Indones. J. Math. Nat. Sci. 47(2) 2024
Indonesian Journal of Mathematics and Natural Sciences
https://journal.unnes.ac.id/journals/JM/index

Optimalisasi Algoritma Naïve Bayes untuk Klasifikasi Tweet Berbahasa Indonesia dalam Mengatasi *Hate Speech* di Platform X

Muhammad Haikal*, Alamsyah

Program Studi Teknik Informatika, Fakultas MIPA, Universitas Negeri Semarang
Gedung D5 Lt.2, Kampus Sekaran Gunungpati, Semarang 50229

E-mail: muhammad.haikal494972@gmail.com

Abstrak

Ujaran kebencian merupakan bentuk ekspresi yang digunakan untuk menyatakan kebencian dan sering kali bersifat destruktif, dengan tujuan menentang individu atau kelompok tertentu atas berbagai alasan. Kasus ujaran kebencian sangat sering ditemukan di media sosial, terutama menjelang pemilihan umum yang rutin diadakan setiap musim pemilu. Untuk mengatasi ujaran kebencian, diperlukan tindakan pemantauan dengan menyensor kata-kata yang berpotensi menyinggung dan menyerang unsur pribadi, seperti suku, agama, ras, dan antargolongan. Penelitian sebelumnya hanya berfokus pada analisis sentimen tweet untuk mengukur bobot positif atau negatifnya. Penelitian ini melanjutkan studi terkait ujaran kebencian dengan pengembangan melalui penerapan algoritma Naive Bayes, khususnya varian Multinomial dan Gaussian, serta sistem sensor otomatis yang bertujuan meningkatkan akurasi klasifikasi. Sistem ini diimplementasikan pada media sosial dengan harapan dapat mengurangi jumlah ujaran kebencian secara signifikan. Dari 13.169 tweet yang diambil sebagai dataset, data diklasifikasikan ke dalam 12 kategori dengan tingkat akurasi tertinggi 90%. Data hasil pengujian disimpan dalam bentuk kamus ujaran kebencian yang memuat kata-kata tidak layak, sehingga algoritma dapat mendeteksi dan secara otomatis melakukan sensor pada tweet yang mengandung ujaran kebencian.

Kata kunci: Deteksi ujaran kebencian, naïve bayes, pengklasifikasian tweet.

Abstract

Hate speech is a form of expression used to express hatred and is often destructive, aimed at opposing individuals or certain groups for various reasons. Cases of hate speech are frequently found on social media, especially during election seasons, which occur regularly. To combat hate speech, monitoring actions are needed by censoring words that have the potential to offend and attack personal elements, such as ethnicity, religion, and race. Previous research has generally focused only on sentiment analysis of tweets to determine their positive or negative weights. This research continues the study on hate speech by developing the application of the Naive Bayes algorithm, specifically the Multinomial and Gaussian variants, along with an automatic censorship system aimed at improving classification accuracy. This system is implemented on social media with the hope of significantly reducing the amount of hate speech. From 13,169 tweets collected as the dataset, the data was classified into 12 categories with the highest accuracy rate being 90%. The test results are stored in the form of a hate speech dictionary that contains inappropriate words, allowing the algorithm to detect and automatically censor tweets containing hate speech.

Keywords: *Hate speech detection*, naïve bayes, *tweet classification*.

How to cite:
Haikal M., Alamsyah A. (2024). Optimalisasi algoritma naïve bayes untuk klasifikasi tweet berbahasa Indonesia dalam mengatasi hate speech di platform X. *Indonesian Journal of Mathematics and Natural Sciences*, 47(2), 109-114.

PENDAHULUAN

Ujaran kebencian merupakan bentuk komunikasi yang bertujuan untuk merendahkan kelompok tertentu, seperti ras, warna kulit, etnis, jenis kelamin, orientasi seksual, kebangsaan, agama, atau karakteristik lainnya, dengan meremehkan individu atau kelompok atas dasar tertentu, yang dapat memicu kemarahan dan/atau diskriminasi verbal bersifat destruktif.

