

KLASIFIKASI EMOSI DENGAN MENGGUNAKAN ALGORITMA NAÏVE BAYES BERDASARKAN TWEET DENGAN BAHASA SUNDA

Hendri Prabowo¹, Waskito Budi Utomo², Dwi Septiani Juandaputri³.

Program Studi Manajemen Informatika¹²³

POLITEKNIK GEOLOGI DAN PERTAMBANGAN AGP

prabowo.sun3@gmail.com¹, ito.waskitobudiutomo@gmail.com², dwi21septia@gmail.com³.

(*) Corresponding Author : prabowo.sun3@gmail.com¹

Abstract— In the current digital era, social media platforms like Twitter have become the primary means for people to express their feelings and emotions. In the context of local culture and language, such as Sundanese, this platform provides valuable insights into the unique and distinctive emotional dynamics of its speakers. The challenges in detecting Sundanese emotions written on Twitter include several aspects, such as the complexity of the language and dialectal variations. Sundanese language has many dialects and idiomatic expressions that can have different meanings depending on the context. Additionally, the use of mixed language or code-switching between Sundanese and Indonesian language in a single tweet can add challenges to emotion classification. Other factors include the limited data available for training machine learning models, as well as the need to develop natural language processing tools specific to Sundanese. These challenges require careful and innovative approaches to produce accurate and reliable emotion detection. In summary, the evaluation results indicate that the model has a low accuracy rate, only around 26.35%, highlighting the need for improvement to provide more accurate predictions in identifying emotions. The prediction results also indicate that the model tends to classify emotions other than happiness as happiness, indicating a significant bias in the model and reinforcing the need for increased accuracy and differentiation in emotion prediction.

Keywords: Accuracy, Sundanese, Naive Bayes Algorithm, Wonderful Indonesia.

Abstrak— Dalam era digital saat ini, media sosial seperti Twitter telah menjadi sarana utama bagi masyarakat untuk mengekspresikan perasaan dan emosi mereka. Dalam konteks budaya dan bahasa lokal, seperti Bahasa Sunda, platform ini memberikan wawasan berharga mengenai dinamika emosi yang unik dan khas dari penuturnya. Permasalahan dalam mendeteksi emosi Bahasa Sunda yang ditulis di Twitter mencakup beberapa aspek, antara lain kompleksitas bahasa dan variasi dialek. Bahasa Sunda memiliki banyak ragam dialek dan ekspresi idiomatik yang bisa berbeda makna tergantung pada konteksnya. Selain itu, penggunaan bahasa campuran atau kode switching antara Bahasa Sunda dan Bahasa Indonesia dalam satu tweet dapat menambah tantangan dalam klasifikasi emosi. Faktor lain termasuk keterbatasan data yang tersedia untuk pelatihan model machine learning, serta kebutuhan untuk mengembangkan alat pemrosesan bahasa alami yang spesifik untuk Bahasa Sunda. Tantangan ini memerlukan pendekatan yang cermat dan inovatif untuk menghasilkan deteksi emosi yang akurat dan dapat diandalkan. Secara ringkas, hasil evaluasi menunjukkan bahwa model memiliki tingkat akurasi yang rendah, hanya sekitar 26.35%, menunjukkan perlunya perbaikan agar dapat memberikan prediksi yang lebih akurat dalam mengidentifikasi emosi. Hasil prediksi juga menunjukkan bahwa model cenderung mengklasifikasikan emosi selain gembira sebagai gembira, menunjukkan adanya bias yang signifikan dalam model dan memperkuat kebutuhan akan peningkatan akurasi dan diferensiasi dalam prediksi emosi.

Kata Kunci : Akurasi, Sunda, Algoritma Naove Bayes, Wonderfull Indonesia.

INTRODUCTION

Dalam era digital saat ini, media sosial seperti Twitter telah menjadi sarana utama bagi masyarakat untuk mengekspresikan perasaan dan emosi mereka. Dalam konteks budaya dan bahasa lokal, seperti Bahasa Sunda, platform ini memberikan wawasan berharga mengenai dinamika emosi yang unik dan khas dari

penuturnya. Penelitian tentang emosi dalam Bahasa Sunda yang diungkapkan melalui tweet dapat mengungkapkan pola-pola ekspresi emosi yang berbeda dari bahasa lain, memperkaya pemahaman kita tentang keragaman emosi manusia. Melalui analisis teks dan klasifikasi emosi, kita dapat mengeksplorasi bagaimana komunitas Sunda mengartikulasikan perasaan mereka dalam

lingkungan digital, membuka peluang untuk penelitian lebih lanjut dalam bidang linguistik dan psikologi sosial.(Emosi et al., 2022)

Permasalahan dalam mendeteksi emosi Bahasa Sunda yang ditulis di Twitter mencakup beberapa aspek, antara lain kompleksitas bahasa dan variasi dialek. Bahasa Sunda memiliki banyak ragam dialek dan ekspresi idiomatik yang bisa berbeda makna tergantung pada konteksnya. Selain itu, penggunaan bahasa campuran atau kode switching antara Bahasa Sunda dan Bahasa Indonesia dalam satu tweet dapat menambah tantangan dalam klasifikasi emosi. Faktor lain termasuk keterbatasan data yang tersedia untuk pelatihan model machine learning, serta kebutuhan untuk mengembangkan alat pemrosesan bahasa alami yang spesifik untuk Bahasa Sunda. Tantangan ini memerlukan pendekatan yang cermat dan inovatif untuk menghasilkan deteksi emosi yang akurat dan dapat diandalkan.

