Methotika : Jurnal Ilmiah Teknik Informatika
E-ISSN : 2776 – 5792
Vol. 6, No. 1, April 2026, pp. 135-144
https://ejurnal.methodist.ac.id/index.php/methotika/

Analisis Sentimen Tanggapan Masyarakat di Media Sosial X terhadap Piala Asia U-23 AFC 2024 Menggunakan Metode DBSCAN dan Naive Bayes

Thomy Franklin Panjaitan[1], Sri Agustina Rumapea[2], Arina Prima Silalahi[3]
[123]Fakultas Ilmu Komputer, Universitas Methodist Indonesia

Info Artikel

History Artikel:

Received, Feb 21 , 2026
Revised, Mar 17, 2026
Accepted, April 08, 2026

Keywords:

Piala Asia U-23 AFC 2024,
X (*Twitter*),
DBSCAN,
Text Mining,
Analisis Sentimen,
Naïve Bayes TF-IDF.

ABSTRAK

Percakapan di X/Twitter tentang Piala Asia U-23 AFC 2024 berlangsung cepat, heterogen, dan sarat noise, sehingga membutuhkan pendekatan komputasional untuk memetakan topik dan membaca sentimen publik secara andal. Penelitian ini bertujuan menyusun pipeline penambangan teks yang mengombinasikan DBSCAN untuk klasterisasi topik dan Naïve Bayes (NB) untuk klasifikasi sentimen. Data dikumpulkan berbasis kata kunci/hashtag relevan, kemudian dipraproses (pembersihan, normalisasi, stopword removal, stemming), direpresentasikan dengan TF-IDF, diklaster menggunakan DBSCAN (metrik cosine, parameter utama: ε = 11,5; min_samples = 10), lalu diklasifikasikan sentimennya (positif/Netral/negatif) dengan NB. Hasil menunjukkan DBSCAN efektif memisahkan tema percakapan, dengan ±84,65% tweet teridentifikasi sebagai core points dan 2,88% sebagai border points; sisanya noise. Distribusi sentimen didominasi netral, disusul positif, sedangkan negatif relatif kecil. Model NB mencapai akurasi sekitar 60%; performa kelas negatif masih menantang akibat code-mixing, slang, ketidakseimbangan kelas, dan nuansa sarkasme. Disimpulkan bahwa kombinasi "DBSCAN→peta topik" dan "NB→label sentimen" layak sebagai baseline pemantauan opini publik berbasis media sosial, sembari membuka arah peningkatan melalui perluasan leksikon slang, penanganan code-mixing, penyeimbangan kelas, dan eksplorasi model berbasis konteks.

Penulis Koresponden:

Thomy Franklin Panjaitan
Fakultas Ilmu Komputer,
Universitas Methodist Indonesia, Medan,
Jl. Hang Tuah No.8, Medan - Sumatera Utara.
Email: thomypanjaitan5570@gmail.com

1. PENDAHULUAN

Media sosial, khususnya X/Twitter, telah menjadi salah satu sumber utama untuk mengekspresikan opini publik terhadap isu-isu aktual. Karakteristik data yang bersifat singkat, rawan noise, dan sering mengandung campuran bahasa (code-mixing) menuntut penerapan pendekatan text mining yang sistematis untuk mengekstraksi topik dan menentukan polaritas sentimen secara akurat. Beberapa penelitian terdahulu telah mengembangkan model analisis sentimen dan klasterisasi topik menggunakan data Twitter berbahasa Indonesia.

Pendekatan supervised learning dengan Naive Bayes (NB) efektif menggambarkan kecenderungan opini (positif, negatif, netral). Sementara penerapan metode unsupervised seperti klasterisasi dapat membantu merapikan heterogenitas topik sebelum proses klasifikasi. yang menggunakan algoritma DBSCAN untuk klasterisasi percakapan publik di Twitter dan melaporkan hasil Silhouette Score yang lebih tinggi dibanding K-Means, menunjukkan keunggulan DBSCAN dalam mengidentifikasi outlier serta menangani kepadatan data yang tidak seragam[1].

Pada sisi analisis sentimen, beberapa studi menunjukkan bahwa Naïve Bayes memiliki kinerja stabil dan efisien pada data berbahasa Indonesia setelah melalui tahapan praproses dan pembobotan TF-IDF. NB mampu memberikan akurasi tinggi, dan performanya dapat ditingkatkan dengan penerapan seleksi fitur, seperti metode Chi-Square. Pendekatan hibrida yang menggabungkan pemetaan topik dan klasifikasi sentimen untuk memperoleh gambaran lebih menyeluruh terhadap opini publik di media sosial[2].

