Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 3, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i3.8627

A Supervised Learning Model for Sentiment Analysis Based on Regional Dialects in Tourism-Related Issues

Tb Ai Munandar
Informatics Department, Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya, Jl Tirtayasa, DKI Jakarta, 12160, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-03-29
Revised 2025-07-12
Accepted 2025-07-17

Abstract – *Indonesia has an exceptionally rich diversity of regional languages, one of which is the Bekasi dialect, often used in social media communication. The uniqueness of this dialect presents specific challenges in extracting public opinion, especially in text-based sentiment analysis. This study aims to develop a sentiment analysis framework that incorporates regional dialects from social media data and evaluate the effectiveness of various supervised learning algorithms. Data were collected from the Facebook group "Explore Bekasi Tourism," totaling 1,257 posts and comments, which were filtered down to 1,000 relevant instances. A manual validation process was conducted by linguistic experts to convert non-standard terms and regional dialects into standardized Indonesian, followed by translation into English for annotation purposes. The analysis method involved preprocessing steps (tokenizing, case folding, stemming), feature weighting using TF-IDF, and sentiment classification using four algorithms: Naive Bayes, K-Nearest Neighbor, Support Vector Machine, and Decision Tree. The evaluation results show that Naive Bayes achieved the best performance with an accuracy of 76%, followed by K-Nearest Neighbor (67.5%), SVM (65.5%), and Decision Tree (28%). These findings highlight the crucial role of expert judgment in processing dialect-based data to ensure accurate sentiment classification. The study recommends developing a broader annotated corpus of regional dialects and exploring deep learning methods in future research to enhance classification performance and generalizability*

Keywords: *Facebook; Regional Dialect; Sentiment Analysis; Supervised Learning; TF-IDF*

Corresponding Author:

Tb Ai Munandar

Email: tbaimunandar@gmail.com

Abstrak – *Indonesia memiliki keragaman bahasa daerah yang sangat tinggi, salah satunya dialek Bekasi yang kerap muncul dalam komunikasi di media sosial. Keunikan dialek ini menghadirkan tantangan tersendiri dalam proses ekstraksi opini publik, khususnya dalam analisis sentimen berbasis teks. Penelitian ini bertujuan untuk mengembangkan kerangka kerja analisis sentimen yang mempertimbangkan keberadaan dialek lokal dalam data media sosial, serta menguji efektivitas beberapa algoritma pembelajaran terawasi. Data diperoleh dari grup Facebook "Explore Bekasi Tourism" dengan total 1.257 unggahan dan komentar yang kemudian diseleksi menjadi 1.000 data yang relevan. Proses validasi dilakukan secara manual oleh ahli bahasa guna mengkonversi istilah tidak baku dan dialek daerah ke dalam Bahasa Indonesia baku, lalu diterjemahkan ke dalam Bahasa Inggris untuk keperluan anotasi. Metode analisis meliputi tahapan preprocessing (tokenizing, case folding, stemming), pembobotan fitur menggunakan TF-IDF, serta klasifikasi sentimen dengan empat algoritma: Naive Bayes, K-Nearest Neighbor, Support Vector Machine, dan Decision Tree. Hasil evaluasi menunjukkan bahwa algoritma Naive Bayes memberikan performa terbaik dengan akurasi 76%, diikuti oleh K-Nearest Neighbor (67,5%), SVM (65,5%), dan Decision Tree (28%). Temuan ini menunjukkan bahwa keterlibatan ahli bahasa dalam pemrosesan data berbasis dialek sangat penting untuk memastikan keakuratan klasifikasi sentimen. Penelitian ini merekomendasikan pengembangan korpus dialek lokal serta penerapan teknik deep learning untuk riset lanjutan.*

Kata Kunci*: Analisis Sentimen, Dialek Daerah, Facebook, Pembelajaran Terawasi, TF-IDF*

I. PENDAHULUAN

Indonesia, sebagai negara kepulauan, memiliki kekayaan budaya yang luar biasa, tercermin dalam keragaman bahasanya. Terdapat lebih dari 652 bahasa daerah yang tersebar di berbagai provinsi, belum termasuk berbagai dialek dan sub-dialek yang digunakan antardaerah, bahkan dalam satu provinsi sekalipun [1]. Kekayaan linguistik ini memperkaya warisan budaya Indonesia, menghadirkan beragam aksen, kebiasaan, dan ekspresi khas dari setiap wilayah. Salah satu contohnya adalah dialek Bekasi yang memadukan unsur Betawi dan Sunda dalam pengucapan bahasa nasional. Keunikan ini, meskipun memperkaya interaksi sosial, menciptakan tantangan tersendiri dalam ekstraksi pengetahuan dari teks, khususnya di media sosial.

Platform seperti Facebook menjadi wadah ekspresi masyarakat [2], [3] terutama warga Bekasi dalam menyampaikan opini tentang isu-isu lokal, termasuk tantangan pariwisata pasca-COVID-19. Pandemi telah memberikan dampak signifikan terhadap sektor pariwisata Indonesia, termasuk Bekasi, namun pasca-pandemi, sektor ini kembali menggeliat dan menjadi penggerak pertumbuhan ekonomi. Seiring dengan itu, media sosial

berkembang menjadi medium utama penyampaian opini publik yang dapat memengaruhi persepsi terhadap destinasi wisata. Opini tersebut sangat subjektif, bercampur antara sentimen positif dan negatif, seringkali dinyatakan dengan dialek lokal. Hal ini memperumit proses analisis karena penggunaan dialek menyebabkan kesulitan dalam anotasi dan penerjemahan ke dalam bahasa resmi [4].

Analisis sentimen sebenarnya bukanlah hal baru. Dalam waktu sepuluh tahun terakhir, sejumlah penelitian telah dilakukan terhadap opini publik di media elektronik, misalnya terkait penghapusan ujian nasional [5], [6], analisis berita berbasis deret waktu untuk tujuan pelaporan ekonomi dan keuangan [7], [8], klasifikasi sentimen berbasis data Twitter dan Tumblr [9]-[11], hingga isu kesejahteraan guru honorer [12], [13]. Namun, sebagian besar dari penelitian tersebut belum menyinggung secara khusus ekspresi dalam dialek daerah. Penelitian terkait analisis sentimen yang mempertimbangkan dialek lokal masih jarang dilakukan [14], terutama karena keterbatasan anotasi bahasa yang sesuai untuk dialek daerah. Padahal, penggunaan dialek di media sosial semakin menjadi tren, bahkan sering dikombinasikan dengan bahasa asing seperti Inggris.

