Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.20241127404
Vol. 11, No. 2, April 2024, hlm. 337-348 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

OTOMATISASI PENDETEKSI KATA BAKU DAN TIDAK BAKU PADA DATA TWITTER BERBASIS KBBI

M. Irfan Raif*[1], Nuraisa Novia Hidayati[2],Tekad Matulatan[3]

[1,3] Universitas Maritim Raja Ali Haji, Tanjung Pinang
[2] Badan Riset dan Inovasi Nasional, Jakarta
Email: [1]2001020056@student.umrah.ac.id, [2]nuraisa.novia.hidayati@brin.go.id, [3]tekad.matulatan@umrah.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 26 Juni 2023, diterima untuk diterbitkan: 04 April 2024)

Abstrak

Penelitian ini berfokus pada pengembangan sistem deteksi otomatis untuk membedakan kata baku dan tidak baku pada data Twitter, berdasarkan Kamus Besar Bahasa Indonesia (KBBI). Karena Twitter merupakan *platform* media sosial yang sering menggunakan kata-kata yang tidak baku, penelitian ini penting untuk memastikan komunikasi yang efektif. Melalui normalisasi kata-kata tidak baku, penelitian ini berkontribusi signifikan terhadap pra-pemrosesan dan analisis tweet, yang merupakan langkah penting dalam klasifikasi teks media sosial. Sistem otomatis yang dikembangkan tidak hanya membantu peneliti dengan mudah mengidentifikasi penggunaan kata-kata slang atau tidak baku, namun juga meningkatkan kualitas komunikasi dan pemahaman pesan dalam *tweet* yang mencerminkan tren bahasa terkini. Pendekatan yang dilakukan dalam penelitian ini meliputi langkah-langkah seperti pengumpulan data, *preprocessing*, identifikasi bahasa tidak baku, penghapusan kata berimbuhan, identifikasi slang, dan penggunaan metode *lexicon-based* untuk kamus opini. Pendekatan ini efektif dalam mendukung analisis sentimen pada teks mining dan memastikan hasil klasifikasi sentimen pada data Twitter lebih akurat. Hasil percobaan menunjukkan bahwa langkah *preprocessing* tersebut berhasil meningkatkan akurasi metode penentuan polarisasi, dengan tingkat akurasi InSet sebesar 66,66% dan F1-score sebesar 61,40%.

Kata kunci: *analisis sentimen, twitter, InSet, lexicon-based, pendeteksi kata tidak baku, preprocessing, normalisasi.*

AUTOMATION OF STANDARD AND NON-STANDARD WORD DETECTION IN KBBI-BASED TWITTER DATA

Abstract

This research focuses on developing an automatic detection system to distinguish between standard and nonstandard words in Twitter data, based on the Kamus Besar Bahasa Indonesia (KBBI). As Twitter is a social media platform that often uses nonstandard words, this research is important to ensure effective communication. Through the normalization of nonstandard words, this research contributes significantly to the pre-processing and analysis of tweets, which is an important step in social media text classification. The automated system developed not only helps researchers easily identify the use of slang or nonstandard words, but also improves the quality of communication and message understanding in tweets that reflect current language trends. The approach taken in this research includes steps such as data collection, preprocessing, nonstandard language identification, removal of affixed words, slang identification, and the use of lexicon-based methods for opinion dictionaries. This approach is effective in supporting sentiment analysis in text mining and ensures more accurate sentiment classification results on Twitter data. Experimental results show that these preprocessing steps successfully improve the accuracy of the polarization determination method, with an InSet accuracy rate of 66.66% and F1-score of 61.40%.

Keywords: *Sentiment analysis, twitter, InSet, lexicon-based, Out-of-vocabulary word detector, preprocessing, normalization.*

1. PENDAHULUAN

Media sosial seperti Twitter saat ini telah menjadi platform yang paling populer untuk berkomunikasi dan berinteraksi dengan pengguna lain di seluruh dunia. Twitter adalah salah satu media sosial terpopuler yang digunakan oleh jutaan

pengguna untuk berbagai tujuan, termasuk berbagi informasi dan berkomunikasi. Dalam Twitter, pengguna dapat menulis, membaca, dan berbagi teks pendek yang disebut sebagai *tweet*. Batasan karakter untuk setiap *tweet* adalah 280 huruf. (Meftah et al., 2018) (Kumar and Gruzd, 2019).

Pengguna Twitter sering menggunakan kata-kata tidak baku yang dapat menyulitkan pemahaman pesan dan merusak kualitas komunikasi. Kata baku mengacu pada kata yang sesuai dengan kaidah baku, sedangkan kata tidak baku merujuk pada kata yang lazim digunakan dalam percakapan sehari-hari atau tidak mengikuti prinsip *good governance* (EYD) (Bengi Ruhamah, Adnan, 2018).

Normalisasi kata diperlukan untuk mengubah kata tidak baku menjadi kata baku dalam kalimat atau *tweet*. Ini membantu dalam menangani masalah seperti mendeteksi singkatan, bahasa gaul atau slang, kesalahan ejaan, dan penggunaan bahasa yang tidak pantas. Normalisasi kata meningkatkan pemrosesan dan analisis *tweet* secara efektif (Ivan and Adikara, 2019).

Ada beberapa penelitian sebelumnya melakukan tahap *preprocessing* yang meliputi *tokenizing*, *cleansing*, *case folding*, perbaikan kata tidak baku, *filtering*, dan *stemming*. *Preprocessing* ini penting untuk mengurangi kata-kata yang tidak relevan dan meningkatkan akurasi klasifikasi. Dalam penelitian ini, metode *Levenshtein Distance* digunakan untuk mengubah kata tidak baku menjadi kata baku, dan pengklasifikasi *Naïve Bayes* digunakan untuk proses klasifikasi. Hasil pengujian menunjukkan tingkat akurasi tertinggi dengan perbaikan kata tidak baku adalah sebesar 98.33% Dan untuk *precision*, *recall*, dan *f- measure* adalah 96.77%, 100%, dan 98.36%. (Antinasari, Perdana and Fauzi, 2017).

Penelitian sebelumnya yang menggunakan metode *Levenshtein Distance* untuk mengatasi salah ejaan dan menggunakan metode *Naïve Bayes Classifier* untuk melihat tingginya akurasi, ada beberapa tahap *preprocessing* yang dilakukan peneliti meliputi *case folding, cleansing, tokenizing, stemming,* dan *stopword*. Penelitian ini menggunakan data *training* sebanyak 450 data dengan masing-masing kategori sebanyak 150 positif, 150 negatif dan 150 netral. Hasil pengujian pada 100 data uji dari penelitian sebelumnya menghasilkan tingkat akurasi sebesar 67.05% menggunakan *Levenshtein Distance* dan 63.83% tanpa *Levenhstein Distance* pada proses klasifikasi menggunakan *Naïve Bayes* (Rozi, Ardiansyah and Rebeka, 2019).

Perbedaan penelitian ini dengan penelitian sebelumnya adalah penggunaan KBBI sebagai sumber refrensi untuk mendeteksi kata-kata baku dan tidak baku dengan berfokus kepada *preprocessing* dengan menggunakan *lexicon-based* untuk kamus opini dan menggunakan Metode InSet untuk analisis sentimen *Confusion Matrix* untuk menghitung nilai sentimen.

