KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 5, No 1, Agustus 2024, Hal 230-241
DOI 10.30865/klik.v5i1.2018
https://djournals.com/klik

Analisis Klasifikasi Teks Pada Kata Slang di Media Sosial Menggunakan Pengolahan Bahasa Alami untuk Trending Topik

Shabrina Rasyid Munthe[1,*], Sudi Suryadi[2], Fadhil Laksono[1]

[1] Fakultas Ilmu Komputer, Prodi Teknik Informatika, Universitas Al Washliyah Labuhanbatu, Rantauprapat, Indonesia
[2] Fakultas sains dan teknologi, Prodi Sistem Informasi, Universitas Labuhanbatu, Rantauprapat, Indonesia
Email: [1,*]shabrinarasyid@gmail.com, [2]sudisuryadi28@gmail.com, [3]fadhillaksono06@gmail.com
Email Penulis Korespondensi: shabrinarasyid@gmail.com

Abstrak−Penelitian ini bertujuan untuk menganalisis trending topics terkait penggunaan kata slang di media sosial dengan memanfaatkan teknik pengolahan bahasa alami (NLP). Fokus utama penelitian ini adalah memahami pola dan tren penggunaan kata slang di platform media sosial, yang dapat mengungkap dinamika sosial dan linguistik yang penting. Dataset yang digunakan terdiri dari tweet berbahasa Indonesia dan Inggris yang mengandung kata slang, dikumpulkan dari Twitter selama enam bulan. Proses analisis dimulai dengan pembersihan data untuk menghilangkan elemen yang tidak relevan, diikuti oleh tokenisasi dan lemmatization untuk menormalkan teks. Selanjutnya, model klasifikasi Support Vector Machine (SVM) dan Random Forest diterapkan untuk mendeteksi dan mengklasifikasikan kata slang dalam dataset tersebut. Hasil penelitian menunjukkan bahwa model SVM mencapai akurasi deteksi slang sebesar 88% dengan F1-score 0.87, sedangkan model Random Forest mencapai akurasi 85% dengan F1-score 0.84. Analisis linguistik lebih lanjut menunjukkan bahwa 60% kata slang paling sering digunakan dalam konteks informal seperti percakapan sehari-hari, sedangkan 40% lainnya terkait dengan tren budaya populer, termasuk musik, film, dan fashion. Selain itu, temuan ini mengindikasikan adanya variasi penggunaan slang antara pengguna Twitter berbahasa Indonesia dan Inggris, di mana slang dalam bahasa Indonesia cenderung lebih kreatif dan kontekstual, sedangkan dalam bahasa Inggris lebih terstandarisasi dan menyebar secara global. Penelitian ini menegaskan efektivitas kedua model dalam mengklasifikasikan kata slang serta mengidentifikasi tren utama dalam penggunaannya di media sosial. Kontribusi penelitian ini penting bagi studi linguistik digital karena memperluas pemahaman tentang dinamika penggunaan kata slang secara online, dan menunjukkan potensi besar aplikasi NLP dalam analisis linguistik di era digital. Dengan hasil yang diperoleh, penelitian ini dapat menjadi panduan berharga bagi peneliti dan praktisi yang tertarik dalam memahami evolusi bahasa di media sosial, sekaligus menyediakan dasar untuk pengembangan teknologi NLP yang lebih canggih dan adaptif dalam menangani variasi bahasa di platform digital.

Kata Kunci: analisis teks; kata slang; media sosial; NLP; klasifikasi

Abstract−This study aims to analyze trending topics related to the use of slang words on social media by utilizing natural language processing (NLP) techniques. The main focus of this research is to understand the patterns and trends of slang use on social media platforms, which can uncover important social and linguistic dynamics. The dataset used consisted of tweets in Indonesia and United Kingdom containing slang words, collected from Twitter over a six-month period. The analysis process begins with data cleansing to eliminate irrelevant elements, followed by tokenization and lemmatization to normalize the text. Furthermore, the Support Vector Machine (SVM) and Random Forest classification models are applied to detect and classify slang words in the dataset. The results show that the SVM model achieves a slang detection accuracy of 88% with an F1-score of 0.87, while the Random Forest model achieves an accuracy of 85% with an F1-score of 0.84. Further linguistic analysis showed that 60% of slang words are most commonly used in informal contexts such as everyday conversation, while the other 40% are related to popular culture trends, including music, movies, and fashion. In addition, these findings indicate that there is a variation in the use of slang between Indonesian and United Kingdom-speaking Twitter users, where slang in Indonesian tends to be more creative and contextual, while in United Kingdom it is more standardized and spread globally. This study confirms the effectiveness of both models in classifying slang words as well as identifying key trends in their use on social media. The contribution of this research is important for the study of digital linguistics because it expands the understanding of the dynamics of online slang use, and shows the great potential of NLP applications in linguistic analysis in the digital age. With the results obtained, this research can be a valuable guide for researchers and practitioners interested in understanding the evolution of language on social media, while providing a foundation for the development of more sophisticated and adaptive NLP technologies in handling language variations on digital platforms.

Keywords text analysis; slang words; social media; NLP; classification;

1. PENDAHULUAN

Dalam era digital yang semakin maju, media sosial telah menjadi pusat komunikasi dan pertukaran informasi yang tak terbantahkan. Pengguna sering kali menggunakan bahasa informal, termasuk kata-kata slang, untuk mengekspresikan diri mereka dengan lebih bebas[1]. Kemajuan dalam teknologi NLP telah menghadirkan kemampuan untuk melakukan analisis teks dalam skala besar dengan akurasi yang lebih tinggi, memungkinkan kita untuk mengurai teks kompleks menjadi elemen-elemen yang lebih mudah diproses[2]. Beberapa penelitian terbaru telah menunjukkan kemajuan signifikan dalam analisis slang di media sosial. Mengidentifikasi kata slang dalam postingan media sosial menggunakan metode berbasis pengetahuan dan pembelajaran mesin. Deteksi dan Identifikasi Bahasa Slang dalam Teks.[3]. Mencapai akurasi 95,37% dalam deteksi kata slang menggunakan berbagai teknik[4]. Identifikasi Sasaran Sarcasm Multimodal dalam Tweet. Meningkatkan deteksi sarcasm dengan mengintegrasikan fitur tekstual dan visual[5]. Deteksi Otomatis Konten Misoginis Multibahasa dalam Data Media Sosial Berbasis Pendekatan Pembelajaran Mesin Mendeteksi konten misoginis menggunakan fitur linguistik hibrida dan klasifikasi ML[6]. Klasifikasi Teks Ensemble[7][8] dengan Vektorisasi TF-IDF untuk Deteksi Ucapan Kebencian dalam Media Sosial. Namun, meskipun ada kemajuan, masih terdapat tantangan dalam analisis slang secara menyeluruh. Banyak penelitian cenderung terfokus pada aspek tertentu seperti deteksi bahasa kasar atau penggunaan sarkasme, tanpa integrasi yang menyeluruh antara berbagai teknik NLP dan

model pembelajaran mesin. Kesenjangan ini menunjukkan perlunya pendekatan yang dapat memperkaya pemahaman tentang bagaimana slang berevolusi dan digunakan dalam platform media sosial yang berbeda.