Penelitian ketiga yang dilakukan oleh dian rezky wulandari dkk pada e-proceeding of engineering tahun 2022 dengan judul deteksi emosi berbasis teks untuk menganalisis kuliah daring selama masa pandemi menggunakan algoritme naïve bayes menjelaskan bahwa pandemi covid – 19 muncul pertama kali pada tahun 2020 di indonesia. Hal ini menyebabkan perubahan di berbagai aspek salah satunya pendidikan dan sosial. Media sosial banyak digunakan untuk mengekspresikan banyak hal seperti mengungkapkan emosi mereka terkait kuliah daring seperti pada twitter. Penelitian ini bertujuan untuk mengetahui bagaimana emosi yang timbul terkait kuliah daring. Pada tugas akhir ini digunakan algoritme naïve bayes. Berdasarkan penelitian yang dilakukan oleh shaver, emosi tersebut terdiri dari lima kelas yaitu senang, sedih, marah, takut, dan cinta. Pembagian data dibagi menjadi empat kelas emosi (senang, marah, takut, dan cinta) dan tiga kelas emosi (marah, senang, dan cinta). Data yang digunakan yaitu tweets yang diambil pada github dan juga dilakukan web scraping pada twitter yang diimplementasikan kedalam website. Pemodelan yang telah dibuat dilakukan pengujian menggunakan metode confusion matrix untuk mengetahui apakah model dapat dikatakan baik. Hasil menunjukkan bahwa sistem pendeteksi emosi berbasis teks mendapatkan akurasi sebesar 80% untuk tiga kelas emosi dengan test size 0.5, 73.20% untuk empat kelas emosi dengan test size 0.1, dan 60.87% untuk lima kelas emosi dengan test size 0.1.

Penelitian terdahulu oleh Dian Rezky Wulandari dkk. pada tahun 2022 fokus pada deteksi emosi berbasis teks untuk menganalisis kuliah daring selama pandemi menggunakan algoritme Naïve Bayes, dengan data tweets yang sebagian besar

berbahasa Indonesia. Penelitian ini mengelompokkan emosi ke dalam tiga hingga lima kelas dan mendapatkan akurasi tertinggi sebesar 80% untuk tiga kelas emosi. Sedangkan penelitian yang akan dilakukan berjudul "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda" akan fokus pada deteksi emosi dalam tweet yang secara spesifik menggunakan Bahasa Sunda. Perbedaan utama terletak pada bahasa yang digunakan dalam tweet, yaitu Bahasa Sunda, yang memiliki kompleksitas dan nuansa budaya yang berbeda dari Bahasa Indonesia. Penelitian ini akan mengeksplorasi bagaimana algoritme Naïve Bayes dapat diadaptasi untuk klasifikasi emosi dalam konteks bahasa dan budaya yang berbeda, serta mengatasi tantangan yang terkait dengan dialek dan kode switching.(Basuki et al., 2020; Gultom & Yando Tantra, 2019; Nurzilla, 2024)

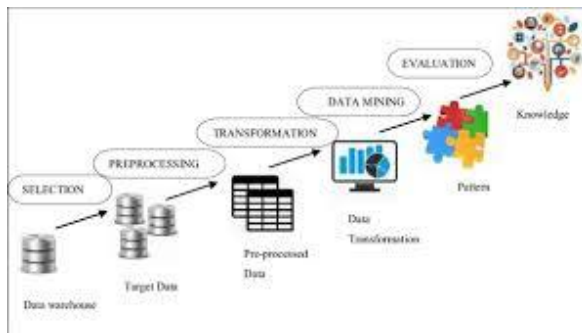
Tujuan dari penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda" adalah untuk mengembangkan dan menguji model klasifikasi emosi yang dapat mendeteksi dan mengklasifikasikan emosi dalam tweet yang ditulis dalam Bahasa Sunda. Penelitian ini bertujuan untuk memahami lebih dalam bagaimana emosi diekspresikan dalam Bahasa Sunda di media sosial, khususnya Twitter, dan untuk mengatasi tantangan linguistik dan budaya yang terkait dengan analisis teks dalam bahasa daerah. Dengan memanfaatkan algoritma Naïve Bayes, penelitian ini berupaya untuk menghasilkan model yang akurat dan andal dalam mengidentifikasi berbagai emosi, seperti senang, sedih, marah, takut, dan cinta, yang diungkapkan dalam tweet-tweet berbahasa Sunda. Hasil penelitian ini diharapkan dapat berkontribusi pada pengembangan teknologi pemrosesan bahasa alami (NLP) yang lebih inklusif dan efektif untuk bahasa-bahasa daerah di Indonesia.