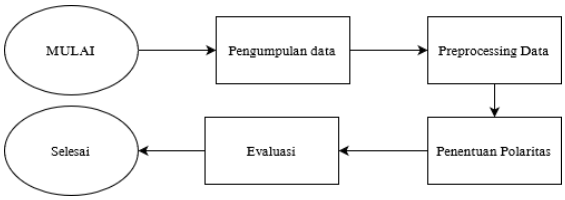
Penelitian ini mencakup langkah-langkah *preprocessing* teks, termasuk *Stopword Removal*, yang bertujuan untuk menghapus kata-kata yang tidak relevan berdasarkan daftar *stopword*. Metode *Naïve Bayes* digunakan untuk analisis sentimen. Dalam penelitian ini, digunakan daftar stopword yang dibentuk melalui algoritme *Term Based Random Sampling* dengan parameter X, Y, dan L. Hasil evaluasi menunjukkan bahwa penggunaan *stopword Term Based Random Sampling* dengan parameter tertentu menghasilkan akurasi tertinggi. Pengujian menunjukkan bahwa penggunaan *Term Based Random Sampling* dalam *Stopword Removal* mampu mencapai akurasi yang lebih tinggi dibandingkan dengan penggunaan *stopword* atau tanpa proses *Stopword Removal*. (Rinandyaswara, Sari and Furqon, 2022).

Penelitian ini fokus pada *preprocessing* dan deteksi kata-kata baku dan non-baku dalam data Twitter menggunakan Kamus Besar Bahasa Indonesia (KBBI) dengan tujuan mengembangkan sistem otomatis yang akurat. Deteksi otomatis kata-kata baku dan non-baku dalam data Twitter berbasis KBBI sangat penting dalam penelitian, memungkinkan peneliti untuk dengan mudah menemukan kata-kata non-baku atau slang yang relevan dalam penelitian mereka.

2. METODE PENELITIAN

Analisis sentimen adalah cabang penelitian text mining yang berhubungan dengan bidang lebih luas seperti pemrosesan Bahasa alami, linguistik komputasional, dan *text mining*. Tujuannya adalah untuk menganalisis sentimen, pendapat, sikap, evaluasi, penilaian, dan perasaan seseorang terhadap subjek tertentu, produk, layanan, organisasi, dan aktivitas. (Bhatia, Sharma and Bhatia, 2018) (A. and Sonawane, 2016) (Rasool et al., 2019).

Penelitian ini mengusulkan pendekatan terstruktur untuk mengotomatisasi deteksi kata-kata baku dan non-baku dalam data Twitter berdasarkan KBBI.



Gambar 1. Langkah penelitian

Gambar 1 ini adalah bentuk langkah-langkah yang digunakan dalam penelitian ini untuk mencapai tujuan.

2.1 Pengumpulan data

Penelitian ini menggunakan data yang sudah ada dari penelitian sebelumnya (Pebiana et al., 2022), berasal dari platform Twitter dengan keyword "IKN (Ibu Kota Negara)". Terdapat 12.520 data *tweet*, dengan 3 label sentimen yaitu label *negative*, *positive*, dan *neutral*. Label *negative* berjumlah 8734 data

tweet, label *positive* berjumlah 1940 data *tweet*, dan label *neutral* berjumlah 1846 data *tweet*.

2.2 *Preprocessing* data

Preprocessing merupakan langkah awal dalam pemrosesan untuk mengonversi kata pada data teks komentar tweet, dengan tujuan menghasilkan kata-kata yang lebih singkat yang mencerminkan sentimen. Langkah ini melibatkan seleksi dan eliminasi kata-kata yang tidak diperlukan, sehingga diperlukan penghapusan beberapa komponen dari data teks komentar *tweet* untuk menyaring *tweet* (Darwis, Pratiwi and Pasaribu, 2020).

Preprocessing adalah tahap penting dalam pengolahan teks yang terdiri dari kata-kata, kalimat, dan paragraf. *Preprocessing* melibatkan pengolahan sekumpulan karakter yang memiliki makna menjadi teks yang lebih baik untuk disediakan kepada algoritma pembelajaran mesin. Teknik *preprocessing* digunakan untuk mencapai bentuk yang optimal dari data teks asli (Ramachandran and Parvathi, 2019).

Data *tweet* yang terkumpul perlu dilakukan *preprocessing* untuk menghasilkan data yang jelas dan terstruktur sehingga dapat memberikan hasil klasifikasi sentimen yang lebih akurat.

2.2.1 Data *Cleaning*

Pembersihan data mengacu pada mengidentifikasi elemen data yang tidak lengkap, tidak akurat, tidak tepat, atau tidak relevan, kemudian mengganti, memodifikasi, atau menghapus data kotor. Dalam tulisan ini, 3.000 *tweet* diekstraksi dari Twitter untuk menganalisis sentimen pengguna Twitter tentang layanan kesehatan online dan beberapa penyakit menggunakan R. Kemudian, *tweet* tersebut diubah menjadi bingkai data dan dilakukan proses pembersihan data karena banyak data yang tidak diperlukan. itu tidak akan berguna. digunakan dalam analisis sentimen sebagai tanda baca, spasi, dll (Saini et al., 2019).

Cleaning merupakan langkah awal dalam text *preprocessing* yang dilakukan untuk membersihkan atau menghilangkan *noise* pada data. Proses *cleaning* pada penelitian terdiri dari beberapa Langkah yaitu:

1. *Handling Hashtag* adalah proses mengelola atau memanfaatkan penggunaan *hashtag* dalam konteks tertentu, seperti media sosial atau kempane pemasaran *hashtag* adalah *hashtag* (#) diikuti kata kunci atau frasa yang digunakan untuk mengindeks dan mengatur konten terkait.
2. *Handling* karakter *NON-ASCII* adalah pemrosesan atau penanganan karakter yang tidak ada dalam rangkaian karakter *ASCII* standar. *ASCII* adalah standar pengkodean karakter yang terdiri dari 128 karakter, umumnya digunakan dalam Bahasa inggris dan beberapa Bahasa lain yang menggunakan alfabet latin.
3. *Handling URLs, Mentions, and hashtag* adalah proses pengelolaan dan mengelola *Uniform Resource Location* (URL), *mention*, dan *hashtag* dalam konteks penggunaannya di media sosial atau teks terkait sistem atau konten yang dapat ditautkan.
4. *Handling Remove whitespace leading & trailing* adalah proses memanipulasi atau mengubah teks untuk menghapus spasi di awal (*leading*) dan akhir (*trailing*) string. *Whitespace* adalah karakter *non-printable* seperti spasi, tab, atau karakter baris baru.
5. *Handling Remove single char* adalah proses pemrosesan atau manipulasi teks untuk menghapus karakter tunggal dalam sebuah string.
6. *Handling Remove special characters and digits* adalah proses pemrosesan atau manipulasi teks untuk menghilangkan karakter dan digit khusus dari sebuah string. Karakter khusus termasuk tanda baca, symbol, dan karakter khusus lainnya yang bukan alfabet atau numerik.
7. *Handling lowercase* adalah proses pengolahan atau manipulasi teks untuk mengubah semua huruf dalam sebuah string menjadi huruf kecil.
8. *Handling Menghapus tanda baca dari setiap kata* adalah proses pengolahan atau manipulasi teks untuk menghilangkan tanda baca yang muncul di akhir setiap kata dalam sebuah string.
9. Tokenisasi atau *parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada dasarnya proses tokenisasi adalah pemenggalan kalimat menjadi kata, filter tokens adalah tahap dimana menghilangkan kata yang dikonfigurasikan untuk dihilangkan berdasarkan jumlah hurufnya (Utama et al., 2019).