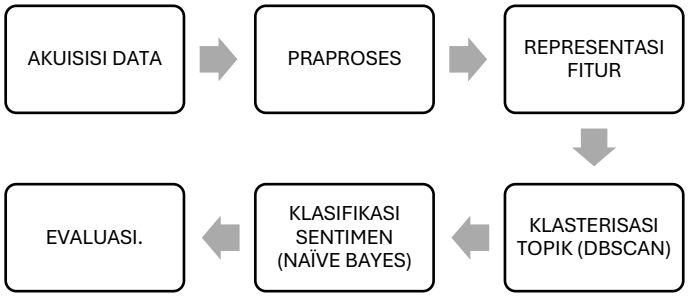
Penelitian pada aplikasi Satu Sehat yang menggunakan algoritma Naïve Bayes (NB) dan Support Vector Machine (SVM) menunjukkan bahwa SVM memperoleh akurasi sedikit lebih tinggi dibanding NB, meskipun keduanya mampu melakukan klasifikasi sentimen dengan baik. Mayoritas tweet yang dianalisis memperlihatkan sentimen positif, yang mengindikasikan penerimaan publik terhadap aplikasi tersebut. Selain itu, temuan tersebut menegaskan bahwa pendekatan machine learning berbasis TF-IDF efektif dalam menilai persepsi publik di media sosial serta memperkuat posisi NB dan SVM sebagai model pembanding yang relevan dalam analisis sentimen Twitter berbahasa Indonesia[3].

Secara keseluruhan, penelitian-penelitian terdahulu tersebut menunjukkan bahwa kombinasi DBSCAN untuk klasterisasi topik dan Naïve Bayes untuk klasifikasi sentimen memiliki potensi tinggi untuk diterapkan dalam konteks percakapan publik yang kompleks, seperti topik olahraga nasional. Namun, kajian terapan yang secara khusus mengombinasikan kedua pendekatan tersebut pada domain spesifik, seperti wacana tentang Piala Asia U-23 AFC 2024, masih terbatas. Oleh karena itu, penelitian ini mengisi celah tersebut dengan merancang pipeline yang memadukan klasterisasi topik menggunakan DBSCAN dan klasifikasi sentimen menggunakan NB untuk menganalisis percakapan publik di X/Twitter.

2. METODE PENELITIAN

2.1. Framework Penelitian

Penelitian ini merancang pipeline text mining untuk memetakan topik percakapan di X/Twitter tentang Piala Asia U-23 AFC 2024 dan melabeli polaritas sentimennya. Alur ringkasnya:



Gambar 1. Framwork Penelitian

1. kuisisi Data. Mengambil tweet sesuai kata kunci/hashtag dan rentang waktu turnamen. Hanya bahasa Indonesia/ID–EN (campur kode) sesuai fokus studi, serta menghapus duplikasi/retweet.
2. Praproses Teks. Pembersihan URL, mention, emoji; normalisasi slang/varian ejaan; penanganan code-mixing ID–EN; stopword removal; dan stemming bahasa Indonesia (mis. Sastrawi) untuk menurunkan variasi bentuk kata dan menstabilkan fitur. Studi di Indonesia menunjukkan campur-kode di Twitter lazim terjadi, dan pemilihan stemmer berpengaruh nyata pada kualitas fitur. UIR Press Journal +1.
3. Representasi Fitur. Dokumen diproyeksikan ke vektor TF-IDF (unigram–bigram) agar token penting di konteks korpus mendapat bobot proporsional dan siap untuk pemodelan.
4. Klasterisasi Topik (Unsupervised). DBSCAN pada ruang TF-IDF dengan metrik cosine untuk membentuk klaster topik dan menandai outlier/noise. DBSCAN dipilih karena tahan terhadap kepadatan tak seragam dan mampu mengidentifikasi noise eksplisit—terbukti unggul (Silhouette Score lebih tinggi) dibanding K-Means pada data teks Indonesia. Penentuan parameter dilakukan

lewat k-distance plot (menaksir eps) dan aturan praktis min_samples (≥5). E-Journal Universitas Negeri Surabaya.