Konsep penyelesaian untuk penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda" mengikuti kerangka kerja Knowledge Discovery in Databases (KDD). Proses dimulai dengan pengumpulan data tweet berbahasa Sunda melalui web scraping atau API Twitter, diikuti dengan pra-pemrosesan untuk membersihkan dan normalisasi teks. Data kemudian diubah menjadi representasi numerik menggunakan teknik seperti TF-IDF. Setelah itu, subset data dipilih dan dibagi menjadi set pelatihan dan pengujian. Model klasifikasi emosi dibangun menggunakan algoritma Naïve Bayes dan dilatih dengan data pelatihan yang telah diberi label emosi. Kinerja model dievaluasi menggunakan metode seperti confusion matrix, akurasi, presisi, recall, dan F1-score. Hasil evaluasi dianalisis untuk

interpretasi lebih lanjut, dan model yang telah divalidasi diimplementasikan untuk mendeteksi emosi dalam tweet berbahasa Sunda secara real-time. Akhirnya, model akan dipantau dan diperbarui secara berkala untuk memastikan kinerja optimal sesuai perubahan bahasa dan ekspresi emosi di media sosial.

MATERIALS AND METHODS

Dalam membangun model dengan judul Klasifikasi Emosi Dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet Dengan Bahasa Sunda dapat dilakukan dengan mengikuti beberapa langkah berikut :



Gambar 1. Metode Penelitian

Konsep penyelesaian dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda" mengacu pada kerangka kerja Knowledge Discovery in Databases (KDD). Tahapan penyelesaiannya adalah sebagai berikut:

1. Pengumpulan Data : Mengumpulkan data tweet berbahasa Sunda dari Twitter menggunakan teknik web scraping atau API Twitter atau Kaggle.
2. Pra-pemrosesan Data : Membersihkan dan mempersiapkan data tweet, termasuk penghapusan tanda baca, URL, emoji, dan karakter khusus. Normalisasi teks dilakukan dengan menghapus kata-kata umum (stop words), stemming, dan tokenisasi.
3. Transformasi Data : Mengubah data teks yang telah diproses menjadi bentuk yang dapat diolah oleh algoritma Naïve Bayes, seperti representasi vektor fitur menggunakan metode TF-IDF atau word embeddings. Memilih subset data yang relevan untuk pelatihan dan pengujian model klasifikasi emosi.
4. Pemodelan : Membangun model klasifikasi emosi menggunakan algoritma Naïve Bayes dengan menggunakan data pelatihan yang telah diolah.

5. Evaluasi Model : Menguji kinerja model menggunakan data pengujian dan metrik evaluasi seperti confusion matrix, akurasi, presisi, recall, dan F1-score. Menganalisis hasil evaluasi untuk mengevaluasi kinerja model dan mengidentifikasi faktor-faktor yang mempengaruhi hasil klasifikasi. Mengimplementasikan model klasifikasi emosi yang telah diuji ke dalam aplikasi yang dapat digunakan untuk menganalisis tweet berbahasa Sunda secara real-time. Memantau kinerja model secara berkala dan melakukan pembaruan sesuai dengan perubahan bahasa dan ekspresi emosi di media sosial.

Sumber Data

Sumber data yang bersumber dari data open, seperti yang tersedia di Kaggle.com, menjadi kunci dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda". Data yang dapat diakses secara publik ini memberikan kesempatan untuk mengumpulkan sampel yang representatif dari tweet berbahasa Sunda, memungkinkan pengembangan model klasifikasi emosi yang lebih luas dan bersifat inklusif link data : <https://www.kaggle.com/datasets/rabieelkharoua/sundanese-twitter-dataset>

Teknik Pengumpulan data

Pengumpulan data untuk penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda" dilakukan melalui dua sumber utama, yaitu Kaggle.com dan Twitter. Data tweet berbahasa Sunda yang tersedia di Kaggle.com digunakan sebagai salah satu sumber data, yang kemungkinan besar telah melalui proses pengumpulan, pemrosesan, dan penyimpanan sebelumnya. Selain itu, data juga dikumpulkan secara langsung dari Twitter menggunakan teknik web scraping atau melalui API Twitter untuk mengumpulkan tweet-tweet terbaru yang menggunakan bahasa Sunda. Data yang dikumpulkan dari kedua sumber ini kemudian digunakan untuk analisis emosi dan pengembangan model klasifikasi emosi berbasis teks. Proses pengumpulan data ini menjadi langkah awal yang penting dalam penelitian ini, karena kualitas dan representativitas data akan berpengaruh pada akurasi dan validitas hasil penelitian.

Teknik yang digunakan dalam penelitian ini yaitu teknik data sekunder.

