Jurnal Informatika dan Sistem Informasi (JIFoSI)
Vol. 5, No. 3. 2024
e-ISSN : 2722-130
DOI: https://doi.org/10.33005/jifosi.v5i3.464

DETEKSI UJARAN KEBENCIAN MENGGUNAKAN MULTINOMIAL NAÏVE BAYES UNTUK PUNDIT SEPAKBOLA DI MEDIA SOSIAL X

Dade Reindra Firdaus[1], Hari Soetanto[2*]
E-mail : [1)]2011500879@student.budiluhur.ac.id, [2)]hari.soetanto@budiluhur.ac.id
[1,2] Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

(Naskah masuk: 4 Desember 2024, diterima untuk diterbitkan: 31 Desember 2024)

Abstrak

Dalam era digital, media sosial menjadi platform utama untuk berbagi pendapat dan informasi. Twitter (sekarang dikenal sebagai X) memungkinkan interaksi cepat, namun juga memfasilitasi penyebaran ujaran kebencian yang dapat berdampak negatif bagi individu maupun komunitas. Sepak bola, sebagai olahraga populer di Indonesia, memiliki pundit seperti Bung Towel yang aktif memberikan analisis dan komentar. Ungkapan-ungkapan ini sering menimbulkan respons beragam, mulai dari dukungan hingga kritik, termasuk ujaran kebencian. Ujaran kebencian merupakan pernyataan yang menyebarkan kebencian berdasarkan suku, agama, ras, dan karakteristik lainnya. Penyebarannya di media sosial sangat cepat dan luas, sehingga berpotensi merusak hubungan antar kelompok, memicu diskriminasi, bahkan konflik sosial. Penelitian ini bertujuan untuk mendeteksi ujaran kebencian terhadap pundit sepak bola Indonesia, khususnya Bung Towel, dengan menerapkan text mining menggunakan metode Multinomial Naive Bayes. Data diperoleh melalui Tweet Harvest pada bulan Maret 2024, menghasilkan 400 tweet yang telah dilabeli secara manual. Proses klasifikasi menggunakan pembobotan kata TF-IDF dan algoritma Naive Bayes. Hasil pengujian menunjukkan akurasi sebesar 86,76%, presisi 82,93%, recall 94,44%, dan F1-Score 88,31%. Hasil ini membuktikan bahwa pendekatan ini efektif dalam mendeteksi ujaran kebencian di media sosial terhadap figur publik seperti Bung Towel.
Kata kunci; *Text Mining, ujaran kebencian, Naïve Bayes, TF - IDF*

1. PENDAHULUAN

Dalam era digital ini, media sosial telah menjadi platform utama bagi masyarakat untuk berbagi pendapat dan informasi. Twitter sering dimanfaatkan karena kemudahan penggunaannya untuk memperoleh informasi yang bernilai [1]. Meski demikian, media sosial tidak hanya membawa keuntungan, tetapi juga bisa memberikan efek buruk yang merugikan beberapa pihak, hingga menimbulkan kasus-kasus kriminal seperti fitnah dan ujaran kebencian [2].

Sepakbola adalah olahraga favorit di berbagai lapisan masyarakat. Hingga kini, popularitas sepakbola tetap tinggi, yang dapat dilihat dari banyaknya turnamen yang diadakan di berbagai tingkatan, mulai dari daerah, nasional, hingga internasional [3]. Pertandingan sepak bola biasanya dipandu oleh komentator yang bertugas untuk mengulas setiap pertandingan yang berlangsung [4]. Bung Towel, sebagai salah satu pundit sepak bola terkenal di Indonesia, sering kali menjadi sorotan di media sosial. Komentar dan analisisnya tidak hanya mendapatkan dukungan tetapi juga kritik, yang kadang kala berbentuk ujaran kebencian.