5. Klasifikasi Sentimen (Supervised). Multinomial Naïve Bayes (NB) melabeli sentimen {positif, netral, negatif} per dokumen/klaster. Literatur Indonesia menunjukkan NB konsisten kuat/efisien pada tweet setelah praproses & TF-IDF; beberapa studi mencatat akurasi tinggi dan peningkatan performa dengan seleksi fitur (Chi-Square). Untuk kontrol, SVM dapat dipakai sebagai pembanding. SciSpace +2 UNDIP E-Journal +2.
6. Evaluasi & Analisis. Validasi silang (k-fold) dengan metrik akurasi, precision, recall, F1, serta confusion matrix. Analisis lanjut menelaah distribusi sentimen per klaster (topik), kata kunci dominan, dan implikasi praktis bagi penyelenggara/klub.

3. HASIL DAN PEMBAHASAN

3.1 Preprocessing Data

Preprocessing bertujuan untuk mempersiapkan teks yang belum terstruktur menjadi teks yang terstruktur sehingga dapat digunakan untuk proses selanjutnya[4], [5].

3.1.1 Case Folding

Case folding merupakan tahap perubahan huruf dari huruf kapital menjadi huruf kecil. Hasil dari penerapannya dapat dilihat pada Tabel 1.

Tabel 1. Hasil Case Folding

No	Komentar	Case Folding
1	@idextratime @bliblidotcom tiba-tiba keinget arhan mulangin rombongan hyung korea u23 di afc u23 (tendangan penalti penentuan) eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama... mana habis ini juga pulang indo garuda callin	@idextratime @bliblidotcom tiba-tiba keinget arhan mulangin rombongan hyung korea u23 di afc u23 (tendangan penalti penentuan) eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama... mana habis ini juga pulang indo garuda callin

3.1.2 Cleaning

Cleaning adalah proses menghilangkan tanda baca dan kata-kata yang tidak perlu. Hasil dari penerapannya dapat dilihat pada Tabel 2

Tabel 2. Hasil Cleaning

No	Case Folding	Cleaning
1	@idextratime @bliblidotcom tiba-tiba keinget arhan mulangin rombongan hyung korea u23 di afc u23 (tendangan penalti penentuan) eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama... mana habis ini juga pulang indo garuda callin	tiba tiba keinget arhan mulangin rombongan hyung korea u di afc u tendangan penalti penentuan eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama mana habis ini juga pulang indo garuda callin

3.1.3 Stopword/Filtering

Stopword removal merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil parsing deskripsi apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak. Hasil dari penerapannya dapat dilihat pada Tabel 3

Tabel 3. Hasil filtering

No	Cleaning	Filtering
1	tiba tiba keinget arhan mulangin rombongan hyung korea u di afc u tendangan penalti penentuan eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama mana habis ini juga pulang indo garuda callin	tiba tiba keinget arhan mulangin rombongan hyung korea u di afc u tendangan penalti penentuan eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama mana habis ini juga pulang indo garuda callin

3.1.4 Tokenizing

Tokenizing adalah proses memecah dokumen menjadi kumpulan kata. Hasil dari penerapannya dapat dilihat pada Tabel 4

Tabel 4. Hasil Tokenizing

No	Filtering	Tokenizing
1	tiba tiba keinget arhan mulangin rombongan hyung korea u di afc u tendangan penalti penentuan eh sekarang debut arhan di korea langsung dikeluarin dari lapangan ga pake lama mana habis ini juga pulang indo garuda callin	['tiba', 'tiba', 'keinget', 'arhan', 'mulangin', 'rombongan', 'hyung', 'korea', 'u23', 'di', 'afc', 'u23', 'tendangan', 'penalti', 'penentuan', 'eh', 'sekarang', 'debut', 'arhan', 'di', 'korea', 'langsung', 'dikeluarin', 'dari', 'lapangan', 'ga', 'pake', 'lama', 'mana', 'habis', 'ini', 'juga', 'pulang', 'indo', 'garuda', 'callin']

3.2. Pembobotan dengan TF-IDF

Setelah tahap prapemrosesan, Proses pembobotan kata dilakukan dengan menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency) untuk memberikan bobot pada setiap kata[6]. Pembobotan ini mencerminkan frekuensi kemunculan kata dalam sebuah dokumen. dan pentingnya kata itu dalam seluruh kumpulan dokumen.