2.2.2 Normalisasi teks

Normalisasi teks adalah proses mengubah format teks untuk memenuhi tujuan tertentu. Prosedur normalisasi bervariasi tergantung pada jenis teks, format keluaran yang diinginkan, tujuan standarisasi, dan metode yang digunakan. Penting untuk mempertimbangkan karakteristik ini agar dapat mengartikan "normalisasi teks" dengan jelas dalam konteks yang spesifik (Duran, Avanço and Nunes, 2015). Istilah yang digunakan untuk menyampaikan gagasan dengan mengubah format teks untuk memenuhi tujuan tertentu.

A. Identifikasi bahasa tidak baku

Kata tidak baku atau slang seperti "OMG" dan "LOL" seringkali digunakan dalam interaksi media sosial, terutama di Twitter yang memiliki batasan jumlah karakter. Dengan demikian, diharapkan bahwa memasukkan kata-kata serapan tersebut dalam perhitungan sentimen keseluruhan dari tweet dapat meningkatkan akurasi analisis (Ray, 2017).

Kata-kata yang tidak sesuai dengan KBBI akan di sebut kata-kata yang tidak baku dan akan di masukan kedalam 1 file yang berisikan kata-kata yang tidak baku, KBBI yang di gunakan adalah KBBI *offline* yang di dapat dari github, dibawah ini adalah link untuk menuju ke KBBI di github.
https://github.com/andrisetiawan/lexicon.git

B. Menghapus kata yang berimbuhan

Setelah di identifikasi kata-kata yang tidak baku, selanjutnya akan dilakukan penghapusan kata-kata berimbuhan terdiri dari beberapa kata:

- Kata *Prefiks* (awalan), yaitu afiks yang ditambahi di kiri bentuk dasar atau di awal bentuk dasar (Herawati, Juansah and Tisnasari, 2019).
  Contoh:
  'ber', 'me', 'di', 'ter', 'ke', 'se', 'pe'
- Kata *sufiks* (akhiran), yaitu afiks yang ditambahi di kanan bentuk dasar (Herawati, Juansah and Tisnasari, 2019).
  Contoh:
  'kan', 'an', 'lah', 'nya'.
- Kata *infiks* (sisipan), yaitu afiks yang ditambahkan di tengah bentuk kata (Herawati, Juansah and Tisnasari, 2019).
  Contoh:
  'el', 'em'
- Kata *Konfiks*, yaitu afiks yang ditambahi di kiri dan kanan bentuk dasar (Herawati, Juansah and Tisnasari, 2019).
  Contoh:
  'me', 'mem', 'men', 'meng', 'meny', 'pe', 'pem', 'pen', 'peng', 'peny'

C. Proses identifikasi kata *slang*

Metode *Lexicon* adalah metode klasifikasi kalimat yang menggunakan pendapat tentang relevansi kata-kata dalam dataset Kamus Bahasa Indonesia. Kata-kata tersebut diberi nilai polarisasi untuk menentukan emosi yang terkandung dalam kalimat. Hal ini membantu mengidentifikasi apakah sebuah kalimat memiliki emosi positif, negatif, atau netral (Arief and Imanuel, 2019).

Metode *lexicon-based* menggunakan kamus opini untuk mengidentifikasi apakah suatu kalimat mengandung opini. Metode learning-based, di sisi lain, menggunakan machine learning untuk mengklasifikasikan teks opini. Metode ini memanfaatkan data training yang telah di klasifikasikan secara manual untuk melakukan klasifikasi otomatis. (Azhar, 2018).

Lexicon based merupakan faktor yang menentukan sentimen suatu kalimat sehingga kalimat dapat digolongkan ke dalam kelas-kelas. Dimana setiap kata dalam kamus mempunyai skor polaritas yang diberi nilai dari -1 (untuk kelas negatif) hingga +1 (untuk kelas positif). Pada library Text Blob developer dapat menggunakan property sentiment polarity untuk melihat skor sentimen dari suatu kata atau kalimat (Nafan and Amalia, 2019).

2.3 Penentuan Polaritas

Preprocessing adalah tahap penting dalam pengolahan teks yang terdiri dari kata-kata, kalimat, dan paragraf. *Preprocessing* melibatkan pengolahan sekumpulan karakter yang memiliki makna menjadi teks yang lebih baik untuk disediakan kepada algoritma pembelajaran mesin. Teknik *preprocessing* digunakan untuk mencapai bentuk yang optimal dari data teks asli (Ramachandran and Parvathi, 2019).

A. InSet

InSet adalah metode berbasis leksikal yang dibangun Pada tahun 2017, metode leksikal bernama InSet dikembangkan untuk mengumpulkan *tweet* tahun 2016 dalam Bahasa Indonesia. Dalam pengumpulan data tersebut, sekitar 10.000 *tweet* berhasil dikumpulkan. Data *tweet* kemudian melalui proses *preprocessing*, termasuk penghapusan iklan berulang, penggunaan huruf kecil, penghilangan URL dan objek Twitter, serta penghapusan karakter khusus dan *stopwords*. Kamus dibentuk menggunakan n-gram dan menghilangkan kata dengan frekuensi 1. Saat ini, terdapat 12.503 kata positif dan 13.164 kata negative.

Kata-kata dalam *tweet* akan dibandingkan dengan kata-kata dalam *lexicon* untuk menghitung *polarity score. Polarity score* dihitung dengan menambahkan bobot kata yang terdeteksi dalam kalimat. Selain itu, data *tweet* akan diklasifikasikan kedalam kategori sentimen menggunakan algoritma yang diimplementasikan. Secara umum, proses ini dapat dijelaskan dengan algoritma berikut:

If sentiment score > 0 then Sentimen Positif (1)
If sentiment score < 0 then Sentimen Netral (2)
If sentiment score = 0 then Netral (3)

Pengklasifikasian kalimat *tweet* menjadi sentimen positif, negatif, dan netral ditentukan berdasarkan pembobotan score polaritas (sentimen *score*) yang diperoleh. Kalimat *tweet* tergolong kelas positif jika titik polaritasnya lebih besar dari 0 dan tergolong negatif jika titik polaritasnya kurang dari 0. Sedangkan *tweet* dengan titik polaritasnya 0 akan tergolong kelas netral (Koto and Rahmaningtyas, 2018).

2.4 Evaluasi

Confusion Matrix adalah sumber informasi untuk mengevaluasi kinerja model. Nilai TP (*True Positive*) dan TN (*True Negative*) mengindikasikan jumlah prediksi yang benar oleh model, sedangkan nilai FP (*False Positive*) dan FN (*False Negative*) menunjukkan jumlah prediksi yang salah oleh model. Untuk mengevaluasi kinerja pemodelan, dapat dihitung nilai akurasi, presisi, recall, dan *F1-Score* menggunakan rumus yang terdapat pada persamaan (1), (2), (3), dan (4) (Fadli and Hidayatullah, 2019).

