Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

Perbandingan Performa Algoritma NBC, C4.5, dan KNN dalam Analisis Sentimen Masyarakat terhadap Krisis Petani Muda pada Media Sosial Facebook

Nurkholis, Inggih Permana, Febi Nur Salisah, Mustakim, M Afdal*

Fakultas Sains dan Teknologi, Program Studi Sistem Informasi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]112050316023@students.uin-suska.ac.id, [2]inggihpermana@uin-suska.ac.id, [3]febinursalisah@uin-suska.ac.id, [4]mustakim@uin-suska.ac.id, [5,*]m.afdal@uin-suska.ac.id
Email Penulis Korespondensi: 12050316023@students.uin-suska.ac.id
Submitted: **16/10/2024**; Accepted: **01/12/2024**; Published**: 03/12/2024**

Abstrak−Di Indonesia, petani muda menghadapi berbagai tantangan dan krisis yang menghambat pertumbuhan dan keberlanjutan sektor pertanian. Mereka menghadapi kendala seperti kurangnya akses terhadap modal, teknologi yang terbatas, perubahan iklim, dan rendahnya harga jual hasil panen. Selain itu, mereka juga sering kali menghadapi masalah dalam memperoleh informasi yang akurat dan relevan dalam upaya memudahkan pengambilan keputusan lebih baik pada usaha pertanian, sehingga minat para pemuda sekarang untuk menjadi petani semakin berkurang. Penelitian bertujuan untuk Perbandingan Performa Algoritma NBC, C4.5, dan KNN dalam Analisis Sentimen Masyarakat terhadap Krisis Petani Muda pada Media Sosial Facebook. Penerapan metode K-Fold Cross Validation yaitu (K=10). Analisis sentimen dilakukan dengan 3 label (positif, negatif, dan netral). Data yang dipakai dalam pembuatan model klasifikasi (data hasil preprocessing kolom stemming) menggunakan (Google Colab) berjumlah 4.878 data dengan sentiment Positif 43.13% (2.104), Netral 39.59% (1.931), Negatif 17.28% (843) dari data awal tanpa komentar bersarang yaitu 4.981 dan jumlah data facebook keseluruhan yaitu 2.900 suka, 6.700 komentar, dan 3,3 juta penonton. Hasil akurasi algoritma NBC sebesar 57.32%, algoritma C4.5 sebesar 98.42%, dan algoritma KNN (K=19) sebesar 97.33%. Dapat disimpulkan bahwa hasil perbandingan performa ketiga algoritma menggunakan (Rapidminer10.3), algoritma C4.5 mendapatkan akurasi lebih tinggi sebesar 98.42% dan lebih unggul karena menghasilkan pohon keputusan.

Kata Kunci: NBC; C4.5; KNN; Krisis Petani Muda; Analisis Sentimen

Abstract−In Indonesia, young farmers face various challenges and crises that hinder the growth and sustainability of the agricultural sector. They face obstacles such as lack of access to capital, limited technology, climate change, and low selling prices for their crops. In addition, they also often face problems in obtaining accurate and relevant information in an effort to facilitate better decision-making in agricultural businesses, so that the interest of young people today to become farmers is decreasing. The study aims to Compare the Performance of NBC, C4.5, and KNN Algorithms in the Analysis of Public Sentiment towards the Young Farmer Crisis on Facebook Social Media. The application of the K-Fold Cross Validation method is (K = 10). Sentiment analysis is carried out with 3 labels (positive, negative, and neutral). The data used in making the classification model (data from preprocessing the stemming column) using (Google Colab) amounted to 4,878 data with Positive sentiment of 43.13% (2,104), Neutral 39.59% (1,931), Negative 17.28% (843) from the initial data without nested comments, which is 4,981 and the total number of Facebook data is 2,900 likes, 6,700 comments, and 3.3 million viewers. The accuracy of the NBC algorithm is 57.32%, the C4.5 algorithm is 98.42%, and the KNN algorithm (K = 19) is 97.33%. It can be concluded that the results of the comparison of the performance of the three algorithms using (Rapidminer10.3), the C4.5 algorithm gets a higher accuracy of 98.42% and is superior because it produces a decision tree.

Keywords: NBC; C4.5; KNN; Young Farmer Crisis; Sentiment analysis

1. PENDAHULUAN

Di Indonesia, petani muda menghadapi berbagai tantangan dan krisis yang menghambat pertumbuhan dan keberlanjutan sektor pertanian. Mereka menghadapi kendala seperti kurangnya akses terhadap modal, teknologi yang terbatas, perubahan iklim, dan rendahnya harga jual hasil panen. Selain itu, mereka juga sering kali menghadapi masalah dalam memperoleh informasi yang akurat dan relevan dalam upaya memudahkan pengambilan keputusan lebih baik pada usaha pertanian [1].

Menurut Maihani, dkk. (2021) dan Arvianti, dkk. (2019) untuk keberhasilan pertanian berkelanjutan sangat ditentukan oleh kualitas sumber daya manusianya (SDM) [2][3]. Berdasarkan Kementerian Pertanian, (2021) untuk mendapatkan SDM berkelanjutan di sektor pertanian perlu dilakukan perekrutan para petani muda dikarenakan 97,67% yang bersumber dari Badan Statistik Pusat (BPS) (2022), para petani sekarang umurnya sekitar 50 tahun keatas dengan jumlahnya yang semakin meningkat, sedangkan tenaga kerja berusia muda sebagai petani semakin berkurang [4]. Jika saja angka penghasilan / pendapatan bekerja sebagai tani di tingkatkan, mungkin minat tenaga kerja berusia muda akan semakin meningkat, namun kebanyakan orang tua sekarang menginginkan anak-anaknya mendapatkan pekerjaan dibidang lain dan sebaliknya banyak anak muda sekarang yang tidak mau bekerja melanjutkan pekerjaan orang tuanya sebagai petani karena menginginkan pekerjaan yang bergengsi, mudah dan instan seperti menjadi konten kreator youtube dan sebagainya [2][3].

