5th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
7 September 2024 – Jakarta, Indonesia
Volume 3, Nomor 2, September 2024 - ISSN 2962-8628 (*online*)

DETEKSI UJARAN KEBENCIAN PADA MEDIA SOSIAL X DALAM KASUS PENGUNGSI ROHINGYA MENGGGUNAKAN METODE *MULTINOMIAL NAÏVE BAYES*

Deam Dharma Agung[1*], Achmad Solichin[2]

[1,2] Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1*]2011502305@student.budiluhur.ac.id, [2*]achmad.solichin@budiluhur.ac.id
(* : corresponding author)

Abstrak- Dalam era digital saat ini, media sosial telah menjadi *platform* utama bagi masyarakat untuk berbagi pendapat dan informasi. *Platform* seperti Twitter, yang sekarang dikenal sebagai X memfasilitasi interaksi yang mudah dan cepat namun juga menghadirkan tantangan seperti penyebaran ujaran kebencian yang berdampak negatif bagi individu maupun komunitas. Kasus pengungsi Rohingya telah menjadi perhatian global karena kekerasan sistematis oleh militer Myanmar terhadap etnis minoritas yang tidak diakui kewarganegaraannya. Banyak masyarakat global yang mengecam perlakuan Myanmar terhadap Rohingya, yang dianggap melanggar hak asasi manusia (HAM). Komentar dan analisis tentang kasus ini sering mendapatkan dukungan dan kritik, kadang berupa ujaran kebencian (*hate speech*). Ujaran kebencian (*hate speech*) adalah tindakan menyebarkan kebencian berdasarkan suku, agama, ras, dan lainnya, yang dapat memicu perpecahan, diskriminasi, kekerasan, dan konflik sosial. Media sosial sering digunakan untuk menyebarkan pesan kebencian dengan cepat dan luas, yang merusak hubungan antar kelompok dalam masyarakat, menyebabkan ketidakstabilan sosial, dan mengancam perdamaian. Penelitian ini berfokus pada penerapan *Text Mining* untuk mendeteksi ujaran kebencian menggunakan Metode *Multinomial Naive Bayes* terhadap kasus pengungsi Rohingya di media sosial Twitter atau X. Dataset diperoleh menggunakan *Tweet Harvest* dari bulan Januari hingga Maret 2024, dengan 436 data yang dikumpulkan dan diberi label otomatis oleh *vader lexicon* serta divalidasi secara manual oleh pakar dengan tujuan mengembangkan model pendeteksi ujaran kebencian pada media sosial X. Kata kunci yang digunakan berkaitan dengan Rohingya dan beberapa kata yang mengandung unsur kebencian. Penggunaan pembobotan kata TF-IDF dan Algoritma Multinomial Naive Bayes dalam mendeteksi ujaran kebencian menghasilkan nilai akurasi sebesar 69%, presisi 93%, dan recall 51%. Hasil ini menunjukkan bahwa sistem dapat mendeteksi *tweet* yang mengandung ujaran kebencian (*hate speech*) dengan metode *Multinomial Naive Bayes*. Kontribusi utama dari penelitian ini adalah penerapan dan evaluasi *Multinomial Naïve Bayes* dalam konteks bahasa Indonesia untuk deteksi ujaran kebencian, yang memberikan wawasan baru dalam menangani isu sosial kritis ini di media sosial, serta menjadi landasan bagi pengembangan lebih lanjut dalam analisis teks berbahasa Indonesia.

Kata Kunci: *Multinomial Naïve Bayes, Rohingya, Text Mining, TF-IDF, Ujaran Kebencian*

HATE SPEECH DETECTION ON X SOCIAL MEDIA IN THE CASE OF ROHINGYA REFUGEES USING MULTINOMIAL NAÏVE BAYES METHOD

Abstract- *In today's digital era, social media has become a major platform for people to share opinions and information. Platforms such as Twitter, now known as X, facilitate easy and fast interactions but also present challenges such as the spread of hate speech that has a negative impact on individuals and communities. The case of Rohingya refugees has become a global concern due to the systematic violence by the Myanmar military against ethnic minorities who are not recognized as citizens. Many in the global community have condemned Myanmar's treatment of the Rohingya, which is considered a violation of human rights. Comments and analysis on this case often receive support and criticism, sometimes in the form of hate speech. Hate speech is the act of spreading hatred based on ethnicity, religion, race, and others, which can trigger division, discrimination, violence, and social conflict. Social media is often used to spread hate messages quickly and widely, which damages relations between groups in society, causes social instability, and threatens peace. This study focuses on the application of Text Mining to detect hate speech using the Multinomial Naive Bayes Method for the case of Rohingya refugees on social media Twitter or X. The dataset was obtained using Tweet Harvest from January to March 2024, with 436 data collected and automatically labeled by the vader lexicon and manually validated by experts with the aim of developing a hate speech detection model on social media X. The keywords used are related to Rohingya and several words that contain elements of hatred. The use of TF-IDF word weighting and the Multinomial Naive Bayes Algorithm in detecting hate speech produces an accuracy value of 69%, precision of 93%, and recall of 51%. These results indicate that the system can detect tweets containing hate speech using the Multinomial Naive Bayes method. The main contribution of this study is the application and evaluation of Multinomial Naïve Bayes in the Indonesian context for hate speech detection, which provides new insights in dealing with this critical social issue on social media, as well as being a foundation for further development in Indonesian text analysis.*

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

1. PENDAHULUAN

Kasus pengungsi Rohingya telah menjadi perhatian globak karena kekerasan sistematis yang dilakukan oleh militer Myanmar, yang menganggap mereka sebagai etnis minoritas tanpa kewarganegaraan. Banyak masyarakat dunia mengecam perlakuan ini sebagai pelanggaran HAM. Dalam era digital, penyebaran informasi tentang pengungsi Rohingya terjadi sangat cepat melalui *platform* sosial media seperti Twitter (X), yang memiliki fitur cukup lengkap untuk menyalurkan aspirasi pengguna.