RESULTS AND DISCUSSION

A. Data Selection

Pada tahap seleksi data dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda", data yang telah dikumpulkan melalui web scraping dari Twitter dan data yang diambil dari Kaggle.com harus disiapkan untuk proses pemodelan. Pertama, data mentah yang telah melalui proses pra-pemrosesan (seperti pembersihan, normalisasi, dan tokenisasi) akan dibagi menjadi dua subset: set pelatihan dan set pengujian. Pemisahan ini dilakukan secara acak namun seimbang untuk memastikan bahwa kedua subset tersebut memiliki distribusi kelas emosi yang serupa, sehingga model dapat dilatih secara efektif dan diuji secara adil. Set pelatihan digunakan untuk melatih model klasifikasi emosi dengan algoritma Naïve Bayes, di mana model ini akan mempelajari pola-pola dan karakteristik dari tweet berbahasa Sunda yang mengandung berbagai emosi. Set pengujian, di sisi lain, digunakan untuk mengevaluasi kinerja model dengan menguji seberapa baik model dapat mengklasifikasikan emosi pada data yang belum pernah dilihat sebelumnya. Dengan pemilihan dan pemisahan data yang tepat, penelitian ini berupaya memastikan bahwa model yang dihasilkan tidak hanya akurat tetapi juga dapat diandalkan ketika diaplikasikan pada data tweet berbahasa Sunda yang baru dan beragam.

Tabel 1 Data Set

label	data
anger	sok geura leungit atuh sia teh corona, matak gelo yeuh aing unggal poe gogoleran
anger	Nu katoel katuhu nu nyerina kenca, goblog wasitna
joy	Bingah pisan patepang sareng pangerasa. Sing katampi kalayan pinuh midulur...
fear	asa hariwang kieu.. lalakon hirup teh asa nyorangan.. asa ieu mah..
anger	Orang mana sih anying, sampis pisan. Bunuh ae lah bunuh
sadness	nya prihatin siih tp kedah kitu at pak Bozz hapuntn abi tos kmh wntu
joy	sae nya kang ai tos ningal kompak kitu teh bingah ningal nage...-D sok janten kangen ai emut bndung teh
sadness	Cai soca ngeclak mun emut kana dosa, rumasa raga lamokot dosa, ngungudag dunya taya reureuhna, teu kabayang kumaha engke di alam baka
anger	Gokar! Punteun, buat yg masih pd nongkrong yg masih jalan2 gajelas,

punteun ieu mah, ari sia goblog2 teuing sih:((Sia terpapar korona di luar nu di imah ge beunang! Cik karunya ka diri sorangan teh gobloggg Cicing heula saminggu mah dripda cicing di r.isolasi teu kaluar2 sateh.

anger kumah pernahnya naik angkot dari Cihampelas trs turun di baso bujangan, kan 2rb Aja ya, manalagi aku pake seragam mah osok murah.. tpi akuteh ngasih 5rb da lagi bageur.. Eh si mamang na kalah ngomong "belegug budak teh, kalah mere loba." aNjing PISAN EUY GEUS DIBERE TEH

anger Sorry kang, abdi mah TEU PERCAYA ka HUKUM nu kieu PATUT. HUKUM MILIK BURUY KECEBONG. nepi ka iraha oge teu rido teu sudi dilicikan. Teu hayang damai mun can jujur jeung adil. Akang teu kudu ngajak2 kabeh rakyat kudu sabalong jeung si ABU JANDA, embung geuleuh!!

joy Bingah ninggal bajaj di jakarta aya sticker Pikiran Rakyat, jadi hoyong geura uwih ka Bandung

sadness rada nyeredet kanu manah ngadangu suara bebedegan petasan teu tiasa ngempeul sreng kluarga..

anger Anjing pisan eta budak letik kenek wani ka awewe si anjing setan budak haram hungkul kesel

fear baru aja update tentang corona :(bener aja kan tuh di depok 2 orang udah positif :(sieun

B. Preprocessing

Pada tahap pra-pemrosesan data dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda", sejumlah langkah penting dilakukan untuk mempersiapkan data mentah agar siap digunakan dalam model klasifikasi. Proses ini dimulai dengan pembersihan data, di mana elemen-elemen yang tidak relevan seperti tanda baca, URL, emoji, dan karakter khusus dihapus untuk memastikan bahwa teks bersih dan fokus pada kata-kata yang bermakna.

Selanjutnya, normalisasi teks dilakukan dengan mengubah semua karakter menjadi huruf kecil untuk mengurangi variasi yang disebabkan oleh perbedaan kapitalisasi. Pada tahap ini juga, kata-kata umum atau stop words yang tidak memberikan informasi penting dihapus dari teks. Kemudian, proses tokenisasi dilakukan untuk memecah teks

menjadi unit-unit kata (token), yang akan digunakan sebagai dasar untuk representasi fitur.

Setelah tokenisasi, teknik stemming diterapkan untuk mengubah kata-kata ke bentuk dasarnya, yang membantu dalam mengurangi variasi kata yang memiliki makna serupa. Selain itu, proses penanganan kode switching dilakukan untuk mengidentifikasi dan menangani kata-kata yang mungkin ditulis dalam bahasa lain selain Bahasa Sunda.

Hasil dari pra-pemrosesan ini adalah teks yang sudah bersih, dinormalisasi, dan terstruktur, siap untuk diubah menjadi representasi numerik menggunakan teknik seperti TF-IDF atau word embeddings. Dengan pra-pemrosesan yang tepat, data yang dihasilkan akan lebih konsisten dan lebih mudah dianalisis oleh algoritma Naïve Bayes, meningkatkan akurasi dan efektivitas model klasifikasi emosi yang dikembangkan dalam penelitian ini.