Ujaran kebencian merupakan aksi menyebarkan kebencian terhadap orang atau kelompok berdasarkan suku, agama, ras, dan karakteristik lainnya yang bisa menyebabkan

diskriminasi, kekerasan, serta konflik sosial [5]. Tindakan ini seringkali dilakukan melalui berbagai media, termasuk media sosial, yang dapat memperluas jangkauan pesan kebencian tersebut secara cepat dan luas. Dampaknya bisa menciptakan pola pikir buruk bagi seseorang, merusak persaudaraan dan bahkan bisa mengganggu ketenangan suatu bangsa [6].

Sebelumnya, telah ada beberapa studi mengenai deteksi ujaran kebencian. Salah satunya, penelitian oleh Rija Muhamad Yazid dengan judul Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia. Penelitian tersebut menggunakan 11.809 *tweet*, menggunakan ekstraksi fitur N Gram dan TF-IDF untuk fitur ekstraksi dan mencapai akurasi 64,95% [7]. Penelitian lainnya dilakukan oleh Noor Aliyah Susanti yang berjudul Klasifikasi Data Tweet Ujaran Kebencian Di Media Sosial Menggunakan Naive Bayes. Penelitian tersebut menggunakan 5.521 *tweet*, menggunakan TF-IDF untuk fitur ekstraksi dan mencapai akurasi mendekati 70% [8].

Perbedaan utama antara penelitian ini dengan penelitian Rija Muhamad Yazid adalah penggunaan Multinomial Naïve Bayes yang secara khusus dirancang untuk data teks, sedangkan penelitian Rija Muhamad Yazid menggunakan Naïve Bayes umum dengan kombinasi fitur N-Gram. Dibandingkan dengan penelitian Noor Aliyah Susanti, penelitian ini juga menggunakan Multinomial Naïve Bayes, tetapi dengan fokus pada kasus khusus yakni ujaran kebencian terhadap pundit sepakbola Indonesia di media 13 sosial X. Penelitian ini juga diharapkan dapat mencapai akurasi yang lebih baik atau setidaknya sebanding dengan penelitian Noor Aliyah Susanti.

Penelitian ini berbeda dari studi-studi sebelumnya karena fokus utamanya pada ujaran kebencian yang ditujukan kepada pundit sepakbola Indonesia, seperti Bung Towel, di media sosial X. Berbeda dengan penelitian lain yang biasanya membahas ujaran kebencian dalam konteks yang lebih umum, penelitian ini menyoroti kasus yang spesifik. Kasus ini secara khusus mengangkat isu dalam dunia sepakbola Indonesia. Pendekatan ini memberikan perhatian khusus pada figur publik dalam olahraga yang sering menjadi target ujaran kebencian.

Penelitian ini bertujuan untuk mengembangkan model deteksi ujaran kebencian menggunakan metode Naïve Bayes yang dioptimalkan untuk teks berbahasa Indonesia. Model ini akan diterapkan untuk menganalisis tweet yang ditujukan kepada Bung Towel di Media Sosial X. Dengan menerapkan teknik *text mining* dan metode Naïve Bayes, penelitian ini diharapkan dapat memberikan solusi yang efektif untuk mendeteksi ujaran kebencian dan membantu menjaga etika komunikasi di media sosial. Penelitian ini juga diharapkan dapat memberikan kontribusi akademis dalam literatur deteksi ujaran kebencian serta memberikan alat praktis bagi pengelola media sosial dan komunitas sepak bola di Indonesia.

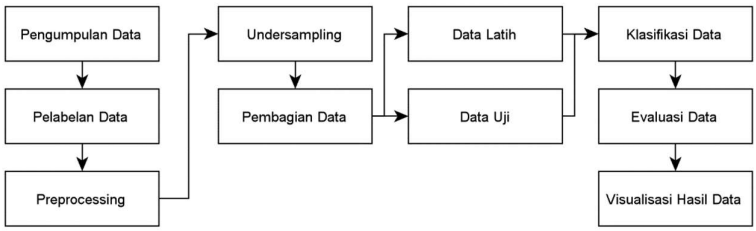
2. METODOLOGI

2.1 Data Penelitian

Dataset atau data dalam penelitian ini berasal dari media sosial Twitter atau X. Data yang digunakan adalah *tweet* yang dikumpulkan dari Maret 2024. Data tersebut diperoleh menggunakan Tweet Harvest. Kata kunci yang digunakan untuk mengumpulkan *dataset* adalah yang berkaitan dengan Bung Towel, dengan total sebanyak 400 data.