Pengungsi Rohingya yang melarikan diri dari kekerasan dari Myanmar mencari perlindungan di berbagai negara, termasuk Indonesia. Beberapa wilayah di Indonesia, seperti Aceh, menjadi tujuan utama karena jarak geografis yang dekat. Kedatangan pengungsi ini memicu berbagai reaksi, mulai dari simpati hingga penolakan. Ujaran kebencian terhadap pengungsi Rohingya sering muncul di media sosial dapat memperburuk keadaan.

Twitter atau X sebagai *platform* media sosial memfasilitasi penyebaran informasi dan aspirasi pengguna, namun juga menimbulkan tantangan seperti penyebaran ujaran kebencian. Ujaran kebencian didefinisikan sebagai ucapan yang menyulut kebencian terhadap kelompok tertentu berdasarkan ras, agama, etnisitas, dan lainnya [1]. Ujaran kebencian ini bisa terjadi secara lisan atau tulisan, termasuk di media digital seperti Twitter maka dari itu dirasa butuh suatu sistem untuk memfilter suatu teks yang mengandung unsur kebencian.

Penelitian ini bertujuan mengembangkan model deteksi ujaran kebencian terhadap etnis Rohingya di Twitter menggunakan teknik *Text Mining* dan algoritma *Multinomial Naïve Bayes*. Tantangan utamanya adalah mendeteksi ujaran kebencian dalam teks bahasa Indonesia yang sering menggunakan kata gaul, singkatan, imbuhan, dan kesalahan ejaan. Teknik *Text Mining* akan mengekstraksi informasi dari teks, sementara *Multinomial Naïve Bayes* digunakan untuk klasifikasi teks. Kontribusi utama dari penelitian ini adalah penerapan dan evaluasi *Multinomial Naïve Bayes* dalam konteks bahasa Indonesia untuk deteksi ujaran kebencian, yang memberikan wawasan baru dalam menangani isu sosial kritis ini di media sosial, serta menjadi landasan bagi pengembangan lebih lanjut dalam analisis teks berbahasa Indonesia.

Penelitian sebelumnya yang dilakukan pada tahun 2023 dengan judul "Implementasi Metode *Multinomial Naïve Bayes* untuk mendeteksi *Hate Speech* pada Twitter" [2] proses labeling dilakukan secara manual sedangkan penelitian saat ini dilakukan secara otomatis menggunakan kamus *lexicon*, serta data yang digunakan lebih sedikit dibandingkan dengan penelitian sebelumnya yang menyebabkan ada perbedaan hasil pengujian.

Implementasi *Multinomial Naïve Bayes* dipilih karena efisiensi dan kecepatan dalam mengolah data teks. Pada penelitian ini, akan menggunakan metode *Multinomial Naïve Bayes* karena metode ini sangat efisien dalam pembelajaran, terutama pada dataset yang relatif besar [3]. Algoritma ini efektif untuk klasifikasi teks, mudah diimplementasikan, dan memerlukan sumber daya komputasi yang minimal. *Multinomial Naïve Bayes* mampu memberikan hasil yang akurat dalam klasifikasi teks multikategori dengan pendekatan sederhana, menjadikannya pilihan optimal untuk mendeteksi ujaran kebencian dan aplikasi serupa lainnya. Penelitian ini memberikan solusi untuk mendeteksi ujaran kebencian di media sosial, meningkatkan kesadaran masyarakat tentang etika berkomunikasi, dan menawarkan panduan bagi pembuat kebijakan untuk menangani ujaran kebencian di media sosial secara lebih efektif.

Dalam proses menjalankan penelitian ini, penulis perlu merujuk pada literatur penelitian sebelumnya yang terkait dengan fokus penelitian. Tujuannya untuk mendapatkan pemahaman yang mendalam mengenai topik penelitian yang sedang dijalankan. Penelitian tersebut akan menjadi sumber referensi dan panduan dalam pengembangan penelitian ini. Penelitian yang dijadikan rujukan dengan judul "Implementasi Metode Naïve Bayes untuk Mendeteksi Hate Speech pada Twitter" mendapatkan akurasi sebesar 71% [2]. Tabel 1 menunjukan studi literatur yang digunakan dalam penelitian ini.

Tabel 1. Studi Literatur

Nama	Tahun	Tujuan Penelitian	Metode	Hasil Penelitian
[2]	2023	Mendeteksi ujaran kebencian pada Twitter	*Naïve Bayes*	Hasil akurasi sebesar 71%.
[4]	2019	Klasifikasi *Hate Speech* Berbahasa Indonesia di Twitter	*Naïve Bayes*	Dengan data sebanyak 250 *record* menghasilkan akurasi 98%, presisi 100%, *recall* 96,15%, dan *f-measure* 98,03%.
[5]	2022	Analisis sentimen dengan Rapidminer	*Multinomial Naïve Bayes*	Pengguna dapat menganalisis sentimen dengan mudah.
[6]	2022	Klasifikasi *tweet* yang mengandung ujaran kebencian pada sosial media twitter	*Multinomial Naïve Bayes*	Data berjumlah 5.221 yang didapatkan dari situs *Kaggle* tentang ujaran kebencian menghasilkan akurasi 70%
[7]	2019	Klasifikasi ujaran	*Naïve Bayes,*	Nilai akurasi tertinggi sebesar