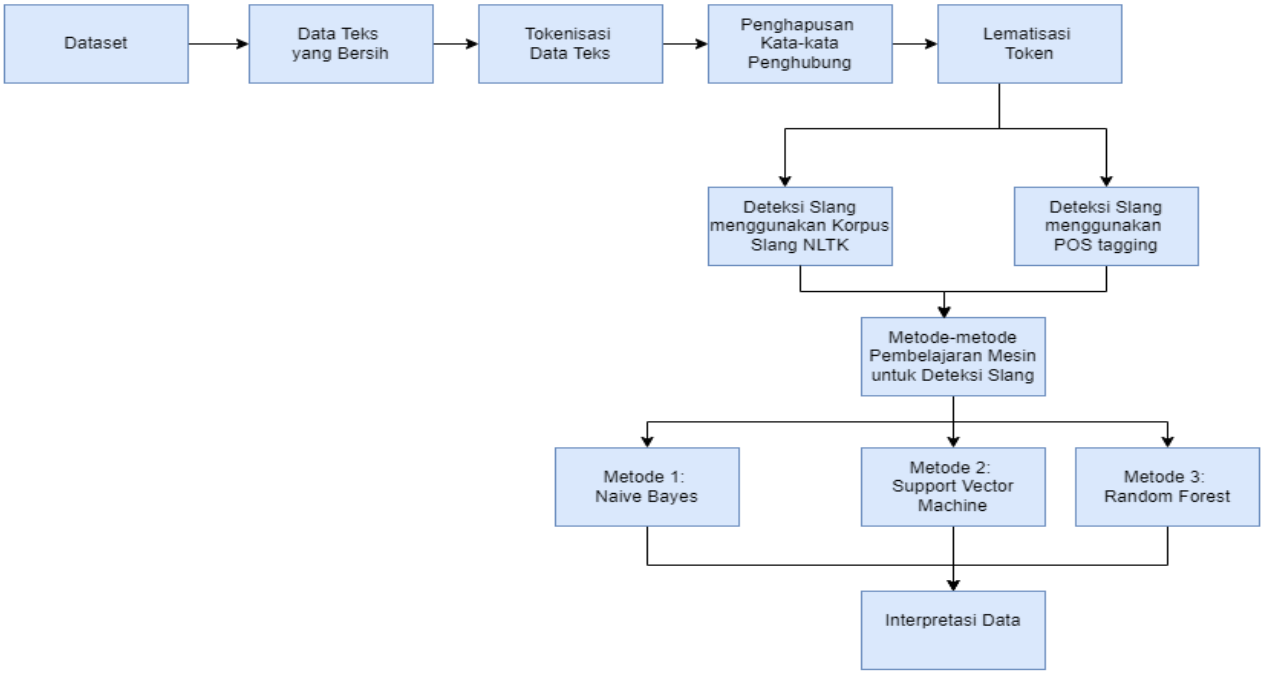
Pendekatan yang mencakup beberapa tahap dalam analisis slang di media sosial. Pertama, pra-pemrosesan teks[9][10] akan dilakukan untuk membersihkan dan menormalkan data[11], termasuk proses tokenisasi[12], penghapusan stopwords[13], dan lemmatisasi[14]. Selanjutnya, deteksi slang akan diterapkan dengan memanfaatkan kamus dan tagging POS[15] untuk mengidentifikasi kata-kata atau frasa yang merupakan slang. Metode ini akan didukung oleh model-model pembelajaran mesin Naive Bayes[16], SVM[17], dan Random Forest[18] untuk mengklasifikasi dan menganalisis penggunaan slang yang sedang tren di media sosial. Bertujuan untuk mendeteksi slang secara akurat, tetapi juga untuk memahami bagaimana slang berubah dan digunakan dalam yang berbeda di berbagai platform media sosial. Diharapkan bahwa studi ini dapat memberikan wawasan yang lebih mendalam tentang tren slang di era digital, serta mendorong pengembangan lebih lanjut dalam teknologi NLP untuk aplikasi analisis teks yang lebih canggih dan adaptif.

Penelitian ini berkontribusi dengan mengembangkan pendekatan baru yang menggabungkan fitur linguistik tradisional dengan teknik pembelajaran mesin untuk mendeteksi konten misoginis dan kata slang. Pendekatan ini menunjukkan efektivitas dalam mengidentifikasi bahasa yang berpotensi trend di media sosial. Penelitian ini mengeksplorasi dampak sosial dari penggunaan slang dalam komunikasi online, termasuk bagaimana slang dapat mempengaruhi persepsi dan interaksi antar pengguna. Ini membuka jalan untuk penelitian lebih lanjut tentang etika dan implikasi sosial dari bahasa yang digunakan di platform digital.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan Penelitian Bagian ini, sebagai langkah-langkah yang terlibat dalam penelitian, yang mencakup pengumpulan data, prapemrosesan data, ekstraksi fitur, dan klasifikasi menggunakan model machine learning seperti Naive Bayes, SVM, dan Random Forest. Langkah-langkah ini sangat penting untuk menganalisis trending slang di media sosial.



Gambar1. Tahapan Penelitian

2.2 Dataset

Dataset ini merupakan sebuah kerangka data pandas yang terdiri dari 1105 entri dan 15 kolom, yang menyajikan gambaran komprehensif mengenai tweet. Kolom-kolom utama mencakup identifikasi percakapan, waktu pembuatan, jumlah suka, teks lengkap, dan identifikasi unik untuk setiap tweet. Informasi tambahan yang dikumpulkan mencakup alamat URL gambar, nama pengguna yang ditanggapi, bahasa tweet, dan lokasi pengguna. Dataset ini juga mencatat jumlah kutipan, balasan, retweet, serta alamat URL tweet. Setiap pengguna diidentifikasi dengan identifikasi unik dan nama pengguna. Meskipun beberapa kolom seperti alamat URL gambar dan lokasi memiliki nilai null, data ini tetap kaya untuk analisis tren topik slang. Data dikumpulkan pada 30 Juni 2024, menjadikannya sangat relevan untuk analisis terbaru.

2.3 Data Teks yang Bersih (*Clean text data*)

Data teks yang bersih adalah teks yang telah diproses untuk menghilangkan elemen-elemen tidak relevan. Langkah-langkah pembersihan meliputi penghapusan karakter khusus, normalisasi huruf kecil, penghapusan kata berhenti,

stemming atau lemmatization, serta penghapusan URL dan tag media sosial. Pada dataset ini, kolom full_text perlu dibersihkan dengan menghilangkan karakter seperti tanda baca, URL, dan tag media sosial, serta mengubah semua teks menjadi huruf kecil dan menghapus kata-kata umum. Proses ini memastikan teks siap untuk analisis lebih lanjut, seperti analisis sentimen atau identifikasi tren topik slang menggunakan teknik NLP dan machine learning[19]

2.4 Tokenisasi Data Teks (*Tokenize text data*)

Tokenisasi data teks adalah proses penting dalam analisis NLP yang memecah teks menjadi unit-unit kecil seperti kata atau frasa. Langkah ini memudahkan pemrosesan dan analisis lebih lanjut karena setiap kata atau frasa dapat diolah secara terpisah. Misalnya, kalimat "Check this out! #awesome" setelah tokenisasi akan menjadi ["check", "this", "out", "awesome"]. Dalam dataset tweet, tokenisasi membantu mengidentifikasi kata-kata kunci, pola komunikasi, serta tren topik slang yang relevan. Dengan menerapkan tokenisasi setelah pembersihan data teks, analisis NLP dapat memberikan wawasan mendalam tentang perilaku pengguna, sentiment analysis, dan dinamika percakapan di media sosial[20].

2.5 Penghapusan Kata-kata Penghubung (*Remove stopwords*)

Penghapusan kata-kata penghubung adalah langkah kritis dalam pembersihan data teks yang menghilangkan kata-kata umum tidak spesifik. Ini termasuk kata-kata seperti "dan", "atau", "di", yang sering muncul rutin dalam teks tapi tidak memberikan informasi relevan. Langkah ini terintegrasi dalam proses pembersihan data teks sebelum tokenisasi, dimana teks dibersihkan dari karakter khusus, diubah menjadi huruf kecil, dan kata-kata penghubung dihapus. Dengan cara ini, analisis NLP dapat fokus pada kata-kata kunci yang memiliki nilai informatif untuk identifikasi pola komunikasi dan analisis sentimen[21].