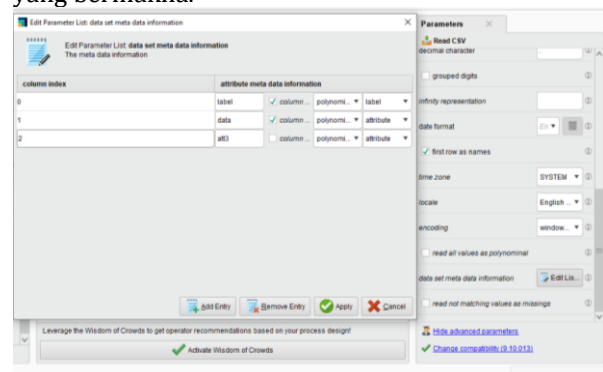
Tabel 2. Preprocessing

label	data
anger	sok geura leungit atuh sia teh corona, matak gelo yeuh aing unggal poe gogoleran
anger	Nu katoel katuhu nu nyerina kenca, goblog wasitna
joy	Bingah pisan patepang sareng pangerasa. Sing katampi kalayan pinuh midulur...
fear	asa hariwang kieu.. lalakon hirup teh asa nyorangan.. asa ieu mah..
anger	Orang mana sih anying, sampis pisan. Bunuh ae lah bunuh
sadness	nya prihatin siih tp kedah kitu at pak Bozz hapuntn abi tos kmh wntu
joy	sae nya kang ai tos ningal kompak kitu teh bingah ningal nage...-D sok janten kangen ai emut bndung teh
sadness	Cai soca ngeclak mun emut kana dosa, rumasa raga lamokot dosa, ngungudag dunya taya reureuhna, teu kabayang kumaha engke di alam baka
anger	Gokar! Punteun, buat yg masih pd nongkrong yg masih jalan2 gajelas, punteun ieu mah, ari sia goblog2 teuing sih:((Sia terpapar korona di luar nu di imah ge beunang! Cik karunya ka diri sorangan teh gobloggg Cicing heula saminggu mah dripda cicing di r.isolasi teu kaluar2 sateh.

anger	kumah pernahnya naik angkot dari Cihampelas trs turun di baso bujangan, kan 2rb Aja ya, manalagi aku pake seragam mah osok murah.. tpi akuteh ngasih 5rb da lagi bageur.. Eh si mamang na kalah ngomong "belegug budak teh, kalah mere loba." aNJING PISAN EUY GEUS DIBERE TEH
anger	Sorry kang, abdi mah TEU PERCAYA ka HUKUM nu kieu PATUT. HUKUM MILIK BURUY KECEBONG. nepi ka iraha oge teu rido teu sudi dilicikan. Teu hayang damai mun can jujur jeung adil. Akang teu kudu ngajak2 kabeh rakyat kudu sabalong jeung si ABU JANDA, embung geuleuh!!
joy	Bingah ninggal bajaj di jakarta aya sticker Pikiran Rakyat, jadi hoyong geura uwih ka Bandung
sadness	rada nyeredet kanu manah ngadangu suara bebededagan petasan teu tiasa ngempeul sreng keluarga..
anger	Anjing pisan eta budak letik kenek wani ka awewe si aning setan budak haram hungkul kesel
fear	baru aja update tentang corona :(bener aja kan tuh di depok 2 orang udah positif :(sieun

C. Transformation

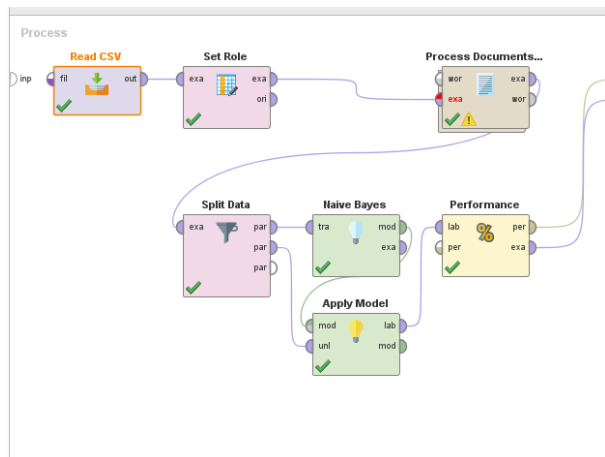
Pada tahap transformasi dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda", data teks yang telah melalui proses pra-pemrosesan diubah menjadi bentuk numerik yang dapat diolah oleh algoritma Naïve Bayes. Proses transformasi ini melibatkan beberapa teknik utama untuk mengonversi teks menjadi representasi fitur yang bermakna.



Gambar 2. Transformation

D. Data Mining.

Pada tahap data mining dalam penelitian "Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda", fokus utama adalah membangun dan melatih model klasifikasi emosi menggunakan algoritma Naïve Bayes. Proses ini melibatkan beberapa langkah kunci yang penting untuk memastikan model yang dihasilkan akurat dan andal dalam mendeteksi emosi dari tweet berbahasa Sunda.