Ekstraksi dengan fitur tf-idf (Term Frequency and Inverse Document Frequency) merupakan salah satu proses teknik ekstraksi fitur dengan memberikan nilai pada masing-masing. Hasil TF- IDF yang diperoleh dapat dilihat pada Tabel 5.

Tabel 5. Hasil TF-IDF

TOKEN	TERM FREQUENCY (TF)													DF	D/DF	IDF Log(D/DF)
	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	..	D21			
idextratime	1	1	0	0	0	0	0	0	0	0	0		0	2	11	1,041392685
bliblidotcom	1	1	0	0	0	0	0	0	0	0	0		0	2	11	1,041392685
keinget	1	0	0	0	0	0	0	0	0	0	0		0	1	22	1,342422681
arhan	2	0	0	0	0	0	0	0	0	0	0		0	2	11	1,041392685
mulangin	1	0	0	0	0	0	0	0	0	0	0		0	1	22	1,342422681
rombong	1	0	0	0	0	0	0	0	0	0	0		0	1	22	1,342422681
hyung	1	0	0	0	0	0	0	0	0	0	0		0	1	22	1,342422681
korea	2	0	0	0	0	0	0	0	0	0	1		0	2	11	1,041392685
u23	2	1	1	1	1	1	1	1	0	1	0		1	19	1,157894737	0,06366908
afc	1	1	1	1	1	1	1	1	0	1	0		1	19	1,157894737	0,06366908
……..																
pake	1	0	0	0	0	0	0	0	0	0	0		0	1	22	1,342422681

3.3 Term weighting (w)

W (TF × IDF) adalah bobot yang diberikan pada suatu kata dalam dokumen berdasarkan kombinasi frekuensi kemunculannya di dokumen tertentu (TF) dan kepentingannya di seluruh

koleksi dokumen (IDF)[7]. Adapun hasil term weighting dalam penelitian ini dapat dilihat pada tabel 6.

Tabel 6. Hasil term weighting

TOKEN	W (TF X IDF)												
	D0	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**….**	**D21**
idextratime	1,04	1,04	0	0	0	0	0	0	0	0	0		0
bliblidotcom	1,04	1,04	0	0	0	0	0	0	0	0	0		0
keinget	1,34	0	0	0	0	0	0	0	0	0	0		0
arhan	2,08	0	0	0	0	0	0	0	0	0	0		0
mulangin	1,34	0	0	0	0	0	0	0	0	0	0		0
rombong	1,34	0	0	0	0	0	0	0	0	0	0		0
hyung	1,34	0	0	0	0	0	0	0	0	0	0		0
korea	2,08	0	0	0	0	0	0	0	0	0	1,04		0
u23	0,12	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0	0,06	0		0,06
afc	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0,06	0	0,06	0		0,06
……													
pake	1,34	0	0	0	0	0	0	0	0	0	0		0

3.4 Implementasi Metode DBSCAN

3.4.1 Tahapan metode Implementasi metode DBSCAN

Dalam tahap analisis menggunakan metode DBSCAN, setelah pembobotan TF-IDF dilakukan dan nilai bobot data diperoleh, data diproses menggunakan DBSCAN melalui tahapan berikut[8]:

1. Tentukan nilai Epsilon dan MinPoints
   - Epsilon adalah radius maksimum untuk menentukan tetangga suatu titik. Dalam penelitian ini ditetapkan sebesar 9.5, artinya semua titik dalam jarak Euclidean 9.5 dianggap tetangga.
   - MinPoints adalah jumlah minimum titik dalam radius epsilon agar dianggap core point. Nilai MinPoints ditetapkan 8, sehingga titik dengan <8 tetangga dianggap noise atau border point.

2. Pilih titik pusat awal yang belum dikunjungi
   Titik pusat dipilih secara acak, dan secara manual ditetapkan pada dokumen 0 sebagai titik awal.