		kebencian pada cuitan Bahasa Indonesia	*SVM*, dan *Logistic Regresion*	98% dan terendah sebesar 80%.
[8]	2021	Mendeteksi cuitan yang mengandung *cyberbullying*	*Multinomial Naïve Bayes, Linear SVM, Logistic Regresion* dan KNN.	Akurasi MNB sebesar 96%, *Logistic Regresion* sebesar 99%, SVM sebesar 99%, dan KNN sebesar 91%.
[9]	2023	Deteksi ujaran kebencian pada twitter berbahasa Indonesia	*Naïve Bayes*	Hasil akurasi sebesar 64,957% dengan kombinasi metode ekstraksi fitur seperti *word unigram, word bigram* dan *character quadram.*
[10]	2021	Deteksi ujaran ancaman pada postingan di Twitter	*Naïve Bayes*	Klasifikasikan *tweet* dan *mention* di twitter yang berupa ujaran ancaman mendapatkan akurasi sebesar 66%.
[11]	2019	Klasifikasikan ujaran kebencian pada twitter	*Naïve Bayes*	Hasil akurasi sebesar 84%, presisi 92%, *recall* 79,31%, dan *f-measure* 85,18%.
[12]	2020	Deteksi ujaran kebencian pada masa pemilu 2024 di sosial media	*Naïve Bayes*	Data berasal dari komentar pada *feeds* akun Instagram para bakal calon Presiden RI mendapatkan akurasi sebesar 85%, presisi sebesar 90%, dan *recall* sebesar 81%.

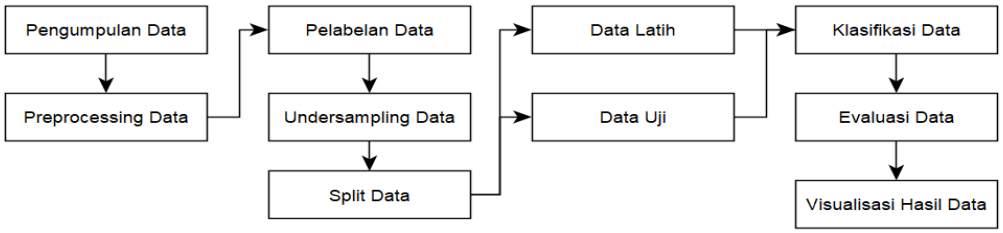
2. METODE PENELITIAN

2.1 Data Penelitian

Data atau dataset yang digunakan dalam penelitian ini bersumber dari media sosial X atau Twitter. Data tersebut dikumpulkan sejak Januari 2024 hingga Maret 2024 dan diperoleh menggunakan *Tweet Harvest.* Data yang digunakan sebanyak 436 cuitan yang bersumber dari kata kunci Rohingya. Berikut adalah contoh hasil dari penarikan data cuitan di media sosial Twitter / X

2.2 Penerapan Metode

Dalam penelitian ini, metode *Multinomial Naïve Bayes* digunakan untuk membangun sistem deteksi ujaran kebencian. Ada beberapa tahapan yang dilakukan untuk mencapai tujuan penelitian dan menjalankan sistem secara keseluruhan. Tahapan-tahapan tersebut dijelaskan dalam Gambar 1.



Gambar 1. Tahapan Metode

Tahap pertama yaitu mengumpulkan data *tweet* yang akan digunakan sebagai *dataset* melalui proses *crawling*. Data yang terkumpul kemudian akan dilakukan tahapan *preprocessing* dan diberi label secara otomatis oleh *Vader Lexicon* ke dalam dua kategori, yaitu *hate speech* dan *non hate speech*. Untuk analisis data *tweet* dalam penelitian ini, perlu dilakukan validasi data menggunakan pendekatan makna linguistik secara konseptual

dengan merujuk pada makna kata atau kalimat berdasarkan makna tanpa memperhatikan konteks. Dalam proses pelabelan ini, arahan diberikan oleh Bu Saskia Lydiani, S.Pd., M.Si dari Universitas Budi Luhur Jakarta.

Preprocessing bertujuan untuk mengubah data dari yang tidak terstruktur menjadi terstruktur guna mempermudah proses analisis. Merujuk pada penelitian yang telah dilakukan [13] dan juga [14], maka penelitian ini akan dilakukan beberapa tahapan *preprocessing* yang terdiri dari *cleansing*, *case folding*, *remove stopword* dan *slangword*, *tokenization* dan *stemming*.

a. *Cleansing*

Cleansing adalah langkah untuk menghilangkan tanda baca dan karakter yang tidak mempengaruhi dalam penelitian dihilangkan dari teks [15]. Karakter yang dibersihkan seperti tanda baca, situs URL, hashtag dll.

b. *Case Folding*

Case folding adalah tahap dimana semua huruf dalam dokumen diubah menjadi huruf kecil agar teks berada dalam format standar [13].

c. *Tokenization*

Tokenization adalah proses memecah sepotong teks menjadi token individu, biasanya berupa kata-kata atau tanda baca. Tokenisasi adalah langkah pra-pemrosesan umum saat bekerja dengan data teks, karena memungkinkan teks diproses dan dianalisis dengan lebih mudah (yang sebelumnya kalimat kemudian dipecah menjadi kata perkata) [16].

d. *Replace Slangword*

Replace slangword adalah proses kata yang termasuk *slangword* atau kata tidak baku (bahasa gaul) diubah ke dalam bentuk baku sesuai KBBI. Yang termasuk dalam kategori *slangword* adalah kata tidak baku, singkatan-singkatan, atau salah penulisannya [17].

e. *Remove Stopword*

Remove Stopword yaitu suatu proses terjadi guna mengeliminasi semua kata penghubung dan kata yang tidak dibutuhkan dalam *dataset* tersebut. Pada tahap ini, kata-kata yang tidak memiliki makna, seperti kata sambung, akan dihapus dari data [13].

f. *Stemming*

Stemming adalah proses pengubahan sebuah kata ke bentuk dasarnya dengan menghilangkan imbuhan yang terdiri dari awalan, akhiran, awalan dan akhiran, dan sisipan [18].

