JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10585

PENINGKATAN KINERJA METODE LEXICON-BASED DENGAN KOREKSI EJAAN JARO-WINKLER PADA ANALISIS SENTIMEN MEDIA SOSIAL

Virgiawan Asegaf[1], Kemal Farouq Mauladi[2], Nur Qomariyah Nawafilah[3]

[1,2,3]Universitas Islam Lamongan; Jalan Veteran No. 53A, Jetis, Kecamatan Lamongan, Kabupaten Lamongan, Jawa Timur

Keywords:
Analisis Sentimen; Lexicon Based; Jaro-Winkler; Ijazah Palsu.

Corespondent Email:
virgiawanasegaf02@gmail.com

Abstrak. Opini publik di media sosial mengenai isu sensitif seperti dugaan ijazah palsu tokoh publik memiliki dampak besar terhadap dinamika kepercayaan masyarakat. Namun, analisis sentimen berbasis kamus pada teks komentar Instagram kerap terhambat oleh tingginya ketidakteraturan bahasa dan kesalahan pengetikan (typo), yang menurunkan sensitivitas deteksi sistem konvensional. Penelitian ini menerapkan pendekatan analisis sentimen berbasis leksikon yang dioptimasi dengan algoritma jarak string Jaro-Winkler untuk mengevaluasi 5.485 komentar bersih dari akun Instagram @tvonenews. Hasil komputasi menunjukkan bahwa integrasi Jaro-Winkler dengan batas panjang kata ≥ 4 karakter dan ambang batas kemiripan ≥ 0,90 berhasil mendeteksi dan menyelamatkan 570 komentar bermuatan opini yang sebelumnya keliru diklasifikasikan sebagai opini netral oleh leksikon murni. Klasifikasi akhir menghasilkan distribusi sentimen yang didominasi oleh kelas Negatif sebesar 43,8% (2.405 komentar), disusul oleh sentimen Netral sebesar 36,9% (2.025 komentar), dan Positif sebesar 19,2% (1.055 komentar). Penelitian ini membuktikan bahwa kombinasi Leksikon dan Jaro-Winkler sangat efisien meningkatkan kinerja deteksi pada kosakata tidak baku tanpa memerlukan proses pelatihan model data, sekaligus merepresentasikan tingginya kritik dan tuntutan masyarakat akan transparansi hukum.

Abstract. *Public opinion on social media regarding sensitive issues, such as alleged fake diplomas of public figures, significantly impacts the dynamics of societal trust. However, dictionary-based sentiment analysis on Instagram comments is frequently impeded by high linguistic irregularity and typographical errors (typos), which degrade the detection sensitivity of conventional systems. This study applies a lexicon-based sentiment analysis approach optimized with the Jaro-Winkler string distance algorithm to evaluate 5,485 clean comments from the @tvonenews Instagram account. The computational results demonstrate that the Jaro-Winkler integration with a word length limit of ≥ 4 characters and a similarity threshold of ≥ 0.90 successfully detected and recovered 570 opinionated comments that were previously misclassified as neutral by the pure lexicon method. The final classification yielded a distribution dominated by Negative sentiment at 43.8% (2,405 comments), followed by Neutral at 36.9% (2,025 comments), and Positive at 19.2% (1,055 comments). This research proves that combining Lexicon and Jaro-Winkler efficiently enhances detection performance on non-standard vocabulary without requiring model training data, while also highlighting the public's strong criticism and demand for legal transparency.*

1. PENDAHULUAN

Perkembangan pesat media sosial di Indonesia telah mengubah platform daring khususnya Instagram, menjadi ruang publik utama bagi masyarakat untuk menyuarakan opini dan reaksi terhadap berbagai fenomena

sosial, politik, maupun isu hukum yang sedang berkembang [1]. Salah satu topik yang memicu perhatian dan perdebatan intensif di kalangan warganet adalah isu pemberitaan mengenai dugaan ijazah palsu tokoh publik yang diunggah oleh akun media berita resmi @tvonenews. Besarnya volume komentar yang masuk pada unggahan tersebut memuat beragam opini, mulai dari dukungan, keraguan, hingga tuntutan transparansi hukum. Untuk memahami kecenderungan opini publik secara masif dan cepat, diperlukan penerapan teknologi *Natural Language Processing* (NLP) melalui analisis sentimen.