3. Hitung jarak Euclidean
   Jarak antara semua data dengan titik pusat yang belum dikunjungi dihitung menggunakan rumus Euclidean, misalnya dari dokumen 0 ke dokumen 1[9]. Hasilnya sebagai berikut

Tabel 7. Perhitungan Jarak Data Ke Titik Pusat

Jarak Dokumen	$\sum_{i=0}^{n}(D_0 - D_1)^2$	$\sqrt{\sum_{i=1}^{n}(D0 - D1)^2}$
D0-D0	58,3601	0
D0-D1	82,0134	7,639378247

D0-D2	69,5787	9,056124999
D0-D3	82,5626	8,341384777
D0-D4	131,9974	9,086396425
D0-D5	69,5696	11,48901214
D0-D6	96,93	8,340839286
D0-D7	84,7985	9,845303449
D0-D8	62,7387	9,208610101
D0-D9	51,4178	7,920776477
D0-D10	65,5691	7,170620615
……..		
D0-D18	64,8232	9,593862621

4. Setelah nilai Epsilon = 9,5 dan MinPoints = 8 ditentukan, langkah selanjutnya adalah memilih titik pusat awal yang belum dikunjungi. Dalam perhitungan manual ini, titik pusat awal ditetapkan pada D0 (dokumen 0). Selanjutnya dilakukan perhitungan jarak Euclidean antara D0 dan seluruh titik data lainnya. Titik yang memiliki jarak ≤ Epsilon dikategorikan sebagai tetangga dan dimasukkan ke dalam kluster yang sama, sedangkan titik dengan jarak > Epsilon tidak termasuk. Hasil perhitungan iterasi pertama disajikan pada Tabel 3.8, di mana titik seperti D1, D2, D3, D4, D6, D8, dan seterusnya termasuk ke dalam kluster D0, sementara titik D5, D7, D14, dan D16 tidak memenuhi kriteria. Titik terjauh yang masih berada dalam radius Epsilon dipilih sebagai pusat baru untuk iterasi berikutnya. Berdasarkan hasil perhitungan, titik pusat berikutnya adalah D8, sebagaimana ditampilkan pada Tabel 3.9.

Tabel 8. Perhitungan Iterasi Pertama

ITERASI 1		
Jarak D0	**Nilai**	**<=EPSILON**
D0-D0	0	YA
D0-D1	7,639378247	YA
D0-D2	9,056124999	YA
D0-D3	8,341384777	YA
D0-D4	9,086396425	YA
D0-D5	11,48901214	TIDAK
D0-D6	8,340839286	YA
D0-D7	9,845303449	TIDAK
D0-D8	9,208610101	YA
D0-D9	7,920776477	YA
D0-D10	7,170620615	YA
……		
D0-D21	8,03967661	YA

Tabel 9. Titik Pusat Baru yang Terpilih

Epsilon	**MinPoints**	**Total Ya**	**Titik Pusat Terpilih**
9,5	8	17	D0
			D8

5. Proses iterasi dilanjutkan secara berulang hingga semua titik data diklasifikasikan, baik ke dalam suatu kluster maupun sebagai noise. Pada iterasi terakhir, hasil perhitungan ditunjukkan pada Tabel 3.10, di mana titik D17 teridentifikasi sebagai pusat kluster karena memiliki jarak 0 dari dirinya sendiri dan memenuhi syarat MinPoints (≥8 titik tetangga dalam radius 9,5). Sementara itu, titik D5 memiliki jarak sebesar 10,3781983 dari D17, yang lebih besar daripada nilai Epsilon, sehingga tidak termasuk dalam radius titik pusat manapun. Dengan demikian, D5 dikategorikan sebagai noise, karena tidak dapat menjadi core point maupun border point. Hal ini menunjukkan bahwa DBSCAN mengelompokkan data berdasarkan kedekatan jarak antar titik, di mana titik-titik yang terisolasi di luar radius Epsilon didefinisikan sebagai noise.

Tabel 10. Iterasi Terakhir

ITERASI 21			
Jarak D17	**Nilai**	**Keterangan**	**Titik Pusat Terpilih**
D17-D0	7,203221502	Ya	D0
D17-D1	6,246863213	Ya	D8
D17-D2	7,556176282	Ya	D6
D17-D3	6,647789106	Ya	D7
D17-D4	7,674594452	Ya	D11
D17-D5	10,3781983	Tidak	D14
D17-D6	6,775426186	Ya	D20
D17-D7	8,648861197	Ya	D18
D17-D8	7,494964977	Ya	D3
D17-D9	6,35058265	Ya	D2
D17-D10	5,551522314	Ya	D4
.....			
D17-D21	6,208896842	Ya	

3.4.2 Implementasi DBSCAN untuk Klasterisasi Tweet Menggunakan Python

Penelitian ini mengimplementasikan algoritma DBSCAN untuk klasterisasi data tweet terkait Piala Asia U-19 menggunakan bahasa pemrograman Python. Proses implementasi terdiri dari tiga tahap utama: pra-pemrosesan data, ekstraksi fitur TF-IDF, dan klasterisasi menggunakan DBSCAN.