Setelah tahap *preprocessing*, dilakukan tahap *labeling* dan *undersampling*. Proses ini bertujuan untuk memilih data yang seimbang guna mencegah bias atau ketidakseimbangan jumlah data pada model. Langkah-langkah *undersampling* dimulai dengan mengidentifikasi jumlah data pada setiap kelas dan menentukan kelas dengan jumlah data terendah sebagai acuan. Setelah itu, jumlah data pada kelas lainnya disesuaikan agar seimbang dengan kelas acuan. Hal ini memastikan model tidak berat sebelah atau mayoritas pada salah satu kelas, sehingga meningkatkan akurasi dan keandalan hasil prediksi.

Data *tweet* akan akan dibagi menjadi dua bagian dengan rasio: 80% untuk data latih dan 20% untuk data uji dari keseluruhan data. Data latih akan digunakan untuk melatih model, sementara data uji akan digunakan untuk mengevaluasi model dan mengukur akurasi algoritma yang digunakan.

Selanjutnya dilakukan ekstraksi fitur menggunakan *Term Frequenxy – Inverse Document Frecuency*. TF-IDF merupakan algoritma yang digunakan untuk menghitung bobot kata dengan cara mempertimbangkan ada berapa banyak kata muncul (frekuensi kata) dan berapa banyak kata tersebut ditemukan dalam dokumen. Oleh karena itu TF-IDF dapat digunakan untuk mengontrol bobot dari kata. Sehingga Ketika kata tersebut selalu ada dalam setiap file atau kalimat maka bisa dikatakan bahwa kata tersebut tidak penting atau kata umum [19].
Rumus perhitugnan TF-IDF:

$$TF\ (d,t) = \frac{f(d,t)}{n(d)} \quad (1)$$

$$IDF\ (t) = \log\ (\frac{N}{df(t)}) \quad (2)$$

$$TF - IDF(d,t) = TF(d,t)\ x\ IDF(t) \quad (3)$$

Keterangan :
d = Dokumen.

t = *Term* (kata) yang sedang dievaluasi.
$f(d,t)$ = Frekuensi kemunculan *term* t dalam dokumen d.
$n(d)$ = Jumlah total *term* dalam dokumen d.
N = Jumlah total dokumen dalam koleksi.
$df(t)$ =Jumlah dokumen di dalam koleksi yang mengandung *term* t.

Tahapan berikutnya adalah *Multinomial Naïve Bayes.* Algoritma *Multinomial Naïve Bayes* merupakan salah satu metode pembelajaran probabilistik didasarkan pada teorema *Bayes* yang digunakan dalam *Natural Language Processing* (NLP). Algoritma ini bekerja pada konsep *term frequency* yang berarti berapa kali kata tersebut muncul dalam sebuah dokumen. Model ini menjelaskan dua fakta yaitu apakah kata tersebut muncul dalam sebuah dokumen atau tidak serta frekuensinya kemunculan dalam dokumen [20].
Rumus *Multinomial Naïve Bayes sebagai berikut:*

$$P(c|d) = P(c) \times \prod_{i=1}^{n} P(t_i|c) \quad (4)$$

Nilai $P(c)$ yang merupakan probabilitas *prior* dihitung dengan menggunakan persamaan sebagai berikut:

$$P(c) = \frac{Nc}{N} \quad (5)$$

Seleksi fitur merupakan langkah kritis dalam proses klasifikasi karena berdampak langsung pada kinerja model. Berbagai jenis fitur tersedia untuk digunakan, namun, dalam penelitian ini, fokus diberikan pada metode pembobotan yang dikenal sebagai *Term Frequency-Inverse Document Frequency* (TF-IDF).

Berikut adalah formula untuk menghitung *likelihood* atau probabilitas kemunculan kata ke-i saat menerapkan metode pembobotan TF-IDF:

$$P(t_i \mid c) = \frac{W_{ct} + 1}{\sum_{t} W_{ct} + V} \quad (6)$$

Pada penelitian ini, pengujian dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* adalah sebuah matriks yamg memuat data klasifikasi yang dilakukan oleh sistem klasifikasi baik secara aktual maupun prediktif. Dengan mengevaluasi data pada matriks akan diketahui bagaimana performa suatu model [21]. Ada 4 (empat) istilah dalam *confusion matrix* yang menjelaskan hasil pengukuran kinerja klasifikasi, yaitu *True Negative* (TN), *False Positive* (FP), *True Positive* (TP), dan *False Negativ*e (FN) [20]. Tabel 2 menunjukan *confusion matrix*.

Tabel 2. *Confusion Matrix*

Actual		*Prediction*	
		True	*False*
True		*True Positive* (TP)	*False Negative* (FN)
False		*False Positive* (FP)	*True Negative* (TN)

Kinerja *confusion matrix* dapat diukur menggunakan dengan nilai TP, FP, FN, dan TN. *True Positive* merupakan data positif yang diprediksi benar. *True Negative* adalah data negatif yang diprediksi benar. *False Positive* adalah data negatif namun diprediksi sebagai data positif. *False Negative* adalah data positif namun diprediksi sebagai data negatif [20].