Gambar 3 Desain

Berdasarkan gambar 3 merupakan diagram alir proses data mining yang digunakan untuk klasifikasi emosi menggunakan algoritma Naïve Bayes. Berikut adalah penjelasan detail dari setiap komponen dalam diagram tersebut:

1. Read CSV Fungsi ini digunakan untuk membaca file CSV yang berisi dataset tweet berbahasa Sunda. Ini adalah langkah pertama dalam proses, di mana data mentah diimpor ke dalam sistem untuk diproses lebih lanjut. Output dari langkah ini adalah dataset yang akan diteruskan ke langkah berikutnya.
2. Set Role Langkah ini menentukan peran dari kolom-kolom dalam dataset. Misalnya, menentukan kolom mana yang merupakan fitur (teks tweet) dan kolom mana yang merupakan label (kelas emosi). Setelah peran ditetapkan, data dikirim ke langkah pra-pemrosesan.
3. Process Documents Langkah ini mencakup berbagai tahapan pra-pemrosesan teks, seperti tokenisasi, stemming, penghapusan stop words, dan lainnya. Tujuannya adalah untuk membersihkan dan menyiapkan teks agar dapat digunakan dalam model machine learning. Hasil dari proses ini adalah dataset yang telah diproses yang siap untuk dibagi menjadi set pelatihan dan set pengujian.
 - a) Tokenize, Langkah pertama dalam pra-pemrosesan adalah tokenisasi, di mana teks dibagi menjadi unit-unit kata (token). Proses

ini memecah teks menjadi kata-kata individual atau token yang lebih kecil yang akan dianalisis lebih lanjut. Output dari langkah ini adalah dokumen yang terdiri dari token-token individual.

- b) Tokenize (2), Langkah tokenisasi kedua ini mungkin digunakan untuk memastikan bahwa tokenisasi dilakukan secara menyeluruh, atau untuk memproses bagian-bagian tertentu dari teks yang memerlukan tokenisasi tambahan. Outputnya tetap berupa dokumen yang berisi token-token individual.
 - c) Transform Cases, Pada langkah ini, semua huruf dalam teks diubah menjadi huruf kecil (lowercase). Transformasi ini membantu dalam konsistensi dan mengurangi variasi yang disebabkan oleh perbedaan kapitalisasi, sehingga kata-kata seperti "Bahasa" dan "bahasa" diperlakukan sebagai kata yang sama. Output dari langkah ini adalah dokumen dengan token yang semuanya dalam huruf kecil
 - d) Filter Tokens (by Length), Langkah ini menyaring token berdasarkan panjangnya. Token yang terlalu pendek (misalnya, satu atau dua huruf) yang biasanya tidak memiliki makna signifikan dalam analisis teks dihapus. Output dari langkah ini adalah dokumen yang berisi token dengan panjang yang sesuai dengan kriteria yang ditentukan.
 - e) Filter Stopwords (Dictionary), Langkah ini menghapus kata-kata umum atau stop words yang tidak memberikan informasi penting dalam analisis emosi. Stop words seperti "dan", "yang", "di", dll., dihapus dari teks untuk fokus pada kata-kata yang lebih bermakna. Output dari langkah ini adalah dokumen yang bebas dari stop words.
4. Split Data Dataset yang telah diproses dibagi menjadi dua subset: set pelatihan (training set) dan set pengujian (test set). Ini penting untuk memvalidasi kinerja model secara objektif. Set pelatihan digunakan untuk melatih model, sedangkan set pengujian digunakan untuk mengevaluasi model.
 5. Naïve Bayes Pada langkah ini, algoritma Naïve Bayes digunakan untuk melatih model klasifikasi emosi berdasarkan set pelatihan. Model ini mempelajari pola dan karakteristik data untuk dapat mengklasifikasikan emosi dengan benar. Output dari langkah ini adalah model Naïve Bayes yang terlatih.
 6. Apply Model Model Naïve Bayes yang telah dilatih diterapkan pada set pengujian untuk melakukan prediksi kelas emosi. Ini memungkinkan evaluasi kinerja model dengan

menggunakan data yang belum pernah dilihat sebelumnya oleh model.

7. Performance Langkah ini mengevaluasi kinerja model berdasarkan hasil prediksi pada set pengujian. Metrik evaluasi seperti akurasi, presisi, recall, dan F1-score dihitung untuk memberikan gambaran tentang seberapa baik model bekerja Hasil evaluasi ini digunakan untuk memahami kekuatan dan kelemahan model, serta untuk mengidentifikasi area yang memerlukan perbaikan.

E. Evaluation

Hasil dari penerapan diatas maka dapat hasil sebagai berikut :

accuracy: 26.35%

	true anger	true joy	true fear	true sadness	class precis
pred. anger	2	0	0	0	100.00%
pred. joy	154	162	153	157	25.88%
pred. fear	0	0	0	0	0.00%
pred. sadness	0	0	0	2	100.00%
class recall	1.28%	100.00%	0.00%	1.26%	

Gambar 4. Hasil Akurasi

Berdasarkan gambar 4 menunjukkan matriks untuk klasifikasi emosi, dengan empat kategori emosi: anger (marah), joy (gembira), fear (takut), dan sadness (sedih). Berikut adalah penjelasan dan uraian detailnya:

1. Accuracy (Akurasi): 26.35 %

Akurasi adalah persentase prediksi yang benar dari keseluruhan prediksi. Dalam hal ini, model memiliki akurasi sebesar 26.35%.