1. Pra-pemrosesan Data

Langkah awal melibatkan:

- Import library seperti Sastrawi, NLTK, Pandas, NumPy, dan Scikit-learn.
- Membaca dataset tweet melalui read_csv.
- Case folding untuk mengubah teks menjadi huruf kecil
- Tokenisasi menggunakan RegexpTokenizer
- Stemming dengan Sastrawi untuk memperoleh bentuk dasar kata
- Filtering stopwords untuk menghapus kata-kata tidak bermakna

2. Ekstraksi Fitur TF-IDF

Setelah filtering, dilakukan konversi teks ke dalam bentuk numerik menggunakan metode TF-IDF:

- TF (Term Frequency) menghitung frekuensi kata dalam dokumen.
- IDF (Inverse Document Frequency) menilai pentingnya kata berdasarkan keberadaannya dalam dokumen lain.
- Dihasilkan matriks TF-IDF yang menggambarkan bobot kata .

3. Klasterisasi dengan DBSCAN

Dengan parameter eps = 11.5 dan min_samples = 10, DBSCAN diterapkan untuk mengelompokkan dokumen berdasarkan kepadatan data:

- Klaster yang terbentuk dianalisis berdasarkan titik inti, tepi, dan noise.
- Jumlah titik per kategori (inti, tepi, noise)
- Seluruh hasil diekspor ke file Excel.

4. Visualisasi dengan PCA

Untuk memudahkan interpretasi, dilakukan reduksi dimensi menggunakan PCA, sehingga hasil klasterisasi dapat divisualisasikan dalam bentuk dua dimensi.

3.4.3 Hasil Implementasi Clustering

Penelitian ini mengimplementasikan algoritma DBSCAN untuk mengelompokkan tanggapan masyarakat di media sosial X terkait Piala Asia U-23 AFC 2024. Data tanggapan terlebih dahulu direpresentasikan ke dalam bentuk numerik menggunakan metode TF-IDF, guna memudahkan proses klasterisasi berbasis kepadatan. Hasil klasterisasi menunjukkan bahwa dari keseluruhan data, sebanyak 1.588 data poin (84,65%) tergabung dalam satu klaster utama (Klaster 0), sementara 288 data (15,35%) dikategorikan sebagai noise. Di antara anggota klaster, 1.539 data (97,12%) termasuk core points, sedangkan 49 data (2,88%) merupakan border points.

Visualisasi klaster dua dimensi menunjukkan pemisahan yang jelas antara Klaster 0 dan data noise. Klaster ditampilkan dengan warna oranye, sedangkan noise diwakili warna biru, menggambarkan dominasi data dalam satu kelompok dengan karakteristik yang mirip.

Secara keseluruhan, metode DBSCAN terbukti efektif dalam mengelompokkan data dengan pola serupa, serta mampu mengidentifikasi data outlier melalui kategori noise. Representasi fitur menggunakan TF-IDF turut meningkatkan akurasi pengelompokan melalui penekanan pada kata-kata penting. Kesimpulannya, DBSCAN menghasilkan pengelompokan yang memuaskan, dengan proporsi klaster utama yang signifikan. Ke depan, analisis konten klaster dapat dilakukan untuk mengungkap tema dominan, serta pendekatan klasterisasi lain dapat diterapkan sebagai pembanding kinerja.

3.5 Implementasi naive bayes

3.5.1 Pelabelan Otomatis

Pelebelan otomatis dilakukan menggunakan pustaka TextBlob untuk memberikan label sentimen pada data teks secara otomatis. Proses ini terdiri atas tiga tahap utama: preprocessing, translate, dan pelebelan. Tahap preprocessing meliputi pembersihan teks, penghapusan stopwords, tokenisasi, dan normalisasi. Selanjutnya, tahap translate digunakan untuk menerjemahkan teks ke dalam bahasa Inggris, guna meningkatkan efektivitas analisis oleh TextBlob. Terakhir, tahap pelebelan dilakukan dengan mengklasifikasikan teks menjadi sentimen positif, negatif, atau netral berdasarkan skor polaritas. Hasil pelebelan ini digunakan sebagai data pelatihan (training data) untuk model Naive Bayes[10].