Sebuah postingan video di halaman Facebook Seputar Gumelar pada tanggal 30 Agustus 2023 yang memperlihatkan para petani sedang mencangkul di sawah, mendapat tanggapan beragam dari warganet. Ada yang setuju dengan pernyataan bahwa Indonesia sudah mulai krisis petani muda, ada juga yang tidak setuju dan ada yang

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

bersikap biasa saja / netral. Hal tersebut perlu dipikirkan untuk keberlanjutan usaha tani di masa yang akan mendatang, karena berdasarkan Kementerian Pertanian, (2021) yang bersumber dari Badan Statistik Pusat (BPS) (2018) menunjukkan bahwa penghasilan utama penduduk Indonesia di 73 ribu desa (87%) berasal dari sektor pertanian dan sebaran penghasilan di sektor pertanian berdasarkan komoditas tertingginya yaitu yang pertama padi dan yang kedua yaitu palawija dengan persentase padi 44% dan persentase palawija 16% [4].

Penelitian ini bertujuan untuk Perbandingan Performa Algoritma NBC, C4.5, dan KNN dalam Analisis Sentimen Masyarakat terhadap Krisis Petani Muda pada Media Sosial Facebook. Analisis sentimen memungkinkan pemahaman yang lebih dalam tentang opini dan pandangan masyarakat [5]. Analisis sentimen pada penelitian ini akan menggunakan tiga label yaitu positif, negatif, dan netral, [6][7]. Menurut Oktavia, dkk. (2023) analisis sentimen mampu mendeteksi emosi dan opini dari seseorang terhadap suatu topik tertentu[8].

Untuk mendapatkan model klasifikasi bisa dilakukan dengan banyak cara, salah satu dari banyaknya cara, kita bisa menggunakan teknik machine learning. Dalam penelitian ini, kita akan menggunakan tiga algoritma klasifikasi data yang bekerja secara relatif dengan cara yang lebih sederhana dibandingkan dengan metode klasifikasi data yang lainnya, adapun algoritmanya yaitu Naïve Bayes Classifier (NBC), C4.5, dan K-Nearest Neighbor (K-NN). Algoritma-algoritma ini dipilih karena memiliki kinerja cukup tinggi dalam klasifikasi data dengan keunggulannya yaitu perhitungan komputasinya sederhana dan mempunyai tingkat akurasi yang baik. Hasil dari ketiga algoritma klasifikasi ini dibandingkan untuk mengetahui mana yang mendapatkan performa terbaik [9].

Algoritma NBC diperkenalkan oleh Thomas Bayes pada tahun 1763 [10]. Algoritma ini adalah metode yang menggunakan statistik sederhana berdasarkan Teorema Bayes untuk menentukan kelas sebuah data observasi dengan menggunakan atribut-atribut yang ada dan prosesnya yang cepat [11]. Algoritma C4.5 diperkenalkan oleh Quinlan tahun (1996) [12]. C4.5 adalah algoritma pembelajaran mesin berbasis pohon keputusan yang menggunakan algoritma ID3 untuk membangun pohon keputusan [13] [14]. Induksi decision tree hanya bisa dilakukan menggunakan fitur bertipe kategorikal (nominal / ordinal), sedangkan fitur bertipe numerik (interval / rasio) tidak bisa digunakan, dengan demikian pembeda Algoritma C4.5 sebagai versi perbaikan dari ID3 yaitu, Algoritma C4.5 dapat menangani fitur bertipe numerik [15]. Algoritma KNN adalah algoritma yang dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing [14]. Algoritma KNN ini dapat memproses data yang besar dengan komputasi kinerja tinggi, kinerja yang baik, dan merupakan algoritma yang sederhana dan mudah dipelajari [11]. Alasan kenapa penelitian pada kali ini memilih 3 metode ini, karena pada penelitian terdahulunya memperoleh akurasi yang tinggi [11][16][9][17][6][7].

Beberapa penelitian terdahulu telah menggunakan Algoritma NBC, C4.5 dan KNN sebagai algoritma untuk klasifikasi dan sentimen, seperti Warraihan, dkk. (2023) dalam Analisis Sentimen Pengguna Transportasi Online Maxim Pada Instagram Menggunakan Naïve Bayes Classifier dan K-Nearest Neighbour, penelitian tersebut menunjukkan KNN menghasilkan akurasi yang baik, yaitu sebesar 84% [11]. Kedua Pramana, dkk. (2023) dalam Analisis Sentimen Terhadap Pemindahan Ibu Kota Negara Menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neightbors, penelitian tersebut menunjukkan NBC menghasilkan akurasi yang baik, yaitu sebesar 69.23% [16]. Ketiga Dina, dkk. (2023) dalam Perbandingan Algoritma NBC, KNN, dan C4. 5 Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan, penelitian tersebut menunjukkan C4.5 menghasilkan akurasi yang baik, yaitu sebesar 80.16% [9]. Keempat Alfandi Safira & Hasan, (2023) dalam analisis sentimen masyarakat terhadap perilaku korupsi pejabat pemerintah pada sosial media Twitter menggunakan Algoritma NBC, penelitian tersebut menunjukkan NBC menghasilkan akurasi yang baik, yaitu sebesar 89.73% [17]. Kelima penelitian yang dilakukan Fersellia, dkk. (2023) dalam analisis sentimen kepuasan pengguna Aplikasi Shopee Food menggunakan C4.5, penelitian tersebut menunjukkan Algoritma C4.5 menghasilkan akurasi yang baik, yaitu sebesar 88% [6]. Keenam penelitian yang dilakukan Amardita, dkk. (2022) dalam analisis sentimen terhadap ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung menggunakan Algoritma KNN, penelitian tersebut menunjukkan KNN menghasilkan akurasi yang baik, yaitu sebesar 88.29% [7].

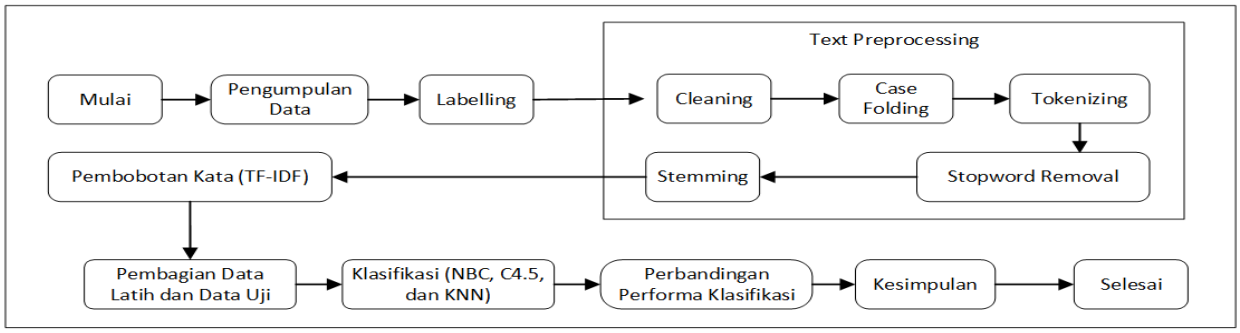
Adapun terdapat perbedaan penelitian ini dan terdahulunya. Pada penelitian kali ini akan menggunakan perbandingan 3 algoritma untuk klasifikasi (NBC, C4.5, dan KNN). Analisis sentimen dilakukan dengan 3 pelabelan yaitu positif, negatif dan netral. Data penelitian diperoleh dari komentar-komentar salah satu postingan halaman Seputar Gumelar pada media sosial facebook.