1. *Accuracy* merupakan ukuran seberapa dekat prediksi yang dihasilkan oleh *classifier* dengan nilai sebenarnya. Hal ini diperoleh dengan membandingkan jumlah data yang diklasifikasikan dengan benar dengan jumlah keseluruhan data yang ada.

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+TN+FN} \times 100\% \quad (1)$$

2. *Precision* adalah rasio antara jumlah dokumen terkait dengan jumlah keseluruhan dokumen yang ditemukan oleh sistem.

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

3. *Recall* adalah rasio jumlah dokumen terkait dengan jumlah total dokumen dalam kumpulan dokumen yang dianggap relevan, biasanya dinyatakan dalam persentase (%).

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

4. *F1-SCORE* adalah membandingan presisi dan perolehan rata-rata tertimbang (Hidayat, Ardiansyah and Setyanto, 2021).
5.

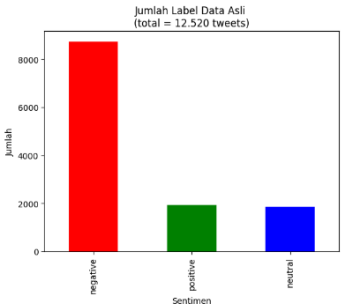
$$\text{F1} - \text{score} = \frac{2 \times Precission \times Recall}{Precission + Recall} \times 100\% \quad (4)$$

keterangan:
TP = Jumlah data kelas positif (0) diprediksi benar sebagai kelas positif (0).
FN = Jumlah data kelas positif (0) diprediksi salah sebagai kelas negatif (1).
TN = Jumlah data kelas negatif (1) diprediksi benar sebagai kelas negatif (1).
FP = Jumlah data kelas negatif (1) diprediksi salah sebagai kelas positif (0).

3. Hasil dan Pembahasan

3.1 Komposisi Data

Data yang digunakan dalam penelitian ini adalah data dari penelitian [3]. Data tersebut terdapat 12.520 data *tweet*, yang terdiri dari 3 label yaitu *negative*, *positive*, dan *neutral*.



Gambar 2. Label data asli

Gambar 2 di atas jumlah label *negative* terdapat 8734 data *tweet*, label *positive* terdapat 1940 data *tweet*, dan label *neutral* terdapat 1846 data *tweet*.

3.2 Hasil *Preprocessing*

Data *Cleaning*
Proses *cleaning* yang terdiri dari beberapa langkah yaitu:

1. *Handling Hashtag* dibawah ini adalah contoh dari *handling hashtag* melakukan penambahan spasi setiap kata hashtag yang ada
2. *Handling* karakter NON-ASCII Dalam contoh ini, tidak ada karakter NON-ASCII untuk diproses.
3. *Handling URLs, Mentions, and hashtag* Dalam contoh ini, tidak ada perubahan yang dilakukan pada URL atau penyebutan. Namun, tagar "#Gak" tetap dipertahankan.
4. *Handling remove single char* adalah proses menghapus satu karakter dalam teks. Dalam contoh ini, karakter "B" otonomi dihilangkan.
5. *Handling remove special characters and digits* adalah penghapusan karakter dan angka khusus dari teks. Dalam contoh ini, digits ("2024") dihilangkan.
6. *Handling* menghapus tanda baca dari setiap kata adalah proses menghilangkan tanda baca pada akhir setiap kata dalam sebuah teks. Dalam contoh ini, tanda baca dihapus di akhir setiap kata.
7. *Handling lowercase* adalah proses mengubah semua huruf dalam teks menjadi huruf kecil. Dalam contoh ini, semua huruf diubah menjadi huruf kecil.
8. Tokenisasi biasanya dilakukan dengan memisahkan kata dalam kalimat berdasarkan spasi atau tanda baca sebagai pembatas.

Tabel 1. Data *cleaning*

Sebelum	Sesudah
#GakHeran dah terindikasi lewat penempatan wakil kepala ikn	Gak Heran dah terindikasi lewat penempatan wakil kepala ikn
Rasa-rasanya IKN masih hutan aja si Tjahjo udah dikeluarin duluan sama Izroâ€™il Profâ€¦	rasa-rasanya ikn masih hutan aja si tjahjo udah dikeluarin duluan sama izro il prof
B... Yg di ikn bloon	... Yg di ikn bloon
Hayoo..ASN, 2024 masih pilih yg memaksa ke IKN ?? yg di ikn bloon	hayoo asn masih pilih yg memaksa ke ikn ['yg', 'di', 'ikn', 'bloon']

Tabel 1 menunjukkan proses data cleaning yang dilakukan pada data *tweet* berbahasa Indonesia yang berkaitan dengan IKN (Ibukota Negara). Data cleaning adalah proses menghapus atau memodifikasi data yang tidak relevan, tidak akurat, tidak lengkap, atau tidak sesuai dengan tujuan analisis. Tabel 1 menampilkan contoh data sebelum dan sesudah data cleaning.

Normalisasi teks

1. Identifikasi Bahasa tidak baku
   Kata tidak baku adalah ungkapan nonformal yang digunakan dalam Bahasa sehari-hari. Identifikasi kata tidak baku melibatkan perbandingan dengan kamus kata baku (KBBI). Jika kata tidak terdapat dalam kamus, maka dianggap tidak baku.

Tabel 2. Kata tidak baku

Hasil kata tidak baku
lho, yg, ikn, otorita, bebani, modarlah, memindahkan, asn, sdh, nya, mengeluh

Tabel 2 merupakan hasil identifikasi kata tidak baku dari perbandingan kamus kata baku (KBBI).

Dalam melakukan identifikasi kata tidak baku, peneliti berhasil mendapatkan sekitar 8198 kata tidak baku, termasuk kata-kata tidak baku atau singkatan yang sering muncul. Dibawah ini adalah tabel dari top 8 kata tidak baku atau singkatan yang sering digunakan pada data ikn.

Tabel 3. Kata yang sering muncul

Kata sering muncul	Jumlah kata sering muncul	kelas sintaktik
ikn	13310	NN (Noun)
yg	3920	SC (Conjunction)
gak	1143	NEG (Negative Particle)
ga	911	NEG (Negative Particle)
utk	523	IN (Preposition)
tdk	493	NEG (Negative Particle)
dgn	442	IN (Preposition)
jd	411	NN (Noun)

Berdasarkan tabel 3. Dari 8 kata tidak baku yang disebutkan sebelumnya, dapat disimpulkan bahwa kelas sintaktik yang sering muncul adalah kelas NEG (*Negative Particle*). Kelas sintaktik ini digunakan untuk kata-kata seperti “gak” dengan 1143 kata yang sering digunakan, “ga” dengan 911 kata, dan “tdk” dengan 493 kata, memiliki arti yang sama yaitu “tidak” dan berfungsi sebagai partikel *negative* dalam bahasa Indonesia.

2. Menghapus kata yang berimbuhan

Hasil penghapusan kata yang berimbuhan

Tabel 4. hasil penghapusan kata yang berimbuhan

Sebelum	Sesudah
lho, yg, ikn, otorita, bebani, modarlah, memindahkan, asn, sdh, nya, mengeluh	lho, yg, ikn, otorita, bebani, asn, sdh

Tabel 4 menunjukkan hasil penghapusan kata yang berimbuhan yang dilakukan pada data *tweet* berbahasa Indonesia yang berkaitan dengan IKN (Ibu kota Negara). Penghapusan kata berimbuhan adalah proses menghilangkan imbuhan, seperti awalan, akhiran, atau sisipan, dari kata yang ditambahkan dengan imbuhan untuk membentuk kata baru yang memiliki makna berbeda. Penghapusan kata berimbuhan bertujuan untuk mengubah kata menjadi bentuk dasarnya atau kata baku.