Kondisi ini mendorong pentingnya riset untuk menjawab kebutuhan analisis sentimen berbasis dialek daerah. Beberapa studi telah dilakukan pada dialek-dialek Arab [15], [16] seperti Aljazair [17], Kurdi [18], Yordania [19], Tunisia, Riyadh, Dammam, dan Jeddah [20]-[22], serta pada dialek dari negara lain termasuk penggunaan script Romania [23], Burma [4], India [24]-[26], dialek Arab-Maroko [27]-[29], bahasa Dravida-Kannada [30], dan Tagalog-Filipina [31]. Pendekatan yang digunakan dalam penelitian tersebut beragam, termasuk pembangunan korpus khusus seperti dalam penelitian pada dialek Yordania [19]. Namun, secara umum, masih banyak dialek yang belum terakomodasi, dan sebagian besar korpus yang tersedia masih didominasi oleh bahasa Inggris.

Dialek sendiri merupakan variasi bahasa yang digunakan oleh kelompok penutur di wilayah tertentu, dikenal pula sebagai dialek regional [32]. Dialek ini sering kali memiliki ciri khas yang digunakan dalam konteks informal dan bercampur dengan unsur humor [33]. Beberapa dialek telah memiliki nama tradisional yang menjadi penanda identitas geografis penuturnya [34]. Fenomena ini menarik perhatian tidak hanya dalam bidang linguistik, tetapi juga dalam komputasi, khususnya dalam pengembangan sistem analisis sentimen.

Seiring berkembangnya bidang analisis sentimen, terutama dalam konteks bisnis, penggalian opini publik menjadi aspek penting dalam memahami perilaku konsumen dan menyusun strategi pemasaran. Platform seperti Facebook, Twitter, Reddit, hingga Instagram menjadi sumber data utama dalam penelitian analisis sentimen [35], [36]. Secara umum, analisis sentimen terdiri dari lima tahap utama, yaitu: pengumpulan data, pra-pemrosesan, ekstraksi fitur, klasifikasi sentimen, dan evaluasi sistem [17].

Namun, hingga saat ini, belum tersedia kerangka kerja standar yang dapat secara efektif mentransformasikan dialek daerah ke dalam bahasa nasional, apalagi ke bahasa internasional seperti Inggris. Peran ahli bahasa sangat krusial dalam proses validasi transformasi ini. Oleh karena itu, penelitian ini bertujuan untuk (1) mengusulkan kerangka kerja baru untuk analisis sentimen yang mempertimbangkan dialek daerah, (2) menganalisis sentimen opini masyarakat terkait pariwisata Bekasi pasca-COVID-19 di media sosial, dan (3) menguji efektivitas beberapa algoritma pembelajaran terawasi dalam mengklasifikasikan opini pariwisata berdasarkan data media sosial.

Kebaruan dari penelitian ini terletak pada penggabungan pendekatan linguistik dan komputasional untuk membangun sistem analisis sentimen berbasis dialek lokal, yang belum banyak dijelajahi dalam konteks bahasa daerah Indonesia. Studi ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan teknologi pemrosesan bahasa alami yang lebih inklusif terhadap keberagaman linguistik, serta mendukung pengambilan keputusan yang lebih baik dalam sektor pariwisata berbasis opini masyarakat. Tabel 1 memperlihatkan ulasan sistematis terhadap penelitian-penelitian sebelumnya yang relevan dengan topik utama penelitian yang dilakukan.

TABEL 1
ULASAN PENELITIAN SEBELUMNYA

No.	Penulis dan Tahun	Fokus Penelitian	Bahasa / Dialek	Metode / Model	Kelebihan	Kekurangan
1	Erlin et al., 2022	Sentimen publik terhadap penghapusan Ujian Nasional	Indonesia (baku)	SVM	Fokus pada isu nasional	Tidak menangani dialek lokal
2	Sidorov et al., 2018	Analisis berita ekonomi berbasis deret waktu	Inggris	Fractality & time-series	Komputasional dan temporal	Tidak fokus pada media sosial atau dialek
3	Jiawei & Murata, 2019	Prediksi pasar saham menggunakan analisis sentimen	Inggris	LSTM Neural Network	Deep learning untuk tren pasar	Tidak membahasa minoritas / dialek

4	Kumar & Jaiswal, 2017	Analisis Twitter dan Tumblr	Inggris	Soft Computing	Komparasi platform media	Tidak membahas ekspresi regional
5	Lazuardi et al., 2020	Opini publik tentang guru honorer	Indonesia (baku)	Naive Bayes	Studi sosial kontekstual	Dialek lokal tidak ditangani
6	Rohini et al., 2016	Analisis domain sentiment berbasis bahasa daerah	Kannada	Rule-based	Spesifik ke regional	Pendekatan sederhana, tidak teruji secara ML
7	Al Suwaidi et al., 2016	Sentimen Twitter berbahasa Emirati	Arab (dialek)	Naive Bayes, klasikal	Spesifik ke dialek Arab	Tidak menggunakan anotasi multibahasa
8	Duwairi, 2015	Sentimen dialek bahasa Arab	Arab (dialek)	Rule-based	Pionir untuk dialek Arab	Skala data terbatas
9	Chader et al., 2019	Arabizi (Algerian Dialect)	Arab campuran (Aljazair)	Text normalization + ML	Menangani variasi Arab Latin	Fokus pada transliterasi
10	Amin et al., 2022	Tantangan SA pada Bahasa Kurdi	Kurdi	Diskusi kualitatif	Mengidentifikasi hambatan dialek	Tanpa eksperimen praktis
11	Atoum & Nouman, 2019	Analisis tweet Bahass Yordania	Arab (Yordania)	Dataset khusus + model SA	Membuat korpus dialek	Dialek Arab, bukan Asia Tenggara
12	Hdioud & Tirari, 2022	Sentimen dialek Maroko menggunakan DL	Arab (Maroko)	Deep Learning (DL)	Menggunakan AraBERT	Tidak fokus pada konversi linguistik
13	Ranjitha & Bhanu, 2021	SA dialek Kannada dengan Decision Tree	Kannada	Decision Tree + Kamus Lokal	Gunakan kamus dialek	Validasi manual terbatas

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan teknik sentiment analysis berbasis machine learning untuk mengkaji opini masyarakat terhadap destinasi wisata di Bekasi pasca-COVID-19. Proses analisis dimulai dengan pengumpulan data dari platform media sosial Facebook menggunakan Facebook API. Data yang dikumpulkan terdiri dari seribu entri, berupa unggahan dan komentar yang menggunakan bahasa resmi maupun dialek lokal Bekasi. Data diperoleh dari grup Facebook bernama "Explore Bekasi Tourism" yang memiliki 92.100 anggota, dengan periode pengambilan data antara Januari hingga Mei 2022.