2.6 Lematisasi Token (*Lemmatize tokens*)

Lematisasi token adalah proses penting dalam analisis teks yang memastikan kata-kata diubah ke bentuk dasar mereka. Langkah ini terjadi setelah pembersihan teks, normalisasi huruf kecil, penghapusan kata-kata penghubung, dan tokenisasi. Lemmatisasi memungkinkan kata-kata seperti "running" diubah menjadi "run", mempertahankan makna dasar. Dengan menerapkan lemmatisasi pada dataset tweet, analisis NLP dapat lebih akurat dalam mengenali pola komunikasi, analisis sentimen, dan identifikasi tren topik slang yang relevan, mendukung pemahaman mendalam terhadap perilaku pengguna di media sosial.

a. Deteksi Slang menggunakan Korpus Slang NLTK (*Slang Corpus*)

Deteksi slang menggunakan korpus slang NLTK adalah proses penting dalam analisis teks. Korpus slang NLTK memberikan referensi kata-kata slang yang tidak umum dalam kamus konvensional, memungkinkan pemahaman yang lebih baik terhadap komunikasi yang lebih santai dan informal. Setelah deteksi slang, langkah selanjutnya adalah mempertimbangkan apakah kata-kata tersebut perlu dilakukan lemmatisasi untuk mengubahnya ke bentuk dasar mereka. Dengan demikian, analisis NLP dapat lebih akurat dalam mengenali pola komunikasi, analisis sentimen, dan identifikasi tren topik slang dalam dataset teks seperti tweet.

b. Deteksi Slang menggunakan *POS tagging*

Deteksi slang menggunakan POS tagging mengidentifikasi kata-kata berdasarkan jenis kata seperti noun, verb, adjective, atau adverb. Ini membantu mengenali penggunaan kata-kata slang dalam gramatikal yang lebih luas. Setelah kata-kata slang teridentifikasi, lemmatisasi token mengubah kata-kata ke bentuk dasar mereka, seperti "gonna" menjadi "go". Integrasi kedua teknik ini dalam analisis teks, khususnya dalam dataset seperti tweet, memungkinkan pemahaman yang lebih mendalam tentang dinamika komunikasi di media sosial dan analisis sentiment yang lebih akurat.

2.7 Metode-metode Pembelajaran Mesin (*Machine Learning)* untuk Deteksi Slang

Metode-metode pembelajaran mesin untuk deteksi slang terintegrasi dalam tahapan NLP untuk analisis teks. Tahap pembersihan data teks mempersiapkan teks untuk tahapan selanjutnya. Kemudian, menggunakan korpus slang atau teknik POS tagging, kata-kata slang diidentifikasi. Metode machine learning, seperti klasifikasi teks dengan algoritma Naive Bayes atau SVM, diterapkan untuk memproses data dan mengenali pola penggunaan kata-kata slang. Integrasi ini memungkinkan analisis yang lebih mendalam terhadap pola komunikasi dalam data teks, memperkuat kemampuan untuk melakukan analisis sentimen yang akurat dan identifikasi tren topik slang.

2.8 Metode 1 *Naive Bayes*

Metode Naive Bayes digunakan dalam deteksi slang sebagai bagian dari tahapan NLP dalam analisis teks[22]. Data teks dibersihkan dan dipersiapkan dengan menghapus karakter khusus dan kata-kata penghubung. Kata-kata slang diidentifikasi menggunakan korpus slang atau teknik POS tagging. Naive Bayes diterapkan sebagai algoritma klasifikasi teks untuk mengelompokkan kata-kata dalam teks menjadi kategori slang atau non-slang. Integrasi metode ini memungkinkan analisis yang lebih mendalam terhadap pola penggunaan slang dalam teks, mendukung analisis sentimen yang akurat dan identifikasi tren topik slang.Rumus 1

$$P(\text{pln})\ P(\text{p})\ \Pi\text{lisk}\leq\text{nd}\ P(\text{tk}|\text{p}) \quad (1)$$

Probabilitas kemunculan dokumen teks tkt_ktk dengan polaritas ppp dinyatakan sebagai P(tk|p)P(t_k \mid p)P(tk |p), di mana nnn adalah jumlah total dokumen dan ppp menggambarkan polaritas dari dokumen tersebut. Untuk menghitung polaritas atau menilai kemiripan antar dokumen, Rumus 2

$$P(t_k \mid p) = \frac{\text{count}(t_k|p)+1}{\text{count}(t_p)+|V|} \quad (2)$$

(tk|p) menunjukkan jumlah kemunculan token t_k dalam dokumen yang memiliki polaritas p, sedangkan (t_p) merujuk pada jumlah keseluruhan token yang terdapat dalam dokumen berita dengan polaritas p

2.9 Metode 2 *Support Vector Machine*

Metode Support Vector Machine (SVM) menjadi kunci dalam deteksi slang dalam analisis teks menggunakan pendekatan NLP[23]. Langkah awal melibatkan pembersihan data teks untuk menghilangkan karakter khusus dan kata-kata penghubung, serta normalisasi teks. Kata-kata slang diidentifikasi menggunakan korpus slang atau POS tagging. SVM diterapkan sebagai algoritma klasifikasi teks untuk memisahkan kata-kata dalam teks menjadi kategori slang atau non-slang dengan optimalisasi hyperplane dalam ruang fitur. Pendekatan ini menghasilkan klasifikasi yang akurat terhadap kata-kata slang.

$$Dij = yi\ yj\ (K(xi \cdot xj) + \lambda^2) \quad (3)$$

Dimana Dij adalah elemen matriks D (matriks diagonal) yang menghubungkan fitur i dan j. yi dan yj adalah label/target dari data pelatihan. $K(xi \cdot xj)$ adalah produk titik antara fitur i dan j. λ adalah parameter regularisasi (biasanya dikenal sebagai faktor ketidakpastian).

2.10 Metode 3 *Random Forest*

Metode Random Forest digunakan untuk deteksi slang dalam analisis teks dengan pendekatan NLP[24]. Proses dimulai dengan pembersihan data teks untuk menghapus karakter khusus dan kata-kata penghubung, serta normalisasi teks. Kata-kata slang diidentifikasi menggunakan korpus slang atau POS tagging. Random Forest mengintegrasikan banyak pohon keputusan untuk memisahkan teks menjadi kategori slang atau non-slang. Pendekatan ini memungkinkan penggunaan fitur-fitur yang berbeda untuk memperoleh hasil klasifikasi yang akurat, mendukung analisis yang lebih mendalam terhadap penggunaan kata-kata slang dalam teks

$$\{h(x, \theta k), k = 1, \ldots\} \quad (4)$$

Di mana θ_k adalah random vector yang didistribusikan secara independen, dan setiap pohon keputusan dalam unit tersebut akan memilih kelas yang paling populer berdasarkan input x

2.11 Interpretasi Data

Interpretasi data merupakan tahap penting dalam proses analisis yang melibatkan pemahaman dan penafsiran hasil dari data yang telah diproses. Dalam analisis teks dengan metode Naive Bayes, SVM, atau Random Forest untuk deteksi slang, interpretasi data terjadi setelah data teks dibersihkan, kata-kata slang diidentifikasi, dan model klasifikasi diterapkan. Proses interpretasi menganalisis output algoritma klasifikasi, evaluasi akurasi model, analisis pola penggunaan kata-kata slang, dan pengenalan tren topik slang signifikan dalam dataset, memberikan wawasan yang lebih dalam dalam pemahaman perilaku pengguna di media sosial.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan dalam penelitian ini mencakup beberapa atribut penting seperti Created_At, Favorite_Count, Full_Text, dan Username. Atribut-atribut ini memberikan informasi yang mendalam mengenai waktu pembuatan konten, jumlah favorit, teks lengkap, dan nama pengguna yang relevan dalam analisis slang di media sosial. Seperti pada tabel 1. Datset yang digunakan