3. Kata seharusnya dikecualikan

Dalam proses penelitian ini, beberapa kata harus dikecualikan agar tidak dianggap sebagai kata yang tidak baku. Contoh: nama negara, provinsi, kabupaten/kota, kecamatan, desa. Kata benda ini harus dikecualikan karena tidak termasuk dalam kategori non-standar. Juga, penting untuk dicatat bahwa nama tidak memiliki nilai sentimen. Dibawah ini adalah contoh kata yang seharusnya dikecualikan agar tidak dianggap sebagai kata tidak baku.

Tabel 5. Kata pengecualian

teks	Kata yang dikecualikan
woi bego buat dibiayakan ibu kota negara kumpulan satu kalimantan bisa kok setoran nya tiap tahun dari tahun lebih dari sekarang sudah lebih tahun apa perlu ga usah disetor ke jakarta biar orang kaya kamu mampus	Kalimantan, dan jakarta

Tabel 5 adalah daftar kata pengecualian yang tidak dipertimbangkan dalam analisis atau pemrosesan teks tertentu. Dalam contoh kalimat, kata-kata "Kalimantan" dan "jakarta" dikecualikan, sehingga tidak akan memengaruhi hasil analisis atau pemrosesan teks tersebut.

4. Proses normalisasi kata tidak baku

Dalam penulisan orang di Twitter, terdapat pola yang cukup umum dalam menggunakan singkatan dan kata-kata yang diketik. Berikut adalah beberapa contoh singkatan yang biasa digunakan dalam penelitian ini :

Tabel 6. Normalisasi kata tidak baku

Sebelum	Sesudah Normalisasi
yg	yang
sdh	sudah
gak	tidak
tdk	tidak
jg	juga
utk	untuk

Tabel 6 adalah normalisasi kata tidak baku yang menunjukkan kata-kata tidak baku beserta bentuk yang telah dinormalisasi. Normalisasi kata dilakukan untuk menyamakan penggunaan kata-kata sehingga lebih sesuai dengan aturan tata bahasa yang benar.

Selain itu, ada juga kecenderungan kata-kata yang salah eja, yaitu kesalahan dalam penulisan kata yang dapat disebabkan oleh pengetikan yang cepat atau kurangnya perhatian pada kesalahan ketik. Misalnya, pengguna dapat menulis "ambyaaaar" daripada "ambyar" atau “prilaku” daripada “perilaku”.

5. Kata yang dihapus *stopword* memiliki sentimen

Dalam penelitian ini, peneliti sebenarnya tidak menggunakan *stopword* karena ada beberapa kata yang memiliki sentimen dihapus, contoh kata yang tidak memiliki sentimen.

Tabel 7. Kata dihapus *stopword*

teks	Kata yang di hapus
['semoga', 'hanya', 'diberi', 'angin', 'surga', 'saja', 'wkwkwkwkwk', 'ambyar', 'itu', 'proyek', 'ibu', 'kota', 'negara', 'kurang', 'dana']	hanya, dan kurang

Tabel 7 menunjukkan kata-kata yang dihapus sebagai stopword dari sebuah teks. Kata-kata yang dihilangkan meliputi: 'hanya', 'kurang', Dengan kata lain, stopword yang dihapus terdiri dari kata-kata tersebut.

3.3 Hasil klasifikasi

Dalam analisis yang dilakukan dengan metode *Lexicon Based* menggunakan InSet *Lexicon*, metode ini menghasilkan model klasifikasi yang dibagi menjadi tiga kelas sentiment yaitu *positive*, *negative* dan *neutral*. Proses analisis ini akan dibagi menjadi dua percobaan dari setiap 8734 data *tweet* dengan 100% label negative, dari 12.520 data *tweet* yang terdapat label negative 8734 data *tweet*, label positive 1940 data *tweet*, dan label neutral 1846 data *tweet*. Dari Setiap data *tweet* akan diperlakukan sama yaitu menganalisis sebaran data dan mengevaluasi nilai matriks data sebelum melalui tahapan *preprocessing* dan setelah *preprocessing*. Berikut ini adalah percobaan pertama menggunakan 100% label negative. Berikut ini adalah hasil *polarity* dari InSet menggunakan data 100% label negative.

Tabel 8. Hasil penentuan *polarity* dari InSet

text	*preprocessing*	Polarity score	polarity
Lho!!!! Dia Yg Minta IKN , Kok Otorita. Di Bebani!! Ya Modarlah?	lho dia yang minta ibu kota negara kok otoritas di beban iya modarlah	-15	negative
Belum jadi kok sudah mau memindahkan ASN Pak? Yakin kalau IKN sdh jadi bapak masih jadi Menteri?	belum jadi kok sudah mau memindahkan asn pak yakin kalau ibu kota negara sudah jadi bapak masih jadi menteri	0	neutral
Lu yakin IKN bakal mulus..?	kamu yakin ibu kota negara bakal mulus halu	1	positive
Halu lu sama keq junjungan lu..	kamu sama seperti junjungan kamu		
Merehabilitasi hutan???.bukannya hutannya ditebang mau dijadikan IKN..	merehabilitasi hutan bukannya hutannya ditebang mau dijadikan ibu kota negara	3	positive

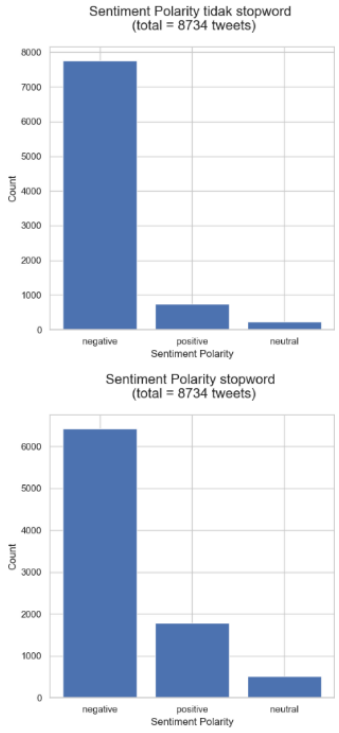
Berdasarkan tabel 8 terlihat bahwa metode InSet mampu melakukan analisis sentiment, akan tetapi dari hasil *polarity* dari InSet diatas terdapat beberapa *tweet* yang tidak sesuai sama label yang ditentukan yaitu negative, karena hasil analisa dari beberapa *tweet* yang tidak sesuai akan diidentifikasi, kenapa hasil dari *polarity* di atas berbeda-beda..? karena menurut peneliti penurunan ini ialah dikarenakan *stopword* dapat menguragi informasi dan mengubah makna tweet yang diproses sehingga tweet tersebut kehilangan sentimennya.