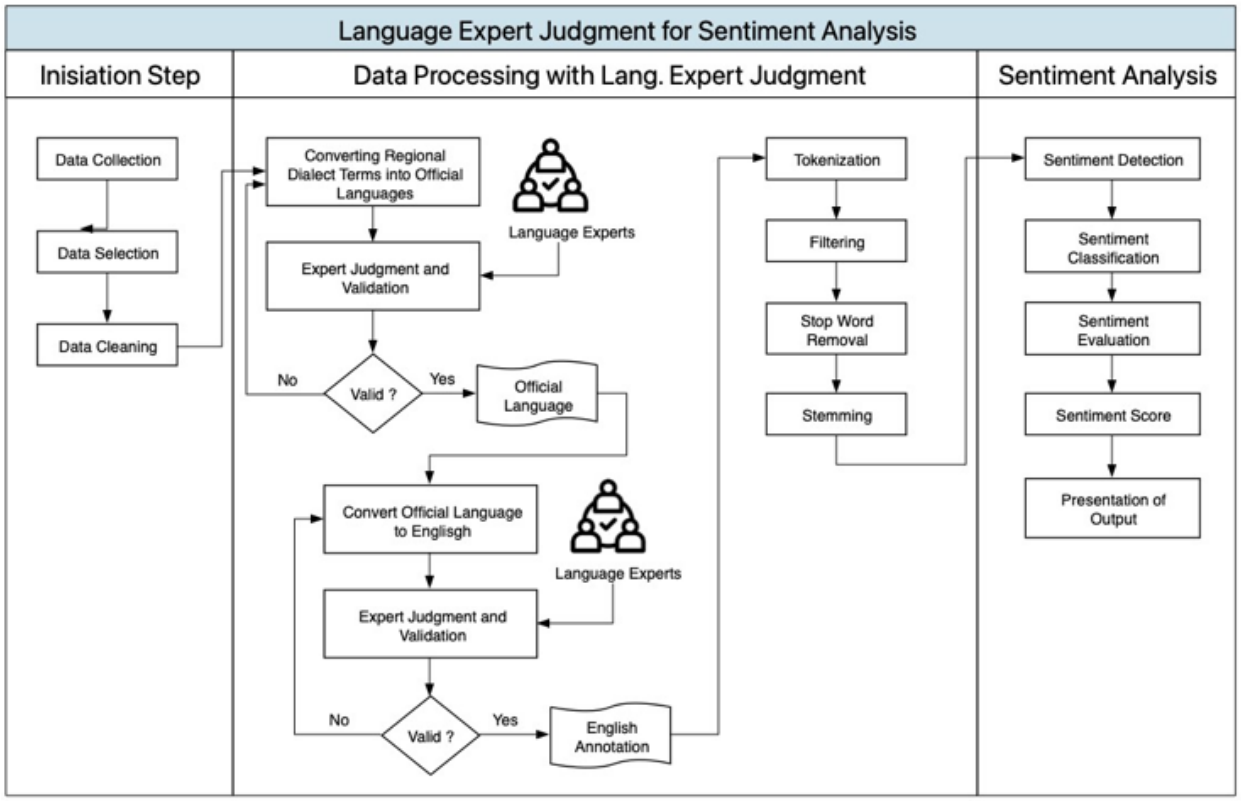
Tahapan dalam analisis sentimen pada penelitian ini mengacu pada alur kerja umum yang meliputi: pengumpulan data, pemrosesan data, ekstraksi fitur, klasifikasi sentimen (positif, negatif, atau netral), dan penilaian sentimen. Namun, alur ini tidak sepenuhnya dapat diterapkan pada data dengan dialek lokal karena kendala ketersediaan anotasi bahasa. Oleh karena itu, penelitian ini mengusulkan penyesuaian khusus pada tahap data processing. Tahap ini melibatkan proses konversi istilah dari dialek daerah ke dalam bahasa resmi, dan selanjutnya ke dalam bahasa Inggris agar sesuai dengan korpus anotasi yang tersedia. Untuk memastikan akurasi, validasi istilah dilakukan oleh ahli bahasa yang memiliki kompetensi dalam dialek lokal, bahasa Indonesia, dan bahasa Inggris. Penyesuaian ini bertujuan agar alur kerja dapat diadopsi secara lebih luas dalam penelitian analisis sentimen berbasis dialek lokal.

Pada tahap ekstraksi fitur, metode yang digunakan adalah *Term Frequency–Inverse Document Frequency* (TF-IDF), yang memberikan bobot pada setiap istilah berdasarkan frekuensi kemunculannya dalam dokumen dan distribusinya di seluruh korpus [37]. TF dihitung sebagai frekuensi kemunculan suatu istilah dalam dokumen dibandingkan total istilah dalam dokumen tersebut. Sementara itu, IDF diperoleh dari logaritma jumlah dokumen dalam korpus dibagi jumlah dokumen yang mengandung istilah tersebut [38]. Nilai TF-IDF akhir dihitung dengan mengalikan nilai TF dan IDF.

Setelah fitur diekstraksi, proses klasifikasi sentimen dilakukan menggunakan lima algoritma pembelajaran terawasi, yaitu: *Naive Bayes Classifier* (NBC), *K-Nearest Neighbor* (k-NN), *Support Vector Machine* (SVM), dan *Decision Tree. Algoritma* NBC bekerja berdasarkan *Teorema Bayes* dan asumsi independensi antar atribut, dengan rumus probabilitas kelas berdasarkan atribut input [39]-[41]. Metode k-NN membandingkan pola baru dengan k tetangga terdekat dalam ruang berdimensi-n, dan menggunakan matriks jarak, seperti *Euclidean distance*, untuk menemukan kemiripan [39]. SVM digunakan untuk memisahkan data

dalam dimensi tinggi dengan mencari hyperplane terbaik yang memisahkan dua kelas data menggunakan fungsi kernel [39], [42]. Sementara itu, algoritma *Decision Tree* menyusun pohon keputusan berdasarkan nilai information gain, yang dihitung melalui entropi dari masing-masing atribut [39], [41].

Seluruh algoritma akan diuji pada dataset yang sama dan dibandingkan kinerjanya untuk menentukan model klasifikasi terbaik. Evaluasi dilakukan berdasarkan akurasi klasifikasi, dan hasil dari masing-masing metode akan dianalisis untuk melihat sejauh mana efektifitas pendekatan yang diusulkan dalam menangani data yang mengandung dialek lokal. Diagram alur penelitian seperti ditunjukkan pada Gambar 1.



Gambar 1. Kerangka Penelitian Sentimen Analisis Dengan Expert Judgment

III. HASIL DAN PEMBAHASAN

Bagian ini terdiri dari dua subbagian. Subbagian pertama menyajikan hasil yang diperoleh dari penelitian yang telah dilakukan, sedangkan subbagian kedua membahas temuan, termasuk interpretasi dari hasil klasifikasi.