Dalam ranah analisis sentimen, pendekatan berbasis kamus atau *Lexicon-Based* menjadi salah satu metode yang paling diminati karena efisiensinya. Berbeda dengan pendekatan *Supervised Machine Learning* yang membutuhkan proses komputasi pelatihan model yang berat, pendekatan *Lexicon-Based* bekerja secara langsung dengan cara mencocokkan setiap token kata di dalam teks dengan kamus leksikon yang telah memiliki bobot polaritas positif atau negatif [2]. Efektivitas pendekatan leksikon ini telah dibuktikan oleh sejumlah penelitian terdahulu.

Penelitian yang dilakukan oleh [3]. membuktikan bahwa pendekatan *Lexicon-Based* mampu mendeteksi sentimen dan melakukan ekstraksi informasi secara baik pada ulasan pembelajaran daring, penelitian ini berhasil memetakan polaritas persepsi siswa dengan efisiensi komputasi yang tinggi tanpa bergantung pada pemodelan algoritma latih yang kompleks. Penelitian lain oleh [4] juga menerapkan metode *Lexicon-Based* untuk membedah sentimen masyarakat Indonesia terhadap lembaga Danantara. Hasil penelitian tersebut menunjukkan bahwa pendekatan leksikon sangat akurat dalam mengklasifikasikan tanggapan masyarakat ke dalam kelas positif, negatif, dan netral secara cepat pada teks berbahasa Indonesia modern.

Meskipun penelitian terdahulu [3], [4] berhasil membuktikan kehandalan metode *Lexicon-Based*, terdapat kesenjangan penelitian yang mendasar pada arsitektur sistem yang digunakan. Kedua penelitian tersebut masih mengandalkan pencocokan kata berbasis ejaan kaku. Sementara itu, karakteristik utama dari komentar publik di media sosial Indonesia adalah tingginya penggunaan bahasa informal, singkatan, serta kesalahan pengetikan (*typo*) atau variasi ejaan yang ekstrem [5]. Ketika sistem leksikon konvensional dihadapkan pada kosakata tidak baku akibat *typo* tersebut, sistem akan gagal mencocokkannya dengan data kamus, dan secara otomatis membuang bobot sentimen kata tersebut menjadi nilai 0 (Netral). Kelemahan ini memicu tingginya angka kesalahan deteksi opini berbobot menjadi sentimen netral palsu, yang pada akhirnya mendegradasi akurasi dan validitas pemetaan sentimen publik secara keseluruhan.

Berdasarkan analisis kesenjangan tersebut, penelitian ini menawarkan kebaruan dengan mengintegrasikan algoritma pengukuran jarak string Jaro-Winkler Distance ke dalam pendekatan *Lexicon-Based*. Algoritma Jaro-Winkler dirancang untuk mengukur tingkat kemiripan karakter antar-kata dengan memberikan bobot lebih tinggi pada kesamaan karakter di awal kata (*prefix*), sehingga sangat ideal untuk mendeteksi kesalahan ketik [6]. Dengan inovasi ini, sistem mampu mengenali dan menyelamatkan kata-kata bermuatan opini yang *typo* secara dinamis tanpa perlu merombak kamus atau melatih model data baru.

Tujuan utama dari penelitian ini adalah untuk mengimplementasikan dan menguji efektivitas penggabungan metode *Lexicon-Based* dengan algoritma Jaro-Winkler Distance dalam mengklasifikasikan sentimen komentar publik di Instagram @tvonenews terkait dugaan ijazah palsu dan Menganalisis kuantitatif peningkatan sensitivitas deteksi sistem baru dalam menurunkan rasio kesalahan klasifikasi kelas Netral dibandingkan dengan metode leksikon konvensional.

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen merupakan teknik komputasional yang bertujuan untuk mengidentifikasi dan mengekstraksi polaritas emosi dari suatu teks, baik pada tingkat kalimat maupun keseluruhan dokumen. Metode ini bekerja dengan cara membedah kecenderungan opini yang terkandung di dalam objek data, lalu mengklasifikasikannya ke dalam tiga kelas utama, yaitu sentimen positif, negatif, dan netral [7].

2.2Web Scraping

Web scraping merupakan teknik komputasional yang digunakan untuk mengekstraksi data dan informasi dari halaman situs web secara otomatis memanfaatkan skrip atau program komputer. Melalui teknik ini, berbagai elemen data publik seperti teks, gambar, maupun metadata dapat dikumpulkan dari beragam platform daring secara masif dan cepat. Keunggulan utama dari *web scraping* terletak pada otomatisasi proses akuisisi data yang menggantikan metode manual, sehingga memungkinkan integrasi informasi, percepatan proses analisis, serta pelaksanaan penelitian ilmiah berjalan secara lebih efektif dan efisien [8].