2.2 Organisasi Naskah

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 adalah mengumpulkan data *tweet* yang akan digunakan sebagai *dataset* melalui proses *crawling*. Data yang terkumpul kemudian akan diberi label secara manual ke dalam dua kategori, yaitu *hate speech* dan *non hate speech*. Untuk analisis data tweet dalam penelitian ini, digunakan pendekatan makna linguistik dengan konsep konseptual. Konsep konseptual ini merujuk pada makna kata atau kalimat berdasarkan arti gramatikal tanpa memperhatikan konteks. Dalam proses pelabelan ini, arahan diberikan oleh Bu Saskia Lydiani, S.Pd., M.Si dari Universitas Budi Luhur Jakarta.

Proses *preprocessing* bertujuan untuk mengidentifikasi dan mengeluarkan pengetahuan yang berharga dan relevan dari data teks yang belum terstruktur [9]. Merujuk pada penelitian yang telah dilakukan [10] dan juga [11], maka penelitian ini akan dilakukan beberapa tahapan *preprocessing* yaitu *cleansing*, *case folding*, *normalization*, *tokenization* dan *stemming*.

a. Cleansing, merupakan langkah di mana karakter dan tanda baca yang tidak relevan dihapus dari teks [12]. Dalam beberapa situasi, proses cleansing juga dilakukan untuk menghapus atribut yang tidak dibutuhkan dalam analisis. Misalnya, menghilangkan URL, mention pada tweet, hashtag, dan spasi berlebihan.

b. *Case* Folding, adalah proses mengonversi huruf alfabet yang telah melalui tahap pembersihan menjadi huruf kecil [13].

c. *Tokenization*, adalah tahap di mana rangkaian kata dalam kalimat, paragraf, atau halaman dipecah menjadi unit-unit kata yang terpisah [14].

d. *Replace Slangword*, merupakan proses menggantikan kata-kata singkat atau slang dengan ejaan bahasa Indonesia yang tepat [11].

e. *Remove Stopword*, proses penghapusan stopword adalah langkah untuk menghapus istilah yang tidak bermakna atau tidak relevan [15].

f. *Stemming*, adalah teknik yang digunakan untuk mengembalikan kata ke bentuk dasarnya dengan menghilangkan semua imbuhan, termasuk prefiks, sufiks, kombinasi dari keduanya, dan infiks [16].

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

Dataset tweet akan dipisahkan menjadi dua bagian, yaitu 80% untuk pelatihan dan 20% untuk pengujian. Bagian pelatihan akan dipakai untuk melatih model, sedangkan bagian pengujian akan digunakan untuk menilai model dan mengevaluasi kinerja algoritma.

Selanjutnya dilakukan ekstraksi fitur menggunakan *Term Frequenxy – Inverse Document Frecuency.* TF-IDF merupakan metode untuk menghitung atau memberikan bobot pada kata dengan menggunakan teknik preprocessing dan frekuensi kemunculan kata dalam dokumen, yang menandakan sejauh mana kata tersebut penting dalam dokumen tersebut [17]. Metode Pembobotan Term Frequency Inverse Document Frequency (TF-IDF) sering

digunakan untuk menilai relevansi kata (term) dalam dokumen atau kalimat dengan memberikan nilai atau bobot pada setiap kata [18].
Rumus perhitungan TF-IDF pada persamaan (1), (2), dan (3):

