UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

KLASIFIKASI HATE SPEECH DAN EMOSI DALAM TEKS BERBAHASA INDONESIA PADA PENGGUNA TWITTER MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER

Oleh:

Chandra Hary Pratama

Yulian Findawati, ST, M.MT.

Progam Studi Informatika

Universitas Muhammadiyah Sidoarjo

Mei 2024

Pendahuluan

Ujaran kebencian atau hate speech adalah bentuk ekspresi yang menghasut, menyebarkan, atau mempromosikan kebencian, diskriminasi, dan kekerasan terhadap seseorang atau kelompok. Ujaran kebencian sering ditemui di media sosial, seperti Twitter. Media sosial, termasuk Twitter, sering disalahgunakan sebagai tempat ekspresi ujaran kebencian dan emosi. Pada penelitian (Martins *et al.*, 2018) menunjukkan bahwa emosi tertentu seperti kemarahan dan kebencian lebih berkorelasi dengan ujaran kebencian di *Twitter*. Klasifikasi emosi terdiri dari "*Anger*", "*Anticipation*", "*Disgust*", "*Fear*", "*Joy*", "*Sadness*", "*Surprise*" dan "*Trust*".

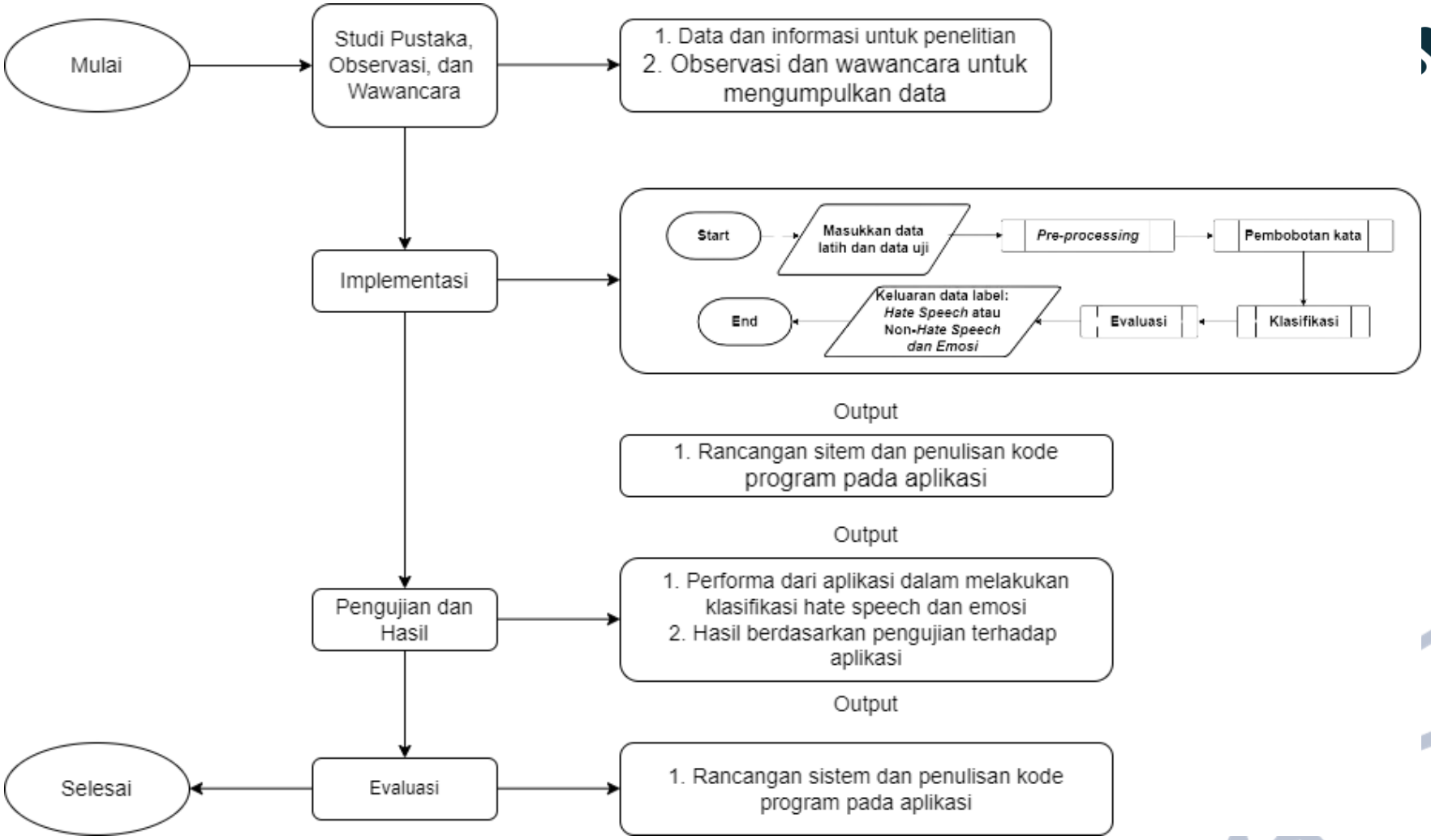
Survei *We Are Social* menyebutkan dalam tinjauan media sosial penduduk Indonesia yang aktif bermain media sosial mencapai 4,62 milyar orang pada tahun 2022, dan jumlah perangkat mobile yang terhubung mencapai 8,28 milyar. Twitter menempati peringkat ke enam dengan presentase 58,3%.