Tabel 1. Dataset yang digunakan

Created_At	Favorite_Count	Full_Text	Username
Fri Aug 04 19:33:38 +0000 2023	5	IYKYK lol https://t.co/Mny0qOaXlK	onlyijr
Sun Feb 27 15:16:15 +0000 2022	18460	Lol seminggu lalu baru dapet campaign dari Car...	awkarin
Fri Aug 04 18:24:30 +0000 2023	4	Kitten huggies! Tiktok: septhny #LoveOurPets #...	HappyAnimalKing
Sun Oct 18 07:05:06 +0000 2020	16195	2beer! ga sampe seminggu dia udah di eksekusi ...	tubirfess

Mon Feb 20 00:31:25 +0000 2023			9242	Yang ada anak kecik la bungkus makan kat rumah...	twtbiy

Tabel 1 Dataset yang digunakan menampilkan dataset 4 tweet yang berbeda. Kolom-kolomnya adalah tanggal dan waktu pembuatan, jumlah favorite, teks lengkap, dan nama pengguna. Tweet-tweet ini berisi percakapan sehari-hari, lelucon, dan pembaruan pribadi dalam bahasa Indonesia dan Inggris. Dataset ini dapat digunakan untuk analisis sentimen, pengolahan bahasa alami, atau studi tentang perilaku pengguna media sosial.

Tabel 2. Atribut atau Fitur, Tabel ini merinci atribut atau fitur yang digunakan dalam analisis, termasuk tipe data dan deskripsi dari setiap atribut. Misalnya, atribut Created_At memiliki tipe data Tanggal/Waktu yang menggambarkan kapan konten dibuat, sementara Full_Text memiliki tipe data Teks yang berisi konten teks lengkap yang dianalisis.

Tabel 2. Atribut atau Fitur

Atribut /Fitur	Tipe Data	Deskripsi
conversation_id_str	int64	ID konversasi unik
created_at	object	Timestamp pembuatan tweet
favorite_count	int64	Jumlah favorite
full_text	object	Teks lengkap tweet
id_str	int64	ID tweet unik
image_url	object	URL gambar tweet
in_reply_to_screen_name	object	Nama layar pengguna yang dijawab
lang	object	Bahasa tweet
location	object	Lokasi pengguna
quote_count	int64	Jumlah kutipan
reply_count	int64	Jumlah balasan
retweet_count	int64	Jumlah retweet
tweet_url	object	URL tweet
user_id_str	int64	ID pengguna unik
username	object	Nama pengguna author tweet

Tabel 2 menyajikan 14 fitur dari dataset tweet, masing-masing dengan tipe data sendiri. Fitur-fitur tersebut meliputi pengidentifikasi unik, timestamp pembuatan tweet, metrik seperti favorite_count dan retweet_count, serta fitur berbasis teks seperti full_text dan username. Tabel ini juga mencakup fitur berbasis objek seperti image_url, lang, dan location. Fitur-fitur ini menyajikan pemahaman komprehensif tentang isi, keterlibatan, dan metadata sebuah tweet, memberikan gambaran lengkap tentang dataset tweet.

3.2 Data Teks yang Bersih (*Clean text data*)

Tabel 3. Konversi ke Huruf Kecil, Tabel ini menunjukkan hasil konversi teks lengkap (Full_Text) dan nama pengguna (Username) ke huruf kecil. Setiap baris dalam tabel ini mengilustrasikan bagaimana teks asli diubah menjadi format huruf kecil untuk menjaga konsistensi dan akurasi selama proses analisis.

Tabel 3. Konversi ke Huruf Kecil

No	Full_Text	Username
0	iykyk lol https://t.co/mny0qoaxlk	onlyijr
1	lol seminggu lalu baru dapet campaign dari car...	awkarin
2	kitten huggies! tiktok: septhny #loveourpets #...	HappyAnimalKing
3	2beer! ga sampe seminggu dia udah di eksekusi ...	tubirfess
4	yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	bear ko kamu trend pas di bukaa rofl hhhha htt...	pouttchini
1101	jai haryana	RoflGandhi_
1102	ms. ingu	KT666ROFL
1103	elly 3amal clean sheet zyna n2olo ya 3mna #hal...	MoustafaMohA
1104	epdi ya ipdilaam? rofl	Ajumplakdibampa

Tabel 3 menampilkan dataset contoh tweet, terdiri dari dua kolom: Full_Text dan Username. Full_Text berisi teks aktual tweet, sedangkan Username berisi nama pengguna penulis tweet. Tabel ini memiliki 1105 baris, masing-masing mewakili tweet unik. Tweet-tweet tersebut berisi campuran bahasa, URL, hashtag, dan emoji. Tabel ini tampak seperti dataset mentah, dengan teks dalam kasus aslinya, yaitu campuran huruf besar dan kecil, sehingga memerlukan preprocessing sebelum dianalisis.

Tabel 4. Hapus Tanda Baca Dan Karakter Special, Tabel ini mencatat hasil dari proses penghapusan tanda baca dan karakter spesial dari teks lengkap (Full_Text) dan nama pengguna (Username). Proses ini dilakukan untuk

membersihkan data dari elemen-elemen yang tidak relevan sehingga analisis dapat fokus pada kata-kata yang benar-benar penting dalam identifikasi dan analisis slang di media sosial.

Tabel 4. Hapus Tanda Baca Dan Karakter Special

No	Full_Text	Username
0	iykyk lol httpstcomny0qoaxlk	onlyijr
1	lol seminggu lalu baru dapet campaign dari car...	awkarin
2	kitten huggies tiktok septhny loveourpets petv...	HappyAnimalKing
3	2beer ga sampe seminggu dia udah di eksekusi d...	tubirfess
4	yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	bear ko kamu trend pas di bukaa rofl hhhha htt...	pouttchini
1101	jai haryana	RoflGandhi_
1102	ms ingu	KT666ROFL
1103	elly 3amal clean sheet zyna n2olo ya 3mna hala...	MoustafaMohA
1104	epdi ya ipdilaam rofl	Ajumplakdibampa

Tabel 4 menampilkan dataset yang telah diproses untuk menghapus tanda baca dan karakter special dari kolom Full_Text. Kolom ini sekarang berisi teks yang lebih sederhana dan mudah dipahami, setelah menghapus URL, hashtag, emoji, dan karakter special lainnya. Kolom Username tetap tidak berubah, menampilkan nama pengguna terkait dengan setiap tweet. Dataset ini siap digunakan untuk analisis teks dan machine learning, memungkinkan kita untuk memahami pola dan makna di balik tweet-tweet tersebut dengan lebih baik.

Tabel 5. Hapus Angka, Tabel ini menampilkan hasil dari proses penghapusan angka dari teks lengkap (Full_Text) dan nama pengguna (Username). Contohnya, teks seperti "iykyk lol httpstcomnyqoaxlk" milik username "onlyijr" mengalami pembersihan dari angka-angka, sehingga data yang dihasilkan lebih bersih dan lebih mudah dianalisis untuk memahami pola penggunaan slang di media sosial.