Dalam melakukan analisis ini peneliti tidak menggunakan *stopword* karena menurut peneliti, menggunakan stopword pada penelitian ini lebeih bekerja optimal pada pengklasifikasian dokumen dibandingkan sentimen (Angelina et al., 2023), peneliti menyatakan bahwa Penurunan nilai f1-score yang dialami ini dapat disebabkan oleh karena penerapan *Stopword*, yang pada dasarnya akan melakukan penghapusan kata-kata pada *stoplist*, sering muncul namun dianggap tidak penting dan berpengaruh signifikan terhadap makna kalimat. tergantung apa yang digunakan, seperti konjungsi dan kata ganti orang. Dibawah ini adalah hasil *polarity* menggunakan *stopword* dan tidak menggunakan *stopword* yang ditentukan oleh InSet untuk menentukan *polarity*.

Tabel 9. Hasil *polarity* dari InSet

	Tidak menggunakan stopword	Stopword
Positive	745	1788
Negative	7761	6428
neutral	228	518

Tabel 9 menunjukkan hasil polaritas dari InSet, dibandingkan antara penggunaan stopword dan tidak menggunakan stopword. Polaritas ini mengukur sentimen dari suatu teks, yang dapat bersifat positif, negatif, atau netral.



Gambar 3. Sentimen *polarity*

Gambar 3 merupakan hasil dari perbandingan sentimen *polarity* dari hasil tidak menggunakan stopword dan menggunakan stopword.

Tabel 10. Hasil dari *polarity stopword*

Hasil stopword	Polarity score	Polarity
['semoga', 'angin', 'surga', 'wkwkwkwkwk', 'ambyar', 'proyek', 'kota', 'negara', 'dana']	3	positive
['ayo', 'asn', 'pilih', 'memaksa', 'kota', 'negara']	1	positive
['diajak', 'rembukan', 'dilibatkan', 'proyek', 'kota', 'negara']	3	positive

Tabel 11. hasil dari *polarity* tidak *stopword*

Hasil no stopword	Polarity _score	Polarity
['semoga', 'hanya', 'diberi', 'angin', 'surga', 'saja', 'wkwkwkwkwk', 'ambyar', 'itu', 'proyek', 'ibu', 'kota', 'negara', 'kurang', 'dana']	-8	negative
['ayo', 'asn', 'masih', 'pilih', 'yang', 'memaksa', 'ke', 'ibu', 'kota', 'negara']	-4	negative
['apa', 'tidak', 'diajak', 'rembukan', 'tidak', 'dilibatkan', 'dalam', 'proyek', 'ibu', 'kota', 'negara']	-7	negative

Berdasarkan tabel 10 dan 11 terlihat bahwa InSet mampu melakukan analisis. Akan tetapi dari 2 tabel tersebut memiliki *polarity* yang berbeda, kenapa *polarity* pada tabel 11 lebih sesuai polarity nya dari pada tabel 10? Karena dalam kalimat "semoga hanya diberi angin surga saja wkwkwkwkwk ambyar itu proyek ibu kota negara kurang dana" kata "kurang" merupakan kata adverbia yang menyampaikan informasi negative terhadap pemindahan IKN. Jika kata "kurang" dianggap sebagai kata penghubung dan dihapus dalam proses penghapusan *stopword*, maka informasi sentimen tersebut akan hilang. Oleh karena itu, penting untuk mengidentifikasi kata adverbia seperti "kurang" dan mempertahankan nilai sentimen yang terkandung dalam kalimat tersebut.

Dalam kasus lain, dari kata "apa tidak diajak rembukan tidak dilibatkan dalam proyek ibu kota negara" kata "tidak" adalah kata adverbia yang memberikan informasi sentimen negative terhadap pemindahan IKN. Jika kata "tidak" dianggap sebagai kata penghubung dan di hapus, maka informasi sentimen negative tersebut juga akan hilang. Oleh karena itu, identifikasi kata adverbia seperti "tidak" sangat penting agar nilai sentimen yang tepat dapat dipertahankan.

Dengan demikian, dalam mengatasi permasalahan *stopword*, penting untuk mengidentifikasi kata-kata adverbia yang memiliki nilai sentimen, seperti "kurang" atau "tidak," dan memperlakukan kata-kata tersebut secara khusus agar informasi sentimen yang relevan tidak terhapus dalam proses penghapusan *stopword*.

Berikut ini adalah hasil dari evaluasi menggunakan *Confusion Matrix* untuk melihat hasil mana yang lebih stabil menggunakan *average macro* dan *average weighted*.

Tabel 12. Hasil evaluasi *Confusion Matrix*

	macro		weighted	
	no stopword	Stop word	no stop word	Stopword
Accuracy	88,86 %	73,6 %	88,86 %	73,6 %
Recall	29,62 %	24,53 %	88,86 %	73,60 %
Precision score	33,33 %	33,33 %	100,0%	100.0 %
F1 score	31,37 %	28,26 %	94,18 %	84,79 %

Tabel 12 merupakan hasil evaluasi model klasifikasi dengan menggunakan Confusion Matrix, membandingkan performa antara penggunaan dan tidak penggunaan stopword. Secara keseluruhan, model tanpa stopword menunjukkan akurasi lebih tinggi, mencapai 88,86%, dibandingkan dengan model yang menggunakan stopword (73,6%). Meskipun precision score tetap konstan pada 33,33% baik dengan atau tanpa stopword, recall menurun dari 29,62% menjadi 24,53% ketika stopword digunakan. Hal ini mengindikasikan penurunan kemampuan model dalam menemukan instance positif. Selain itu, F1 score juga menurun dari 31,37% menjadi 28,26%, menyoroti adanya trade-off antara precision dan recall ketika stopword digunakan. Analisis ini memberikan wawasan tentang dampak penggunaan stopword terhadap kinerja model klasifikasi teks.

Pendekatan *Macro Average* juga digunakan untuk mengukur rata-rata skor akurasi, presisi, recall, dan F1 dari semua kelas. Rata-rata Makro yang besar menunjukkan bahwa algoritma bekerja dengan baik di semua kelas dan sebaliknya (Grandini, Bagli and Visani, 2020).

WAE (*Weighted Average Ensemble*) didefinisikan sebagai menggabungkan model terlatih dan menetapkan bobot berdasarkan kinerjanya untuk meningkatkan akurasi sistem secara keseluruhan (Chakraborty et al., 2023).

3.4 Hasil Evaluasi

Evaluasi performa dari hasil penentuan *polarity* sentimen dari *InSet* akan dilakukan dengan *Confusion Matrix* dengan menggunakan average weighted, alasan penggunaan *weighted* dibandingkan dengan *average* lainnya adalah *weighted* mempertimbangkan ketidakseimbangan kelas, Evaluasi ini juga menggunakan rata-rata tertimbang (*average weighted*) untuk memberikan bobot yang berbeda kepada setiap kelas *tweet* dalam perhitungan nilai rata-rata.

Jumlah data yang digunakan dalam evaluasi ini adalah 12.520 data *tweet*, terdapat 3 label *negative* terdapat 8734 data *tweet*, *positive* terdapat 1940 data *tweet*, dan *neutral* terdapat 1846 data *tweet*, dalam penentuan *polarity* ini peneliti menggunakan 2 perbandingan yaitu nilai *Confusion Matrix* antara tidak menggunakan normalisasi slang dan

menggunakan normalisasi slang dan menggunakan *average weighted* untuk menghitung bobotnya.