Dengan memanfaatkan *confusion matrix*, dapat diestimasikan nilai *accuracy*, *precision*, dan *recall*. *Accuracy* merupakan proporsi dari sampel yang terklasifikasi secara tepat dibandingkan dengan total sampel yang diamati. Untuk menghitung nilai *accuracy*, dapat dilakukan dengan menggunakan formula berikut.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (7)$$

Precision adalah rasio antara jumlah sampel positif yang terklasifikasi dengan benar terhadap total prediksi sampel positif. Untuk menentukan nilai *precision*, formula berikut dapat digunakan:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall merupakan rasio antara jumlah sampel positif yang terklasifikasi dengan benar terhadap total jumlah sampel positif. Untuk menentukan nilai *recall*, dapat digunakan formula berikut:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Penelitian ini menggunakan data yang diambil dari media sosial Twitter atau X sebagai *dataset*. Pengumpulan data *tweet* dilakukan melalui *Tweet Harvest*. Kata kunci yang digunakan untuk meng-*crawling* data dari Twitter atau X meliputi kata yang berkaitan dengan Rohingya serta ujaran kebencianya. Total jumlah tweet yang terkumpul mencapai 436 *tweet*. Data hasil *crawling* ini kemudian disimpan dalam format Excel untuk tahapan *preprocessing* dan dilakukan pelabelan otomatis oleh *Vader Lexicon* serta validasi manual oleh pakar. Tabel 3 menunjukan dataset hasil pengumpulan data.

Tabel 3. *Dataset* Hasil Pengumpulan Data dengan Label Kelas

No.	Tweet	Label
1.	ghtedwined Yg berada d kubunya Prabowo kasi tau tuh Rohingya aj kaga bisa d urus	(*Hate Speech)*
2.	@affordxyz @USAID Ulama larang Rakyat Aceh tolak Rohingya lagi. Fokus urusan lain aja	(Non *Hate Speech)*
3.	Ulama Aceh Minta Pemerintah Segera Relokasi Rohingya ke Tempat Layak https://t.co/VFPcma9M6E https://t.co/jejEXavvUa	(*Non Hate Speech)*
4.	@Sunyisejati @mardigu024 dijaman Prabowo Rohingya bolak balik ke aceh.	(*Hate Speech*)
5.	Ngatur rakyat untuk jangan tolak rohingya Tapi baca pendapat rakyat di kolom komentar aja gak mampu	*(Hate Speech)*

3.2 *Preprocessing*

Pada langkah ini, data mentah yang diperoleh dari proses *crawling* diubah secara signifikan menjadi data yang telah melalui proses pembersihan untuk memungkinkan pengolahan oleh sistem. Informasi lebih lanjut mengenai tahapan *preprocessing* dapat ditemukan untuk memahami detail prosesnya. Tabel 4 menunjukan tahapan *preprocessing*.

Tabel 4. Tahapan *Preprocessing*

Tahapan *Preprocessing*	**Hasil**
Tweet Asli	Ngatur rakyat untuk jangan tolak rohingya Tapi baca pendapat rakyat di kolom komentar aja gak mampu
Case Folding + Cleansing	ngatur rakyat untuk jangan tolak rohingya tapi baca pendapat rakyat di kolom komentar aja gak mampu
Tokenization	[“ngatur”, “rakyat”, “untuk”, “jangan”, “tolak”, “rohingya”, “tapi”, “baca”, “pendapat”, “rakyat”, “di”, “kolom”, “komentar”, “aja”, “gak”, “mampu”]
Replace Slang Word	[“ngatur”, “rakyat”, “untuk”, “jangan”, “tolak”, “rohingya”, “tapi”, “baca”, “pendapat”, “rakyat”, “di”, “kolom”, “komentar”, “saja”, “tidak”, “mampu”]
Remove Stop Word	[“ngatur”, “rakyat”, “tolak”, “rohingya”, “baca”, “pendapat”, “rakyat”, “kolom”, “komentar”, “mampu”]
Stemming	[“atur”, “rakyat”, “tolak”, “rohingya”, “baca”, “dapat”, “rakyat”, “kolom”, “komentar”, “mampu”]

3.3 *Labelling, Undersampling* dan *Split Data*

Setelah data melewati tahapan *preprocessing*, selanjutnya dilakukan tahapan *labeling*, *undersampling* dan *split data*. Tahapan ini meliputi pelabelan data menggunakan *Vader Lexicon* serta validasi data oleh pakar, lalu pemilihan data yang seimbang untuk mencegah ketidakseimbangan jumlah data pada model serta membagi data menjadi data latih dan data uji.

Tahapan *undersampling* dilakukan setelah data diberi label, dengan mengidentifikasi jumlah data pada setiap kelas dan menentukan kelas dengan jumlah data terendah sebagai acuan. Kemudian, jumlah data pada kelas mayoritas dikurangi sehingga seimbang dengan jumlah data pada kelas acuan.

Setelah itu, data yang telah seimbang dibagi menjadi dua bagian: data latih dan data uji. Rasio pembagian jumlahnya sebesar 80% untuk data latih dan 20% untuk data uji. Pembagian dilakukan secara acak untuk menghindari bias. Tahapan ini penting untuk memastikan bahwa model yang akan dibangun tidak terpengaruh oleh ketidakseimbangan data dan dapat diuji secara valid dengan data yang belum pernah dilihat oleh model sebelumnya.