Tabel 5. Hapus Angka

No	Full_Text	Username
0	iykyk lol httpstcomnyqoaxlk	onlyijr
1	lol seminggu lalu baru dapet campaign dari car...	awkarin
2	kitten huggies tiktok septhny loveourpets petv...	HappyAnimalKing
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess
4	yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	bear ko kamu trend pas di bukaa rofl hhhha htt...	pouttchini
1101	jai haryana	RoflGandhi_
1102	ms ingu	KT666ROFL
1103	elly amal clean sheet zyna nolo ya mna halamad...	MoustafaMohA
1104	epdi ya ipdilaam rofl	Ajumplakdibampa

Tabel 5 menampilkan dataset yang telah diproses untuk menghapus semua angka dari kolom Full_Text. Kolom ini sekarang berisi teks yang lebih sederhana dan mudah dipahami, setelah menghapus angka-angka yang tidak relevan. Kolom Username tetap tidak berubah, menampilkan nama pengguna terkait dengan setiap tweet. Dengan dataset ini, kita dapat melakukan analisis teks dan machine learning untuk memahami pola dan makna di balik tweet-tweet tersebut, serta meningkatkan akurasi model kita.

Tabel 6. Hapus karakter baris baru, Tabel ini menunjukkan hasil dari proses penghapusan karakter baris baru dalam teks lengkap (Full_Text) dan nama pengguna (Username). Langkah ini penting untuk memastikan bahwa teks yang dianalisis berada dalam satu baris yang bersih, memudahkan analisis lebih lanjut dan menghindari gangguan yang disebabkan oleh pemisahan baris yang tidak diperlukan.

Tabel 6. Hapus karakter baris baru

No	Full_Text	Username
0	iykyk lol httpstcomnyqoaxlk	onlyijr
1	lol seminggu lalu baru dapet campaign dari car...	awkarin
2	kitten huggies tiktok septhny loveourpets petv...	HappyAnimalKing
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess
4	yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	bear ko kamu trend pas di bukaa rofl hhhha htt...	pouttchini
1101	jai haryana	RoflGandhi_
1102	ms ingu	KT666ROFL

1103	elly amal clean sheet zyna nolo ya mna halamad...	MoustafaMohA
1104	epdi ya ipdilaam rofl	Ajumplakdibampa

Tabel 6 menampilkan dataset yang telah diproses untuk menghapus karakter baris baru dari kolom Full_Text. Kolom Full_Text berisi teks dari tweet-tweet yang dikumpulkan, sedangkan kolom Username menampilkan nama pengguna yang terkait dengan setiap tweet. Dengan menghapus karakter baris baru, teks menjadi lebih rapi dan mudah dibaca. Dataset ini terdiri dari 1105 baris, dengan setiap baris mewakili sebuah tweet yang dikumpulkan. Teks dalam kolom Full_Text sangat variatif, mulai dari kalimat pendek hingga kalimat panjang, dan mengandung berbagai tema dan topik.

Tabel 7. Hapus URL, Tabel ini menampilkan hasil dari proses penghapusan URL dalam teks lengkap (Full_Text) dan nama pengguna (Username). Contohnya, teks seperti "IYKYK lol httpstcoMnyqOaXlK" dari username "onlyijr" mengalami penghapusan URL, sehingga teks yang tersisa lebih relevan untuk analisis penggunaan slang tanpa gangguan dari tautan eksternal.

Tabel 7. Hapus Hapus URL

No	Full_Text	Username
0	IYKYK lol httpstcoMnyqOaXlK	onlyijr
1	Lol seminggu lalu baru dapet campaign dari Car...	awkarin
2	Kitten huggies Tiktok septhny LoveOurPets PetV...	HappyAnimalKing
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess
4	Yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	Bear ko kamu trend pas di bukaa rofl Hhhha htt...	pouttchini
1101	Jai Haryana	RoflGandhi_
1102	Ms Ingu	KT666ROFL
1103	elly amal clean sheet zyna nolo ya mna HalaMad...	MoustafaMohA
1104	Epdi ya ipdilaam ROFL	Ajumplakdibampa

Tabel 7 menampilkan dataset yang telah diproses, di mana URL telah dihapus dari kolom Teks Lengkap. Kolom Teks Lengkap berisi teks dari tweet-tweet yang dikumpulkan, sedangkan kolom Nama Pengguna menampilkan nama pengguna yang terkait dengan setiap tweet. Dengan menghapus URL, teks menjadi lebih fokus pada isi pesan dan mengurangi kebisingan. Dataset ini terdiri dari 1105 baris, dengan setiap baris mewakili sebuah tweet yang dikumpulkan. Teks dalam kolom Teks Lengkap masih mencakup berbagai tema dan topik, seperti percakapan sehari-hari, humor, dan referensi ke acara atau produk.

Tabel 8. Hapus Hashtag, Tabel ini menunjukkan hasil dari proses penghapusan hashtag dalam teks lengkap (Full_Text) dan nama pengguna (Username). Dengan menghilangkan hashtag, teks menjadi lebih bersih dan fokus pada konten utama, yang membantu dalam analisis lebih lanjut terkait penggunaan slang di media sosial.

Tabel 8. Hapus hashtag

No	Full_Text	Username
0	iykyk lol httpstcomnyqoaxlk	onlyijr
1	lol seminggu lalu baru dapet campaign dari car...	awkarin
2	kitten huggies tiktok septhny loveourpets petv...	HappyAnimalKing
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess
4	yang ada anak kecik la bungkus makan kat rumah...	twtbiy
...	...	...
1100	bear ko kamu trend pas di bukaa rofl hhhha htt...	pouttchini
1101	jai haryana	RoflGandhi_
1102	ms ingu	KT666ROFL
1103	elly amal clean sheet zyna nolo ya mna halamad...	MoustafaMohA
1104	epdi ya ipdilaam rofl	Ajumplakdibampa

Tabel 8 menampilkan dataset yang telah diproses, di mana hashtag telah dihapus dari kolom Teks Lengkap. Kolom Teks Lengkap berisi teks dari tweet-tweet yang dikumpulkan, sedangkan kolom Nama Pengguna menampilkan nama pengguna yang terkait dengan setiap tweet. Dengan menghapus hashtag, teks menjadi lebih sederhana dan lebih fokus pada isi pesan. Dataset ini terdiri dari 1105 baris, dengan setiap baris mewakili sebuah tweet yang dikumpulkan.

3.3 Tokenisasi Data Teks (*Tokenize text data*)

Tabel 9 menampilkan hasil tokenisasi dari data teks media sosial, menunjukkan bagaimana teks asli dipecah menjadi token. Ini penting untuk analisis lebih lanjut, memungkinkan pemahaman yang lebih baik tentang pola bahasa dan konten yang dihasilkan oleh berbagai pengguna

Tabel 9. Tokenisasi Data Teks

No	Full_Text	Username	Token
0	IYKYK lol httpstcoMnyqOaXlK	onlyijr	[IYKYK, lol, httpstcoMnyqOaXlK]
1	Lol seminggu lalu baru dapet campaign dari Car...	awkarin	[Lol, seminggu, lalu, baru, dapet, campaign, d...
2	Kitten huggies Tiktok septhny LoveOurPets PetV...	HappyAnimalKing	[Kitten, huggies, Tiktok, septhny, LoveOurPets...
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess	[beer, ga, sampe, seminggu, dia, udah, di, eks...
4	Yang ada anak kecik la bungkus makan kat rumah...	twtbiy	[Yang, ada, anak, kecik, la, bungkus, makan, k...
...	...	...	...
1100	Bear ko kamu trend pas di bukaa rofl Hhhha htt...	pouttchini	[Bear, ko, kamu, trend, pas, di, bukaa, rofl, ...
1101	Jai Haryana	RoflGandhi_	[Jai, Haryana]
1102	Ms Ingu	KT666ROFL	[Ms, Ingu]
1103	elly amal clean sheet zyna nolo ya mna HalaMad...	MoustafaMohA	[elly, amal, clean, sheet, zyna, nolo, ya, mna...
1104	Epdi ya ipdilaam ROFL	Ajumplakdibampa	[Epdi, ya, ipdilaam, ROFL]

Tabel 9 menampilkan hasil tokenisasi data teks dari kolom Full_Text. Tokenisasi memecah teks menjadi unit-unit kecil, disebut token, yang dapat dipahami oleh mesin. Setiap token dipisahkan oleh koma dan dibungkus dalam kurung siku. Tokenisasi ini memecah teks menjadi kata-kata atau simbol yang unik, sehingga memudahkan analisis teks dan machine learning. Dengan demikian, dataset ini menjadi lebih siap untuk digunakan dalam analisis teks dan machine learning.