Platform Twitter (X) merupakan salah satu media sosial paling populer saat ini, dengan jumlah pengguna bulanan mencapai lebih dari 300 juta. Salah satu ketentuan Twitter adalah melarang pengguna mengunggah ancaman kekerasan, pelecehan, dan konten kebencian. Namun, masih banyak pengguna yang melanggar peraturan tersebut dan menggunakan akun Twitter untuk menyebarkan ujaran kebencian serta kata-kata negatif.

Berdasarkan studi Asogwa *et al.* (2022), data ujaran kebencian yang diambil dari Twitter digunakan untuk klasifikasi dengan 12 label berbeda, termasuk individu, kelompok, agama, ras, fisik, gender, lainnya, ujaran kebencian ringan, ujaran kebencian sedang, dan ujaran kebencian berat. Penelitian ini menerapkan algoritma Naïve Bayes (NB) dan *support vector machine* (SVM). Sementara itu Aljero *et al.* (2021) menggunakan *genetic programming*, Fatahillah *et al.* (2017) menggunakan Naive Bayes *classifier*, Ketsbaia *et al.* (2023) menggunakan *multi-stage machine learning*, Oriola dan Kotze (2020) menggunakan teknik evaluasi mesin learning, Plaza *et al.* (2021) menggunakan *multi-task learning*, Sreelakshmi *et al.* (2024) menggunakan *cost-sensitive learning*, dan Obaid *et al.* (2024) dan Zhou *et al.* (2020) menggunakan algoritma *deep learning* untuk mendeteksi ujaran kebencian.

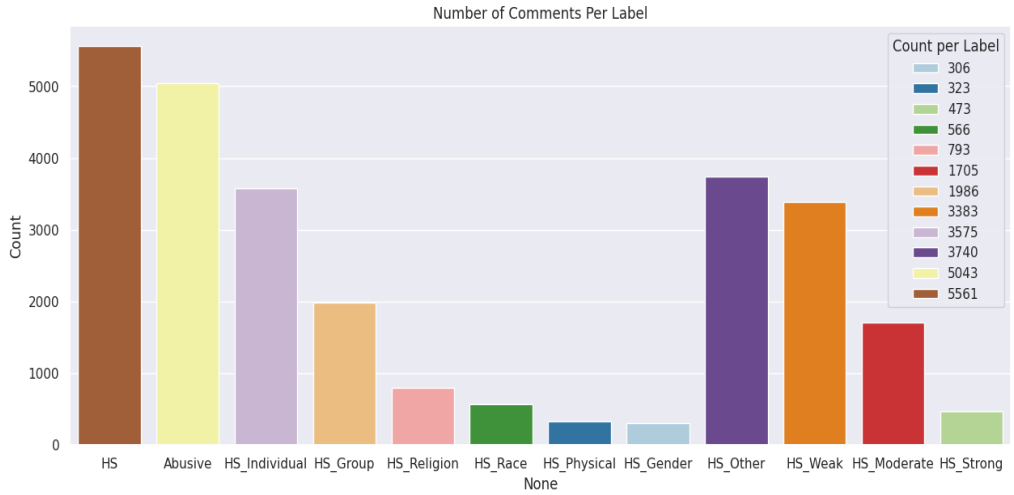
Pada penelitian Ibrohim dan Budi (2019), algoritma *logistic regression* digunakan untuk mendeteksi ujaran kebencian di Twitter berbahasa Indonesia, mencapai akurasi sebesar 79,85%. Penelitian ini mengkategorikan ujaran kebencian berdasarkan beberapa label seperti agama, ras, dan gender, serta tingkat keparahan. *Logistic regression* memberikan peningkatan akurasi klasifikasi dibandingkan dengan penelitian sebelumnya yang menggunakan algoritma lain seperti Naive Bayes dan *support vector machine*.

Berbeda dengan penelitian Ahmad *et al.* (2019) yang mengklasifikasian data terfokus pada 12 kategori berbeda, dengan tujuan mempelajari jumlah serta intensitas ujaran kebencian yang paling sering muncul di berbagai kategori. Hal ini dianggap penting karena seperti disebutkan Ali *et al.* (2021) dan Komnas HAM (2016), bahwa jumlah pengguna media sosial terus meningkat sejalan dengan bertambahnya ujaran kebencian. Oleh karena itu, pengklasifikasian menggunakan beberapa *classifier* dilakukan dalam penelitian ini menggunakan algoritma NB.