2.3Lexicon Based

Pendekatan *Lexicon-Based* merupakan salah satu metode klasifikasi dalam data mining dan pemrosesan teks yang berlandaskan pada pemanfaatan kamus kosakata khusus (leksikon). Setiap entri kata di dalam kamus tersebut telah dianotasi dengan penanda polaritas atau orientasi sentimen tertentu. Melalui pendekatan ini, sistem komputasi dapat mengevaluasi dan mengklasifikasikan muatan emosi di dalam suatu teks secara otomatis ke dalam kategori positif, negatif, maupun netral. Karena efisiensi dan kemudahannya dalam memproses data tanpa memerlukan pelatihan model, metode ini sangat luas diaplikasikan dalam berbagai ranah analisis opini, khususnya pada pemetaan sentimen media sosial dan evaluasi ulasan produk [9].

Bobot polaritas suatu kalimat diperoleh melalui penjumlahan skor polaritas kata positif ($S_{positive}$) dan skor polaritas kata negatif ($S_{negative}$). Perhitungan akumulasi skor untuk masing-masing orientasi sentimen tersebut dirumuskan pada Persamaan (1) dan Persamaan (2) [10].

$$S_{positive} = \sum_{i \in t}^{n} positive\_score_i \quad (1)$$

$$S_{negative} = \sum_{i \in t}^{n} negative\_score_i \quad (2)$$

Keterangan:

- $S_{positive}$ = Total skor polaritas dari kata bermuatan positif dalam satu kalimat
- $S_{negative}$ = Total skor polaritas dari kata bermuatan negatif dalam satu kalimat
- n = Jumlah total kata di dalam kalimat
- $i \in t$ = Token kata ke-i yang cocok dengan entri pada kamus leksikon

Nilai $S_{positive}$ dan $S_{negative}$ yang diperoleh dari kata-kata penyusun kalimat tersebut kemudian digunakan sebagai acuan dalam proses perbandingan. Untuk menentukan klasifikasi orientasi sentimen akhir dari sebuah kalimat, sistem membandingkan jumlah nilai positif dan negatif sesuai Persamaan (3):

$$Sentiment = \begin{cases} positive & \text{if } S_{positive} > S_{negative} \\ neutral & \text{if } S_{positive} = S_{negative} \\ negative & \text{if } S_{positive} < S_{negative} \end{cases} \quad (3)$$

Melalui aturan keputusan pada Persamaan (3), orientasi opini publik dapat diklasifikasikan secara otomatis dan akurat ke dalam kategori positif, netral, maupun negatif berdasarkan dominasi skor polaritas kata yang terkandung di dalamnya.

2.4Jaro Winkler

Jaro-Winkler adalah metode *string matching* yang bertumpu pada pengukuran jarak antar karakter untuk menilai kemiripan dua kata. Algoritma ini beroperasi berdasarkan tiga komponen dasar yaitu penentuan panjang string, identifikasi karakter yang serupa, serta kalkulasi transposisi karakter [11]. Nilai dasar Jaro Distance (dj) dikalkulasi melalui Persamaan (4).

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \quad (4)$$

Keterangan:

d_j = Nilai kemiripan Jaro Distance
M = Jumlah karakter yang cocok
$|s_1|$ = Panjang karakter dari string 1
$|s_2|$ = Panjang karakter dari string 2
t = Jumlah transposisi (*transpositions*)

Untuk menentukan nilai transposisi (t), sistem menghitung jumlah karakter yang berpasangan namun berada pada indeks posisi yang berbeda (n), sebagaimana dinyatakan pada Persamaan (5).

$$t = \frac{n}{2} \quad (5)$$

Keterangan:

n = Jumlah karakter yang cocok namun urutan posisinya tidak sesuai.

Algoritma Jaro-Winkler (dw) kemudian menyempurnakan skor dasar tersebut dengan memanfaatkan skala awalan untuk memberikan bobot penilaian yang lebih tinggi apabila terdapat kesamaan karakter pada awal kata. Rumus akhir Jaro-Winkler Distance dirumuskan pada Persamaan (6).

$$d_w = d_j + \left(l \times p \times \left(1 - d_j\right)\right) \quad (6)$$

Keterangan:

d_w = Nilai kemiripan akhir Jaro-Winkler Distance

d_j = Nilai Jaro Distance antara string 1 dan string 2

l = Panjang awalan (*prefix*) yang sama persis di awal string, dengan batas maksimal 4 karakter awal

p = Konstanta faktor pembesar, di mana nilai standar yang ditetapkan oleh Winkler adalah 0,1.