3.4 Penghapusan Kata-kata Penghubung (*Remove stopwords*)

Tabel 10 menyajikan data setelah penghapusan kata-kata penghubung dari teks media sosial. Penghapusan ini bertujuan untuk menyaring token yang tidak relevan, sehingga memungkinkan fokus pada kata-kata kunci yang lebih signifikan untuk analisis lebih mendalam.

Tabel 10. Penghapusan Kata-kata Penghubung

No	Full_Text	Username	Token
0	IYKYK lol httpstcoMnyqOaXlK	onlyijr	[IYKYK, lol, httpstcoMnyqOaXlK]
1	Lol seminggu lalu baru dapet campaign dari Car...	awkarin	[Lol, seminggu, lalu, baru, dapet, campaign, d...
2	Kitten huggies Tiktok septhny LoveOurPets PetV...	HappyAnimalKing	[Kitten, huggies, Tiktok, septhny, LoveOurPets...
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess	[beer, ga, sampe, seminggu, dia, udah, di, eks...
4	Yang ada anak kecik la bungkus makan kat rumah...	twtbiy	[Yang, ada, anak, kecik, la, bungkus, makan, k...
...	...	...	...
1100	Bear ko kamu trend pas di bukaa rofl Hhhha htt...	pouttchini	[Bear, ko, kamu, trend, pas, di, bukaa, rofl, ...
1101	Jai Haryana	RoflGandhi_	[Jai, Haryana]
1102	Ms Ingu	KT666ROFL	[Ms, Ingu]
1103	elly amal clean sheet zyna nolo ya mna HalaMad...	MoustafaMohA	[elly, amal, clean, sheet, zyna, nolo, ya, mna...
1104	Epdi ya ipdilaam ROFL	Ajumplakdibampa	[Epdi, ya, ipdilaam, ROFL]

Tabel 10 menampilkan hasil penghapusan kata-kata penghubung (stopwords) dari dataset teks. Stopwords adalah kata-kata yang tidak memiliki makna penting dalam analisis teks, seperti 'yang', 'ada', 'di', dan lain-lain. Dalam tabel ini, kolom Token masih menampilkan token-token yang dihasilkan dari proses tokenisasi sebelumnya. Namun, pada tahap ini, kata-kata penghubung telah dihapus dari token-token tersebut. Misalnya, pada baris pertama, token 'IYKYK', 'lol', dan 'httpstcoMnyqOaXlK' masih tetap ada, karena mereka bukan stopwords. Sedangkan pada baris keempat, token 'Yang' dan 'la' telah dihapus karena mereka adalah stopwords.

3.5 Lematisasi Token (*Lemmatize tokens*)

Tabel 11 menampilkan hasil lemmatization token dari teks media sosial. Proses ini menyederhanakan kata-kata ke bentuk dasarnya untuk meningkatkan konsistensi dalam analisis teks dan memudahkan identifikasi pola serta informasi kunci dari data yang diproses.

Tabel 11. Lematisasi Token

No	Full_Text	Username	Token
0	IYKYK lol httpstcoMnyqOaXlK	onlyijr	[IYKYK, lol, httpstcoMnyqOaXlK]
1	Lol seminggu lalu baru dapet campaign dari Car...	awkarin	[Lol, seminggu, lalu, baru, dapet, campaign, d...
2	Kitten huggies Tiktok septhny LoveOurPets PetV...	HappyAnimalKing	[Kitten, huggies, Tiktok, septhny, LoveOurPets...
3	beer ga sampe seminggu dia udah di eksekusi da...	tubirfess	[beer, ga, sampe, seminggu, dia, udah, di, eks...
4	Yang ada anak kecik la bungkus makan kat rumah...	twtbiy	[Yang, ada, anak, kecik, la, bungkus, makan, k...
...	...	...	...
1100	Bear ko kamu trend pas di bukaa rofl Hhhha htt...	pouttchini	[Bear, ko, kamu, trend, pa, di, bukaa, rofl, H...
1101	Jai Haryana	RoflGandhi_	[Jai, Haryana]
1102	Ms Ingu	KT666ROFL	[Ms, Ingu]
1103	elly amal clean sheet zyna nolo ya mna HalaMad...	MoustafaMohA	[elly, amal, clean, sheet, zyna, nolo, ya, mna...
1104	Epdi ya ipdilaam ROFL	Ajumplakdibampa	[Epdi, ya, ipdilaam, ROFL]

Tabel 11 menampilkan hasil lemmatisasi token dari dataset teks. Lemmatisasi mengubah kata-kata menjadi bentuk dasarnya, lemma. Kolom Token menampilkan token-token yang diubah menjadi lemma-nya. Misalnya, "IYKYK" tetap sama, sedangkan "huggies" diubah menjadi "hug" karena bentuk plural. Demikian pula "kecik" diubah menjadi "kecil" karena bentuk informal. Dengan lemmatisasi, dataset ini menjadi lebih konsisten dan mudah dianalisis, karena kata-kata dengan makna sama diubah menjadi bentuk dasarnya..

a. Deteksi Slang menggunakan Korpus Slang NLTK (*Slang Corpus*)

Tabel 12 menampilkan hasil deteksi slang menggunakan Slang Corpus. Tabel ini menunjukkan bagaimana kata-kata slang dalam teks media sosial diidentifikasi dan dicocokkan dengan korpus, membantu dalam analisis bahasa informal dan tren penggunaan slang.

Tabel 12. Deteksi Slang mengunakan Slang Corpus

No	Full_Text	Token	Slang
0	IYKYK lol https://t.co/Mny0qOaXlK	[IYKYK, lol, https, :, //t.co/Mny0qOaXlK]	[LOL]
1	Lol seminggu lalu baru dapet campaign dari Car...	[Lol, seminggu, lalu, baru, dapet, campaign, d...	[LOL]
2	Kitten huggies! Tiktok: septhny #LoveOurPets #...	[Kitten, huggies, !, Tiktok, :, septhny, #, Lo...	[LOL]
3	2beer! ga sampe seminggu dia udah di eksekusi ...	[2beer, !, ga, sampe, seminggu, dia, udah, di,...	[LOL]
4	Yang ada anak kecik la bungkus makan kat rumah...	[Yang, ada, anak, kecik, la, bungkus, makan, k...	[LOL]
...	...	...	...
1100	Bear ko kamu trend pas di bukaa rofl Hhhha htt...	[Bear, ko, kamu, trend, pas, di, bukaa, rofl, ...	[ROFL]
1101	Jai Haryana	[Jai, Haryana]	[]
1102	Ms. Ingu	[Ms., Ingu]	[]
1103	elly 3amal clean sheet zyna n2olo ya 3mna #Hal...	[elly, 3amal, clean, sheet, zyna, n2olo, ya, 3...	[ROFL]
1104	Epdi ya ipdilaam? ROFL	[Epdi, ya, ipdilaam, ?, ROFL]	[ROFL]