A. Hasil Penelitian

1) Penilaian dan Validasi Ahli Bahasa

Sebanyak 1.257 data berhasil dikumpulkan selama proses pengambilan data yang terdiri dari unggahan dan komentar pada platform Facebook. Setiap data mencakup informasi seperti tanggal unggahan, ID pengguna, nama pengguna, isi unggahan, komentar, serta deskripsi berbagi. Sebagian besar data menunjukkan struktur bahasa yang tidak baku dan banyak dipengaruhi oleh dialek lokal Bekasi. Untuk meningkatkan kualitas dataset, dilakukan proses seleksi dan pembersihan data sehingga diperoleh 1.000 data akhir yang layak digunakan. Tahapan ini dilakuakn untuk mempermudah konversi bentuk bahasa yang tidak standar dan dialek daerah.

Proses konversi dilakukan secara manual sehingga membutuhkan waktu dan tenaga. Proses konversi ini dilakukan secara teliti dengan mengidentifikasi dan mengelompokkan kata atau frasa yang tidak standar serta mengandung dialek lokal. Frasa kemudian dipisahkan dari dataset utama untuk memudahkan pengelolaan. Setelah dilakukan konversi, hasilnya kemudian ditinjau dan divalidasi oleh ahli bahasa yang kompeten dalam bahasa Indonesia dan pemahaman tentang dialek lokal. Validasi ini dilakukan secara iteratif hingga diperoleh hasil yang disepakati. Hasil validasi terhadap kata-kata non-standar dan dialek daerah untuk sebagian data ditampilkan pada Tabel 2.

Setelah proses validasi, struktur kalimat asli disesuaikan dengan bentuk bahasa Indonesia yang baku, kemudian diterjemahkan ke dalam bahasa Inggris. Proses ini dilakukan untuk memastikan tidak terjadi kehilangan makna dari kalimat awal. Penyesuaian melibatkan perubahan struktur frasa, penambahan konjungsi yang sesuai, serta modifikasi linguistik lainnya dengan tetap menjaga makna asli. Hasil dari proses ini ditunjukkan dalam Tabel 3 yang menyajikan sebagian kalimat asli, hasil konversi ke bahasa Indonesia, dan terjemahan ke dalam bahasa Inggris.

TABEL 2
BEBERAPA HASIL PENILAIAN DAN VALIDASI AHLI BAHASA TERHADAP KATA TIDAK BAKU DAN DIALEK DAERAH

Regional Dialect	After Adjustment	Expert Judgement Result	Expert Validation Result
Bagen	Biarin	Biarkan	Invalid, fix it according to the suggestions provided
Ge	Aja	Saja	Invalid, fix it according to the suggestions provided
Olog	Boros	Boros	Valid
Ilok	Masa	Betulkah?	Invalid, fix it according to the suggestions provided
Ora Danta	Engga Jelas	Tidak Jelas	Invalid, fix it according to the suggestions provided
Uantri Pool	Antri Banget	Antri Sekali	Invalid, fix it according to the suggestions provided
Kongkow	Berkumpul	Berkumpul	Valid
Nyo	Ayo	Ayo	Valid
Gretong	Gratis	Gratis	Valid
Misquen	Miskin	Miskin	Valid
Now	Sekarang	Sekarang	Valid
Awang	Males	Malas	Invalid, fix it according to the suggestions provided
Sediain	Menyediakan	Sediakan	Invalid, fix it according to the suggestions provided
Pas	Tepat	Sesuai	Invalid, fix it according to the suggestions provided
Kne	Sini	Kesini	Invalid, fix it according to the suggestions provided

TABEL 3
HASIL KONVERSI DARI TEKS ASLI – BAHASA INDONESIA – BAHASA INGGRIS

Text before conversion	Text after converting to Indonesian	Text after converting to English
Bagus ya teh.	Bagus ya kak	It is nice, sis
engga ada musholanya doang. Dan kalau hujan susah neduh dan hasilnya basah kuyup	Tidak ada mushola. Jika hujan turun, sulit mencari tempat berteduh, dan kalaupun ada tetap basah kuyup	There is no prayer room. If it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet
Kalo ga salah ini bekas danau samba cuma beda pintu masuknya di rubah	Kalau tidak salah ini bekas danau Samba, bedanya adalah posisi pintu masuk yang diubah.	If I am not mistaken, this is the former Samba Lake. The difference is that the position of the entrance has been changed.
iya tempat fotonya bagus-bagus mbak	Iya tempat fotonya bagus-bagus kak	Yes, it is a great photo spot
jelek dah engga usah kesana mbak situ mah	Jelek dan tidak perlu kesana kak	It is not interesting and there is no need to go there, sis.

Langkah selanjutnya adalah pelabelan data, yaitu memberikan label sentimen (positif, negatif, atau netral) pada setiap kalimat yang telah melalui proses konversi dan anotasi bahasa Inggris. Proses pelabelan ini dilakukan secara manual dan kemudian ditinjau ulang oleh ahli bahasa untuk memastikan ketepatan label. Contoh hasil pelabelan data ditampilkan pada Tabel 4.

TABEL 4
BEBERAPA HASIL PELABELAN DATASET

Text	Label
It is nice, sis	Positive
There is no prayer room. If it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet	Negative
If I am not mistaken, this is the former Samba lake, the difference is the position of the entrance has been changed.	Neutral
Yes, it is a great photo spot	Positive
It is not interesting and there is no need to go there, sis.	Negative

2) Analisis Sentimen

Setelah seluruh data dianotasi dan dilabeli, tahap analisis sentimen dilanjutkan dengan proses *case folding*, *tokenizing*, dan *stemming. Case folding* dilakukan untuk mengubah semua huruf menjadi huruf kecil sebagaimana diperlihatkan pada Tabel 5. Selanjutnya, proses tokenizing digunakan untuk memisahkan setiap kata dalam kalimat sehingga dapat dianalisis lebih lanjut (Tabel 6). Setelah itu, proses stemming dilakukan untuk menghapus kata-kata tidak bermakna atau mengubah kata ke bentuk dasar, seperti terlihat pada Tabel 7.

TABEL 5
BEBERAPA CONTOH DATASET SETELAH PROSES CASE FOLDING

Before Case Folding	After Case Folding
It is nice, sis	it is nice, sis
There is no prayer room. If it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet	there is no prayer room. if it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet
If I am not mistaken, this is the former Samba lake, the difference is the position of the entrance has been changed.	if i am not mistaken, this is the former samba lake, the difference is the position of the entrance has been changed.
Yes, it is a great photo spot	yes, it is a great photo spot
It is not interesting and there is no need to go there, sis.	it is not interesting and there is no need to go there, sis.