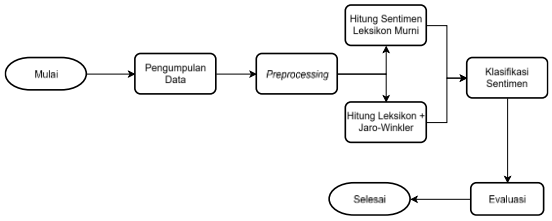
2.5 Preprocessing

Preprocessing merupakan serangkaian tahapan komputasi awal yang wajib dilakukan dalam NLP untuk membersihkan dan menstandardisasi teks mentah agar dapat diproses oleh algoritma komputer secara optimal [12]. Tahapan standar dalam prapemrosesan teks meliputi:

1. *Case Folding*: Proses konversi seluruh karakter huruf kapital di dalam teks menjadi huruf kecil [12].
2. *Cleaning*: Pembersihan teks dari elemen non-alfabet yang tidak relevan, seperti tautan URL, mention username, tanda baca, angka, emoji, dan simbol [13].
3. Normalisasi: Transformasi kata-kata tidak baku, bahasa gaul (*slang*), atau singkatan menjadi kata formal sesuai tata bahasa standar memanfaatkan kamus khusus [14].
4. *Tokenizing*: Pemecahan struktur kalimat utuh menjadi deretan satuan kata (token) [13].
5. *Stopword Removal*: Pengeliminasian kata-kata umum yang memiliki frekuensi kemunculan tinggi namun tidak memiliki bobot polaritas sentimen, seperti kata hubung [14].
6. *Stemming*: Proses pengembalian kata berimbuhan menjadi bentuk kata dasar dengan menghilangkan awalan, sisipan, maupun akhiran [13].

3. METODE PENELITIAN

Fokus utama penelitian adalah mengekstraksi opini publik dari data teks di media sosial serta membandingkan performa dua pendekatan algoritma dalam menangani ketidakteraturan bahasa seperti *typo*. Secara garis besar alur kerja dalam penelitian ini digambarkan melalui bagan alir pada Gambar 1.



Gambar 1. Alur Penelitian

3.1 Pengumpulan Data

Pengumpulan data primer dilakukan melalui teknik *web scraping* menggunakan bahasa pemrograman Python dengan pustaka instagrapi untuk mengekstraksi komentar publik dari akun Instagram @tvonenews terkait isu dugaan ijazah palsu yang diunggah pada bulan November 2025. Keseluruhan data mentah tersebut kemudian diekspor ke dalam berkas berformat CSV.

3.2 Preprocessing

Komentar yang diekstraksi dari Instagram umumnya dipenuhi dengan noise berupa tautan, mention, emoji, angka, serta format huruf yang tidak beraturan. Oleh karena itu, data mentah dibersihkan melalui serangkaian tahapan *preprocessing,* mulai dari *Case Folding*, *Cleaning,* Normalisasi, *Tokenizing, Stopword Removal,* dan *Stemming.*

3.3 Hitung Sentimen Leksikon Murni

Sistem mencocokkan setiap token kata dengan kamus leksikon sentimen (positif dan negatif). Pencocokan pada tahap ini dilakukan secara kaku (*exact match*), yang berarti sebuah kata hanya akan mendapatkan bobot sentimen (positif +1 atau negatif -1) jika ejaan hurufnya identik 100% dengan kata di dalam kamus.

3.4 Hitung Sentimen Leksikon + Jaro-Winkler

Sistem mengevaluasi kata menggunakan algoritma pengukuran jarak string Jaro-Winkler. Untuk mencegah kesalahan deteksi pada kosakata bersuku kata pendek yang padat

makna, perhitungan Jaro-Winkler dikunci dan hanya dieksekusi pada kata yang memiliki panjang minimal 4 karakter. Apabila suatu kata mengalami kesalahan ketik (*typo*) namun skor kemiripannya dengan entri di kamus leksikon mencapai ambang batas (*threshold*) kemiripan ≥ 0.90, maka kata tersebut secara tetap diakui dan diberikan bobot sentimen.

3.5 Klasifikasi Sentimen

Tahap ini merupakan proses agregasi nilai untuk menentukan arah polaritas opini dari satu kalimat komentar secara utuh. Sistem menjumlahkan seluruh bobot skor kata yang ada di dalam satu baris komentar untuk menghasilkan skor sentimen. Keputusan klasifikasi akhir ditarik berdasarkan aturan: nilai total skor > 0 diklasifikasikan ke dalam kelas Positif, nilai skor < 0 diklasifikasikan ke dalam kelas Negatif, dan nilai skor = 0 diklasifikasikan ke dalam kelas Netral.