METODE

Dataset

Penelitian ini menggunakan dataset dari laman Kaggle (https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text), yang terdiri dari 13.169 tweet yang dikelompokkan dalam 12 label berbeda yaitu HS, Abusive, HS_Individual, HS_Group, HS_Religion, HS_Race, HS_Physical, HS_Gender, HS_Other, HS_Weak, HS_Moderate, dan HS_Strong. Dataset ini telah digunakan oleh Ibrohim dan Budi (2019). Isi dataset tersebut dipetakan dalam bentuk histogram pada Gambar 1.



Gambar 1. Histogram persebaran Tweet

Berdasarkan pemetaan (Gambar 1), label 'HS' ('Hate Speech') sebanyak 5561 Tweet, label 'Abusive' sebanyak 5043 Tweet, 'HS_Individual' sebanyak 3575 Tweet, 'HS_Group' sebanyak 1986 Tweet, 'HS_Religion' sebanyak 793 Tweet, 'HS_Race' sebanyak 566 Tweet, 'HS_Physical' sebanyak

323 Tweet, ‘HS_Gender’ sebanyak 306 Tweet, ‘HS_Other’ sebanyak 3740 Tweet, ‘HS_Weak’ sebanyak 3383 Tweet, ‘HS_Moderate’ sebanyak 1705 Tweet, dan ‘HS_Strong’ sebanyak 473 Tweet.

Label yang dimuat pada Gambar 1 berdasarkan kategori dan jenis dari "Hate Speech" yang diutarakan oleh user melalui Tweet. Kategori tersebut diuraikan pada Tabel 1.

Tabel 1. Label dan kategorinya

No	Label	Kategori
1.	HS	Ujaran kebencian secara umum
2.	Abusive	Ujaran kebencian yang bersifat menjelekkan atau merendahkan seseorang dengan menyerang fisik maupun emosional seseorang
3.	HS_Individual	Ujaran kebencian yang ditujukan kepada individu atau personal seseorang
4.	HS_Group	Ujaran kebencian yang ditujukan kepada kelompok tertentu
5.	HS_Religion	Ujaran kebencian yang terkait dengan agama atau kepercayaan
6.	HS_Race	Ujaran kebencian yang terkait dengan ras atau etnisitas
7.	HS_Physical	Ujaran kebencian yang terkait dengan kondisi fisik atau disabilitas seseorang
8.	HS_Gender	Ujaran kebencian yang terkait dengan gender atau orientasi seksual
9.	HS_Other	Ujaran kebencian yang terkait dengan penghinaan atau fitnah lainnya yang tidak termasuk dalam kategori spesifik di atas
10.	HS_Weak	Ujaran kebencian yang memiliki intensitas lemah
11.	HS_Moderate	Ujaran kebencian yang memiliki intensitas sedang
12.	HS_Strong	Ujaran kebencian yang memiliki intensitas kuat

Kategori-kategori ini membantu dalam mengklasifikasikan jenis ujaran kebencian yang ada di dalam dataset, sehingga analisis dapat dilakukan dengan lebih mendetail dan terfokus pada aspek-aspek tertentu dari ujaran kebencian.

Pada tahap pendeteksian, penelitian ini menggunakan empat tahap metode yaitu *preprocessing*, *splitting* dan *balancing*, klasifikasi dan evaluasi, dan menampilkan skor dan visualisasi (Mansur *et al.,* 2023*).*

Prepocessing

Pada tahap *preprocessing*, penelitian ini mengadaptasi metode yang telah digunakan dalam penelitian sebelumnya. Terdapat empat tahapan utama dalam *preprocessing*, yaitu normalisasi menggunakan kamus alay (*alay dictionary*), penghapusan *stopwords*, *tokenisasi* dan *stemming data* pada dataset.