2. Hasil prediksi

- a) True Anger (Marah yang Sebenarnya)
 - i. Pred. Anger : 2 (Dari 156 data yang benar-benar marah, 2 diprediksi dengan benar sebagai marah)
 - ii. Pred. Joy: 154 (154 data yang sebenarnya marah diprediksi sebagai gembira)
 - iii. Pred. Fear : 0 (Tidak ada data yang sebenarnya marah diprediksi sebagai takut)
 - iv. Pred. Sadness : 0 (Tidak ada data yang sebenarnya marah diprediksi sebagai sedih)
- b) True Joy (Gembira yang Sebenarnya) :
 - i. Pred. Anger : 0 (Tidak ada data yang sebenarnya gembira diprediksi sebagai marah)
 - ii. Pred. Joy : 162 (162 data yang benar-benar gembira diprediksi dengan benar sebagai gembira)
 - iii. Pred. Fear : 0 (Tidak ada data yang sebenarnya gembira diprediksi sebagai takut)
 - iv. Pred. Sadness : 0 (Tidak ada data yang sebenarnya gembira diprediksi sebagai sedih)
- c) True Fear (Takut yang Sebenarnya) :

- i. Pred. Anger : 0 (Tidak ada data yang sebenarnya takut diprediksi sebagai marah)
- ii. Pred. Joy : 153 (153 data yang sebenarnya takut diprediksi sebagai gembira)
- iii. Pred. Fear : 0 (Tidak ada data yang sebenarnya takut diprediksi sebagai takut)
- iv. Pred. Sadness : 0 (Tidak ada data yang sebenarnya takut diprediksi sebagai sedih)
- d) True Sadness (Sedih yang Sebenarnya) :
 - i. Pred. Anger : 0 (Tidak ada data yang sebenarnya sedih diprediksi sebagai marah)
 - ii. Pred. Joy : 157 (157 data yang sebenarnya sedih diprediksi sebagai gembira)
 - iii. Pred. Fear : 0 (Tidak ada data yang sebenarnya sedih diprediksi sebagai takut)
 - iv. Pred. Sadness : 2 (Dari 159 data yang benar-benar sedih, 2 diprediksi dengan benar sebagai sedih)

3. Class Precision (Presisi Kelas) :

- a) Anger (Marah) : 100% (Semua prediksi marah benar-benar marah, namun jumlahnya sangat sedikit)
- b) Joy (Gembira) : 25.88% (Dari semua prediksi gembira, hanya 25.88% yang benar-benar gembira)
- c) Fear (Takut) : 0% (Tidak ada prediksi yang benar-benar takut)
- d) Sadness (Sedih) : 100% (Semua prediksi sedih benar-benar sedih, namun jumlahnya sangat sedikit)
4. Class Recall (Recall Kelas) :
 - a) Anger (Marah) : 1.28% (Dari semua data yang benar-benar marah, hanya 1.28% yang diprediksi dengan benar sebagai marah)
 - b) Joy (Gembira) : 100% (Semua data yang benar-benar gembira diprediksi dengan benar sebagai gembira)
 - c) Fear (Takut) : 0% (Tidak ada data yang benar-benar takut diprediksi dengan benar sebagai takut)
 - d) Sadness (Sedih) : 1.26% (Dari semua data yang benar-benar sedih, hanya 1.26% yang diprediksi dengan benar sebagai sedih)

KESIMPULAN

Berdasarkan hasil penjelasan penelitian ini dengan judul Klasifikasi Emosi dengan Menggunakan Algoritma Naïve Bayes Berdasarkan Tweet dengan Bahasa Sunda dapat ditarik kesimpulan yaitu : Hasil evaluasi, dapat disimpulkan bahwa model memiliki tingkat akurasi yang rendah, yaitu sebesar 26.35%. Artinya, dari keseluruhan data yang ada, hanya sekitar seperempatnya yang diprediksi dengan benar oleh model. Hal ini menunjukkan bahwa model perlu

diperbaiki atau disesuaikan agar dapat memberikan prediksi yang lebih akurat dalam mengidentifikasi emosi berdasarkan data yang diberikan. Kesimpulan dari hasil prediksi menunjukkan bahwa model klasifikasi memiliki kinerja yang kurang memadai dalam mengidentifikasi emosi dengan benar. Dari 156 data yang benar-benar menunjukkan marah, hanya 2 yang diprediksi dengan benar, sementara 154 lainnya diprediksi sebagai gembira. Untuk data yang benar-benar gembira, model berhasil mengidentifikasi semua 162 data dengan benar. Namun, untuk data yang benar-benar menunjukkan takut, semua 153 data diprediksi sebagai gembira. Hal serupa terjadi pada data yang benar-benar sedih, di mana dari 159 data, hanya 2 yang diprediksi dengan benar, sementara 157 lainnya diprediksi sebagai gembira. Hasil ini menunjukkan bahwa model cenderung mengklasifikasikan berbagai emosi selain gembira sebagai gembira, mengindikasikan bias yang signifikan dalam model dan menunjukkan perlunya perbaikan dalam akurasi dan kemampuan diferensiasi model untuk prediksi emosi yang lebih tepat.