Berdasarkan latar belakang yang sudah dijelaskan, Penelitian kali ini akan membuat Perbandingan Performa Algoritma NBC, C4.5, dan KNN dalam Analisis Sentimen Masyarakat terhadap Krisis Petani Muda pada Media Sosial Facebook.

2. METODOLOGI PENELITIAN

Ada 8 tahap atau proses yang akan dilakukan pada penilitian kali ini, pertama pengumpulan data, kedua labeling data, ketiga text preprocessing, keempat TF-IDF, kelima pembagian data pelatihan dan data pengujian, keenam pembuatan untuk model klasifikasi tiga algoritma yaitu NBC, C4.5, dan KNN, ketujuh perbandingan performa klasifikasi, terakhir kesimpulan. Tahapan atau proses ditunjukkan di Gambar 1.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082



Gambar 1. Metodologi Penelitian

2.1 Pengumpulan Data

Pengumpulan data / Crawling / Scraping adalah kegiatan mengekstrak bagian yang diperlukan yang diperoleh di internet [18]. Tahap ini dilakukan menggunakan website https://id.exportcomments.com/ (premium $11 / 3 hari). Data yang diambil berupa komentar postingan facebook Seputar Gumelar tanggal 30 agustus 2023 yang berjudul Indonesia Mulai Krisis Petani Muda.

2.2 Labelling

Dataset yang dihasilkan dari Halaman Facebook Seputar Gumelar merupakan dataset yang tidak memiliki label atau masih unsupervised data, sehingga perlu dilakukan pelabelan data agar dataset dapat dipelajari. Pelabelan dapat dilakukan oleh 3 annotator, yaitu seseorang yang bertanggung jawab untuk memberikan label (negatif, netral, dan positif) pada dataset. Pada penelitian kali ini hanya menggunakan 2 orang annotator, 1 orang dibidang Bahasa Indonesia dan 1 orang di bidang Sistem Informasi. Pelabelan penelitian ini dikategorikan jadi 3 (positif, negatif, dan netral) [11]. Menurut Fitri Wulandari, dkk. (2023) kelas positif merupakan komentar yang setuju dengan kebijakan tersebut, label negatif merupakan komentar yang tidak setuju dengan kebijakan tersebut, dan label netral merupakan warganet yang bersikap biasa saja [18].

2.3 Text Preprocessing

Pada tahap Text Preprocessing, tools yang digunakan yaitu google collaboratory dan Bahasa Pemrograman Python. Menurut N. F. Hasan, dkk. (2022) text Preprocessing adalah serangkaian proses untuk membersihan data teks (penghilang dan solusi noise) dilakukan sebelum diproses, supaya mengasilkan perhitungan lebih optimal [19]. Data dari komentar postingan facebook Indonesia mulai krisis petani muda pada text preprocessing penelitian kali ini terbagi 5 tahap proses yaitu [20] :

a. Cleaning, membersihkan dengan cara menghapus atribut yang tidak relevan seperti nama pengguna, jumlah suka, dan tanggal posting. Kami hanya mempertahankan atribut komentar dan labelnya. Selain itu, kami juga menghilangkan noise seperti emoji, angka, dan tanda baca yang tidak diperlukan dalam analisis.
b. Case Folding, membuat teks jadi huruf kecil atau lowercase.
c. Tokenizing, pemisahan kalimat jadi kata per kata.
d. Stopword Removal, penghilang / penghapus kata yang paling sering muncul tetapi dianggap tidak memiliki makna dan tidak terlalu penting.
e. Stemming, pengubah kata ke bentuk dasarnya, penghilang imbuhan yang ada pada kata. Tidak lupa Library Sastrawi diterapkan di stopword serta stemming.

2.4 Pembobotan Kata TF-IDF

Pada tahap Pembobotan Kata (TF-IDF), tools yang digunakan yaitu RapidMiner 10.3. Penggunaan pembobotan kata dengan metode Term Frequency-Inverse Document Frequency (TF-IDF) melibatkan konversi data teks menjadi representasi numerik, memungkinkan pembobotan kata atau fitur secara numerik. TF-IDF berfungsi sebagai alat statistik yang menilai signifikansi sebuah kata dalam suatu dokumen. TF mencerminkan seberapa sering sebuah kata muncul dalam dokumen tertentu, menunjukkan tingkat kepentingan kata tersebut dalam konteks dokumen tersebut. DF mencerminkan seberapa umum kata tersebut dalam kumpulan dokumen. IDF, sebaliknya, merupakan nilai kebalikan dari DF. Hasil pembobotan kata dengan menggunakan metode TF-IDF diperoleh melalui perkalian dari nilai TF dan IDF. Sesudah text preprocessing, dataset perlu dikonversi ke bentuk numerik untuk ke tahap berikutnya yaitu klasifikasi. Operator yang digunakan pada rapidminer 10.3 yaitu Text Vextorization dan Generate TFIDF. Dalam rumus TF-IDF, perhitungan bobot (W) untuk setiap dokumen terhadap kata kunci dijelaskan sebagai berikut [21] :

$$W_{dt} = TF_{dt} IDF_{ft} \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

W_{dt} adalah nilai dokumen ke-d pada kata ke-t, TF_{dt} adalah jumlah kata yang dicari dalam suatu dokumen, IDF_{ft} adalah Inverse Document *Frequency (log log (*$\frac{N}{df}$*))*, N adalah jumlah dokumen, dan Df adalah jumlah dokumen yang mengandung kata yang dicari.

2.5 Pembagian Data Pelatihan dan Pengujian

Untuk pengukuran sampai mana model bisa memprediksi data yang belum pernah terlihat sebelumnya, penerapan metode K-Fold Cross Validation yaitu 10 fold. Dengan menggunakan RapidMiner 10.3, data secara otomatis dibagi menjadi beberapa bagian untuk pelatihan dan pengujian. Hal ini memungkinkan untuk memperoleh hasil kinerja model yang lebih akurat [9].