3.6 Evaluasi

Tahap terakhir dalam metodologi penelitian ini adalah melakukan evaluasi komparatif terhadap hasil klasifikasi dari kedua metode perhitungan. Evaluasi difokuskan pada analisis kuantitatif terhadap selisih jumlah distribusi komentar yang berhasil dideteksi sentimennya oleh sistem Jaro-Winkler dibandingkan dengan sistem Leksikon Murni. Kinerja algoritma dibuktikan berdasarkan kemampuannya mendeteksi dan menyelamatkan kata-kata bermuatan opini dengan tingkat *typo* ekstrem, yang sebelumnya diklasifikasikan secara keliru sebagai opini Netral oleh metode leksikon konvensional.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Pengumpulan Data

Tahap pengumpulan data menggunakan teknik *web scraping* dari akun Instagram @tvonenews berhasil mengekstraksi sebanyak 6.000 komentar mentah. Data tersebut memuat atribut username, teks komentar, serta tautan sumber unggahan. Karena data diambil langsung dari media sosial, format teks masih sangat tidak terstruktur dan dipenuhi oleh karakter khusus, emoji, serta bahasa slang yang memerlukan proses pembersihan sebelum masuk ke tahap komputasi algoritma. Hasil pengumpulan data dapat dilihat pada Gambar 2.

username	komentar	link_postingan
rtri_120	Udah paling bener jadi ateis daripada beragama tapi munafik 😭	https://www.instagram.com/p/DQ3mVSOiB6Z/
yuliari_liaw	Mereka bukan peneliti., mereka para pembenci!	https://www.instagram.com/p/DQwRQYLj0PF/
metrosquare	Termut dredeg... 😭	https://www.instagram.com/p/DQ9KC9jD-Qh/
ki_harjanto	Cah koplak	https://www.instagram.com/p/DRPX0F5jScp/
b412_c1213_0	🗣️ sip besok komisaris	https://www.instagram.com/p/DQ3mVSOiB6Z/
juanrakas123	Aki2 pikun 🤣🤣	https://www.instagram.com/p/DQ3mVSOiB6Z/
endyjaya07	Rakyat temak si anues ya sur... 😭😭😭	https://www.instagram.com/p/DQ_QMyiDQnj/
abdullahansory	Termut juga ternyata dia😭	https://www.instagram.com/p/DQ3mVSOiB6Z/
rama.jetz	Benahi pak , semua butuh kepastian	https://www.instagram.com/p/DQ9KC9jD-Qh/
primaregsi	Hati ibis	https://www.instagram.com/p/DRPX0F5jScp/
triscibalong	Saya pribadi rakyat tidak butuh ijazah yg kami butuh sistim krjnya nyata, tangkap Roy cs	https://www.instagram.com/p/DQwRQYLj0PF/
kusnari.bashyar	Teliti ndasmu jelas2 fitnah pekok	https://www.instagram.com/p/DQwRQYLj0PF/
silo19442	Kebalikan	https://www.instagram.com/p/DRPX0F5jScp/
muztofa_ibrah	🤣🤣🤣orang kalau terlalu pinter ya begini.... 😭😭	https://www.instagram.com/p/DQwRQYLj0PF/
mhaduli	Pikun y	https://www.instagram.com/p/DQ3mVSOiB6Z/
gus_iba002	Caiiiir, Caiiir, caiiiir....	https://www.instagram.com/p/DQ3mVSOiB6Z/
nunung_sam	Allah akan membalaskan.... Mas Roy cs dim lindungan Allah sll.... Balasan itu nyata....	https://www.instagram.com/p/DRSKDPyFSEt/
onehartanto	Muka monyet	https://www.instagram.com/p/DQ3mVSOiB6Z/
abdilah544	Termul2 buat pasukan baru namanya BEM Pesantren, sejak kapan pesantren se Indonesia menduku	https://www.instagram.com/p/DQzBWmdjWe0/
bintang_inet	Ringkus semua	https://www.instagram.com/p/DRHSH1gjXST/

Gambar 2. *Dataset* Komentar

4.2 Hasil *Preprocessing*

Pada tahap ini, 6.000 data komentar mentah diproses melalui enam tahapan pembersihan teks secara berurutan. Selama proses ini, sistem juga melakukan penyaringan otomatis terhadap baris data yang kosong atau data yang hanya berisi elemen noise sehingga menjadi kosong setelah dibersihkan. Dari 6.000 data mentah, tersisa 5.485 data bersih yang siap dievaluasi.