Rumusan Masalah

Berdasarkan dengan penjelasan permasalahan di latar belakang, maka dapat dirumuskan suatu masalah yakni "Bagaimana cara klasifikasi *hate speech* dan emosi serta perhitungan nilai akurasi pada klasifikasi *hate speech* dan emosi dalam teks berbahasa Indonesia pada pengguna Twitter".

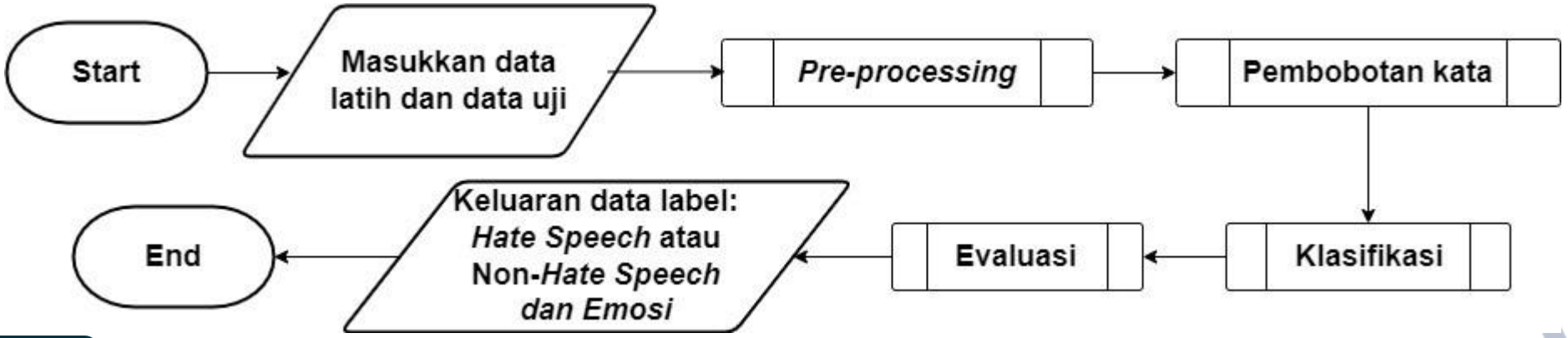
Metode

Beberapa proses penelitian yang terdri dari tahapan studi pustaka dan observasi, tahapan implementasi, tahapan pengujian dan hasil, dan terakhir tahapan evaluasi.