3.4 *Pembobotan TF-IDF*

Setelah data melewati tahapan *undersampling* dan *split data*, selanjutnya dilakukan tahapan pembobotan kata dengan TF-IDF. Tahapan ini meliputi perhitungan TF (*Term Frequency*), perhitungan IDF (*Inverse Document Frequency*), dan kemudian menggabungkan keduanya menjadi nilai TF-IDF. Pembobotan TF-IDF menggunakan data latih. Tabel 5 menunjukan data latih yang akan dilakukan pengujian manual.

Tabel 5. Data Latih

Dokumen	Tweet
DOC1	ada kubu prabowo kasi tahu rohingya tidak urus
DOC2	zaman prabowo rohingya bolak balik aceh
DOC3	ulama aceh minta pemerintah segera relokasi rohingya tempat rakyat
DOC4	ulama larang rakyat aceh tolak rohingya fokus urus saja

Selanjutnya dihitung TF-IDF nya dengan didapat hasil pada Tabel 6 dibawah ini.

Tabel 6. Perhitungan TF-IDF

	TF IDF DOC1	TF IDF DOC2	TF IDF DOC3	TF IDF DOC4
ada	0.075	0	0	0
kubu	0.075	0	0	0
prabowo	0.037	0.05	0	0
kasi	0.075	0	0	0
tahu	0.075	0	0	0
rohingya	0	0	0	0
tidak	0.075	0	0	0
urus	0.037	0	0	0,033
zaman	0	0.1	0	0
bolak	0	0.1	0	0
balik	0	0.1	0	0
aceh	0	0.2	0.013	0.013
ulama	0	0	0.033	0.033
minta	0	0	0.066	0
pemerintah	0	0	0.066	0
segera	0	0	0.066	0
relokasi	0	0	0.066	0
tempat	0	0	0.066	0
layak	0	0	0.066	0
larang	0	0	0	0.066
rakyat	0	0	0	0.066
fokus	0	0	0	0.066
saja	0	0	0	0.066

3.5 *Multinomial Naïve Bayes*

Tahapan Proses klasifikasi dengan menggunakan algoritma *Multinomial Naïve Bayes* dimulai dengan

menghitung probabilitas untuk setiap label atau kelas. Dari contoh data latih yang diberikan sebelumnya, terdapat 4 contoh data latih. Dua di antaranya memiliki label *hate speech* dan dua lainnya memiliki label *non hate speech.*

$$P(Hate\ Speech) = \frac{jumlah\ dokumen\ Hate\ Speech}{Total\ Dokumen} \quad (10)$$

$$P(Hate\ Speech) = \frac{2}{4} = 0.5 \quad (11)$$

$$P(Non\ Hate\ Speech) = \frac{2}{4} = 0.5 \quad (12)$$

Proses pengujian disini menggunakan 1 sampel data uji yang digambarkan pada Tabel 7 berikut:

Tabel 7. Sampel Data Uji

No.	Tweet Bersih	Label Aktual
1.	gaza berani tahu kalau rohingya	*Hate Speech*

Setelah menghitung probabilitas untuk masing-masing label, tahap berikutnya adalah menghitung nilai total TF-IDF berdasarkan masing-masing kelas dari data latih yang akan dijelaskan pada tabel 8 berikut.

Tabel 8. Perhitungan Total TF-IDF

Total TF-IDF	Hasil Total TF-IDF
Non Hate Speech	Total TF-IDF = sum (TF-IDF) *non hate speech* = **0.785**
Hate Speech	Total TF-IDF = sum (TF-IDF) *hate speech* = **0.8422**

Selanjutnya, menghitung probabilitas setiap kata dalam data uji terhadap kelas tertentu. Tabel 9 menunjukkan probabilitas setiap kata dalam data uji terhadap kelas *hate speech*.

Tabel 9. Probabilitas *Likelihood* dan *Posterior* dari *Hate Spech*

Term	**Probabilitas *Likelihood***	**Total (Probabilitas *Posterior*)**
gaza	Prob = (0+1) / (0.8422+23) = 1 / 23.8422 = **0.04194**	0.5 × 0.04194 × 0.04194 × 0.04508 × 0.04194 × 0.04194 = **6.9737699×10−8**
berani	Prob = (0+1) / (0.8422+23) = 1 / 23.8422 = **0.04194**	
tahu	Prob = (0.075+1) / (0. 8422+23) = 1.075 / 23.8422 = **0.04508**	
kalau	Prob = (0+1) / (0. 8422+23) = 1 / 23.8422 = **0.04194**	
rohingya	Prob = (0+1) / (0.8422+23) = 1 / 23.8422 = **0.04194**	

Setelahnya, dilakukan juga perhitungan probabilitas terhadap kelas *non hate speech*. Berikut adalah hasil perhitungan probabilitas data uji terhadap kelas *non hate speech* yang dijelaskan dalam Tabel 10 berikut ini.

Tabel 10. Probabilitas *Likelihood* dan *Posterior* dari *Non Hate Spech*

Term	**Probabilitas *Likelihood***	**Total (Probabilitas *Posterior*)**
gaza	Prob = (0+1) / (0.785+23) = 1 / 23.785 = **0.04204**	0.5 × 0.04204 × 0. 04204 × 0. 04204 × 0. 04204 × 0. 04204 = **6.5657379×10−8**
berani	Prob = (0+1) / (0. 785+23) = 1 / 23. 785 = **0. 04204**	
tahu	Prob = (0+1) / (0. 785+23) = 1 / 23. 785 = **0. 04204**	
kalau	Prob = (0+1) / (0. 785+23) = 1 / 23. 785 = **0. 04204**	
rohingya	Prob = (0+1) / (0. 785+23) = 1 / 23. 785 = **0. 04204**	

Berdasarkan hasil perhitungan, didapatkan bahwa probabilitas data uji terhadap kelas *hate speech* adalah sekitar $\mathbf{6.9739699 \times 10^{-8}}$, sementara probabilitas data uji terhadap kelas *non hate speech* sekitar $\mathbf{6.5657379 \times 10^{-8}}$. Dari perbandingan ini, dapat disimpulkan bahwa data uji tersebut diprediksi termasuk ke dalam kelas *hate speech*.