TABEL 6.
BEBERAPA HASIL TOKENISASI PADA DATASET

Before Tokenizing	After Tokenizing
it is nice, sis	[‘it’, 'is', 'nice', ‘sis’]
there is no prayer room. if it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet	[‘there’, ‘is’,’no’, ‘prayer’, ‘room’, ‘if’, ‘it’, ‘rains’,’it’, ‘is’, ‘difficult’, ‘to’, ‘find’, ‘shelter’, ‘and’, ‘even’, ‘if’, ‘there’, ‘is’ ‘one’, ‘it’, ‘is’, ‘still’, ‘soaking’, ‘wet’]
if i am not mistaken, this is the former samba lake, the difference is the position of the entrance has been changed.	[‘ if’,’i’, ‘am’, ‘not’, ‘mistaken’, ‘this’, ‘is’, ‘the’, ‘former’, ‘samba’, ‘lake’, ‘the’, ‘difference’, ‘is’, ‘the’, ‘position’, ‘of’, ‘the’, ‘entrance’ ‘has’, ‘been’, ‘changed’]
yes, it is a great photo spot	[‘yes’, ‘it’, ‘is’, ‘a’, ‘great’, ‘photo’, ‘spot’]
it is not interesting and there is no need to go there, sis.	[‘it’, ‘is’, ‘not’, ‘interesting’, ‘and’, ‘there’, ‘is’, ‘no’, ‘need’, ‘to’, ‘go’, ‘there’, ‘sis’]

TABEL 7
BEBERAPA HASIL PROSES STEMMING PADA DATASET

Before Stemming	After Stemming
it is nice, sis	nice
there is no prayer room. if it rains, it is difficult to find shelter, and even if there is one, it is still soaking wet	no prayer room rains difficult find shelter one soak wet
if i am not mistaken, this is the former samba lake, the difference is the position of the entrance has been changed.	not mistake former samba lake difference position entrance change
yes, it is a great photo spot	yes great photo shot
it is not interesting and there is no need to go there, sis.	not interest no need go

3) Pembobotan dengan TF-IDF

Tahap berikutnya adalah melakukan pembobotan dengan menghitung nilai *Term Frequency-Inverse Document Frequency* (TF-IDF), yaitu hasil perkalian antara frekuensi kemunculan suatu kata dalam dokumen dan kebalikannya terhadap frekuensi kata dalam seluruh dokumen. Perhitungan diawali dengan menghitung frekuensi kata (*term frequency*), kemudian dilanjutkan dengan menghitung *inverse document frequency* (IDF), dan akhirnya mengalikan keduanya untuk mendapatkan bobot TF-IDF. Contoh perhitungan ditampilkan pada Tabel 8 dan Tabel 9. Hasil akhir pembobotan ditunjukkan dalam Tabel 10.

TABEL 8
BEBERAPA CONTOH DATA DENGAN LABEL KELAS

Text	Class
no prayer room rains difficult find shelter one soak wet	C2
not mistake former samba lake difference position entrance change	C3
yes great photo shot	C1

Hasil bobot TF-IDF selanjutnya digunakan sebagai fitur untuk proses klasifikasi menggunakan empat metode pembelajaran terawasi: *Naive Bayes*, *K-Nearest Neighbor*, *Support Vector Machine*, dan *Decision Tree*. Dataset sebanyak 1.000 data dibagi menjadi dua bagian, yaitu 80% sebagai data latih dan 20% sebagai data uji. Berdasarkan analisis, distribusi data latih terdiri dari 62,9% opini positif, 12,9% opini negatif, dan 24,2% opini netral. Sementara itu, data uji terdiri dari 63,5% opini positif, 14% negatif, dan 22,5% netral.

Evaluasi model dilakukan dengan menghitung akurasi serta membuat confusion matrix untuk memperoleh nilai presisi, recall, dan F1-score. Hasil evaluasi ditunjukkan dalam Tabel 11 dan Tabel 12. Model Naive Bayes menunjukkan performa terbaik dengan akurasi sebesar 76%, diikuti oleh K-Nearest Neighbor (67,5%), Support Vector Machine (65,5%), dan Decision Tree (28%). Dengan demikian, Naive Bayes menjadi metode paling efektif dalam analisis sentimen opini pariwisata masyarakat Bekasi berdasarkan data media sosial Facebook.

TABEL 9
PERHITUNGAN FREKUENSI KEMUNCULAN KATA (TERM FREQUENCY/TF)

Text	TF			TF Normalization			IDF
	C1	C2	C3	C1	C2	C3	
no	1	0	0	0.11	0.00	0.00	0.48
prayer	1	0	0	0.11	0.00	0.00	0.48
room	1	0	0	0.11	0.00	0.00	0.48
rains	1	0	0	0.11	0.00	0.00	0.48
find	1	0	0	0.11	0.00	0.00	0.48
shelter	1	0	0	0.11	0.00	0.00	0.48
one	1	0	0	0.11	0.00	0.00	0.48
soak	1	0	0	0.11	0.00	0.00	0.48
wet	1	0	0	0.11	0.00	0.00	0.48
not	0	1	0	0.00	0.11	0.00	0.48
mistake	0	1	0	0.00	0.11	0.00	0.48
former	0	1	0	0.00	0.11	0.00	0.48
samba	0	1	0	0.00	0.11	0.00	0.48
lake	0	1	0	0.00	0.11	0.00	0.48
difference	0	1	0	0.00	0.11	0.00	0.48
position	0	1	0	0.00	0.11	0.00	0.48
entrance	0	1	0	0.00	0.11	0.00	0.48
change	0	1	0	0.00	0.11	0.00	0.48
yes	0	0	1	0.00	0.00	0.25	0.48
great	0	0	1	0.00	0.00	0.25	0.48
photo	0	0	1	0.00	0.00	0.25	0.48
shot	0	0	1	0.00	0.00	0.25	0.48