2.6 Klasifikasi NBC, C4.5, dan KNN

Pada Tahap Klasifikasi NBC, C4.5, dan KNN, tools yang akan digunakan adalah RapidMiner 10.3. Alasan kenapa penelitian pada kali ini memilih 3 metode ini, karena pada penelitian terdahulunya memperoleh akurasi yang tinggi [11][16][9][17][6][7].

a. Naïve Bayes Classifier.
   Algoritma ini digunakan untuk melakukan klasifikasi dengan asumsi bahwa fitur-fitur dalam data bersifat independen satu sama lain. Parameter Laplace correction digunakan untuk mengatasi masalah ketika ada kelas atau fitur yang tidak muncul dalam data pelatihan / nilai probabilitas 0 (nol). Operator yang digunakan Naïve Bayes, apply model, dan performance [9].
b. C4.5.
   Algoritma ini membangun pohon keputusan untuk melakukan klasifikasi. Parameter pruning dan preprunning digunakan untuk memangkas pohon keputusan agar dapat memberikan hasil akurasi yang lebih baik. Operator yang digunakan Decision Tree, apply model, dan performance [9].
c. K-Nearest Neighbor
   Algoritma ini mengklasifikasi suatu data terbaru dari K data terdekatnya. 10 percobaan nilai K (3, 5, 7, 9, 11, 13, 15, 17, 19, dan 21). Dicoba berbagai nilai K untuk mencari nilai terbaik. Penggunaan Mixed Euclidean Distance merupakan metode perhitungan jarak yang diperuntukkan pada data yang mempunyai kombinasi atribut numerik dan kategorikal. Operator yang digunakan: K-NN, apply model, dan performance [9].

2.7 Perbandingan Performa Klasifikasi

Untuk menilai kinerja model klasifikasi NBC, C4.5, dan K-NN, penelitian akan menggunakan confusion matriks untuk menghitung metrik evaluasi seperti akurasi, presisi, dan recall. Semua perhitungan ini dilakukan menggunakan perangkat lunak RapidMiner 10.3. Dikhususkan algoritma K-NN, eksperimen dilakukan dengan parameter 10 nilai K untuk menentukan nilai optimal. Hasil evaluasi ini kemudian dibandingkan untuk mengetahui dari 3 model algoritma (NBC, C4.5, dan KNN) mana yang memberikan hasil paling bagus. Rumus penghitung akurasi ditunjukkan di persamaan 1. Rumus penghitung presisi ditunjukkan di persamaan 2. Rumus penghitung recall ditunjukkan di persamaan 3 [9].

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \times 100\% \quad (2)$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \quad (3)$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengumpulan Data

Data Komentar tgl 31/08/2024 mencapai 6700 komentar. Data yang di ambil hanya komentar utama saja menggunakan web https://id.exportcomments.com/ (premium $11 / 3 hari) pada tanggal 26/07/2024 jam 13:06 WIB berjumlah 4999 tanpa komentar bersarang. Sampel dataset ditunjukkan di Tabel 1.

Tabel 1. Sampel Dataset Awal

No	Username	Profile ID	Date	Likes	Stars	Comment	(view source)
1	Unasih Unasih	100072090315137	30/08/23 02:54:20	1		Semoga selalu sehat selalu selamat sukses usaha nya selalu berkah rizkinya berlimpah jadi nilai ibadah mendapatkan ridho Allah SWT masuk surga tanpa hisab bersama keluarga 🤲🤲👍👍	view comment

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

2	Memed Memed	100024506482190	30/08/23 03:50:28	9		Gimana gak krisis petani muda . Harga pupuk mahal. Petani bagi hasil. Karena sekarang banyak tuan tanah kembali. Penghasilan tdk sesuai dgn tenaga yg dikeluarkan . Pergi ke kota. Lapangan kerja sulit. Coba mengeruk kekayaan alam itu utk membuat lapangan kerja .	view comment
.....	.....	.....	.....	.....	.....	.....	.....
4.999	Sa'dun	100005629868435	12/11/23 02:23:18			Lokasine syahdu	view comment

Berdasarkan sampel data pada tabel 1 yang telah diperoleh, langkah selanjutnya adalah proses pelabelan pada kolom comment untuk mengetahui komentar mana yang memiliki sentiment positif, negatif dan netral.

3.2 Hasil Labeling

Pelabelan dilakukan oleh 2 orang annotator ditunjukkan di Tabel 2.

Tabel 2. Labeling

No	Username	Comment	Sentiment
1	Unasih Unasih	Semoga selalu sehat selalu selamat sukses usaha nya selalu berkah rizkinya berlimpah jadi nilai ibadah mendapatkan ridho Allah SWT masuk surga tanpa hisab bersama keluarga 👨‍👩‍👧‍👧👍👍	Positif
2	Memed Memed	Gimana gak krisis petani muda . Harga pupuk mahal. Petani bagi hasil. Karena sekarang banyak tuan tanah kembali. Penghasilan tdk sesuai dgn tenaga yg dikeluarkan . Pergi ke kota. Lapangan kerja sulit. Coba mengeruk kekayaan alam itu utk membuat lapangan kerja .	Negatif
.....	.....	.....	.....
4.999	Sa'dun	Lokasine syahdu	Netral

Berdasarkan sampel data pada tabel 2 hasil pelabelan yang telah dilakukan, jumlah data keseluruhan yaitu 4.999 data, dan data berdasarkan sentiment yaitu Positif 43.13% (2104), Netral 39.59% (1931), dan Negatif 17.28% (843). Langkah selanjutnya adalah pembersihan data (text Preprocessing) pada kolom comment.

3.3 Text Preprocessing

Tools yang digunakan yaitu google collaboratory dan Bahasa Pemrograman Python. Hasil Text Preprocessing penelitian ini ditunjukkan di Tabel 3.