REFERENCE

- [1] Ariya Candra, R. (n.d.). *BRIDGE: Jurnal publikasi Sistem Informasi dan Telekomunikasi Emotion Recognition Tentang Ditutupnya TikTok Shop Menggunakan Metode Naive Bayes*.
- [2] Basuki, T. L., Jondri, M. S., & Wisesty, U. N. (2020). Deteksi Polycystic Ovarian Syndrome (PCOS) Menggunakan Klasifikasi Microarray Data dengan Algoritma Artificial Neural Network (ANN) Backpropagation dan Principal Component Analysis. *E-Proceeding of Engineering*.
- [3] Cahyono, Y. (2019). *Analisis Sentiment Tweets Berbahasa Sunda Menggunakan Naive Bayes Classifier dengan Seleksi Feature Chi Squared Statistic*. 4(3). <http://openjournal.unpam.ac.id/index.php/informatika>
- [4] Emosi, A., Sentimen, D., Sinovac, V., Naive, M., Bagus, B., Akbar, M., Taufiq Akbar, A., & Husaini, R. (2022). Analysis of Sentiments and Emotions about Sinovac Vaccine Using Naive Bayes. *Jurnal Informatika Dan Teknologi Informasi*, 19(2), 185–200. <https://doi.org/10.31515/telematika.v19i2.7601>
- [5] Gitacahyani, A., Irma Purnamasari, A., & Ali, I. (2024). KLASIFIKASI ULASAN APLIKASI LINKEDIN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 1).
- [6] Gultom, D. A., & Yando Tantra, W. (2019). Diagnosis Kanker Payudara Menggunakan Jaringan Saraf Tiruan Dengan Metode Backpropagation. In *PSDKU Medan Jurusan Teknik Informatika INFORMATICS ENGINEERING RESEARCH AND TECHNOLOGY*.
- [7] Ilham, A., Azmi Verdikha, N., & Latipah, A. J. (2023). *Jurnal Explore IT/64 Klasifikasi Ujaran Kebencian di Twitter Menggunakan Fitur Ekstraksi Glove dengan Support Vector Machine (SVM)*. <https://doi.org/10.35891/explorit>
- [8] Jurnal, H., Sandi, A. P., & Ningsih, V. W. (2022). JURNAL PUBLIKASI ILMU KOMPUTER DAN MULTIMEDIA IMPLEMENTASI DATA MINING SEBAGAI PENENTU PERSEDIAAN PRODUK DENGAN ALGORITMA FP-GROWTH PADA DATA PENJUALAN SINARMART. *JUPIKOM*, 1(2).
- [9] Karo Karo, I. M., Karo Karo, J. A., Yunianto, Y., Hariyanto, H., Falah, M., & Ginting, M. (2023). Analisis Sentimen Ulasan Aplikasi Info BMKG di Google Play Menggunakan TF-IDF dan Support Vector Machine. *Journal of Information System Research (JOSH)*, 4(4), 1423–1430. <https://doi.org/10.47065/josh.v4i4.3943>
- [10] Nurzilla, N. (2024). PREDIKSI PERTUMBUHAN TUMOR KANKER PAYUDARA MENGGUNAKAN MODEL REGRESI LINEAR BERBASIS MACHINE LEARNING. In *JOURNAL OF ARTIFICIAL INTELLIGENCE APPLICATION (JAIA) P-ISSN : xxxx-xxxx* (Vol. 1, Issue 1).
- [11] Putra, O. V., Musthafa, A., & Wibowo, K. P. (2021). Klasifikasi Ekspresi Teks Berbahasa Jawa Menggunakan Algoritma Long Short Term Memory. *Komputika: Jurnal Sistem Komputer*, 10(2), 137–143. <https://doi.org/10.34010/komputika.v11i1.4616>
- [12] Sanudin, F. T., Irawan, B., & Bahtiar, A. (2024). ANALISIS SENTIMEN ULASAN APLIKASI SAPAWARGA DI PLAYSTORE MENGGUNAKAN ALGORITMA NAÏVE BAYES. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Issue 1).
- [13] Siahaan, N. R., Tiffany, R. Y., Roland, S., Sinaga, E., Nauli, E. V., Fahmi, M. I., Vio, E., & Naibaho, N. B. (n.d.). *ANALISIS SENTIMEN ULASAN APLIKASI MEDIA SOSIAL WHATSAPP MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER*.

- [14] Wulandari, D. R. (n.d.). *Deteksi Emosi Berbasis Teks Untuk Menganalisis Kuliah Daring Selama Masa Pandemi Menggunakan Algoritme Naive Bayes Text Based Emotion Detection For Analysis Online Lecture During Pandemic Using Naive Bayes Algorithm.*