Nomor dan judul gambar ditulis diposisi tengah kolom (*center alignmen*t). Nomor gambar ditulis sesuai dengan urutannya menggunakan angka arab. Judul gambar ditulis dibagian bawah gambar dengan cara *title case*, kecuali untuk kata sambung kata sambung dan kata depan. Judul gambar menggunakan ukuran huruf 8 (delapan). Gambar tidak boleh melebihi batas margin dari tiap kolom, kecuali jika ukuran gambar yang besar tidak cukup dalam 1 kolom, maka dapat melintasi 2 kolom.

Tabel 13. Hasil evaluasi

	No slank	Normalisasi slang
Accuracy	58,31 %	66,66 %
Recall	58,31 %	66,66 %
Precision score	60,1 %	58,85 %
F1 score	58,86 %	61,40 %

Tabel 13 ini menjelaskan tentang hasil penilaian matriks dari hasil klasifikasi sentimen dengan InSet *lexicon*. Berdasarkan tabel diatas terlihat bahwa terjadi kenaikan dari tidak menggunakan slang ke menggunakan slang naik sebesar 8,35 % untuk *accuracy*, 8,35 % untuk *recall* dan 2,54 % *F1 score*. Hasil ini menunjukkan bahwa tahapan *preprocessing* yang diterapkan akan memberikan dampak positif dalam mengoptimalkan data sehingga performa klasifikasi sentiment menjadi lebih baik.

Performa metode InSet cukup baik karena mencapai *accuracy* sebesar 66,66 % dan F1-score sebesar 61,40 %, meskipun hanya terdapat sedikit kenaikan performa dari yang tidak menggunakan slang, hal ini disebabkan kesalahan Dalam pelabelan oleh InSet ditunjukkan pada Tabel 14, kesalahan ini dapat terjadi karena beberapa *tweet* tidak berisi daftar kata *negative* yang termasuk dalam kosakata InSet saat diberi label *negative* oleh annotator.

Terdapat beberapa perbandingan penelitian ini dengan penelitian sebelumnya dari hasil perhitungan *F1-SCORE*. Dalam penelitian ini mendapatkan hasil akhir *F1-SCORE* mencapai 61,40% sedangkan penelitian sebelumnya mendapatkan hasil *F1-SCORE* mencapai 98,36%. Kenapa penelitian sebelumnya lebih tinggi *F1-SCORE* dari pada penelitian ini..? karena penelitian sebelumnya menggunakan kamus tidak baku yang di buat sendiri. Hal ini bisa mempengaruhi kualitas kualitas normalisasi kata-kata tidak baku, yang berdampak pada representasai teks dan akurasi klasifikasi, penelitian ini juga melakukan perbandingan normalisasi dengan proses perbaikan kata tidak baku menggunakan kamus tidak baku dan normalisasi *Levenshtein Distance*. Perbandingan hasil akhir dari penelitian ini adalah menggunakan normalisasi *Levenshtein Distance* yang mendapatkan perbaikan kata tidak baku adalah sebesar 98.33% Dan untuk *precision*, *recall*, dan *f- measure* adalah 96.77%, 100%, dan 98.36%. (Antinasari, Perdana and Fauzi, 2017).

Perbandingan dengan penelitian sebelumnya menunjukkan perubahan fokus dan metode dalam mengatasi tantangan analisis teks di media sosial. Penelitian sebelumnya, yang fokus pada evaluasi opini masyarakat terhadap layanan transportasi publik, mencapai akurasi 67.05% dengan Naïve Bayes dan normalisasi kata menggunakan Levenshtein Distance. Sementara itu, penelitian ini lebih fokus pada pengembangan sistem otomatis membedakan kata baku dan tidak baku di Twitter dengan metode lexicon-based dan normalisasi kata menggunakan KBBI. Meskipun akurasi mencapai 66,66%, perbedaan dalam fokus, metode preprocessing, dan bahasa mungkin menjadi penyebabnya. Penelitian ini memberikan kontribusi pada pemahaman bahasa tidak baku di media sosial, meskipun dengan tantangan berbeda dari penelitian sebelumnya yang lebih menekankan evaluasi opini masyarakat terhadap layanan spesifik (Rozi, Ardiansyah and Rebeka, 2019).

3.5 Kesimpulan

Penelitian ini berhasil melakukan otomatisasi pendeteksi kata baku dan tidak baku pada Twitter. Beberapa contoh kata tidak baku yang teridentifikasi adalah "lho", "yg", "ikn", "otorita", "bebani", "modarlah", "memindahkan", dan "mengeluh". Selanjutnya, kata-kata dengan imbuhan seperti "memindahkan", "modarlah", dan "mengeluh" dihapus. Selain itu, dilakukan proses normalisasi seperti mengganti "yg" menjadi "yang", "sdh" menjadi "sudah", dan "tdk" menjadi "tidak".

Normalisasi kata tidak baku meningkatkan pemahaman dan makna *tweet* dengan mengubahnya ke bentuk standar. Hal ini juga meningkatkan kualitas analisis teks secara keseluruhan dengan mengurangi variasi dan kompleksitas kata tidak baku. Dengan demikian, peneliti dapat lebih fokus pada aspek inti analisis seperti sentimen, tema, atau tren yang terdapat dalam *tweet*.

Dalam melakukan normalisasi kata tidak baku penting untuk meningkatkan pemahaman, konsistensi, dan kualitas analisis data *tweet*. Dengan mengonversi slang ke bentuk yang lebih standar, Anda dapat mengurangi ambiguitas, dan memastikan hasil analisis yang lebih akurat dan relevan.

Berdasarkan hasil dari tahap pengujian yang telah dilakukan terhadap data *tweet* dengan menggunakan *InSet lexicon*, Sebagai metode penentuan *polarity*, hasil awal sejalan dengan hasil akhir bahwa langkah *preprocessing* meningkatkan performa metode penentuan *polarity*, yang menghasilkan *accuracy* InSet sebesar 66,66 % F1 score sebesar 61,40 %.

Pada penelitian selanjutnya, disarankan untuk menerapkan filtering yang lebih ketat terhadap kata-kata adverbia. Dengan filtering yang sistematis dan kriteria yang jelas, adverbia yang digunakan dalam analisis akan memiliki relevansi yang kuat dan memberikan informasi penting. Dalam upaya

memperoleh hasil yang lebih akurat dan terfokus, kami merekomendasikan penerapan filtering pada kata-kata adverbia. Hal ini akan meningkatkan kualitas dan kejelasan hasil penelitian, serta memberikan kontribusi yang lebih signifikan.

DAFTAR PUSTAKA

A., V. and SONAWANE, S.S., 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*, 139(11), pp.5–15. https://doi.org/10.5120/ijca2016908625.

ANGELINA, S.J., BIJAKSANA, A., NEGARA, P. and MUHARDI, H., 2023. Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia. [online] 02(1), pp.165–173. https://doi.org/10.26418/juara.v2i1.69680.

ANTINASARI, P., PERDANA, R.S. and FAUZI, M.A., 2017. Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, [online] 1(12), pp.1718–1724. Available at: <http://j-ptiik.ub.ac.id>.

ARIEF, R. and IMANUEL, K., 2019. ANALISIS SENTIMEN TOPIK VIRAL DESA PENARI PADA MEDIA SOSIAL TWITTER DENGAN METODE LEXICON BASED. *Jurnal Ilmiah Matrik*, [online] 21(3), pp.242–250. https://doi.org/10.33557/jurnalmatrik.v21i3.727.