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

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

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

Tahapan berikutnya adalah *Multinomial Naïve Bayes.* Metode *Multinomial Naïve Bayes*, yang didasarkan pada teorema Bayes, banyak digunakan dalam *Natural Language Processing* (NLP). Algoritma ini menerapkan konsep *term frequency*, yang mengukur frekuensi kemunculan suatu kata dalam sebuah dokumen. Model ini menangani dua aspek: apakah kata tersebut ada dalam dokumen dan seberapa sering kemunculannya dalam dokumen [19].
Rumus *Multinomial Naïve Bayes* persamaan (4) sebagai berikut*:*

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

Nilai $P(c)$ yang merupakan probabilitas *prior* dihitung dengan menggunakan persamaan (5) 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 persamaan (6) adalah formula untuk menghitung *likelihood* atau probabilitas kemunculan kata ke-i saat menerapkan metode pembobotan TF-IDF:

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

Pada penelitian ini, pengujian dilakukan dengan menggunakan *confusion matrix* Tabel 1. *Confusion matrix* adalah sebuah tabel yang menunjukkan jumlah data uji yang diklasifikasikan dengan benar serta jumlah data uji yang diklasifikasikan secara salah [20]. 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) [21].

Tabel 1. *Confusion Matrix*

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

Confusion matrix dinilai berdasarkan nilai TP, FP, FN, dan TN. *True Positive* (TP) adalah jumlah data positif yang teridentifikasi dengan benar. *True Negative* (TN) merujuk pada data negatif yang diprediksi dengan benar. *False Positive* (FP) adalah data negatif yang

keliru diklasifikasikan sebagai positif. *False Negative* (FN) adalah data positif yang salah diidentifikasi sebagai negatif [22].
Dengan memanfaatkan *confusion matrix*, dapat diestimasikan nilai *accuracy*, *precision*, *recall*, dan *F1-Score*. *Accuracy* merupakan proporsi dari sampel yang terklasifikasi secara tepat dibandingkan dengan total sampel yang diamati. Nilai *accuracy* dapat dihitung dengan menggunakan formula (7) berikut.

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

Precision adalah perbandingan antara jumlah sampel positif yang benar-benar diklasifikasikan dengan tepat dengan total sampel yang diprediksi sebagai positif. Nilai *precision* dapat dihitung menggunakan formula (8) berikut.

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

Recall adalah perbandingan antara jumlah sampel positif yang berhasil diklasifikasikan dengan tepat terhadap total jumlah sampel positif. Formula (9) berikut dapat digunakan untuk menghitung nilai *recall:*

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

F1-Score adalah sebuah metrik yang mengkombinasikan precision dan recall dalam satu nilai tunggal. Untuk menghitung nilai *F1-Score*, dapat menggunakan rumus (10) berikut:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recal} \quad (10)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data dan Pelabelan

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 Bung Towel serta ujaran kebencianya. Total jumlah tweet yang terkumpul mencapai 400 tweet (Tabel 2). Data hasil crawling ini kemudian disimpan dalam format Excel untuk proses pelabelan manual oleh pakar, sebelum dilanjutkan ke tahap preprocessing. Pelabelan kelas menggunakan nilai 1 untuk *Hate Speech* dan 0 untuk *Non-Hate Speech.*

Tabel 2. *Dataset* Hasil Pengumpulan Data dengan Label Kelas

No.	Tweet	Label
1.	@GOAL_ID Towel anjing orang kaya gini jgn disorot terus biar pelan2 mati sendiri	1 (*Hate Speech)*
2.	@indosupporter Towel bangsat sampah!!	1 (*Hate Speech)*
3.	@ainurohman klo saya sih masih melihat apa yang bung towel sampaikan sebagai saran (klo tidak bisa disebut kritik) positif. wajar saja sih menurutku	0 (*Non Hate Speech)*
4.	Jujur gada Bung Towel gada oposisi.	0 (*Non Hate Speech*)
5.	@SiaranBolaLive @GOAL_ID Ketik 1 biar bung towel cepet mati	1 (*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 Tabel 3 dapat ditemukan untuk memahami detail prosesnya.