Tabel 3. Text Preprocessing

No	Cleaning	Case Folding	Tokenizing	Stopword	Stemming
1	Semoga sehat selalu bahagia selalu sukses selalu usahanya selalu dilimpahkan rezekinya...	semoga sehat selalu bahagia selalu sukses selalu usahanya selalu dilimpahkan rezekinya...	['semoga', 'sehat', 'selalu', 'bahagia', 'selalu', 'sukses', 'selalu', 'usahanya', 'selalu', 'dilimpahkan', 'rezekinya', ...	['semoga', 'sehat', 'bahagia', 'sukses', 'usahanya', 'dilimpahkan', 'rezekinya', ...	['moga', 'sehat', 'bahagia', 'sukses', 'usaha', 'limpah', 'rezeki', ...
2	Bagaimana dengan krisis petani muda Harga pupuk mahal...	bagaimana dengan krisis petani muda harga pupuk mahal...	['bagaimana', 'dengan', 'krisis', 'petani', 'muda', 'harga', 'pupuk', 'mahal', ...	['krisis', 'petani', 'muda', 'harga', 'pupuk', 'mahal', ...	['krisis', 'tani', 'muda', 'harga', 'pupuk', 'mahal', ...
.....	.....	.....	.....	.....	.....
4.909	Lokasinya luar biasa	lokasinya luar biasa	['lokasinya', 'luar', 'biasa']	['lokasinya']	['lokasi']

Setelah dilakukan Text Preprocessing pada Tabel 3, wajib dilakukan pengecekan manual data yang terdapat nois pada kolom Stemming. Jumlah data awal berjumlah 4.999, setelah text preprocessing menjadi 4.909 data, dan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

data bersih stemming setelah di hapus noisnya menjadi 4.878 data. Setelah datanya bersih, langkah selanjutnya adalah TF-IDF menggunakan tools rapidminer10.3.

3.4 Pembobotan TF-IDF

Operator yang digunakan pada rapidminer 10.3 yaitu Text Vextorization dan Generate TFIDF. Pada operator Text Vextorization, atribut filter type pilih single, atribute pilih komentar, centang include special attributes, centang store training documents, centang store scoring document, document class attribute pilih sentiment, centang apply pruning (sesuaikan dengan jumlah data bersih), dan pada operator Generate TFIDF, centang calculate term frequencies. Tabel sampel TF mencerminkan seberapa sering sebuah kata muncul di dokumen tertentu, menunjukkan tingkat kepentingan kata tersebut pada konteks dokumen tersebut. DF mencerminkan seberapa umum kata tersebut dalam kumpulan dokumen. IDF, sebaliknya, merupakan nilai kebalikannya DF. Hasil pembobotan kata dengan memakai metode TF-IDF diperoleh melalui perkalian nilai TF dan IDF menggunakan tools rapidminer 10.3 ditunjukkan di Tabel 4 dan 5.

Tabel 4. Daftar TF-IDF Berdasarkan Sentiment

No	Token	Negatif	Netral	Positif
1	Tani	452	948	1271
2	Muda	84	231	835
.....	.....	.....	.....	.....
3.272	Zoonnkk	0	1	0

Berdasarkan sampel data pada tabel 4, jumlah kata yang sering muncul berjumlah 3.272 kata, ada 2 kata yang sering muncul yaitu kata Tani dan kata Muda dengan sentiment tertinggi yaitu positif pada kata (Tani) berjumlah 1271 kemunculan kata dengan sentiment positif dan kata (Muda) berjumlah 835 kemunculan kata dengan sentiment positif.

Tabel 5. Hasil Pembobotan TF-IDF

No	**abad**	**Abah**	**Abai**	**Abang**	**Abdi**	**…**	**Zoldyck**	**zong**	**Zongjadi**	**Zonk**	**Zoonnkk**
1	0,0	0,0	0,0	0,0	0,0	…	0,0	0,0	0,0	0,0	0,0
2	0,0	0,0	0,0	0,0	0,0	…	0,0	0,0	0,0	0,0	0,0
.....	…	…	…	…	…	…	…	…	…	…	…
4.878	0,0	0,0	0,0	0,0	0,0	…	0,0	0,0	0,0	0,0	0,0

Berdasarkan sampel data pada tabel 5, jumlah hasil perkalian antara TF dan IDF yang merupakan data bersih stemming tanpa nois berjumlah 4.878 data. Karena pada sampel data tidak ada nilai yang terlihat, contoh kemunculan kata pada kata abad terdapat di nomor 1674 dengan nilai hasil perkalian TF-IDF yaitu 1,1, dan untuk kata abah terdapat di nomor 1980 dengan nilai hasil perkalian TF-IDF yaitu 0,8.

3.5 Pembagian Data Pelatihan dan Pengujian

Adapun pembagian data ini dilakukan terhadap hasil stemming yang telah diperoleh. Untuk pengukuran sampai mana model bisa memprediksi data yang belum pernah terlihat sebelumnya, penerapan metode K-Fold Cross Validation yaitu 10 fold. Dengan menggunakan RapidMiner 10.3, data secara otomatis dibagi menjadi beberapa bagian untuk pelatihan dan pengujian. Hal ini memungkinkan untuk memperoleh hasil kinerja model yang lebih akurat.

3.6 Klasifikasi Algoritma NBC, C4.5, dan K-NN

Hasil Klasifikasi bertujuan untuk mengetahui performa pada Algoritma NBC, C4.5, dan K-NN.

3.6.1 Klasifikasi Algoritma Naïve Bayes Classifier

Berdasarkan hasil klasifikasi, algoritma NBC mendapatkan akurasi 57.32%, recall 63.90%, dan precision 66.14%. Berdasarkan hasil klasifikasi, performa algoritma NBC kurang baik, hasil ditunjukkan di Tabel 6.

Tabel 6. Hasil Performa Klasifikasi Naïve Bayes Classifier

accuracy: 57.32%
weighted_mean_recall: 63.90%
weighted_mean_precision: 66.14%

		Aktual			Precision
		Positif	Negatif	Netral	
Prediksi	**Positif**	648	98	186	69.53%
	Negatif	1449	743	340	29.34%
	Netral	7	2	1405	99.36%
Recall		30.80%	88.14%	72.76%	

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

3.6.2 Klasifikasi Algoritma C4.5

Berdasarkan hasil klasifikasi, algoritma C4.5 mendapatkan akurasi 98.42%, recall 97.79%, dan precision 97.95%. Berdasarkan hasil klasifikasi, performa algoritma C4.5 sangat baik, hasil ditunjukkan di Tabel 7.

Tabel 7. Hasil Performa Klasifikasi C4.5

accuracy: 98.42%
weighted_mean_recall: 97.79%
weighted_mean_precision: 97.95%

		Aktual			Precision
		Positif	Negatif	Netral	
Prediksi	**Positif**	2069	41	0	98.06%
	Negatif	35	801	0	95.81%
	Netral	0	1	1931	99.95%
Recall		98.34%	95.02%	100.00%	

3.6.3 Klasifikasi Algoritma K-Nearest Neighbor

Berdasarkan hasil klasifikasi, untuk algoritma KNN dengan nilai K percobaan yaitu K19 mendapatkan akurasi 97.33%, recall 97.52%, dan precision 96.47%. Berdasarkan hasil klasifikasi, performa algoritma KNN sangat baik, hasil ditunjukkan di Tabel 8 dan Tabel 9.