Metode

Program menggunakan algoritma Naïve Bayes Classifier untuk mengklasifikasikan teks tweet berbahasa Indonesia dalam dua jenis klasifikasi utama: klasifikasi hate speech dan klasifikasi emosi. Untuk klasifikasi hate speech, program memberikan label apakah sebuah tweet termasuk dalam kategori ujaran kebencian atau tidak. Sedangkan untuk klasifikasi emosi, program mengidentifikasi emosi yang terkandung dalam teks tweet, seperti marah, sedih, bahagia, dsb. Output akhir program berupa data terstruktur yang mencakup teks tweet, label hate speech, dan label emosi, yang dapat digunakan untuk analisis lebih lanjut atau moderasi konten di platform Twitter.



Hasil dan Pembahasan

Dataset

Data dari media sosial Twitter yang berhasil dikumpulkan sebanyak 3.972 tweet. Memiliki 15 atribut antara lain URL, Tanggal, Tweet, ID, Username, Likes, Quotesdan sebagainya yang disimpan dalam format .csv. Data selanjutnya diolah ke tahapan preprocessinguntuk meningkatkan struktur dari data. Tabel 1 menunjukkan hasil crawling tweet.

NO	Tweet
1	@arunariftan makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2 sich ya mampus saja ha ha ha racain.
2	Ha ha ha sigundul penguasa ancol karena selama ini taunya hanya jilat2 gabenar mulai dibuka jeroannya, KPK Kejaksaan Agung Mabes Polri tolong segera turun/selidiki MERDEKA. https://t.co/Zzp3e2IZ67
…..	…..
3972	Buat KADRUN2 nih

Hasil dan Pembahasan

Pre-processing

Tahapan preprocessingdata dilakukan untuk meningkatkan performa masing-masing algoritma klasifikasi dalam melakukan prediksi sehingga didapatkan data yang lebih presisi. Tahapan ini meliputi cleaningdata, Setelah dilakukan proses preprocessingdata jumlah data yang dihasilkan sejumlah 3.972 tweet.

Proses	Tweet
Cleaning data	@arunariftan makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2 sich ya mampus saja ha ha ha racain.
Tokenization	[‘makin’, ‘gila’, ‘lihat’, ‘lbp’, ‘mewarnai’, ‘indonesia’, ‘tercinta’, ‘barisan’, ‘saku’, ‘hati’, ‘kadrun’, ’sich’, ‘ya’, ‘mampus’, ‘saja’, ‘ha’, ‘ha’, ‘ha’, ‘racain’]
Stopword Removal	[‘makin’, ‘gila’, ‘lihat’, ‘lbp’, ‘mewarnai’, ‘indonesia’, ‘tercinta’, ‘barisan’, ‘saku’, ‘hati’, ‘kadrun’, ’sich’, ‘ya’, ‘mampus’, ‘saja’, ‘ha’, ‘ha’, ‘ha’, ‘racain’]
Stemming	makin gila lihat lbp mewarnai indonesia tercinta barisan saku hati kadrun2 sich ya mampus saja ha ha ha racain