Tabel 3. Tahapan *Preprocessing*

Tahapan *Preprocessing*	Hasil
Tweet Asli	@ainurohman klo saya sih masih melihat apa yang bung towel sampaikan sebagai saran (klo tidak bisa disebut kritik) positif. Wajar saja sih menurutku
Case Folding + Cleansing	klo saya sih masih melihat apa yang bung towel sampaikan sebagai saran klo tidak bisa disebut kritik positif wajar saja sih menurutku
Tokenization	[“klo”, “saya”, “sih”, “masih”, “melihat”, “apa”, “yang”, “bung”, “towel”, “sampaikan”, “sebagai”, “saran”, “klo”, “tidak”, “bisa”, “disebut”, “kritik”, “positif”, “wajar”, “saja”, “sih”, “menurutku”]
Replace Slang Word	[“kalau”, “saya”, “sih”, “masih”, “melihat”, “apa”, “yang”, “bung”, “towel”, “sampaikan”, “sebagai”, “saran”, “kalau”, “tidak”, “bisa”, “disebut”, “kritik”, “positif”, “wajar”, “saja”, “sih”, “menurutku”]
Remove Stop Word	[“melihat”, “apa”, “bung”, “towel”, “sampaikan”, “saran”, “disebut”, “kritik”, “positif”, “wajar”, “menurutku”]
Stemming	[“lihat”, “apa”, “bung”, “towel”, “sampai”, “saran”, “sebut”, “kritik”, “positif”, “wajar”, “turut”]

3.3 *Undersampling* dan *Split Data*

Setelah data melewati tahapan *preprocessing*, selanjutnya dilakukan tahapan *undersampling* dan *split data*. Tahapan ini meliputi pemilihan data yang seimbang untuk mencegah bias pada model serta membagi data menjadi data latih dan data uji. Tahapan *undersampling* dimulai dengan mengidentifikasi jumlah data pada setiap kelas dan menentukan kelas dengan jumlah data terendah sebagai acuan. Kemudian, jumlah data pada kelas lain dikurangi sehingga seimbang dengan jumlah data pada kelas acuan. Setelah itu, data yang telah seimbang dibagi menjadi dua bagian: data latih dan data uji. Proporsi pembagian umum adalah 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 4.

Tabel 4. Data Latih

Dokumen	*Tweet*
DOC1	towel anjing kaya begini sorot biar pelan mati sendiri
DOC2	towel bangsat sampah
DOC3	lihat apa bung towel sampai saran sebut kritik positif wajar turut
DOC4	jujur gada bung towel gada oposisi

Selanjutnya dihitung TF-IDF nya dengan didapat hasil pada Tabel 5.

Tabel 5. Perhitungan TF-IDF

	TF IDF DOC1	TF IDF DOC2	TF IDF DOC3	TF IDF DOC4
towel	0	0	0	0
anjing	0.066	0	0	0
kaya	0.066	0	0	0
begini	0.066	0	0	0
sorot	0.066	0	0	0
biar	0.066	0	0	0
pelan	0.066	0	0	0
mati	0.066	0	0	0
sendiri	0.066	0	0	0
bangsat	0	0.200466	0	0
sampah	0	0.200466	0	0
lihat	0	0	0.05418	0
apa	0	0	0.05418	0
bung	0	0	0.02709	0.050267
sampai	0	0	0.05418	0
saran	0	0	0.05418	0
sebut	0	0	0.05418	0
kritik	0	0	0.05418	0
positif	0	0	0.05418	0
wajar	0	0	0.05418	0
turut	0	0	0.05418	0
jujur	0	0	0	0.100534
gada	0	0	0	0.09933
oposisi	0	0	0	0.100534

Setelahnya, nilai TF-IDF berdasarkan dokumen dijumlahkan sesuai masing-masing label yang ditampilkan pada Tabel 6.