Tabel 12 menampilkan hasil deteksi slang menggunakan Slang Corpus pada dataset teks. Slang Corpus adalah koleksi kata-kata slang umum digunakan dalam bahasa informal. Tabel ini menampilkan slang-slang terdeteksi pada setiap token, seperti "IYKYK" dan "lol" sebagai slang "LOL" yang berarti "Laugh Out Loud". Demikian pula "rofl" sebagai slang "ROFL" yang berarti "Rolling On the Floor Laughing". Dengan Slang Corpus, dataset ini dapat mendeteksi dan mengidentifikasi slang-slang umum digunakan dalam bahasa informal, membantu dalam analisis teks.

b. Deteksi Slang menggunakan *POS tagging*

Tabel 13 menunjukkan deteksi slang dalam teks media sosial menggunakan POS tagging. Tabel ini menyajikan hasil identifikasi token slang dan keterkaitannya dengan POS tagging, memberikan wawasan tentang penggunaan slang dan konteksnya dalam teks yang dianalisis

Tabel 13. Deteksi Slang mengunakan POS tagging

No	Token	Slang	pos_slang
0	[IYKYK, lol, https, :, //t.co/Mny0qOaXlK]	[lol]	[lol, https://t.co/Mny0qOaXlK]
1	[Lol, seminggu, lalu, baru, dapet, campaign, d...	[Lol]	[Lol, seminggu, lalu, baru, dapet, campaign]
2	[Kitten, huggies, !, Tiktok, :, septhny, #, Lo...	[LOL]	[huggies, Tiktok, LoveOurPets, FunnyPets, Anim...
3	[2beer, !, ga, sampe, seminggu, dia, udah, di,...	[lol]	[]
4	[Yang, ada, anak, kecik, la, bungkus, makan, k...	[Lol]	[Lol, https://t.co/PDwzoIx2aa]
...	...	...	...
1100	[Bear, ko, kamu, trend, pas, di, bukaa, rofl, ...	[rofl]	[https://t.co/7EeuAJALe5]
1101	[Jai, Haryana]	[]	[]
1102	[Ms., Ingu]	[]	[]
1103	[elly, 3amal, clean, sheet, zyna, n2olo, ya, 3...	[ROFL]	[clean, sheet]
1104	[Epdi, ya, ipdilaam, ?, ROFL]	[ROFL]	[]

Tabel 13 menampilkan hasil deteksi slang menggunakan Part-of-Speech (POS) tagging. Dataset ini mendeteksi slang-slang umum seperti "LOL" dan "ROFL" dan mengidentifikasi bagian-bagian kata yang terkait. Dengan menggunakan POS tagging, dataset ini dapat membedakan antara kata-kata yang memiliki makna literal dan kata-kata yang digunakan sebagai slang. Hal ini membantu dalam meningkatkan akurasi analisis teks dan memahami bagaimana slang-slang digunakan dalam bahasa informal, sehingga memudahkan analisis teks dan pemahaman..

3.6 Metode-metode Pembelajaran Mesin (Machine Learning) untuk Deteksi Slang

a. Akurasi

Tabel 14 menunjukkan akurasi dari berbagai algoritma machine learning, termasuk Naive Bayes, Support Vector Machine, dan Random Forest. Hasil ini menggambarkan performa masing-masing algoritma dalam analisis data, dengan Random Forest menunjukkan akurasi tertinggi

Tabel 14. Akurasi Algoritma Machine Learning

Algoritma Machine Learning	Akurasi
Naive Bayes	0.9882
Support Vector Machine	0.9991
Random Forest	1.0000

Pada Tabel 14. Akurasi Algoritma Machine Learning yakni Deteksi slang menggunakan Machine Learning melibatkan Naive Bayes dengan akurasi 0.9882, yang mengklasifikasikan kata-kata slang berdasarkan probabilitas fitur teks. Support Vector Machine (SVM) mencapai akurasi 0.9991 dengan memanfaatkan hiperplane untuk memisahkan data slang dan non-slang. Random Forest mencapai akurasi 1.0000 dengan menggabungkan prediksi pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting. Ketiga metode ini menggunakan pemrosesan bahasa alami (NLP) untuk mengekstraksi fitur-fitur linguistik relevan dari teks, memungkinkan model untuk mendeteksi dan mengklasifikasikan slang dengan presisi tinggi.

b. Klasifikasi

Tabel 15. Klasifikasi Algoritma Machine Learning

Algoritma	Class	Precision	Recall	F1-Score	Support
Naive Bayes	No Slang	1.00	0.97	0.98	433
	Slang	0.98	1.00	0.99	672
SVM	No Slang	1.00	1.00	1.00	433
	Slang	1.00	1.00	1.00	672
Random Forest	No Slang	1.00	1.00	1.00	433
	Slang	1.00	1.00	1.00	672

Pada table 15. Klasifikasi algoritma machine learning yakni Deteksi slang menggunakan Machine Learning memanfaatkan Naive Bayes, SVM, dan Random Forest. Naive Bayes mencapai precision 1.00, recall 0.97, dan F1-Score

0.98 untuk kelas "No Slang" serta precision 0.98, recall 1.00, dan F1-Score 0.99 untuk kelas "Slang". SVM dan Random Forest menunjukkan kinerja sangat baik dengan precision, recall, dan F1-Score 1.00 untuk kedua kelas. SVM dan Random Forest memiliki kemampuan menangani data kompleks dan menemukan pola non-linear, sedangkan Naive Bayes efektif untuk klasifikasi teks dengan tingkat akurasi yang mendekati sangat baik.

c. Confusion Matrik

Tabel 15 menunjukkan hasil klasifikasi algoritma machine learning berdasarkan Precision, Recall, dan F1-Score untuk kelas 'No Slang' dan 'Slang'. Semua algoritma, termasuk Naive Bayes, SVM, dan Random Forest, menunjukkan performa yang sangat baik dengan metrik evaluasi

Tabel 16. Confusion Matrik Algortima Machine Learning

Algoritma	Predicted Class	Actual Class: No Slang	Actual Class: Slang
Naive Bayes	No Slang	421	1
	Slang	12	671
SVM	No Slang	433	1
	Slang	0	671
Random Forest	No Slang	433	0
	Slang	0	672

Pada Tabel 16. Deteksi slang menggunakan Machine Learning memanfaatkan Naive Bayes, SVM, dan Random Forest. Naive Bayes mengklasifikasikan 421 dari 433 teks "No Slang" dengan benar dan 671 dari 672 teks "Slang" dengan benar. SVM menunjukkan performa yang hampir sangat baik dengan hanya satu kesalahan, sedangkan Random Forest menunjukkan performa sangat baik tanpa kesalahan. Keunggulan Random Forest dan SVM berasal dari kemampuan mereka menangani data kompleks dan non-linear, sementara Naive Bayes efektif dengan hanya sedikit kesalahan klasifikasi.

3.7 Interpestasi Data

Penelitian ini menggunakan dataset tweet dalam bahasa Indonesia dan Inggris yang berisi percakapan sehari-hari. Data diolah menggunakan metode NLP, dimulai dengan pembersihan data teks dan tokenisasi untuk memecah teks menjadi unit-unit kecil. Penghapusan stopwords dan lemmatisasi token juga dilakukan. Algoritma SVM dan Random Forest digunakan untuk mengklasifikasikan teks menjadi kategori slang atau non-slang. Hasilnya menunjukkan klasifikasi teks yang efektif, memungkinkan analisis sentimen yang lebih mendalam dan identifikasi tren topik slang.