Hasil dan Pembahasan

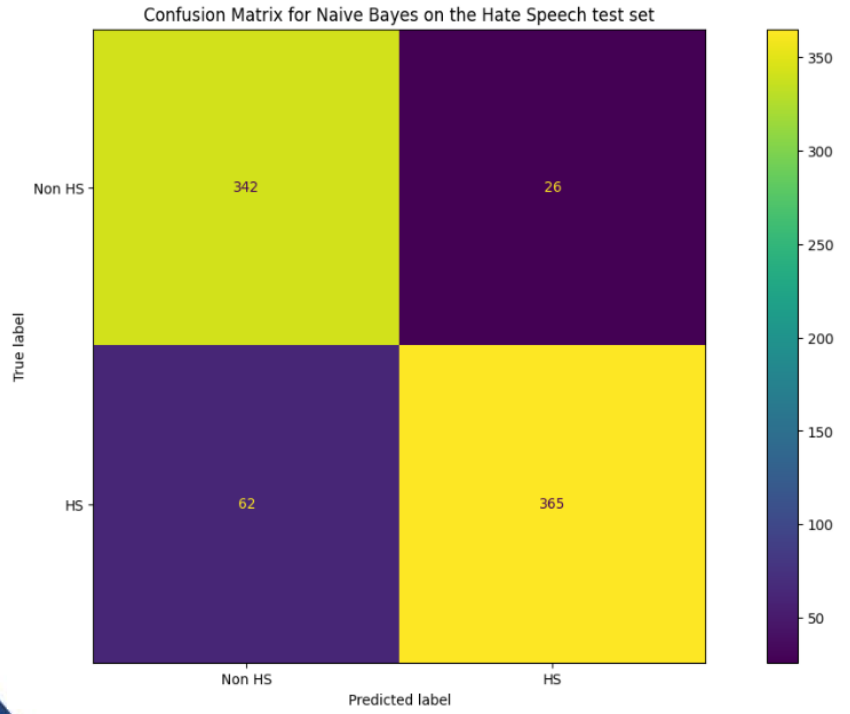
Pembobotan Kata

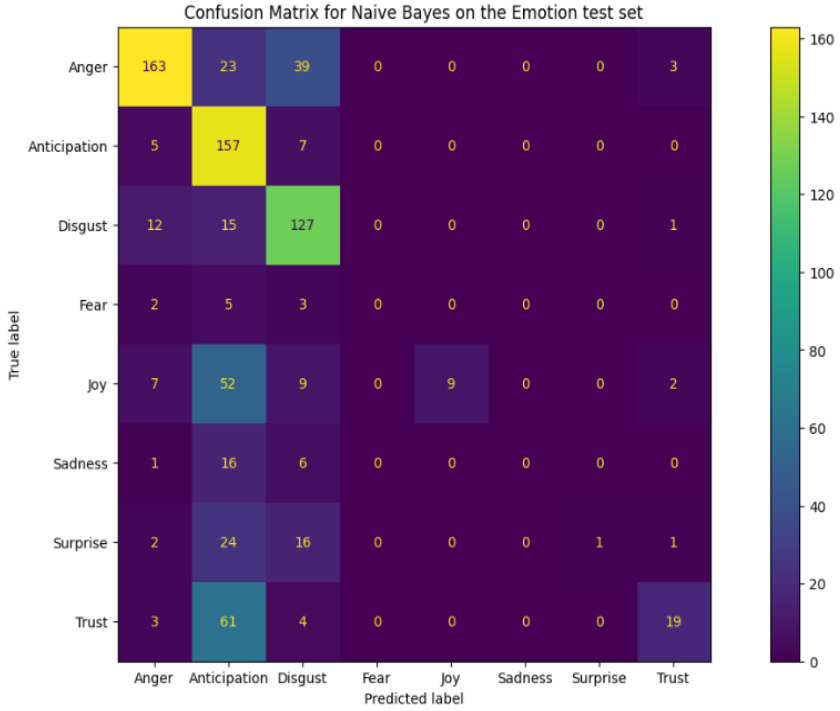
Pembobotan kata (Term Weighting) yaitu suatu mekanisme pemberian nilai pada setiap kata berdasarkan indeks. Proses pembobotan kata ini bertujuan untuk memperoleh jumlah kemunculan kata setelah dilakukan proses pre-processing pada dataset. Pada penelitian ini metode pembobotan kata yang digunakan adalah TF-IDF (Term Frequency Invers Document Frequency). Metode inidigunakan untuk menentukan keterhubungan kata terhadap dokumen dengan cara memberikan bobot pada setiap kata. Pada perhitungan TF-IDF terlebih dahulu menghitung nilai TF pada setiap kata, dengan bobot satu kata adalah 1.