Tabel 6. Perhitungan W

	Hate Speech	***Non Hate Speech***
towel	0	0
anjing	0.066	0
kaya	0.066	0
begini	0.066	0
sorot	0.066	0
biar	0.066	0
pelan	0.066	0
mati	0.066	0
sendiri	0.066	0
bangsat	0.200466	0
sampah	0.200466	0
lihat	0	0.05418
apa	0	0.05418
bung	0	0.077357
sampai	0	0.05418
saran	0	0.05418
sebut	0	0.05418
kritik	0	0.05418
positif	0	0.05418
wajar	0	0.05418
turut	0	0.05418
jujur	0	0.100534
gada	0	0.09933
oposisi	0	0.100534
Total	**0.928932**	**0.865375**

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 diantaranya memiliki label *hate speech* dan dua lainnya memiliki label *non hate speech*.

$$P(Hate\ Speech) = \frac{Jumlah\ dokumen\ Hate\ Speech}{Total\ Dokumen}$$

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

$$P(Non\ Hate\ Speech) = \frac{Jumlah\ dokumen\ Non\ Hate\ Speech}{Total\ Dokumen}$$

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

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

Tabel 7. Sampel Data Uji

No.	Tweet Bersih	Label Aktual
1.	ketik biar bung towel cepat mati	*Hate Speech*

Setelah menghitung probabilitas untuk masing-masing label, tahap berikutnya adalah menghitung nilai total TF-IDF (Tabel 8) berdasarkan masing-masing kelas dari data latih.

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* = **0.865375**
Hate Speech	Total TF-IDF = sum (TF-IDF) *hate speech* = **0.928932**

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*)**
ketik	Prob = (0+1) / (0.928932+24) = 1 /24.928932= **0.040114**	0.5 × 0.040114 × 0.042762× 0.040114 × 0.040114 × 0.042762× 0.040114 = **2.367390515520914×10−9**
biar	Prob = (0.066+1) / (0. 928932+24) = 1.066 /24.928932= **0.042762**	
bung	Prob = (0+1) / (0. 928932+24) = 1 /24.928932= **0.040114**	
towel	Prob = (0+1) / (0. 928932+24) = 1 /24.928932= **0.040114**	
mati	Prob = (0.066+1) / (0. 928932+24) = 1.066 /24.928932= **0.042762**	
cepat	Prob = (0+1) / (0. 928932+24) = 1 /24.928932= **0.040114**	

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*)**
ketik	Prob = (0+1) / (0.865375+24) = 1 /24.865375= **0.040217**	0.5 × 0.040217 × 0.040217 × 0.043327× 0.040217 × 0.040217 × 0.040217 = **2.2791713610813794×10−9**
biar	Prob = (0+1) / (0.865375+24) = 1 /24.865375= **0.040217**	
bung	Prob = (0.077357+1) / (0.865375+24) = 1.077357 /24.865375 = **0.043327**	
towel	Prob = (0+1) / (0.865375+24) = 1 /24.865375= **0.040217**	
mati	Prob = (0+1) / (0.865375+24) = 1 /24.865375= **0.040217**	
cepat	Prob = (0+1) / (0.865375+24) = 1 /24.865375= **0.040217**	

Berdasarkan hasil perhitungan, didapatkan bahwa probabilitas data uji terhadap kelas *hate speech* adalah sekitar 2.367×10^{-9}, sementara probabilitas data uji terhadap kelas *non hate speech* sekitar 2.279×10^{-9}. 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 270 data dan data uji sebanyak 68 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*

	Prediksi Bukan Ujaran Kebencian	Prediksi Ujaran Kebencian
Aktual Bukan Ujaran Kebencian	25 **(TN)**	7 **(FP)**
Aktual Ujaran Kebencian	2 **(FN)**	34 (**TP)**

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 68 sampel, sistem berhasil mengidentifikasi 34 sampel sebagai ujaran kebencian (*True Positive*, TP) dan 25 sampel sebagai bukan ujaran kebencian (*True Negative*, TN). Namun demikian, sistem juga mengalami 7 kesalahan dalam mengklasifikasikan sampel yang sebenarnya bukan ujaran kebencian sebagai ujaran kebencian (*False Positive*, FP), serta 2 kesalahan dalam mengklasifikasikan sampel yang sebenarnya ujaran kebencian sebagai bukan ujaran kebencian (*False Negative*, FN).