Tabel 8. Daftar 10 Nilai K Percobaan

	K = 3	K = 5	K = 7	K = 9	K = 11
Akurasi	90.49%	91.88%	93.81%	94.90%	95.86%
Recall	91.53%	92.69%	94.55%	95.49%	96.30%
Presisi	92.39%	93.03%	94.10%	94.73%	95.44%
	K = 13	K = 15	K = 17	**K = 19**	K = 21
Akurasi	96.27%	96.58%	97.09%	**97.33%**	96.25%
Recall	96.65%	96.86%	97.31%	**97.52%**	96.61%
Presisi	95.67%	95.93%	96.34%	**96.47%**	95.82%

Tabel 9. Hasil Performa Klasifikasi K-Nearest Neighbor

accuracy: 97.33%
weighted_mean_recall: 97.52%
weighted_mean_precision: 96.47%

		Aktual			Precision
		Positif	Negatif	Netral	
Prediksi	**Positif**	2033	14	44	97.23%
	Negatif	71	828	0	92.10%
	Netral	0	1	1887	99.95%
Recall		96.63%	98.22%	97.72%	

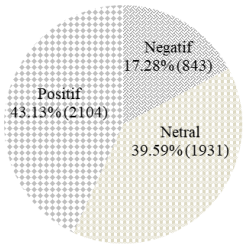
3.7 Visualisasi Perbandingan Performa Klasifikasi

Visualisasi bertujuan untuk mempermudah dalam memahami hasil dari semua proses yang telah dilakukan menggunakan tools RapidMiner 10.3 dan Microsoft Excel 2010.

3.7.1 Visualisasi Data

Visualisasi data bersih berdasarkan sentiment yaitu Positif 43.13% (2104), Netral 39.59% (1931), Negatif 17.28% (843). Visualisasi ditunjukkan di Gambar 2.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

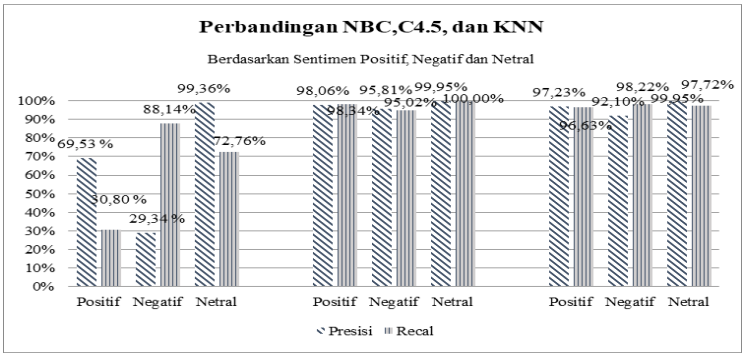


Gambar 2. Visualisasi PIE

Dari visualisasi pada gambar 2, dapat dikatakan bahwa komentar dengan sentiment terbanyak mengenai krisis petani muda yaitu sentiment positif.

3.7.2 Visualisasi Perbandingan Berdasarkan Sentiment

Perbandingan NBC, C4.5, dan KNN berdasarkan sentiment. Algoritma C4.5 mendapatkan Nilai Sentiment Netral tertinggi dengan nilai Presisi 99.95% dan Recall 100%, namun jika membandingkan Positif dan Negatifnya, Sentimen Positif dengan nilai Presisi 98.06% dan Recall 98.34% lebih tinggi dari pada Sentimen Negatif mengenai krisis petani muda. Visualisasi ditunjukkan di Gambar 3.

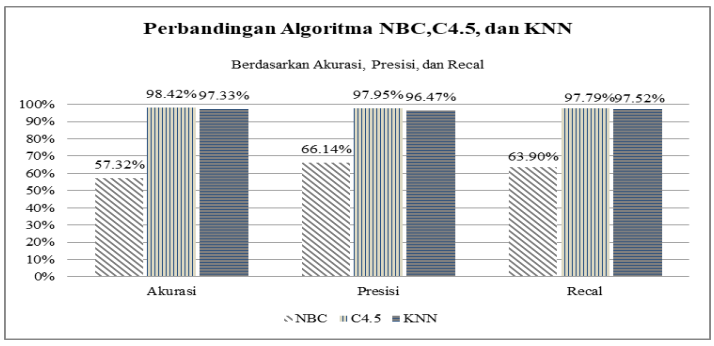


Gambar 3. Hasil Perbandingan Sentimen

Dari visualisasi gambar 3, visualisasi diagram batang yaitu positif, negatif, dan netral bagian pertama milik algoritma NBC, bagian kedua adalah algoritma C4.5, bagian ketiga adalah algoritma KNN. Berdasarkan Presisi dan Recall tertinggi yaitu algoritma C4.5 pada bagian diagram tengah / diagram bagian dua.

3.7.3 Visualisasi Perbandingan Berdasarkan Performa Algoritma

Berdasarkan Visualisasi Hasil Perbandingan Performa Algoritma, Algoritma C4.5 dan KNN mendapatkan performa yang sangat baik, namun algoritma dengan performa tertinggi diperoleh algoritma C4.5 dengan Akurasi 98.42%, Presisi 97.95%, dan Recall 97.79%, dan algoritma dengan performa terendah yaitu algoritma NBC dengan Akurasi 57.32% Visualisasi ditunjukkan di Gambar 4.