3.6 Pengujian

Pengujian sistem adalah tahap penting dalam evaluasi kinerja sebuah aplikasi. Tujuan utamanya adalah untuk menilai seberapa efektif aplikasi dalam mendeteksi dan mengidentifikasi konten yang mengandung ujaran kebencian. Dalam penelitian ini, digunakan pengambilan data latih sebanyak 182 data dan data uji sebanyak 46 data. Model dilatih menggunakan data latih dan diuji menggunakan data uji. Berikut adalah tabel 11 yang memperlihatkan hasil *confusion matrix* untuk evaluasi kinerja aplikasi dalam mendeteksi dan mengidentifikasi konten yang mengandung ujaran kebencian.

Tabel 11. Pengujian *Confusion Matrix*

	Aktual *Hate Speech*	Aktual *Non Hate Speech*
ksi *Non Hate Speech*	14 **(TP)**	1 **(FP)**
Prediksi *Hate Speech*	13 **(FN)**	18 (**TN**)

Hasil dari evaluasi kinerja sistem deteksi ujaran kebencian menggunakan *confusion matrix* menunjukkan gambaran yang mendetail tentang kemampuan sistem dalam mengklasifikasikan konten. Dari data uji yang terdiri dari 46 sampel, sistem berhasil mengidentifikasi 14 sampel sebagai ujaran kebencian (*True Positive*, TP) dan 18 sampel sebagai bukan ujaran kebencian (*True Negative*, TN). Namun demikian, sistem juga mengalami 1 kesalahan dalam mengklasifikasikan sampel yang sebenarnya bukan ujaran kebencian sebagai ujaran kebencian (*False Positive*, FP), serta 13 kesalahan dalam mengklasifikasikan sampel yang sebenarnya ujaran kebencian sebagai bukan ujaran kebencian (*False Negative*, FN). Tabel 12 dibawah ini menunjukan hasil pengujian.

Tabel 12. Pengujian *Evaluasi*

Pengujian		
Accuracy	***Precision***	***Recall***
69%	93%	93%

Berdasarkan pengujian terhadap data uji menggunakan metode algoritma *Multinomial Naïve Bayes*, sistem berhasil mencapai hasil evaluasi yang menggambarkan kemampuannya dalam mengidentifikasi ujaran kebencian. Dengan nilai *accuracy* mencapai 69%, sistem dapat mengklasifikasikan dengan akurat seberapa besar persentase data yang benar dari keseluruhan data yang dievaluasi. *Precision* sebesar 93% menunjukkan ketepatan sistem dalam mengklasifikasikan data kategori *Hate Speech* yang sebenarnya sebagai *Hate Speech* dari total data yang diklasifikasikan demikian. *Recall* mencapai 51%, mengindikasikan kemampuan sistem dalam mendeteksi dan mengenali data kategori *Hate Speech* dari keseluruhan data yang seharusnya terdeteksi.

Dalam penelitian ini, data dari Twitter atau X yang berkaitan dengan kasus pengungsi Rohingya dikumpulkan, dibersihkan, dan dilabeli sebagai *hate speech* atau *non hate speech* menggunakan kamus kata *Vader Lexicon* serta divalidasi oleh pakar. Data diolah dengan metode TF-IDF untuk ekstraksi fitur, lalu diklasifikasikan menggunakan model *Multinomial Naive Bayes* yang dipilih karena efisiensinya dalam menangani data teks. Model ini dilatih dengan 80% data dan diuji dengan 20% sisanya. Setelah melalui pengujian maka akan langsung menampilkan hasil visual dalam bentuk diagram. Meskipun menghadapi tantangan seperti variasi bahasa, model ini terbukti efektif dalam mendeteksi ujaran kebencian di media sosial dalam konteks bahasa Indonesia.

4. KESIMPULAN

Berdasarkan hasil pengujian, evaluasi dan implementasi dari aplikasi yang menggunakan *dataset* dan algoritma yang diusulkan untuk pendeteksian ujaran kebencian, dapat disimpulkan bahwa pendekatan ekstraksi fitur menggunakan TF-IDF dan klasifikasi menggunakan algoritma *Multinomial Naïve Bayes* terhadap pengungsi Rohingya dengan data latih sebanyak 182 data dan data uji sebanyak 46 data, efektif dalam deteksi ujaran kebencian pada data teks. Dengan mencapai tingkat *accuracy* sebesar 69%, *precision* sebesar 93%, *recall* sebesar 51%, sistem mampu mengklasifikasikan dengan baik antara konten yang mengandung *hate speech* dan yang tidak. Penelitian ini memiliki beberapa batasan masalah antara lain data yang digunakan dalam penelitian ini hanya berjumlah 436 data dan bersumber dari Twitter atau X berbahasa Indonesia, dan terdiri dari *tweet* yang dibuat pada rentang waktu Januari 2024 hingga Maret 2024, dengan pelabelan yang dibagi menjadi dua kategori, yaitu *hate speech* (ujaran kebencian) dan *non hate speech* (bukan ujaran kebencian), menggunakan kata kunci yang terkait dengan peristiwa Rohingya, serta kata-kata yang bermakna kebencian seperti bodoh, bunuh, dan pukul. Adapun saran untuk kedepannya yaitu menambahkan jumlah *tweet* yang diolah untuk setiap label, menambahkan daftar