Tabel 12. Pengujian *Evaluasi*

Pengujian			
Accuracy	***Precision***	***Recall***	***F1-Score***
86.76%	82.93%	94.44%	88.31%

Berdasarkan pengujian terhadap data uji menggunakan metode algoritma *Multinomial Naïve Bayes*, sistem berhasil mencapai hasil evaluasi yang menggambarkan kemampuannya dalam mengidentifikasi ujaran kebencian. Dengan Tabel 12, nilai *accuracy* mencapai 86.76%, sistem dapat mengklasifikasikan dengan akurat seberapa besar persentase data yang benar dari keseluruhan data yang dievaluasi. *Precision* sebesar 82.93% menunjukkan ketepatan sistem dalam mengklasifikasikan data kategori *Hate Speech* yang sebenarnya sebagai *Hate Speech* dari total data yang diklasifikasikan demikian. *Recall* mencapai 94.44%, mengindikasikan kemampuan sistem dalam mendeteksi dan mengenali data kategori *Hate Speech* dari keseluruhan data yang seharusnya terdeteksi. *F1-Score* mencapai 88.31%, sebagai rata-rata harmonis dari *precision* dan *recall*, memberikan gambaran menyeluruh tentang performa sistem dalam mengidentifikasi ujaran kebencian dari data yang dievaluasi.

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 *Naïve Bayes* terhadap pundit sepakbola Indonesia dengan data latih sebanyak 283 data dan data uji sebanyak 71 data, efektif dalam deteksi ujaran kebencian pada data teks. Dengan mencapai tingkat akurasi sebesar 86.76%, presisi sebesar 82.93%, *recall* sebesar 94.44%, dan *F1-Score* sebesar 88.31%, sistem mampu mengklasifikasikan dengan baik antara konten yang mengandung ujaran kebencian dan yang tidak. 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.

5. DAFTAR RUJUKAN

[1] Muzaki, A., & Witanti, A. (2021). Sentiment analysis of the community in Twitter to the 2020 election in pandemic COVID-19 by method Naive Bayes Classifier. *Jurnal Teknik Informatika*, *2*(2), 101–107. https://doi.org/10.20884/1.jutif.2021.2.2.51

[2] Af'al, W. (2022). Ujaran kebencian terhadap aktor Arya Saloka di media sosial Twitter: Kajian linguistik forensik pendahuluan. *Jurnal Sinestesia*, *12*(2), 435–444. https://www.sinestesia.pustaka.my.id/journal/article/view/197

[3] Putra, A. T., & A. S. (2020). Kontribusi kelentukan dan daya ledak otot tungkai terhadap heading sepakbola Anton. *[Nama Jurnal Tidak Dicantumkan]*, *2*, 616–626.

[4] Adawiyah, R., Murtadlo, A., & Purwanti. (2021). Analisis jargon Valentino Simanjuntak pada pertandingan sepak bola. *Ilmu Budaya*, *5*(2), 394–403.

[5] Ridwan, R., Hermaliani, E. H., & Ernawati, M. (2024). Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian. *Computer Science*, *4*(1). https://doi.org/10.31294/coscience.v4i1.2990

[6] Anggi, M., Lola, S., & Nabila, N. S. (2024). Ujaran kebencian di era digital dan kontekstualisasi kalimah thayyibah QS. Ibrahim [24] dalam mewujudkan kesolehan sosial. *[Nama Jurnal Tidak Dicantumkan]*, *11*, 77–89.