TABEL 10
HASIL PEMBOBOTAN DENGAN METODE TF-IDF

Text	TF Normalization			IDF	TF-IDF		
	C1	C2	C3		C1	C2	C3
no	0.11	0.00	0.00	0.48	0.05	0.00	0.00
prayer	0.11	0.00	0.00	0.48	0.05	0.00	0.00
room	0.11	0.00	0.00	0.48	0.05	0.00	0.00
rains	0.11	0.00	0.00	0.48	0.05	0.00	0.00
find	0.11	0.00	0.00	0.48	0.05	0.00	0.00
shelter	0.11	0.00	0.00	0.48	0.05	0.00	0.00
one	0.11	0.00	0.00	0.48	0.05	0.00	0.00
soak	0.11	0.00	0.00	0.48	0.05	0.00	0.00
wet	0.11	0.00	0.00	0.48	0.05	0.00	0.00
not	0.00	0.11	0.00	0.48	0.00	0.05	0.00
mistake	0.00	0.11	0.00	0.48	0.00	0.05	0.00
former	0.00	0.11	0.00	0.48	0.00	0.05	0.00
samba	0.00	0.11	0.00	0.48	0.00	0.05	0.00
lake	0.00	0.11	0.00	0.48	0.00	0.05	0.00
difference	0.00	0.11	0.00	0.48	0.00	0.05	0.00
position	0.00	0.11	0.00	0.48	0.00	0.05	0.00
entrance	0.00	0.11	0.00	0.48	0.00	0.05	0.00

change	0.00	0.11	0.00	0.48	0.00	0.05	0.00
yes	0.00	0.00	0.25	0.48	0.00	0.00	0.12
great	0.00	0.00	0.25	0.48	0.00	0.00	0.12
photo	0.00	0.00	0.25	0.48	0.00	0.00	0.12
shot	0.00	0.00	0.25	0.48	0.00	0.00	0.12

TABEL 11
MATRIKS KEBINGUNGAN DARI SEMUA ALGORITMA

Model Classifier		Positive	Negative	Neutral
Naive Bayes	Precision	85,2%	58,9%	64,1%
	Recall	81,8%	82,1%	55,5%
	F1-Score	83,5%	68,6%	59,5%
K-Nearest Neighbor	Precision	77%	44,8%	52,5%
	Recall	79,5%	46,4%	46,6%
	F1-Score	78,2%	45,6%	49,4%
Support Vector Machine	Precision	64,9%	80%	100%
	Recall	99,2%	14,2%	2%
	F1-Score	78,5%	24,2%	4%
Decision Tree	Precision	82,8%	15,8%	100%
	Recall	22,8%	92,8%	22,2%
	F1-Score	35,8%	27%	4%

TABEL 12
AKURASI EMPAT ALGORITMA SUPERVISED LEARNING

Model	Accuracy
Naive Bayes	76.00%
K-Nearest Neighbor	67.50%
Support Vector Machine	65.50%
Decision Tree	28.00%

B. Pembahasan

Hasil penelitian menunjukkan bahwa proses penyesuaian bahasa, mulai dari konversi kata tidak baku dan dialek Bekasi ke dalam Bahasa Indonesia yang baku hingga terjemahan ke Bahasa Inggris, merupakan langkah krusial dalam pengolahan data untuk analisis sentimen berbasis regional dialect. Tahapan ini membutuhkan keterlibatan aktif dari ahli bahasa guna memastikan bahwa setiap konversi tetap menjaga makna dan konteks asli dari opini masyarakat. Validasi dilakukan secara iteratif agar hasil konversi benar-benar akurat dan dapat digunakan dalam tahap selanjutnya. Keberhasilan tahap ini tercermin dalam Tabel 2 dan Tabel 3, yang menunjukkan perubahan signifikan dari kata dan frasa tidak baku menjadi bentuk yang sesuai secara linguistik.

Setelah konversi dan anotasi, data dilabeli berdasarkan sentimen positif, negatif, dan netral. Proses ini penting untuk mendukung pelatihan model klasifikasi dalam tahap supervised learning. Berdasarkan Tabel 4, terlihat bahwa sentimen masyarakat terhadap pariwisata di Bekasi bervariasi. Sebagian besar opini bernada positif, tetapi tetap terdapat komentar negatif dan netral. Dominasi opini positif mengindikasikan bahwa sektor pariwisata mulai mendapatkan respons baik dari masyarakat pasca-pandemi, meskipun masih ada kekurangan yang perlu diperbaiki seperti fasilitas tempat ibadah dan perlindungan terhadap hujan.

Proses preprocessing berupa case folding, tokenizing, dan stemming bertujuan untuk menyederhanakan struktur data teks sebelum dilakukan ekstraksi fitur. Hasil dari tahapan ini disajikan pada Tabel 5 hingga Tabel 7, yang menunjukkan transformasi data menjadi format dasar yang lebih terstruktur dan siap untuk dibobotkan dengan metode TF-IDF. Proses pembobotan TF-IDF menghasilkan fitur yang digunakan dalam pelatihan model klasifikasi. Tabel 8 hingga Tabel 10 memperlihatkan bagaimana frekuensi kata dan bobot dihitung untuk setiap kelas sentimen. Penerapan TF-IDF mampu menyoroti kata-kata yang memiliki kontribusi penting terhadap klasifikasi sentimen.

Selanjutnya, empat algoritma pembelajaran terawasi diuji untuk mengetahui mana yang paling efektif dalam mengklasifikasikan opini masyarakat. Berdasarkan evaluasi performa model melalui confusion matrix (Tabel 11), Naive Bayes menunjukkan performa paling stabil dengan nilai presisi, recall, dan F1-score yang relatif tinggi untuk semua kelas. Hal ini menandakan bahwa Naive Bayes mampu menangkap pola distribusi kata dan sentimen secara proporsional. Sebaliknya, *Support Vector Machine* dan *Decision Tree* menunjukkan kelemahan signifikan, terutama dalam mengenali opini netral. Sementara itu, K-Nearest Neighbor menghasilkan performa sedang, namun tidak lebih unggul dari *Naive Bayes*.

Hasil pengukuran akurasi secara keseluruhan dalam Tabel 12 menunjukkan bahwa *Naive Bayes* mencapai akurasi tertinggi sebesar 76%, diikuti oleh *K-Nearest Neighbor* (67,5%), *Support Vector Machine* (65,5%), dan *Decision Tree* (28%). Performa rendah *Decision Tree* dapat disebabkan oleh overfitting atau ketidakmampuan

model dalam menangani keragaman bahasa dan struktur kalimat dari media sosial. Berdasarkan analisis ini, dapat disimpulkan bahwa *Naive Bayes* merupakan algoritma yang paling efektif untuk analisis sentimen berbasis dialek daerah, khususnya dalam kasus opini masyarakat terhadap pariwisata Bekasi. Keberhasilan model ini menunjukkan pentingnya tahapan validasi linguistik dan preprocessing dalam menciptakan sistem analitik yang andal untuk kebutuhan pengambilan keputusan berbasis opini publik.