Untuk memberikan gambaran jelas mengenai perubahan teks pada setiap prosesnya, hasil *preprocessing* disajikan ke dalam enam tabel sebagai berikut:

1. *Case Folding*
   Tahap pertama adalah *Case Folding* untuk menyeragamkan seluruh huruf kapital pada teks komentar menjadi huruf kecil

Tabel 1. Hasil *Case Folding*

Sebelum	Sesudah
Mereka bukan peneliti., mereka para pembenci!	mereka bukan peneliti., mereka para pembenci!
Termul juga ternyata dia😂	termul juga ternyata dia😂

2. *Cleaning*
   Tahap kedua adalah *Cleaning*, Pada tahap ini, elemen noise seperti sebutan akun (username/mention), tanda baca, angka, simbol, hashtag, tautan URL, hingga emoji dihapus sepenuhnya dari teks.

Tabel 2. Hasil *Cleaning*

Sebelum	Sesudah
mereka bukan peneliti., mereka para pembenci!	mereka bukan peneliti mereka para pembenci
termul juga ternyata dia😂	termul juga ternyata dia

3. Normaliasi
   Tahap ketiga adalah Normalisasi, yakni proses mengoreksi kata-kata *slang*,

singkatan, atau bahasa tidak baku menjadi bentuk kata formal.

Tabel 3. Hasil Normalisasi

Sebelum	Sesudah
mereka bukan peneliti mereka para pembenci	mereka bukan peneliti mereka para pembenci
termul juga ternyata dia	termul juga ternyata dia

4. *Tokenizing*
   Tahap keempat adalah *Tokenizing*, yaitu proses memecah susunan kalimat utuh hasil normalisasi menjadi deretan potongan kata (token).

Tabel 4. Hasil *Tokenizing*

Sebelum	Sesudah
mereka bukan peneliti mereka para pembenci	[mereka, bukan, peneliti, mereka, para, pembenci]
termul juga ternyata dia	[termul, juga, ternyata, dia]

5. *Stopword Removal*
   Tahap kelima adalah *Stopword Removal*, yaitu proses menyaring dan membuang kata-kata umum atau kata hubung yang tidak memiliki muatan informasi atau sentimen.

Tabel 5. Hasil *Stopword Removal*

Sebelum	Sesudah
[mereka, bukan, peneliti, mereka, para, pembenci]	[peneliti, pembenci]
[termul, juga, ternyata, dia]	[termul, ternyata]

6. *Stemming*
   Tahap terakhir adalah *Stemming*, yaitu proses mengubah kata-kata berimbuhan menjadi kata dasar.

Tabel 6. Hasil *Stemming*

Sebelum	Sesudah
[peneliti, pembenci]	teliti benci
[termul, ternyata]	termul nyata

4.3 Perbandingan Kinerja

Fokus utama dalam pembahasan ini adalah mengevaluasi peningkatan performa sistem deteksi leksikal setelah dioptimasi menggunakan algoritma Jaro-Winkler Distance. Pada pendekatan Leksikon Murni, sistem melakukan pencocokan kata secara kaku (*exact match*), yang menyebabkan banyak kata bermuatan sentimen tidak terdeteksi akibat adanya variasi kesalahan pengetikan (*typo*). Hasilnya, sejumlah besar komentar yang seharusnya memiliki polaritas sentimen justru terklasifikasi sebagai opini Netral.

Ketika algoritma Jaro-Winkler diaktifkan untuk mengukur kedekatan *string* dengan batas karakter ≥ 4 dan *threshold* $\geq 0{,}90$, sistem secara dinamis berhasil mengoreksi kata-kata *typo* tersebut. Perbandingan distribusi hasil klasifikasi antara kedua metode ditunjukkan pada Tabel 7.

Tabel 7. Perbandingan Deteksi Sentimen

Sentimen	Leksikon Murni	Leksikon + JW	Peningkatan Deteksi
Positif	864	1055	191
Netral	2595	2025	-570
Negatif	2026	2405	379

Berdasarkan Tabel 7, terlihat jelas adanya pergeseran distribusi data yang signifikan. Pada kolom Peningkatan Deteksi, kelas Netral menunjukkan angka -570. Angka negatif ini bukan merepresentasikan hilangnya data, melainkan membuktikan bahwa Jaro-Winkler berhasil menyelamatkan 570 komentar yang sebelumnya gagal dikenali dan terkategorikan Netral oleh leksikon murni. Ke-570 komentar hasil koreksi tersebut kemudian berhasil dipetakan ke dalam polaritas yang sebenarnya, sehingga memberikan kontribusi penambahan pada kelas Positif sebanyak 191 data dan kelas Negatif sebanyak 379 data.