Hasil dan Pembahasan

Klasifikasi

Setelah tahapan pembobotan kata, data dibagi menjadi data latih dan data uji dengan rasio 60:40. Data latih digunakan untuk membangun model dan mengidentifikasi pola, sementara data uji digunakan untuk evaluasi model. Setelah pembagian data, langkah selanjutnya adalah melakukan klasifikasi menggunakan algoritma Naive Bayes sebagai percobaan pertama..





Hasil dan Pembahasan

Evaluasi

Tahap akhir adalah evaluasi, yang menguji hasil klasifikasi dengan mengukur nilai kebenaran sistem, dengan precision sebagai tolok ukur utama. Accuracy digunakan untuk mengevaluasi ekstraksi fitur, dengan persamaan yang sesuai.

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \times 100\%$$

Model Algoritma	Hate Speech	Emosi
Mean Accuracy	0.889308176100629	0.5987421383647799

Hasil dan Pembahasan

Output

Output yang dihasilkan dari program ini adalah untuk mendeteksi hate speech dan emosi dari sebuah kalimat random. Ada dua jenis output dari program ini yatu berupa keterangan HS/Non HS dan emosi.

Text	Hate Speech	Emosi
gin ngebacot tunjukin muka tampang ditwitter bawa islam islam sadar ideologi indonesia pancasila	True	Anger
komentar dukung setia joko widodo presiden ri cinta rakyat kerja keras	False	Anticipation
makin gila lihat LBP mewarnai Indonesia tercinta, Barisan Saku Hati Kadrun2 sich ya mampus saja ha ha ha racain.	True	Anger
kasihan korban akibat cuci otak kadrun tanggung terima kasih polisi paspampres waspada waspada waspada	True	Anticipation

Temuan Penting Penelitian

Klasifikasi hate speech dan emosi menggunakan algoritma Naive Bayes pada Twitter, temuan pentingnya mencakup dua aspek utama. Pertama, dalam klasifikasi hate speech, model Naive Bayes berhasil mengidentifikasi tweet yang mengandung ujaran kebencian dengan akurasi yang signifikan. Hasil ini menunjukkan bahwa pendekatan Naive Bayes dapat menjadi alat yang efektif dalam memoderasi konten negatif di platform media sosial seperti Twitter. Kedua, dalam klasifikasi emosi, model Naive Bayes mampu mengenali emosi yang terkandung dalam tweet dengan tingkat keakuratan yang memuaskan. Ini menunjukkan potensi algoritma ini dalam menganalisis dan memahami konten emosional dalam skala besar, yang dapat digunakan untuk berbagai tujuan, termasuk analisis sentimen dan pemahaman perilaku pengguna. Kesimpulannya, penggunaan Naive Bayes dalam klasifikasi hate speech dan emosi di Twitter menunjukkan hasil yang menjanjikan dan dapat memberikan kontribusi positif dalam upaya memahami dan mengelola konten yang beragam di platform media sosial tersebut.

Manfaat Penelitian

Penelitian klasifikasi hate speech dan emosi menggunakan algoritma Naive Bayes pada Twitter memiliki manfaat yang signifikan dalam konteks pengelolaan konten dan pemahaman perilaku pengguna. Pertama, kemampuan algoritma Naive Bayes dalam mengidentifikasi ujaran kebencian dapat membantu platform media sosial seperti Twitter untuk secara efektif memoderasi konten negatif dan mengurangi dampaknya terhadap pengguna. Ini dapat meningkatkan pengalaman pengguna secara keseluruhan dan menciptakan lingkungan online yang lebih aman dan inklusif. Kedua, kemampuan model ini dalam mengenali emosi dalam tweet dapat digunakan untuk menganalisis sentimen pengguna dan tren perilaku, yang dapat berguna bagi berbagai pihak, termasuk pemasar, peneliti, dan pengambil kebijakan, dalam membuat keputusan yang lebih tepat dan berbasis data. Dengan demikian, penelitian ini memberikan kontribusi yang berharga dalam memahami dan mengelola konten yang beragam di platform media sosial seperti Twitter.

