengan 312 data pelatihan, dan 78 data pengujian, model hanya mencapai akurasi sebesar 59%. Perbedaan ini menunjukkan bahwa peningkatan jumlah data pelatihan memberikan kontribusi positif terhadap performa model.

Hasil ini juga mengindikasikan bahwa performa metode Multinomial Naïve Bayes dan teknik TF-IDF sangat bergantung pada jumlah data serta variasi kategori dalam *Dataset*. Akurasi yang lebih tinggi pada pengujian kedua dapat disebabkan oleh distribusi data yang lebih seimbang dan jumlah fitur yang lebih representatif, sehingga model dapat menggeneralisasi pola dengan lebih baik. Meskipun akurasi model masih belum optimal, penelitian ini memberikan dasar untuk pengembangan lebih lanjut dengan memanfaatkan *dataset* yang lebih besar dan lebih beragam guna meningkatkan efektivitas deteksi ujaran kebencian pada platform media sosial.

4.2 Saran

Sebagai langkah pengembangan pada penelitian berikutnya adalah penelitian ini dapat mengintegrasikan pelabelan semi-otomatis, di mana model *machine learning* digunakan untuk melakukan pelabelan awal yang kemudian disempurnakan secara manual, sehingga proses menjadi lebih efisien. Selain itu, *dataset* dapat diperluas dengan variasi dialek, slang, dan gaya bahasa khas Indonesia dari berbagai daerah untuk meningkatkan performa model dalam menangani data yang lebih beragam. Penelitian selanjutnya juga dapat berfokus pada pengembangan model NLP khusus bahasa Indonesia dengan arsitektur yang lebih optimal, seperti *Transformer*, yang disesuaikan dengan data lokal. Kemudian, implementasi model ini pada aplikasi nyata, misalnya untuk mendeteksi ujaran kebencian secara real-time di media sosial, dapat memberikan dampak langsung sekaligus validasi terhadap kehandalan model dalam lingkungan dunia nyata.

DAFTAR PUSTAKA

[1] A. C. Sitepu, W. Wanayumini, and Z. Situmorang, "Determining Bullying Text Classification Using Naive Bayes Classification on Social Media," *Jurnal Varian*, vol. 4, no. 2, pp. 133–140, 2021.

[2] "Statista Research Department. 2024. *Leading Countries Based on Number of X (Formerly Twitter) Users as of April 2024*." Available:

https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed: 08-Des-2024].

[3] R. Grimm, and N. Boyon, "*More Than Half of Adults Across 34 Countries Plan to Watch the 2022 Fifa World Cup*." Available: https://www.ipsos.com/sites/default/files/ct/news/documents/2022-11/Ipsos%202022%20FIFA%20World%20Cup%20Global%20Advisor%20Survey%20-%20Global%20Press%20Release.pdf. [Accessed: 08-Des-2024].

[4] I. Firdaus (2023, April 1), "*Kegagalan Indonesia Jadi Tuan Rumah Piala Dunia U20, Keriuhan Warganet dan Gocekan Para Politisi*." Available: https://www.kompas.tv/nasional/393772/kegagalan-indonesia-jadi-tuan-rumah-piala-dunia-u20-keriuhan-warganet-dan-gocekan-para-politisi? [Accessed: [28-Jan-2025].

[5] M. Murni, I. Riadi, and A. Fadlil, "Analisis Sentimen HateSpeech pada Pengguna Layanan Twitter dengan Metode Naïve Bayes Classifier (NBC)," *JURIKOM (Jurnal Riset Komputer)*, vol. 10, no. 2, pp. 566-575, 2023.

[6] V. Geetha, N. Sujatha, L. N. Valli, "Naïve Bayes Classification of Sentiments on Subset using *Tweet*s-during Covid-19," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 21s, pp. 249-255, 2024.

[7] Yuyun, N. Hidayah, and S. Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 820–826, 2021.

[8] K. Karsito and S. Susanti, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naive Bayes Di Perumahan Azzura Residence," *Jurnal SIGMA*, vol. 9, no. 3, pp. 43-48, 2019.

[9] J. Han, M. Kamber, and J. Pei, "*Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems*)," Copyright © 2011 Elsevier Inc. All rights reserved, 2012.

[10] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved Naive Bayes Classification Algorithm for Traffic Risk Management," *EURASIP Journal on Advances in Signal Processing*, pp.1-12, 2021.

[11] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *Journal of Intelligent Computing & Health Informatics*, vol. 1, no. 1, p. 17, 2020.

[12] M. K. Khatami, "*Analisis Sentimen Twitter Menggunakan Naive Bayes dan Support Vector Machine Terhadap KPU pada Pemilihan Umum Presiden 2024*," Prodi TI Sains Teknologi UIN JKT : Jakarta, Ciputat, 2024.

[13] K. A. Lubis, M. T. A. Bangsa, and A. Yudertha, "Analisis Sentimen Opini Masyarakat Terhadap Pindahnya Ibu Kota Indonesia dengan Menggunakan Klasifikasi Naïve Bayes," *Jurnal Teknoinfo*, vol. 18, no. 1, pp. 226-238, 2024.

[14] R. Prasetya, "Penerapan Teknik Data Mining dengan Algoritma Classification Tree untuk Prediksi Hujan", *Jurnal Widya Climago*, Vol.2 No.2, pp. 13-23, 2020.

[15] A. S. Sedghpour, M. R. S. Sedghpour, "*Web Document Categorization Using Naive Bayes Classifier and Latent Semantic Analysis*."2020.

[16] M. H. Humaidi, Sutrisno, and P. W. Laksono, "Implementation of Machine Learning for Text Classification Using the Naive Bayes Algorithm in Academic Information Systems at Sebelas Maret University Indonesia," in *E3S Web of Conferences ICEMECE 2023*, pp. 1-5, 2023

[17] S. Samsir, *et al*., "Naives Bayes Algorithm for Twitter Sentiment Analysis," *in Journal of Physics: Conference Series*, pp. 1-6, 2021.