IV. SIMPULAN

Penelitian ini menunjukkan bahwa proses konversi dan validasi dialek lokal, khususnya dialek Bekasi, merupakan tahap fundamental dalam analisis sentimen berbasis teks media sosial. Dengan melibatkan ahli bahasa, kata-kata tidak baku dan ekspresi lokal berhasil diubah menjadi bentuk baku yang dapat dianotasi dalam Bahasa Inggris tanpa kehilangan makna aslinya. Proses ini memungkinkan pelabelan sentimen dilakukan secara tepat dan sistematis. Tahap preprocessing seperti case folding, tokenizing, dan stemming, serta pembobotan dengan TF-IDF, mampu menghasilkan fitur yang relevan untuk proses klasifikasi. Di antara empat algoritma pembelajaran terawasi yang diuji, Naive Bayes terbukti paling efektif dengan akurasi tertinggi sebesar 76%, serta performa presisi, recall, dan F1-score yang lebih stabil dibandingkan algoritma lainnya. Berdasarkan temuan dan keterbatasan yang diidentifikasi dalam penelitian ini, terdapat beberapa arah pengembangan riset yang dapat dijadikan fokus dalam studi lanjutan. Pertama, perlu dilakukan pembangunan korpus anotasi yang lebih luas dan representatif yang mencakup berbagai dialek daerah di Indonesia. Keberadaan korpus semacam ini akan mendukung pelatihan model analisis sentimen yang lebih akurat dan kontekstual, khususnya dalam menangkap nuansa linguistik yang khas dari masing-masing wilayah. Kedua, disarankan untuk mengeksplorasi pengembangan sistem otomatisasi deteksi dialek (dialect detection automation) guna mengurangi ketergantungan terhadap proses validasi manual oleh ahli bahasa. Implementasi sistem ini dapat dilakukan dengan memanfaatkan pendekatan berbasis pembelajaran mesin atau representasi semantik melalui word embedding dan teknik klasifikasi berbasis karakteristik fonologis maupun morfologis dari teks berbahasa dialek. Ketiga, integrasi arsitektur jaringan saraf modern (neural architectures) seperti Long Short-Term Memory (LSTM) dan Bidirectional Encoder Representations from Transformers (BERT) direkomendasikan untuk meningkatkan kapabilitas klasifikasi model, terutama dalam menghadapi kompleksitas struktur sintaksis dan makna kontekstual yang tersembunyi dalam ujaran berbahasa dialek. Penggunaan model-model ini diharapkan dapat menghasilkan performa yang lebih stabil dan generalisasi yang lebih baik pada data tidak terstruktur. Keempat, penelitian lanjutan dapat diarahkan pada pengembangan sistem deteksi dan translasi dialek daerah secara otomatis dan real-time, yang dapat diintegrasikan ke dalam platform pemantauan opini publik berbasis media sosial. Sistem semacam ini akan sangat bermanfaat dalam mendukung pengambilan keputusan berbasis data linguistik dalam berbagai sektor, seperti pariwisata, pemerintahan, layanan publik, maupun pengembangan kebijakan berbasis masyarakat. Dengan demikian, perluasan cakupan penelitian melalui pendekatan linguistik-komputasional yang lebih canggih menjadi penting guna mewujudkan sistem analitik yang adaptif terhadap keberagaman bahasa dan budaya lokal Indonesia.

DAFTAR PUSTAKA

[1] Kemendikbud, "Badan Bahasa Petakan 652 Bahasa Daerah di Indonesia," *https://www.kemdikbud.go.id/main/blog/2018/07/badan-bahasa-petakan-652-bahasa-daerah-di-indonesia*, Jakarta, Jul. 2018.

[2] Jam'ul Ihsan Bambang *et al.*, "Kebebasan Berbicara di Media Sosial: Antara Regulasi dan Ekspresi," *Student Research Journal*, vol. 3, no. 1, pp. 87–96, Jan. 2025, doi: 10.55606/srj-yappi.v3i1.1692.

[3] H. Siregar, "Analisis Pemanfaatan Media Sosial Sebagai Sarana Sosialisasi Pancasila," *Pancasila: Jurnal Keindonesiaan*, pp. 71–82, Apr. 2022, doi: 10.52738/pjk.v2i1.102.

[4] M.L. Phyu and K. Hashimoto, "Sentiment Analysis of the Burmese Language Using n-Gram-Based Words," *CIC Express Letters*, vol. 13, no. 3, pp. 217–224, 2019, Accessed: Sep. 27, 2024. [Online]. Available: http://www.icicel.org/ell/contents/2019/3/el-13-03-06.pdf

[5] I. Erlin, H. Suliani, L. Asnal, Suryati, and R. Efendi, "Sentiment Analysis for Abolition of National Exams in Indonesia using Support Vector Machine," *Engineering Letters*, vol. 30, no. 4, pp. 1342–1352, 2022, Accessed: Sep. 27, 2024. [Online]. Available: https://www.engineeringletters.com/issues_v30/issue_4/EL_30_4_19.pdf

[6] F. Ahluna, C. Joines Tutuarima, and I. Santoso, "Metode K-Nearest Neighbor Untuk Analisis Sentimen Tentang Penghapusan Ujian Nasional," *Jurnal IKRAITH-INFORMATIKA*, vol. 7, no. 2, pp. 1–6, 2023, [Online]. Available: https://www.instagram.com/p/B599fcnFxXR

[7] S. Sidorov, A. Faizliev, and V.Balash, "Fractality and Multifractality Analysis of News Sentiments Time Series," *IAENG International Journal of Applied Mathematics*, vol. 48, no. 1, pp. 90–97, 2018, Accessed: Sep. 27, 2024. [Online]. Available: https://www.iaeng.org/IJAM/issues_v48/issue_1/IJAM_48_1_13.pdf

[8] X. Jiawei and T. Murata, "Stock Market Trend Prediction with Sentiment Analysis based on LSTM Neural Network," in *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019*, Hongkong, 2019, pp. 475–479. Accessed: Sep. 27, 2024. [Online]. Available: https://www.iaeng.org/publication/IMECS2019/IMECS2019_pp475-479.pdf

[9] A. Kumar and A. Jaiswal, "Empirical Study of Twitter and Tumblr for Sentiment Analysis using Soft Computing Techniques," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2017*, San Fransisco, 2017, pp. 472–476.

[10] S. Elbagir and J. Yang, "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment," in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019*, International Association of Engineers, 2019, p. 575. Accessed: Sep. 27, 2024. [Online]. Available: https://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf

[11] Z. C. Dwinnie and R. Novita, "Penerapan Machine Learning Pada Analisis Sentimen Twitter Sebelum dan Sesudah Debat Calon Presiden dan Wakil Presiden Tahun 2024," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 758, Apr. 2024, doi: 10.30865/mib.v8i2.7504.

[12] D. R. Lazuardi, T. A. Munandar, H. Harsiti, Z. Mutaqin, and R. N. Hays, "Sentiment analysis of public opinions on the welfare of honorary educators using Naive Bayes," *IOP Conf Ser Mater Sci Eng*, vol. 830, no. 3, p. 032018, Apr. 2020, doi: 10.1088/1757-899X/830/3/032018.