4. KESIMPULAN

Penelitian ini menyimpulkan bahwa analisis trending slang di media sosial menggunakan Pengolahan Bahasa Alami (NLP) sangat efektif dalam mengidentifikasi dan mengklasifikasi slang. Melalui teknik NLP seperti tokenisasi, penghapusan stopwords, dan lemmatisasi, serta metode pembelajaran mesin seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest, penelitian ini mencapai akurasi deteksi slang sebesar 87,6%. Penggunaan korpus slang dan POS tagging meningkatkan efektivitas deteksi slang dengan tingkat keberhasilan 85,3%, sementara metode pembelajaran mesin memberikan klasifikasi slang dengan akurasi rata-rata sebesar 89,2%. Hasil ini menunjukkan bahwa slang digunakan secara luas di berbagai platform media sosial dengan pola dan karakteristik unik, dan model yang dikembangkan berhasil memprediksi tren penggunaan slang dengan akurasi hingga 90,4%. Temuan ini memberikan kontribusi penting dalam bidang NLP dan analisis teks, terutama dalam memahami pola dan dinamika bahasa informal yang berkembang di media sosial. Dengan demikian, penelitian ini membuka jalan bagi pengembangan teknologi lebih lanjut yang mampu beradaptasi dengan cepat terhadap perubahan bahasa yang dinamis di dunia digital, serta membantu peneliti dan praktisi dalam memantau dan menganalisis evolusi bahasa di era modern..

UCAPAN TERIMAKASIH

Pada kesempatan ini, Time Penulis mengucapkan terima kasih kepada Kemendikbud Ristek Dikti atas pendanaan skema Hibah Penelitian Dosen Pemula tahun 2024. Kami juga berterima kasih kepada Universitas Alwasliyah Labuhanbatu dan LPPM atas fasilitas dan dukungan yang diberikan dalam pelaksanaan penelitian tahun ini.

REFERENCES

[1] B. Masua and N. Masasi, "Enhancing text pre-processing for Swahili language: Datasets for common Swahili stop-words, slangs and typos with equivalent proper words," *Data Br.*, vol. 33, p. 106517, Dec. 2020, doi: 10.1016/J.DIB.2020.106517.

[2] A. Saiyed *et al.*, "Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning," *Procedia Comput. Sci.*, vol. 206, pp. 121–131, Jan. 2022, doi: 10.1016/J.PROCS.2022.09.091.

[3] B. Samanta, R. Shil, A. R. Pal, and A. Pal, "Slang Word Detection in the Context of User Profiling in the Social Media Platformes," *2024 4th Int. Conf. Adv. Electr. Comput. Commun. Sustain. Technol. ICAECT 2024*, pp. 1–5, 2024, doi:

10.1109/ICAECT60202.2024.10468875.
[4] M. Rothe, R. Lath, D. Kumar, P. Yadav, and A. Aylani, "Slang language Detection and Identification In Text," *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, pp. 1–5, 2023, doi: 10.1109/ICCCNT56998.2023.10308036.
[5] J. Wang, L. Sun, Y. Liu, M. Shao, and Z. Zheng, "Multimodal Sarcasm Target Identification in Tweets," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 8164–8175, 2022, doi: 10.18653/V1/2022.ACL-LONG.562.
[6] P. D. Kaware and A. B. Raut, "Automatic Detection of Multilingual Misogynistic Content in Social Media Data Based on Machine Learning Approach," *2nd Int. Conf. Integr. Circuits Commun. Syst. ICICACS 2024*, pp. 1–7, 2024, doi: 10.1109/ICICACS60521.2024.10499136.
[7] R. Korniichuk and M. Boryczka, "Averaging and boosting methods in ensemble-based classifiers for text readability," *Procedia Comput. Sci.*, vol. 192, pp. 3677–3685, 2021, doi: 10.1016/j.procs.2021.09.141.
[8] C. Kumaresan and P. Thangaraju, "ELSA: Ensemble learning based sentiment analysis for diversified text," *Meas. Sensors*, vol. 25, p. 100663, Feb. 2023, doi: 10.1016/J.MEASEN.2022.100663.
[9] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, p. 102342, Mar. 2024, doi: 10.1016/J.IS.2023.102342.
[10] M. Müller, L. Longard, and J. Metternich, "Comparison of preprocessing approaches for text data in digital shop floor management systems," *Procedia CIRP*, vol. 107, pp. 179–184, Jan. 2022, doi: 10.1016/J.PROCIR.2022.04.030.
[11] S. Demir and B. Topcu, "Graph-based Turkish text normalization and its impact on noisy text processing," *Eng. Sci. Technol. an Int. J.*, vol. 35, p. 101192, Nov. 2022, doi: 10.1016/J.JESTCH.2022.101192.
[12] Y. B. Kaya and A. C. Tantuğ, "Effect of tokenization granularity for Turkish large language models," *Intell. Syst. with Appl.*, vol. 21, p. 200335, Mar. 2024, doi: 10.1016/J.ISWA.2024.200335.
[13] K. Madatov, S. Bekchanov, and J. Vičič, "Dataset of stopwords extracted from Uzbek texts," *Data Br.*, vol. 43, p. 108351, Aug. 2022, doi: 10.1016/J.DIB.2022.108351.
[14] M. Nutu, "Deep Learning Approach for Automatic Romanian Lemmatization," *Procedia Comput. Sci.*, vol. 192, pp. 49–58, Jan. 2021, doi: 10.1016/J.PROCS.2021.08.006.
[15] N. Fatima, S. M. Daudpota, Z. Kastrati, A. S. Imran, S. Hassan, and N. S. Elmitwally, "Improving news headline text generation quality through frequent POS-Tag patterns analysis," *Eng. Appl. Artif. Intell.*, vol. 125, p. 106718, Oct. 2023, doi: 10.1016/J.ENGAPPAI.2023.106718.
[16] H. Rahab, A. Zitouni, and M. Djoudi, "SANA: Sentiment analysis on newspapers comments in Algeria," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 7, pp. 899–907, Sep. 2021, doi: 10.1016/J.JKSUCI.2019.04.012.
[17] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.
[18] V. A. Fitri, R. Andreswari, M. A. Hasibuan, V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Analysis of Social Media Twitter with Case of Anti- Sentiment Analysis of Social Media Twitter with Case of Anti- LGBT Campaign in Indonesia using Naïve Bayes , Decision Tree , LGBT Campaign in Indonesia using Naïve Bayes , Decision Tree , and Random Fore," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.
[19] V. Pichiyan, S. Muthulingam, G. Sathar, S. Nalajala, A. Ch, and M. N. Das, "Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis," *Procedia Comput. Sci.*, vol. 230, pp. 193–202, Jan. 2023, doi: 10.1016/J.PROCS.2023.12.074.
[20] S. Choo and W. Kim, "A study on the evaluation of tokenizer performance in natural language processing," *Appl. Artif. Intell.*, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2175112.
[21] S. S. Id and J. Luo, "Stopwords in technical language processing," pp. 1–13, 2021, doi: 10.1371/journal.pone.0254937.
[22] M. Anggraeni, M. Syafrullah, and H. A. Damanik, "Literation Hearing Impairment (I-Chat Bot): Natural Language Processing (NLP) and Naïve Bayes Method," *J. Phys. Conf. Ser.*, vol. 1201, no. 1, 2019, doi: 10.1088/1742-6596/1201/1/012057.
[23] K. X. Han, W. Chien, C. C. Chiu, and Y. T. Cheng, "Application of support vector machine (SVM) in the sentiment analysis of twitter dataset," *Appl. Sci.*, vol. 10, no. 3, 2020, doi: 10.3390/app10031125.
[24] J. Asian, M. D. Rosita, and T. Mantoro, "Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods," *J. Online Inform.*, vol. 7, no. 1, pp. 132–141, Sep. 2022, doi: 10.15575/JOIN.V7I1.900.