Normalisasi merupakan tahapan untuk menghilangkan kata-kata slang/gaul dalam bahasa Indonesia dan menggantinya dengan kata-kata yang sesuai dengan Kamus Besar Bahasa Indonesia (KBBI). Proses normalisasi ini juga diterapkan pada penelitian Ibrohim dan Budi (2019) untuk memperoleh hasil klasifikasi yang lebih optimal. Kamus alay (*Alay Dictionary*) yang digunakan telah disediakan dalam penelitian sebelumnya (Ibrohim & Budi, 2019) dan dapat ditemukan di forum Kaggle. *Stopwords* adalah kata-kata dalam kalimat yang tidak relevan untuk analisis teks, sehingga biasanya dihapus untuk meningkatkan efisiensi analisis. *Stopwords* bervariasi tergantung pada bahasa dan negara (Buntoro *et al.,* 2016). *Tokenisasi* merupakan proses memecah kalimat lengkap menjadi elemen-elemen individu seperti simbol, kata kunci, dan frasa yang disebut token (Buntoro *et al.,* 2016). Dalam tokenisasi, karakter seperti tanda seru dan titik koma biasanya akan dihapus. *Stemming* adalah proses mengkonsolidasikan berbagai bentuk kata menjadi satu bentuk dasar. Teknik ini digunakan dalam penarikan informasi (IR) saat memproses teks berdasarkan isi dokumen.

Splitting dan Balancing pada dataset

Pada penelitian ini dilakukan *splitting* data pada dataset dengan membaginya menjadi dua bagian tiap label. Memisahkan antara yang bernilai 0 dan bernilai 1 menjadi dua bagian yang berbeda. Kemudian dilakukan penggabungan kembali dataset yang telah dipisahkan ke dalam dataset baru. *Balance* didapatkan dengan memastikan penggabungan dilakukan dengan menggabungkan baris-baris secara berurutan berdasarkan baris sehingga dihasilkan jumlah baris dan kolom yang telah seimbang.

Klasifikasi dan Evaluasi

Penelitian ini menggunakan algoritma NB yang sudah di *enhance* dengan *frequency–inverse document frequency* (TF-IDF) (Aizawa, 2003) untuk menilai seberapa penting sebuah kata pada data dalam dokumen, koleksi, ataupun korpus sebagaimana disajikan pada Persamaan (1).

$$w_{i,j} = tf_{i,j}(log\left(\frac{N}{df_i}\right)) \quad (1)$$

TF IDF dapat dihitung dengan Persamaan (1) dengan mengalikan $tf_{i,j}$ yang merupakan banyaknya kata ke i^{th}pada dokumen ke j^{th} yang dilakukan operasi perkalian dengan hasil pembagian antara total jumlah dokumen N dan jumlah dokumen yang mengandung kata ke i^{th}. Perhitungan ini menghasilkan nilai *inverse document frequency* (IDF). TF-IDF bekerja untuk menentukan frekuensi relatif dari suatu kata dalam sebuah dokumen, dan membandingkannya dengan proporsi kata tersebut dalam keseluruhan dokumen (Robertson, 2004).

Dalam penerapannya, terdapat 4 jenis NB yang digunakan dalam penelitian, yaitu *compliment*, *multinomial*, *gaussian,* dan *bernoulli*. NB adalah algoritma *machine learning* simpel yang menggunakan pedoman Bayes dengan asumsi yang kuat bahwa atribut-atribut tersebut bersifat independen secara kondisonal berdasarkan kelasnya. Meskipun asumsi ini seringkali mengganggu saat pelatihan, tetapi seringkali NB menyajikan hasil akurasi yang dapat bersaing. NB menerapkan aturan Bayes dengan asumsi kuat bahwa atribut-atribut bersifat independen secara kondisional berdasarkan kelasnya. NB menyediakan mekanisme untuk memanfaatkan informasi dari data sampel guna memperkirakan probabilitas posterior $P\ (x \mid y)$ dari setiap kelas y jika diberi objek x sebagaimana sdisajikan pada Persamaan (2). Dengan pendekatan ini, NB memprediksi kelas berdasarkan probabilitas terbesar yang dihitung dari fitur-fitur yang ada.

$$P(x \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \quad (2)$$

Persamaan (2) berlaku untuk data nilai atribut, dimana x_i merupakan nilai dari atribut ke i^{th}pada x dan n merupakan jumlah atributnya. Pengklasifikasi yang dihasilkan menggunakan model linier, setara dengan yang digunakan oleh Logistic Regression, hanya berbeda dalam cara pemilihan parameternya.