AZHAR, Y., 2018. Metode Lexicon-Learning Based Untuk Identifikasi Tweet Opini Berbahasa Indonesia. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, 6(3), p.237. https://doi.org/10.23887/janapati.v6i3.11739.

BENGI RUHAMAH, ADNAN, H., 2018. KEMAMPUAN SISWA DALAM MEMBEDAKAN KATA BAKU DAN KATA TIDAK BAKU DI KELAS V SDNEGERI 3 BANDA ACEHNo Title. *Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, 3, pp.160–163.

BHATIA, S., SHARMA, M. and BHATIA, K.K., 2018. Sentiment Analysis and Mining of Opinions. *Studies in Big Data*, 30(May), pp.503–523. https://doi.org/10.1007/978-3-319-60435-0_20.

CHAKRABORTY, G.S., BATRA, S., SINGH, A., MUHAMMAD, G., TORRES, V.Y. and MAHAJAN, M., 2023. A Novel Deep Learning-Based Classification Framework for COVID-19 Assisted with Weighted Average Ensemble Modeling. *Diagnostics*, 13(10). https://doi.org/10.3390/diagnostics13101806.

DARWIS, D., PRATIWI, E.S. and PASARIBU, A.F.O., 2020. Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Edutic - Scientific Journal of Informatics Education*, 7(1), pp.1–11. https://doi.org/10.21107/edutic.v7i1.8779.

DURAN, M.S., AVANÇO, L. and Nunes, M.G.V., 2015. A Normalizer for UGC in Brazilian Portuguese. *ACL-IJCNLP 2015 - Workshop on Noisy User-Generated Text, WNUT 2015 - Proceedings of the Workshop*, (October 2017), pp.38–47. https://doi.org/10.18653/v1/w15-4305.

FADLI, H.F. and HIDAYATULLAH, A.F., 2019. Identifikasi Cyberbullying Pada Media Sosial Twitter Menggunakan Metode Klasifikasi Random Forest. *Automata*.

GRANDINI, M., BAGLI, E. and VISANI, G., 2020. Metrics for Multi-Class Classification: an Overview. [online] pp.1–17. https://doi.org/https://doi.org/10.48550/arXiv.2008.05756.

HERAWATI, R., JUANSAH, D.E. and TISNASARI, S., 2019. Analisis Afiksasi Dalam Kata-Kata Mutiara Pada Caption Di Media Sosial Instagram Dan Implikasinya Terhadap Pembelajaran Bahasa Indonesia Di Smp. *Membaca Bahasa dan Sastra Indonesia*, [online] 4(1), pp.45–50. https://doi.org/http://dx.doi.org/10.30870/jmbsi.v4i1.6236.

HIDAYAT, W., ARDIANSYAH, M. and SETYANTO, A., 2021. Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb. *Edumatic: Jurnal Pendidikan Informatika*, 5(1), pp.11–20. https://doi.org/10.29408/edumatic.v5i1.3125.

IVAN, Y.A.S. and ADIKARA, P.P., 2019. Classification of Indonesian Hate Speech on Twitter Using NaïVe Bayes and Selection of Information Gain Feature with Word Normalization. *Journal of Information Technology Development and Computer Science*, 3(5), pp.4914–4922.

KOTO, F. and RAHMANINGTYAS, G.Y., 2018. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, 2018-Janua(December), pp.391–394. https://doi.org/10.1109/IALP.2017.8300625.

KUMAR, P. and GRUZD, A., 2019. Social media for informal learning: A case of #Twitterstorians. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019-Janua, pp.2527–2535. https://doi.org/10.24251/hicss.2019.304.

MEFTAH, S., SEMMAR, N., SADAT, F. and HX, K.A., 2018. A neural network model for part-of-speech tagging of social media texts. *In*

Proceedings of the Eleventh International Conference on Language Resources and Evaluation, [online] pp.2821–2828. Available at: <https://aclanthology.org/L18-1446>.

NAFAN, M.Z. and AMALIA, A.E., 2019. Kecenderungan Tanggapan Masyarakat terhadap Ekonomi Indonesia berbasis Lexicon Based Sentiment Analysis. *Jurnal Media Informatika Budidarma*, 3(4), p.268. https://doi.org/10.30865/mib.v3i4.1283.

PEBIANA, S., HIDAYATI, N.N., AFRA, D.I.N., NURFADHILAH, E., PRAFITIA, H.A., PRIHANTORO, J., FAJRI, R., ULINIANSYAH, M.T., SANTOSA, A., AINI, L.R., SAHREZA, Y., SUBEKTI, A.H.K.M., PINEM, J.G., ALFIN, M.R., SEPTADI, A., SHALEHA, S., WIBOWANTO, G.S., JARIN, A., GUNARSO, LATIEF, A.D. and RIZA, H., 2022. Experimentation of Various Preprocessing Pipelines for Sentiment Analysis on Twitter Data about New Indonesia's Capital City Using SVM and CNN. *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2022 - Proceedings*. https://doi.org/10.1109/O-COCOSDA202257103.2022.9997982.

RAMACHANDRAN, D. and PARVATHI, R., 2019. Analysis of Twitter Specific Preprocessing Technique for Tweets. *Procedia Computer Science*, 165(2019), pp.245–251. https://doi.org/10.1016/j.procs.2020.01.083.

RASOOL, A., TAO, R., MARJAN, K. and NAVEED, T., 2019. Twitter Sentiment Analysis: A Case Study for Apparel Brands. *Journal of Physics: Conference Series*, 1176(2). https://doi.org/10.1088/1742-6596/1176/2/022015.

RAY, D., 2017. Lexicon Based Sentiment Analysis of Twitter Data. *International Journal for Research in Applied Science and Engineering Technology*, V(X), pp.910–915. https://doi.org/10.22214/ijraset.2017.10130.

RINANDYASWARA, R., SARI, Y.A. and FURQON, M.T., 2022. Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi). *Jurnal Teknologi Informasi dan Ilmu Komputer*, 9(4), p.717. https://doi.org/10.25126/jtiik.2022934707.

ROZI, I.F., ARDIANSYAH, R. and REBEKA, N., 2019. Penerapan Normalisasi Kata Tidak Baku Menggunakan Levenshtein Distance pada Analisa Sentimen Layanan PT. KAI di Twitter. *Seminar Informatika Aplikatif*, [online] pp.106–112. Available at: <http://jurnalti.polinema.ac.id/index.php/SIAP/article/view/563>.

SAINI, S., PUNHANI, R., BATHLA, R. and SHUKLA, V.K., 2019. Sentiment Analysis on Twitter Data using R. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, (April 2020), pp.68–72. https://doi.org/10.1109/ICACTM.2019.8776685.

UTAMA, H.S., ROSIYADI, D., ARIDARMA, D. and PRAKOSO, B.S., 2019. Sentimen Analisis Kebijakan Ganjil Genap Di Tol Bekasi Menggunakan Algoritma Naive Bayes Dengan Optimalisasi Information Gain. *Jurnal Pilar Nusa Mandiri*, 15(2), pp.247–254. https://doi.org/10.33480/pilar.v15i2.705.

Halaman ini sengaja dikosongkan.