[13] T. Meidiyanti Fadli and A. S. Puspaningrum, "Analisis Sentimen terhadap Kenaikan Gaji Guru Honorer Menggunakan Algoritma Naïve Bayes," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 3, pp. 635–644, Mar. 2025, doi: 10.52436/1.jpti.689.

[14] V. Rohini, M. Thomas, and C.A. Latha, "Domain Based Sentiment Analysis in Regional Languange-Kannada," *International Journal Of Engineering Research & Technology (IJERT)*, vol. 4, no. 22, 2016.

[15] H. Al Suwaidi, T.R. Soomro, and K. Shaalan, "Sentiment Analysis for Emiriti Dialects in Twitter," *Sindh University Research Journal (Science Series*, vol. 48, no. 4, pp. 707–710, 2016.

[16] R. M. Duwairi, "Sentiment analysis for dialectical Arabic," in *2015 6th International Conference on Information and Communication Systems (ICICS)*, IEEE, Apr. 2015, pp. 166–170. doi: 10.1109/IACS.2015.7103221.

[17] A. Chader, D. Lanasri, L. Hamdad, M. Belkheir, and W. Hennoune, "Sentiment Analysis for Arabizi: Application to Algerian Dialect," in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, SCITEPRESS - Science and Technology Publications, 2019, pp. 475–482. doi: 10.5220/0008353904750482.

[18] M.H.S.M. Amin, O. Al-Rassam, and Z.S. Faeq, "Kurdish Language Sentiment Analysis: Problems and Challenges," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 4, pp. 3282–3293, 2022.

[19] J. O. Atoum and M. Nouman, "Sentiment Analysis of Arabic Jordanian Dialect Tweets," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019, doi: 10.14569/IJACSA.2019.0100234.

[20] S. Medhaffar, F. Bougares, Y. Estève, and L. Hadrich-Belguith, "Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 55–61. doi: 10.18653/v1/W17-1307.

[21] A. B. Braiek and Z. N. Ben Salem, "Sentiment Analysis Classification for Text in Social Media: Application to Tunisian Dialect," *International Journal on Cybernetics & Informatics*, vol. 12, no. 2, pp. 313–326, Mar. 2023, doi: 10.5121/ijci.2023.120223.

[22] A. Alqarni and A. Rahman, "Arabic Tweets-Based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 16, Jan. 2023, doi: 10.3390/bdcc7010016.

[23] N. Khurana, "Sentiment Analysis of Regional Languages Written in Roman Script on Social Media," *Data Science and Intelligent Applications, Lecture Notes on Data Engineering and Communications Technologies*, vol. 52, pp. 113–119, 2021, doi: 10.1007/978-981-15-4474-3_13.

[24] P. Shah, P. Swaminarayan, M. Patel, and N. Patel, "Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm," *International Journal of Engineering Trends and Technology*, vol. 70, no. 1, pp. 313–326, Jan. 2022, doi: 10.14445/22315381/IJETT-V70I1P236.

[25] K. Rakshitha, R. H M, M. Pavithra, A. H D, and M. Hegde, "Sentimental analysis of Indian regional languages on social media," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 414–420, Nov. 2021, doi: 10.1016/j.gltp.2021.08.039.

[26] M.B. Shelke and S.N. Deshmukh, "Recent Advances in Sentiment Analysis of Indian Languages," *International Journal of Future Generation Communication and Networking* , vol. 13, no. 4, pp. 1656–1675, 2020, Accessed: Sep. 27, 2024. [Online]. Available: http://sersc.org/journals/index.php/IJFGCN/article/view/32962/18184

[27] B. Hdioud and M. E. H. Tirari, "Sentiment Analysis of Moroccan Dialect Using Deep Learning," in *Proceedings of the 5th International Conference on Big Data and Internet of Things. BDIoT 2021. Lecture Notes in Networks and Systems, Vol 489*, Springer, 2022, pp. 457–466. doi: 10.1007/978-3-031-07969-6_34.

[28] M. Dahbi, R. Saadane, and S. Mbarki, "Citizen Sentiment Analysis in Social Media Moroccan Dialect as Case Study," in *Innovations in Smart Cities Applications Edition 3*, 2020, pp. 16–29. doi: 10.1007/978-3-030-37629-1_2.

[29] Y. Matrane, F. Benabbou, and N. Sael, "Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case," in *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*, IEEE, Jun. 2021, pp. 80–87. doi: 10.1109/ICDATA52997.2021.00024.

[30] P. Ranjitha and K. N. Bhanu, "Improved Sentiment Analysis for Dravidian Language-Kannada Using Dicision Tree Algorithm With Efficient Data Dictionary," *IOP Conf Ser Mater Sci Eng*, vol. 1123, no. 1, p. 012039, Apr. 2021, doi: 10.1088/1757-899X/1123/1/012039.

[31] A.J.V. Boquiren, R.A. Garcia, C.J.D. Hungria, and J.C. de Goma, "Tagalog Sentiment Analysis Using Deep Learning Approach With Backward Slang Inclusion," in *Proceedings of the International Conference on Industrial Engineering and Operations Management Nsukka*, Nigeria, Apr. 2022.

[32] A. Chaer and I. Agustina, *Sosiolinguistik: Perkenalan Awal*, 1st ed. Jakarta: Rineka Cipta Publisher, 1995.

[33] G. Yule, *The Study of Language: Third Edition*. NewYork: Cambridge University Press, 2006.

[34] hawaii.edu, "Language Varieties: Definisions of Different Kinds of Languate Varieties," https://www.hawaii.edu/satocenter/langnet/definitions/index.html#regional, access at August 05, 2024.

[35] M. Farhadloo and E. Rolland, "Fundamentals of Sentiment Analysis and Its Applications," in *Sentiment Analysis and Ontology Engineering* , 2016, pp. 1–24. doi: 10.1007/978-3-319-30319-2_1.

[36] Z. Drus and H. Khalid, "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review," *Procedia Comput Sci*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[37] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Comput Sci*, vol. 8, p. e914, Mar. 2022, doi: 10.7717/peerj-cs.914.

[38] F. Karabiber, "Term Frequency-Inverse Document Frequency," https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency, accessed on August 07, 2024.

[39] J. Han and M. Kamber, *Data Mining Concepts and Techniques* , Third Edition. San Francisco: Morgan Kaufmann Publishers, 2012.

[40] J. Han and M. Kamber, *Data Mining: Concepts and Technique*, Second Edition. San Francisco: Morgan Kaufmann Publishers, 2006.
[41] D.T. Larose, *Data Mining Methods and Models* . New Jersey: John Wiley & Sons, Inc, 2006.
[42] S. Styawati and K. Mustofa, "A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 3, p. 219, Jul. 2019, doi: 10.22146/ijccs.41302.