Referensi

- Afif, M.F.A., Nurhamidah, Y. and Mashuri, M.F. (2021) ‘Kematangan emosi dalam perilaku ujaran kebencian pada kebijakan politik’, *Cognicia*, 9(1), pp. 25–30. Available at: https://doi.org/10.22219/cognicia.v9i1.14234

- Ahmad Gozali, H. and Alfan Rosid, M. (2020) ‘Classification of Student Complaints with Naive Bayes and Literary Methods Klasifikasi Keluhan Mahasiswa dengan Metode Naive Bayes dan Sastrawi’, *Network, and Computer Science)* |, 3(1), pp. 22–26.

- Ahmad Wildan Attabi, Lailil Muflikhah and Mochammad Ali Fauzi (2018) ‘Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain’, *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(11), pp. 4548–4554.

- Al Baqi, S. (2015) ‘Ekspresi Emosi Marah’, *Buletin Psikologi*, 23(1), p. 22. Available at: https://doi.org/10.22146/bpsi.10574.

- Deolika, A., Kusrini, K. and Luthfi, E.T. (2019) ‘Analisis Pembobotan Kata Pada Klasifikasi Text Mining’, *Jurnal Teknologi Informasi*, 3(2), p. 179. Available at: https://doi.org/10.36294/jurti.v3i2.1077.

- Dwitama, A.P.J. and Hidayat, S. (2021) ‘Identifikasi Ujaran Kebencian Multilabel Pada Teks Twitter Berbahasa Indonesia Menggunakan Convolution Neural Network’, *Jurnal Sistem Komputer dan Informatika (JSON)*, 3(2), p. 117. Available at: https://doi.org/10.30865/json.v3i2.3610.

- Fanesya, F., Wihandika, R.C. and Indriati (2019) ‘Deteksi Emosi pada Twitter Menggunakan Metode Naive Bayes dan Kombinasi Fitur’, *jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(7), p. 3.

- Ghassani Saskia, T. (2021) ‘Klasifikasi Hate Speech Dan Abusive LanguagePada Twitter Bahasa Indonesia Dengan MetodeNaive Bayes Classifier’.

Referensi

- Hadna, N.M.S., Santosa, P.I. and Winarno, W.W. (2016) 'Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter', *Seminar Nasional Teknologi Informasi dan Komunikasi*, 2016(March), pp. 57–64.
- Hakiem, M., Fauzi, M.A. and Indriati (2019) 'Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(3), pp. 2443–2451. Available at: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682.
- Hidayat, T.E. and Rosid, A. (2020) 'Analysis of Community Sentiments Regarding Plans to Relocate National Capital Using the Naïve Bayes Method Analisa Sentimen Masyarakat Tentang Rencana Pemindahan Ibukota Negara Dengan Metode Naïve Bayes', *Network, and Computer Science)* |, 3(2), pp. 43–49.
- Kumari, A. (2014) 'Study on Naive Bayesian Classifier and its relation to Information Gain', *International Journal on Recent and Innovation Trends in Computing and Communication*, 2(3), pp. 601–602.
- Liu, I. and Sari, Y.A. (2019) 'Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(5), pp. 4914–4922.
- Martins, B. *et al.* (2018) 'Age Differences in Emotion Regulation Choice: Older Adults Use Distraction Less Than Younger Adults in High-Intensity Positive Contexts', *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 73(4), pp. 603–611. Available at: https://doi.org/10.1093/geronb/gbw028.
- Socrates, I.G.A. *et al.* (2016) 'Optimasi Naive Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio', *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*,7(1), p. 22. Available at: https://doi.org/10.24843/lkjiti.2016.v07.i01.p03.


UMSIDA
DARI SINI PENCERAHAN BERSEMI