Penelitian ini juga menambahkan sistem *censorship* di akhir *modelling* untuk melakukan *censoring* pada kata yang dinilai memiliki bobot sentimen negatif berdasarkan dataset yang telah dipakai.

Menampilkan Skor dan Visualisasi

Data hasil proses algoritma yang dijalankan akan ditampilkan menggunakan matrik *confusion matrix* (Narkhede, 2018) dan skor akurasi. Matrik *confusion* merupakan matriks yang memberikan informasi mengenai hasil perbandingan klasifikasi pada kumpulan data uji yang sudah diketahui nilai sebenarnya. Matrik c*onfusion* terdiri dari empat kombinasi utama, yaitu *true positive* (TP), *false positive* (FP), *false negative* (FN), dan *true negative* (TN). Matriks ini digunakan untuk mengevaluasi performa model klasifikasi, dengan masing-masing elemen memberikan gambaran mengenai prediksi yang benar atau salah.

HASIL DAN PEMBAHASAN

Pemrosesan yang dilakukan dalam riset ini dibagi menjadi dua tahapan utama. Tahap pertama adalah persiapan, meliputi proses *Data Cleaning* dan *Pre-Processing*. Tahap kedua adalah pengujian algoritma terhadap dataset yang digunakan.

Untuk mendapatkan hasil dan akurasi terbaik dalam mengidentifikasi ujaran kebencian dari sebuah tweet, termasuk target ujaran kebencian, kategori, dan tingkat keparahan yang terkandung dalam tweet tersebut, dilakukan serangkaian eksperimen dengan setiap jenis ekstraksi fitur terlebih dahulu. Setelah itu, hasil eksperimen untuk jenis fitur terbaik disimpan, berdasarkan rata-rata akurasi yang diperoleh dari pengujian menggunakan berbagai algoritma pengklasifikasi dan metode transformasi data.

Worldcloud dari Ujaran Kebencian yang Paling Umum

Worldcloud yang ditampilkan pada Gambar 2 menunjukkan kata-kata kasar yang paling umum ditemukan dalam dataset. Representasi visual ini membantu memvisualisasikan frekuensi kata-kata kasar tertentu. Ukuran kata yang lebih besar menunjukkan frekuensi kemunculan yang lebih tinggi. Kata-kata "komunis," "asing," dan "cebong" muncul dalam ukuran terbesar, yang menunjukkan bahwa istilah-istilah tersebut paling sering digunakan dalam dataset ujaran kebencian.

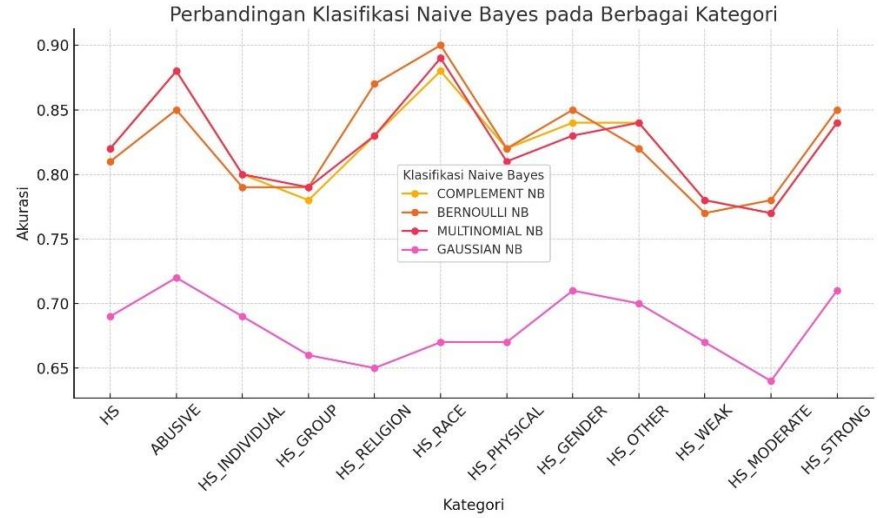
Gambar 2. Kata-kata kasar yang sering muncul