3.5.2 Analisis Sentimen Naïve bayes

Setelah pelebelan, dilakukan analisis sentimen menggunakan algoritma Multinomial Naïve Bayes. Teks yang telah diproses dikonversi ke dalam bentuk numerik menggunakan metode TF-IDF, kemudian dibagi menjadi data latih (80%) dan data uji (20%). Model dilatih untuk mempelajari hubungan antara teks dan sentimen, dan selanjutnya digunakan untuk melakukan prediksi pada. Evaluasi model dilakukan dengan menghitung akurasi serta menghasilkan classification report (precision, recall, f1-score). Prediksi sentimen dari model dibandingkan dengan hasil TextBlob untuk melihat kesesuaian dan distribusi sentimen. Seluruh hasil disimpan dalam file Excel untuk keperluan analisis lanjutan dan visualisasi.

3.5.3 Hasil Analisis Sentimen

Hasil analisis dalam penelitian ini menunjukkan bahwa penerapan algoritma DBSCAN sebelum klasifikasi sentimen berperan penting dalam menyaring data berdasarkan kemiripan pola sentimen, serta mengidentifikasi outlier yang tidak konsisten. Hal ini meningkatkan kualitas data pelatihan untuk model Naïve Bayes. Klasifikasi sentimen dengan Naïve Bayes menghasilkan akurasi sebesar 60%. Sentimen netral menunjukkan performa terbaik dengan precision 0.60, recall 0.89, dan f1-score 0.71, menunjukkan kecenderungan model untuk mengklasifikasikan komentar sebagai netral. Untuk sentimen positif, precision mencapai 0.59 dan recall 0.45, menandakan kemampuan moderat dalam mengenali komentar positif.

Namun, klasifikasi negatif menunjukkan performa rendah dengan recall hanya 0.06, meskipun precision mencapai 0.75, mengindikasikan bahwa model kesulitan mengenali komentar

negatif secara konsisten. Distribusi hasil klasifikasi menunjukkan dominasi kategori netral sebanyak 1.002 data, diikuti oleh positif (483 data), dan negatif (54 data). Ketimpangan ini mencerminkan tantangan model dalam membedakan ekspresi sentimen secara seimbang, yang kemungkinan disebabkan oleh variasi pola bahasa informal di media sosial.

4. KESIMPULAN & SARAN

4.1 Kesimpulan

Penelitian ini membangun pipeline text mining untuk memetakan topik dan membaca sentimen percakapan X/Twitter tentang Piala Asia U-23 AFC 2024 dengan menggabungkan DBSCAN (klaster topik) dan Naïve Bayes (klasifikasi sentimen). Setelah praproses dan pembobotan TF-IDF, DBSCAN dengan parameter eps = 11,5 dan min_samples = 10 mampu memisahkan tema percakapan secara berarti: sekitar 1.588 tweet terklasifikasi sebagai core points (84,65%) dan 49 sebagai border points (2,88%), sementara sisanya dikenali sebagai noise. Struktur ini menunjukkan DBSCAN efektif mengurai percakapan yang padat dan tidak seragam kepadatannya sehingga topik dominan dapat terlihat jelas.

Pada lapisan sentimen, hasil klasifikasi menunjukkan kecenderungan netral sebagai yang paling dominan, disusul sentimen positif, dan negatif yang relatif lebih sedikit. Pola ini selaras dengan karakter diskursus olahraga yang banyak diisi komentar informatif/reaktif serta dukungan kepada tim, sementara keluhan atau kritik tersebar lebih tipis. Kinerja model Naïve Bayes berada di kisaran akurasi ±60%. Performa untuk kelas netral tergolong moderat, tetapi recall untuk kelas negatif masih rendah—indikasi adanya tantangan khas data media sosial Indonesia seperti campur kode (ID–EN), slang/variasi ejaan, ketidakseimbangan kelas, serta nuansa sarkasme yang sulit tertangkap oleh fitur leksikal sederhana.

Secara praktis, kombinasi “DBSCAN untuk peta topik” dan “NB untuk label sentimen” layak dijadikan baseline pemantauan opini publik berbasis media sosial: peta klaster membantu pihak terkait memetakan isu prioritas, sementara distribusi sentimen memberi arah persepsi publik atas tiap isu. Ke depan, kualitas deteksi—terutama pada sentimen negatif—berpotensi meningkat melalui normalisasi slang yang lebih kaya, penanganan code-mixing yang lebih baik, penyeimbangan kelas, serta eksplorasi model berbasis konteks (misalnya SVM sebagai pembanding atau model transformer berbahasa Indonesia). Optimasi parameter DBSCAN dan perluasan fitur ke n-gram yang lebih lebar atau embedding juga layak dipertimbangkan.