kata-kata *stop word dan slang word*, termasuk kosakata yang umum digunakan dalam bahasa sehari-hari dan singkatan yang sering muncul dalam percakapan informal, dan menambahkan lebih banyak data dalam dataset untuk meningkatkan kemampuan sistem dalam melakukan klasifikasi secara lebih akurat dan presisi serta diharapkan dapat digunakan oleh masyarakat dalam rangka meningkatkan kesadaran akan pentingnya beretika dalam berkomunikasi dan menyuarakan pendapat secara lebih bijak khususnya pada platform media sosial.

DAFTAR PUSTAKA

[1] A. Sepima, G. T. P. Siregar, and S. A. Siregar, "Penegakan Hukum Ujaran Kebencian Di Republik Indonesia," *J. Retentum*, vol. 2, no. 2, 2020, doi: 10.46930/retentum.v2i2.908.
[2] U. Surapati and A. Y. Zulkarnain, "Implementasi Metode Naïve Bayes Untuk Mendeteksi Hate Speech Pada Twitter," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 830–837, 2023, doi: 10.31539/intecoms.v6i2.7678.
[3] D. S. Mahendra, B. Rahmat, and R. Mumpuni, "Implementasi Metode Multinomial Naive Bayes dalam Klasifikasi Judul Berita Clickbait," *Neptunus J. Ilmu Komput. Dan Teknol. Inf.*, vol. 2, no. 3, pp. 303–316, 2024.
[4] Ivan, Y. A. Sari, and P. P. Adikara, "Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4914–4922, 2019, [Online]. Available: http://j-ptiik.ub.ac.id
[5] A. Putri and A. Muzakir, "Analisis Sentimen Cyberbullying Kpop Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *הארץ*, vol. 7, no. 8.5.2017, pp. 2003–2005, 2022.
[6] N. A. Susanti, M. Walid, and H. Hoiriyah, "Klasifikasi Data Tweet Ujaran Kebencian Di Media Sosial Menggunakan Naive Bayes Classifier," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 538–543, 2022, doi: 10.36040/jati.v6i2.5174.
[7] K. Antariksa, Y. S. Purnomo WP, and E. Ernawati, "Klasifikasi Ujaran Kebencian pada Cuitan dalam Bahasa Indonesia," *J. Buana Inform.*, vol. 10, no. 2, p. 164, 2019, doi: 10.24002/jbi.v10i2.2451.
[8] N. Abdulloh and A. F. Hidayatullah, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter," *Automata*, vol. Vol 1, no. 1, pp. 1–5, 2021.
[9] R. M. Yazid, F. R. Umbara, and P. N. Sabrina, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia," *Informatics Digit. Expert*, vol. 4, no. 2, pp. 46–52, 2023, doi: 10.36423/index.v4i2.894.
[10] A. R. Rizkirobby, M. Nasrun, and R. A. N, "Deteksi Ujaran Ancaman Berbasis Website Pada Media Sosial Twitter Menggunakan Metode Support Vector Machine Website Based Detection Of Threats In Social Media Twitter Using Support Vector Machine Method," *e-Proceeding Eng.*, vol. 8, no. 2, pp. 500–505, 2021.
[11] M. Hakiem, M. A. Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2443–2451, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682
[12] I Putu Putrayana Wardana, "e-ISSN 2716-2753 Journal of Informatics Engineering and Technology (," vol. 01, no. 1, pp. 42–49, 2020.
[13] D. Wijaya, R. A. Saputra, and F. Irwiensyah, "KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Ulasan Aplikasi Samsat Digital Nasional Pada Google Playstore Menggunakan Algoritma Naïve Bayes," *Media Online*, vol. 4, no. 4, 2024, doi: 10.30865/klik.v4i4.1738.
[14] M. M. Effendi, Z. Mustofa, and A. Turmudi, "Analisis Sentimen Masyarakat Indonesia Dalam Konflik Rusia-Ukraina Di Twitter," *Bull. Inf. Technol.*, vol. 3, no. 4, pp. 355–366, 2022, doi: 10.47065/bit.v3i4.418.
[15] H. Dhery, A. Assyam, and F. N. Hasan, "Analisis Sentimen Twitter Terhadap Perpindahan Ibu Kota Negara Ke IKN Nusantara Menggunakan Orange Data Mining," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 341–349, 2023, doi: 10.30865/klik.v4i1.957.
[16] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.
[17] A. R. Isnain, H. Sulistiani, B. M. Hurohman, A. Nurkholis, and S. Styawati, "Analisis Perbandingan Algoritma LSTM dan Naive Bayes untuk Analisis Sentimen," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 2, p. 299, 2022, doi: 10.26418/jp.v8i2.54704.
[18] Yeni Anistyasari and Eko Hariadi, "Algoritma baru pembentukan kata dasar," *Pros. SNRT (Seminar Nas. Ris. Ter.*, vol. 5662, no. November, pp. 70–76, 2019.
[19] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," *Pros. Semin. Nas. Sist. Inf. dan Teknol.*, pp. 145–149, 2022.
[20] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 820–826, 2021, doi: 10.29207/resti.v5i4.3146.
[21] D. Musfiroh, U. Khaira, P. E. P. Utomo, and T. Suratno, "Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 24–33, 2021, doi: 10.57152/malcom.v1i1.20.