Hasil Pengklasifikasian dan Pelatihan Data

Hasil pengklasifikasian dan pelatihan data dianalisis dengan cara membandingkan nilai *F*1-Score-nya. Tabel 2 menyajikan hasil rataan kinerja berbagai model klasifikasi yang digunakan untuk mendeteksi ujaran kebencian dan bahasa kasar (kata dengan sentiment *negative*) dalam tweet berbahasa Indonesia. Model yang dievaluasi meliputi Complement NB, Bernoulli NB, Multinomial NB, dan Gaussian NB. Data Tabel 2 dapat representasikan dalam bentuk grafis seperti pada Gambar 3.

Tabel 2. F1-score tiap algoritma pada setiap label

Label	Complement NB	Bernoulli NB	Multinomial NB	Gaussian NB
HS	0,82	0,81	0,82	0,69
Abusive	0,88	0,85	0,88	0,72
HS_Individual	0,80	0,79	0,80	0,69
HS_Group	0,78	0,79	0,79	0,66
HS_Religion	0,83	0,87	0,83	0,65
HS_Race	0,88	0,90	0,89	0,67
HS_Physical	0,82	0,82	0,81	0,67
HS_Gender	0,84	0,85	0,83	0,71
hs_other	0,84	0,82	0,84	0,70
hs_weak	0,78	0,77	0,78	0,67
HS_Moderate	0,77	0,78	0,77	0,64
HS_Strong	0,84	0,85	0,84	0,71



Gambar 3. Perbandingan klasifikasi NB pada berbagai kategori

Berdasarkan rata-rata hasil algoritma yang telah dijalankan, seperti yang ditunjukkan pada Tabel 2 dan Gambar 3, diperoleh *F*1 *Score* untuk setiap kategori dengan menggunakan algoritma yang berbeda. Perolehan tertinggi dicapai pada kategori HS_Race oleh Bernoulli NB dengan skor 90%, sedangkan perolehan terendah diperoleh pada kategori HS_Moderate oleh algoritma Gaussian NB dengan skor 64%. Pada penelitian Ibrohim dan Budi (2019), algoritma *logistic regression* digunakan untuk mendeteksi ujaran kebencian di Twitter berbahasa Indonesia, mencapai akurasi sebesar 79,85%. Deteksi ujaran kebencian juga pernah dilakukan Aljero *et al.* (2021) menggunakan *genetic programming*, Fatahillah *et al.* (2017) menggunakan Naive Bayes *classifier*, Ketsbaia *et al.* (2023) menggunakan *multi-stage machine learning*, Oriola dan Kotze (2020) menggunakan teknik evaluasi mesin learning, Plaza *et al.* (2021) menggunakan *multi-task learning*, Sreelakshmi *et al.* (2024) menggunakan *cost-sensitive learning*, dan Obaid *et al.* (2024) dan Zhou *et al.* (2020) menggunakan algoritma *deep learning*.

Pada penelitian Ibrohim dan Budi (2019), algoritma *logistic regression* digunakan untuk mendeteksi ujaran kebencian di Twitter berbahasa Indonesia, mencapai akurasi sebesar 79,85%. Secara keseluruhan, Complement NB dan Multinomial NB menunjukkan kinerja yang konsisten dan tinggi di hampir semua kategori, sementara Bernoulli NB memberikan hasil yang kurang kompetitif dibandingkan dengan algoritma lainnya. Variasi kinerja diantara model-model ini menekankan pentingnya pemilihan model yang tepat berdasarkan kategori ujaran kebencian yang dideteksi.