4.2 Saran

Berdasarkan temuan yang diperoleh, beberapa saran untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Perbandingan metode klasterisasi disarankan untuk membandingkan hasil DBSCAN dengan metode klasterisasi lain seperti K-Means, Agglomerative Clustering, atau HDBSCAN, guna mengevaluasi efektivitas DBSCAN dalam menangani data dengan variasi densitas yang tinggi.
2. Peningkatan preprocessing teks Penggunaan Sastrawi dalam preprocessing memiliki keterbatasan terhadap kosakata informal dan slang media sosial. Penelitian selanjutnya dapat mempertimbangkan pustaka yang lebih adaptif seperti PySastrawi, NLTK kustom, atau pendekatan deep learning NLP seperti spaCy dengan model bahasa Indonesia.
3. Pengelolaan Noise berbasis konteks dianjurkan untuk mengembangkan strategi penghapusan noise berbasis analisis linguistik atau pendekatan kontekstual guna meningkatkan kualitas data dan akurasi klasifikasi, khususnya dalam menangani ambiguitas bahasa di media sosial.
4. Penggunaan model klasifikasi lanjutan karena Naïve Bayes memiliki keterbatasan dalam menangani dependensi antar kata, disarankan untuk menggunakan model yang lebih canggih seperti SVM, Random Forest, atau model deep learning (misalnya LSTM atau BERT) untuk meningkatkan performa klasifikasi sentimen.

REFERENSI

[1] A. K. Santoso, “Analisis Sentimen dan Klasterisasi Topik pada Media Sosial Twitter Berbahasa Indonesia.” 2022.

[2] J. E. S. A.Habiba, R Rizal Isnanto, “Pemilihan Fitur Chi Square Pada Algoritma Naïve Bayes dan Pengaruhnya Terhadap Analisis Sentimen Masyarakat Indonesia Tentang Pembelajaran Tatap Muka Pada Masa Pandemi Covid-19.” 2023.

[3] K. Frecis Matheos Sarimole, “Analisis Sentimen Terhadap Aplikasi Satu Sehat Pada Twitter Menggunakan Algoritma Naive Bayes Dan Support Vector Machine.” 2024.

[4] S. A. Ardiyansa, A. N. Guci, J. Febryan, D. Alhusari, H. A. Fajri, and K. Kunci, “Klasifikasi Sentimen Tweet dengan Arsitektur Hybrid Transformers-CNN pada Platform Twitter,” vol. 14, no. 3, pp. 5001–5012, 2025.

[5] H. G. Simanullang, A. P. Silalahi, and G. P. Pangaribuan, “PENGELOMPOKAN KOMENTAR PENGGUNA JKN MOBILE MENGGUNAKAN METODE K-MEDOIDS,” 2025.

[6] L. M. Barus, H. G. Simanullang, and A. P. Silalahi, “Classification of English-Language Racial Hate Speech Using Random Forest ”,” *J. Intell. Comput. Inf. Syst.*, vol. 1, no. 1, pp. 1–17, 2025.

[7] T. R. Baruara, H. G. Manullang, and A. P. Silalahi, “ANALISA ALGORITMA K-MEDOIDS DALAM,” vol. 4, no. 2, pp. 1–10, 2024.

[8] R. Wulandari, W. Yustanti, S. Si, and M. Kom, “Analisis Text Clustering Kebijakan Pembukaan Daerah Wisata pada Masa Pandemi Berbasis Densitas Spasial (DBSCAN),” vol. 03, no. 02, pp. 1–10, 2022.

[9] N. F. Purba, S. A. Rumapea, and A. P. Silalahi, “Analisis Sentimen Dompet Elektronik Pada Twitter Menggunakan Metode K-Means,” *JITTER J. Ilm. Teknol. dan Komput.*, vol. 5, no. 1, p. 2100, May 2024, doi: 10.24843/JTRTI.2024.v05.i01.p04.

[10] R. P. Al Hajju Arafah, “Penerapan Naïve Bayes untuk Analisis Sentimen Kebijakan Publik di Twitter.” 2025.