Gambar 4. Hasil Perbandingan Performa Algoritma

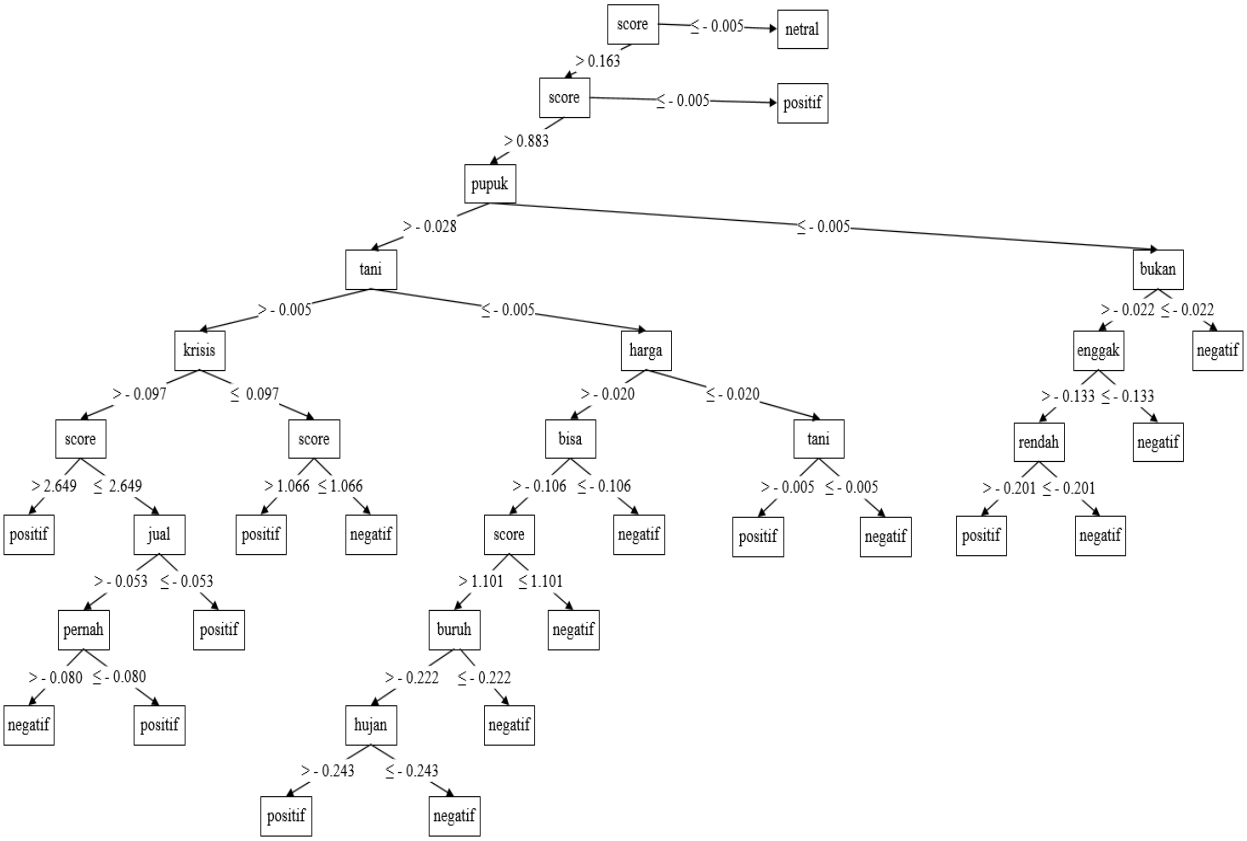
Dari visualisasi pada gambar 4, dapat dikatakan bahwa algoritma yang cocok dan algoritma terbaik dalam pengelolahan pada data komentar facebook mengenai krisis petani muda yaitu algoritma C4.5, dengan performa Akurasi tertinggi mencapai 98.42%, Presisi 97.95%, dan Recall 97.79%.

3.7.4 Visualisasi Pohon Keputusan Algoritma C4.5

Komentar adalah kata kunci atau frasa yang dicari dalam teks. Misalnya, "pupuk", "tani", "krisis", dan sebagainya. Teks ini seperti sebuah pohon terbalik, di mana setiap cabang mewakili kondisi yang berbeda. Misalnya: Cabang

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

Utama: Jika "score > 0.883", maka sentimennya positif. Cabang Cabang: Jika "score > 0.883" dan "komentar:pupuk > -0.026", maka kita akan melihat cabang yang lebih spesifik lagi, dan seterusnya. Kesimpulan Sederhana: Kebanyakan teks memiliki sentimen positif. Sentimen positif seringkali dikaitkan dengan kata-kata seperti "hujan", "rendah" (dalam konteks tertentu), dan kombinasi kata-kata tertentu lainnya. Sentimen negatif sering dikaitkan dengan kata-kata seperti "krisis", "jual", dan "buruh" dalam konteks tertentu. Faktor-faktor yang mempengaruhi sentimen kata-kata seperti "pupuk", "hujan", dan "krisis" memiliki pengaruh yang signifikan terhadap sentimen positif atau negatif. Tanpa konteks yang lebih lengkap, sulit untuk memberikan interpretasi yang sangat akurat. Kata-kata seperti "rendah" bisa memiliki konteks positif (misalnya, harga rendah) atau negatif (misalnya, produksi rendah). Pohon Keputusan ditunjukkan di Gambar 5.



Gambar 5. Pohon Keputusan

4. KESIMPULAN

Berdasarkan hasil pengelolahan data yang telah dilakukan menggunakan tools google collaboratory bahasa pemrograman python sebagai pengelolahan text preprocessing, rapidminer 10.3 sebagai pengelolahan klasifikasi ketiga algoritma NBC, C4.5 dan KNN, serta microsoft excel 2010 sebagai pembenahan data dan visualisasi data. Jumlah data facebook keseluruhan yaitu 2.900 suka, 6.700 komentar, dan 3,3 juta penonton Jumlah data awal tanpa komentar bersarang berjumlah 4.999, setelah text preprocessing menjadi 4.909 data, dan data bersih stemming setelah di hapus noisnya menjadi 4.878 data dengan sentiment Positif 43.13% (2.104), Netral 39.59% (1.931), dan Negatif 17.28% (843). Pembagian data pelatihan dan data pengujian menerapkan metode K-Fold Cross Validation yaitu (K=10). Akurasi algoritma NBC sebesar 57.32%, algoritma C4.5 sebesar 98.42%, dan algoritma KNN (K=19) sebesar 97.33%. Dapat disimpulkan bahwa hasil Perbandingan Performa Algoritma NBC, C4.5, dan KNN dalam Analisis Sentimen Masyarakat terhadap Krisis Petani Muda pada Media Sosial Facebook dengan hasil akurasi tertinggi yaitu algoritma C4.5 mendapatkan akurasi sebesar 98.42%, lebih baik dari algoritma KNN karena akurasinya sebesar 97.33% lebih rendah dari algoritma C4.5, dan lebih baik dari algoritma NBC karena akurasinya hanya sebesar 57.32% lebih rendah dari algoritma C4.5 dan KNN. Selain akurasinya yang tinggi pada algoritma C4.5, adapun terdapat keunggulannya yaitu berupa pohon keputusan dengan akar permasalahan berdasarkan hasil sentiment positif yaitu pupuk, jadi kendala utama yang dialami para petani muda yaitu pupuk.