SIMPULAN

Berdasarkan hasil penelitian ini, berbagai model klasifikasi telah diuji untuk mendeteksi ujaran kebencian dalam tweet berbahasa Indonesia. Hasil dari empat algoritma yang diuji, yaitu Complement Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, dan Gaussian Naïve Bayes, menunjukkan bahwa Bernoulli Naïve Bayes unggul di beberapa kategori dengan akurasi tertinggi 90%. Complement NB dan Multinomial NB juga merupakan pilihan yang sangat baik dengan kinerja yang hampir setara. Meskipun Gaussian NB memiliki kinerja yang lebih rendah, algoritma ini tetap dapat berguna dalam konteks tertentu. Penelitian ini memberikan gambaran mengenai pemilihan model yang tepat untuk tugas deteksi ujaran kebencian, serta dapat digunakan sebagai dasar dalam pengembangan sistem deteksi yang lebih efektif dan akurat di masa depan.

DAFTAR PUSTAKA

Ahmad, M., Octaviansyah, M. F., Kardiana, A., & Prasetyo, K. F. (2019). Sentiment analysis system of Indonesian tweets using lexicon and Naïve Bayes approach. *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, Indonesia, 1-5. https://doi.org/10.1109/ICIC47613.2019.8985930

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1), 45-65.

Ali, M. Z., Ehsan-Ul-Haq, Rauf, S., Javed, K., & Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. *IEEE Access*, 9, 84296-84305. https://doi.org/10.1109/ACCESS.2021.3087827

Aljero, M. K. A., & Dimililer, N. (2021). Genetic programming approach to detect hate speech in social media. *IEEE Access*, 9, 115115–115125. https://doi.org/10.1109/ACCESS.2021.3104535

Asogwa, D. C., Chukwuneke, C. I., Ngene, C. C., & Anigbogu, G. N. (2022). Hate speech classification using SVM and Naive Bayes. *IOSR Journal of Mobile Computing & Application*, 9(1), 27-34. https://doi.org/10.9790/0050-09012734

Buntoro, G. A. (2016). Analisis sentimen hatespeech pada twitter dengan metode naïve bayes classifier dan support vector machine. *Jurnal Dinamika Informatika*, 5(2), 1978-1660.

Fatahillah, N. R., Suryati, P., & Haryawan, C. (2017). Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. *2017 International Conference on Sustainable Information Engineering and Technology* (SIET), Malang, Indonesia, 128–131. https://doi.org/10.1109/SIET.2017.8304122

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. [Online]. Available: https://www.komnasham.go.id/index.php/

Ketsbaia, L., Issac, B., Chen, X., & Jacob, S. M. (2023). A multi-stage machine learning and fuzzy approach to cyber-hate detection. *IEEE Access*, 11, 56046-56065. https://doi.org/10.1109/ACCESS.2023.3282834

Komnas HAM. (2016). Buku Saku Penanganan Ujaran Kebencian. Divisi Hukum Kepolisian Republik Indonesia. [E-Book]. Available at https://www.komnasham.go.id/index.php/publikasi/

Mansur, Z., Omar, N., & Tiun, S. (2023). Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities. *IEEE Access*, 11, 16226-16249. https://doi.org/10.1109/ACCESS.2023.3239375

Narkhede, S. (2018). Understanding confusion matrix. *Towards Data Science*, availabel at https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Obaid, M. H., Elkaffas, S. M., & Guirguis, S. K. (2024). Deep learning algorithms for cyber-bullying detection in social media platforms. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3406595

Oriola, O., & Kotze, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, 8, 21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173

Plaza-Del-Arco, F. M., Molina-Gonzalez, M. D., Urena-Lopez, L. A., & Martin-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9, 112478-112489. https://doi.org/10.1109/ACCESS.2021.3103697

Robertson, S. E. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60, 503-520.

Sreelakshmi, K., Premjith, B., Chakravarthi, B. R., & Soman, K. P. (2024). Detection of hate speech and offensive language code-mixed text in Dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12, 20064-20090. https://doi.org/10.1109/ACCESS.2024.3358811

Wang, W., He, G., & Liu, X. (2019). Text multi-classification based on word embedding and multi-grained cascade forest. IEEE. 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, 13-17, doi: 10.1109/ICCC47050.2019.9064153

Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. *IEEE Access*, 8, 128923-128929. https://doi.org/10.1109/ACCESS.2020.3009244