Sebagai contoh pembuktian kualitatif dari peningkatan tersebut, terdapat komentar warganet "rakyat terbah tiroris .. 😭😭". Setelah melalui tahap *preprocessing*, teks tersebut menyisakan token kata dasar "bah tiroris". Pada pendekatan Leksikon Murni, sistem gagal mendeteksi muatan sentimen pada kalimat ini karena kata "tiroris" tidak terdaftar secara baku di dalam kamus, sehingga komentar ini secara otomatis dilabeli sebagai opini Netral. Namun, ketika dievaluasi menggunakan pendekatan Leksikon + Jaro-Winkler, sistem berhasil mendeteksi adanya kedekatan pola karakter antara token "tiroris" dengan kata baku "teroris" yang terdapat di dalam kamus negatif. Hasil kalkulasi algoritma Jaro-Winkler menghasilkan skor kemiripan

sebesar 0,91. Karena skor tersebut berhasil melampaui threshold minimum 0,90, sistem secara dinamis mengakui kata tersebut sebagai variasi *typo*, memberikan bobot -1, dan mengoreksi klasifikasi kalimat tersebut menjadi sentimen Negatif.

Contoh implementasi keberhasilan algoritma Jaro-Winkler lainnya dalam mengoreksi deteksi leksikal pada kata yang mengalami *typo* dapat dilihat pada Tabel 8.

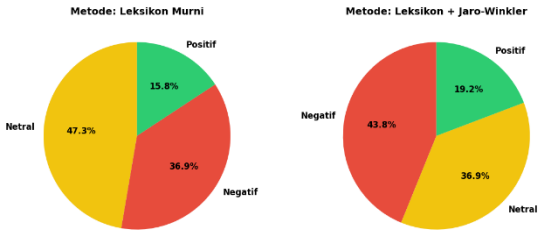
Tabel 8. Hasil Jaro Winkler

Kata *Typo*	Kata Kamus	Bobot
tiroris	teroris	0.91
brisik	berisik	0.96
korup	korupsi	0.94
buzer	buzzer	0.96
bacoct	bacot	0.97

Berdasarkan data pada Tabel 8, terlihat jelas tingkat efektivitas algoritma Jaro-Winkler dalam menangani berbagai bentuk ketidakteraturan leksikal yang sangat umum terjadi di media sosial. Algoritma ini terbukti mampu mengidentifikasi beragam pola kesalahan, mulai dari penghilangan huruf seperti "brisik", “korup” dan "buzer". pergantian huruf seperti "tiroris", hingga penambahan huruf akibat kesalahan pengetikan seperti "bacoct". Keseluruhan variasi tersebut berhasil divalidasi dengan skor kemiripan yang sangat tinggi, yakni merentang dari 0,91 hingga 0,97. Fakta ini membuktikan bahwa integrasi Jaro-Winkler berhasil menutupi kelemahan pencocokan kaku pada metode leksikon, menjadikan sistem analisis sentimen jauh lebih adaptif, presisi, dan sensitif dalam menjaring opini warganet.

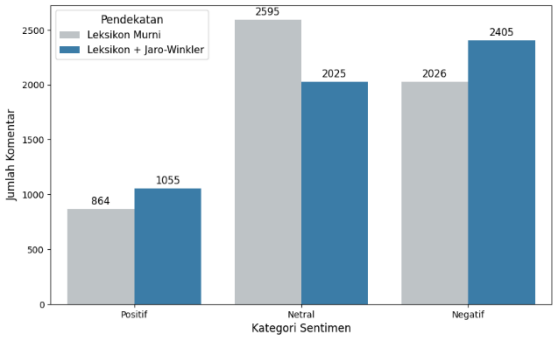
4.4 Hasil Klasifikasi dan Distribusi Sentimen

Berdasarkan agregasi skor menggunakan pendekatan yang telah dioptimasi (Leksikon + Jaro-Winkler), diperoleh hasil klasifikasi akhir terhadap 5.485 data bersih. Persentase pergeseran distribusi opini publik tersebut ditunjukkan pada Gambar 3.