REFERENCES

[1] R. Saleh, I. Oktafiani, and M. Y. Sitohang, “Sulitnya Regenerasi Petani pada Kelompok Generasi Muda,” *J. Stud. Pemuda*, vol. 10, no. 1, p. 1, 2021, doi: 10.22146/studipemudaugm.62533.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1383−1392
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6082

[2] S. Maihani, M. Jamilah, S. Ahmad, and Z. Yamani, "Jurnal Sains Pertanian Krisis tenaga kerja pertanian ' petani muda ' masa depan Future ' young farmers ' agricultural labor crisis," vol. 4, no. 2, pp. 85–91, 2021.
[3] E. Y. Arvianti, M. Masyhuri, L. R. Waluyati, and D. H. Darwanto, "Gambaran Krisis Petani Muda Indonesia," *Agriekonomika*, vol. 8, no. 2, pp. 168–180, 2019, doi: 10.21107/agriekonomika.v8i2.5429.
[4] Kementerian Pertanian, "Rencana Strategis Kementerian Pertanian Tahun 2020-2024," *Salinan Keputusan Menteri Pertan. Republik Indones.*, pp. 1–161, 2021.
[5] M. R. Firdaus, F. M. Rizki, F. M. Gaus, and I. K. Susanto, "Analisis Sentimen Dan Topic Modelling Dalam Aplikasi Ruangguru," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 4, no. 1, p. 66, 2020, doi: 10.30645/j-sakti.v4i1.188.
[6] F. Fersellia, E. Utami, and A. Yaqin, "Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method," *Sinkron*, vol. 8, no. 3, pp. 1554–1563, 2023, doi: 10.33395/sinkron.v8i3.12531.
[7] R. S. Amardita, A. Adiwijaya, and M. D. Purbolaksono, "Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 1, p. 62, 2022, doi: 10.30865/jurikom.v9i1.3793.
[8] D. Oktavia, Y. R. Ramadahan, and M. Minarto, "Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 407–417, 2023, doi: 10.30865/klik.v4i1.1040.
[9] A. Dina, I. Permana, F. Muttakin, and ..., "Perbandingan Algoritma NBC, KNN, dan C4. 5 Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan," *J. Media …*, vol. 7, no. 3, pp. 1079–1087, 2023, doi: 10.30865/mib.v7i3.6316.
[10] S. Mulyani and R. Novita, "Implementation of the Naive Bayes Classifier Algorithm for Classification of Community Sentiment About Depression on Youtube," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1355–1361, 2022, doi: 10.20884/1.jutif.2022.3.5.374.
[11] D. A. Warraihan, I. Permana, Mustakim, R. Novita, M. Afdal, and A. Marsal, "Analisis Sentimen Pengguna Transportasi Online Maxim Pada Instagram Menggunakan Naïve Bayes Classifier dan K-Nearest Neighbour," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1134–1143, 2023, doi: 10.30865/mib.v7i3.6336.
[12] M. Chair, Y. N. Nasution, and N. A. Rizki, "Aplikasi Klasifikasi Algoritma C4.5 (Studi Kasus Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2008)," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 12, no. 1, p. 50, 2017, doi: 10.30872/jim.v12i1.223.
[13] D. Astri Nawangnugraeni *et al.*, "' Jurnal Teknologi Informasi dan Komunikasi C4.5 Algorithm Implementation for Public Sentyment Analysis Covid-19 Vaccine," pp. 151–160, 2022, [Online]. Available: https://doi.org/10.31849/digitalzone.v13i2.11658
[14] F. N. Hasan, N. Hikmah, and D. Y. Utami, "Perbandingan Algoritma C4.5, KNN, dan Naive Bayes untuk Penentuan Model Klasifikasi Penanggung jawab BSI Entrepreneur Center," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 169, 2018, doi: 10.33480/pilar.v14i2.908.
[15] B. G. Gerardo, S. Saifullah, and E. Irawan, "Teknik Data Mining Dalam Penilaian Pengajaran Guru Berdasarkan Indeks Kepuasan Siswa," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 508–514, 2019, doi: 10.30865/komik.v3i1.1634.
[16] D. Pramana, M. Afdal, M. Mustakim, and I. Permana, "Analisis Sentimen Terhadap Pemindahan Ibu Kota Negara Menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neightbors," *J. Media Inform. Budidarma*, vol. 7, no. 3, pp. 1306–1314, 2023, doi: 10.30865/mib.v7i3.6523.
[17] Alfandi Safira and F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *Zo. J. Sist. Inf.*, vol. 5, no. 1, pp. 59–70, 2023, doi: 10.31849/zn.v5i1.12856.
[18] Fitri Wulandari, Elin Haerani, Muhammad Fikry, and Elvia Budianita, "Analisis sentimen larangan penggunaan obat sirup menggunakan algoritma naive bayes classifier," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 1, pp. 88–96, 2023, doi: 10.37859/coscitech.v4i1.4781.
[19] N. F. Hasan, A. Aisyah, R. Rahman, and H. Wonda, "SeHasan, Nur Fitrianingsih, Aisyah Aisyah, Rahman Rahman, and Herlin Wonda. 2022. 'Sentiment Analysis of Public Opinion Regarding Papuan Local Languages Condition Using Data Science Approach.' Digital Zone: Jurnal Teknologi Informasi dan Komunikasi 13(2 S," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 2 SE-Articles, pp. 125–139, 2022, [Online]. Available: http://journal.unilak.ac.id/index.php/dz/article/view/11545
[20] A. ELHAN, M. K. D. HARDHIENATA, H. YENI, S. WIJAYA HARTONO, and J. ADISANTOSO, "Analisis Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia menggunakan Algoritme Random Forest dan BERT Sentiment Analysis of Twitter Users on COVID-19 Vaccines in Indonesia using Random Forest and BERT Algorithms," *J. Ilmu Komput. Agri-informatika*, vol. 9, no. 2, pp. 199–211, 2022, [Online]. Available: https://jurnal.ipb.ac.id/index.php/jika/article/view/44459
[21] I. Yunanto and S. Yulianto, "Twitter Sentiment Analysis Pedulilindungi Application Using Naïve Bayes and Support Vector Machine," *J. Tek. Inform.*, vol. 3, no. 4, pp. 807–814, 2022, doi: 10.20884/1.jutif.2022.3.4.292.