[7] Yazid, R. M., Umbara, F. R., & Sabrina, P. N. (2023). Deteksi ujaran kebencian dengan metode klasifikasi Naïve Bayes dan metode N-Gram pada dataset multi-label Twitter berbahasa Indonesia. *Informatics Digital Expert*, *4*(2), 46–52. https://doi.org/10.36423/index.v4i2.894

[8] Susanti, N. A., Walid, M., & Hoiriyah, H. (2022). Klasifikasi data tweet ujaran kebencian di media sosial menggunakan Naive Bayes Classifier. JATI (Jurnal Mahasiswa Teknik Informatika), 6(2), 538–543. https://doi.org/10.36040/jati.v6i2.5174

[9] Hermawan, L., & Ismiati, M. B. (2020). Pembelajaran text preprocessing berbasis simulator untuk mata kuliah information retrieval. Jurnal Transformatika, 17(2), 188. https://doi.org/10.26623/transformatika.v17i2.1705

[10] Azhari, A. A., Sibaroni, Y., & Prasetiyowati, S. S. (2023). Detection of Indonesian hate speech in the comments column of Indonesian artists' Instagram using the RoBERTa method. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 8(3). https://doi.org/10.29100/jipi.v8i3.3898

[11] Narendra, L. W. (2022). Topic modeling in conversational dialogs for naming intent labels using LDA. JTECS (Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem dan Komputer), 2(1), 65. https://doi.org/10.32503/jtecs.v2i1.1820

[12] Yulita, W. (2021). Analisis sentimen terhadap opini masyarakat tentang vaksin Covid-19 menggunakan algoritma Naïve Bayes Classifier. Jurnal Data Mining dan Sistem Informasi, 2(2), 1. https://doi.org/10.33365/jdmsi.v2i2.1344

[13] Gifari, O. I., Adha, M., Freddy, F., & Durrand, F. F. S. (2022). Analisis sentimen review film menggunakan TF-IDF dan Support Vector Machine. Jurnal Information Technology, 2(1), 36–40. https://doi.org/10.46229/jifotech.v2i1.330

[14] Agustina, N., Adrian, A., & Hermawati, M. (2022). Implementasi algoritma Naïve Bayes Classifier untuk mendeteksi berita palsu pada sosial media. Faktor Exacta, 14(4), 206. https://doi.org/10.30998/faktorexacta.v14i4.11259

[15] Munthe, M. (2022). Penerapan metode TextRank dalam rancangan aplikasi silogisme artikel bahasa Batak. Jurnal Computer & Informatics Research, 1(2), 37.
[16] Anistyasari, Y., & Hariadi, E. (2019). Algoritma baru pembentukan kata dasar. Prosiding Seminar Nasional Riset Terapan (SNRT), 5662(November), 70–76.
[17] Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) untuk mencari relevansi dokumen berdasarkan query. Ilkomnika: Jurnal Computer Science and Applied Informatics, 1(2), 58–64. https://doi.org/10.28926/ilkomnika.v1i2.18
[18] Arifin, N., Enri, U., & Sulistiyowati, N. (2021). Penerapan algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk text classification. STRING (Satuan Tulisan Riset dan Inovasi Teknologi), 6(2), 129. https://doi.org/10.30998/string.v6i2.10133
[19] Yuyun, Hidayah, N., & Sahibu, S. (2021). Algoritma Multinomial Naïve Bayes untuk klasifikasi sentimen pemerintah terhadap penanganan Covid-19 menggunakan data Twitter. JRESTI (Rekayasa Sistem dan Teknologi Informasi), 5(4), 820–826. https://doi.org/10.29207/resti.v5i4.3146
[20] Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier dan Confusion Matrix pada analisis sentimen berbasis teks pada Twitter. J-SAKTI (Jurnal Sains Komputer dan Informatika), 5(2), 697–711.
[21] Rahmad, F., Suryanto, Y., & Ramli, K. (2020). Performance comparison of anti-spam technology using confusion matrix classification. IOP Conference Series: Materials Science and Engineering, 879(1), 012076. https://doi.org/10.1088/1757-899X/879/1/012076
[22] Tangkelayuk, A. (2022). Klasifikasi kualitas air menggunakan metode KNN, Naïve Bayes, dan Decision Tree. JATISI (Jurnal Teknik Informatika dan Sistem Informasi), 9(2), 1109–1119. https://doi.org/10.35957/jatisi.v9i2.2048