Gambar 3. Presentase Sentimen

Berdasarkan Gambar 3, terlihat perubahan pola opini publik yang sangat jelas. Pada metode Leksikon Murni, komentar didominasi oleh kelas Netral (47,3%). Namun, setelah algoritma Jaro-Winkler diterapkan untuk mengoreksi kata-kata yang mengalami salah ketik (*typo*), sentimen Negatif melonjak menjadi kelas yang paling dominan dengan porsi 43,8%, disusul oleh sentimen Netral sebesar 36,9%, dan sentimen Positif menempati porsi terkecil sebesar 19,2%. Dominasi sentimen negatif ini sangat selaras dengan realitas di lapangan, di mana pemberitaan mengenai dugaan ijazah palsu merupakan isu sensitif yang memicu reaksi kritis, keraguan, serta penolakan keras dari masyarakat. Untuk melihat secara lebih rinci jumlah pergeseran data komentar pada setiap kategori, perbandingan analisis ditampilkan pada Gambar 4.



Gambar 4. Distribusi Jumlah Komentar

Berdasarkan data pada Gambar 4, penerapan aturan batas panjang kata ≥ 4 karakter pada Jaro-Winkler memberikan dampak penyelamatan data yang sangat tinggi. Terjadi penurunan jumlah komentar pada kelas Netral secara drastis sebanyak 570 komentar, yaitu dari 2.595 data pada leksikon murni turun menjadi 2.025 data. Angka penurunan ini membuktikan bahwa algoritma Jaro-Winkler berhasil mengenali dan menyelamatkan 570 komentar bermuatan opini yang sebelumnya

gagal terdeteksi dan dianggap netral karena kata dasarnya ditulis dengan ejaan yang salah. Ke-570 komentar hasil koreksi tersebut kemudian berhasil dikelompokkan ke dalam polaritas emosi yang sebenarnya, sehingga memberikan kontribusi penambahan nyata pada dua kelas lainnya. Kelas positif mengalami peningkatan sebanyak 191 komentar, dari awalnya 864 data naik menjadi 1.055 data. Kelas negatif mengalami peningkatan paling tinggi sebanyak 379 komentar, dari awalnya 2.026 data melonjak menjadi 2.405 data. Tingginya peningkatan pada kelas negatif terjadi karena dalam bahasa percakapan di media sosial, sangat banyak kata dasar bermuatan emosi kuat atau umpatan yang hanya terdiri dari 4 hingga 5 huruf. Selanjutnya untuk membedah substansi atau topik utama yang dibicarakan oleh warganet di dalam kolom komentar, sistem memvisualisasikan frekuensi kemunculan kata ke dalam bentuk *Word Cloud* pada Gambar 5.

Gambar 5. *Word Cloud*

Berdasarkan Gambar 5, ukuran teks merepresentasikan tingkat frekuensi atau seberapa sering suatu kata digunakan oleh warganet. Pada kategori sentimen positif kata-kata yang paling mencolok dan mendominasi adalah "ijazah", "asli", "bukti", "benar", "nyata", "dukung", serta "semangat". Kemunculan kata-kata ini memperlihatkan bahwa sentimen positif dari warganet didorong oleh harapan agar kebenaran segera terungkap. Komentar positif banyak berisi dukungan kepada pihak terkait untuk berani membuktikan keaslian ijazah dengan menghadirkan bukti nyata dan transparan di depan publik. Pada kategori sentimen negatif opini publik didominasi oleh kata-kata bernada tuntutan hukum dan kritikan tajam, seperti kata "hukum", "penjara", "fitnah", "palsu", "rusak", dan "salah". Selain itu, ditemukan juga tingginya frekuensi kata umpatan khas bahasa gaul media sosial seperti "termul", "jilat", "panci", dan "bodoh". Pola kata ini secara jelas merepresentasikan tingginya kemarahan, rasa tidak percaya, serta tuntutan warganet agar aparat penegak hukum segera bertindak tegas dalam menangani kasus dugaan ijazah palsu tersebut.

Meskipun sentimen Negatif berhasil mendominasi opini publik (43,8%), jumlah komentar yang terkelompokkan ke dalam label Netral masih tergolong cukup tinggi, yakni menempati urutan kedua dengan persentase 36,9% (2.025 komentar). Masih tingginya angka sentimen netral ini disebabkan oleh keterbatasan metode *lexicon based* yang hanya menilai kata satu per satu sesuai ejaan asli, tanpa bisa memahami maksud di balik susunan kalimat. Pada topik isu ijazah palsu ini, banyak warganet yang menyampaikan pendapat atau kritikannya melalui kalimat sindiran, candaan, atau pertanyaan sarkas tanpa menggunakan kata negatif yang langsung jelas terlihat. Karena sistem tidak menemukan kata kunci negatif di dalam kamus, kalimat sindiran tersebut otomatis dihitung bernilai nol dan masuk ke dalam kelas Netral. Temuan ini sejalan dengan penelitian terdahulu yang menyatakan bahwa ketidakmampuan metode *lexicon based* dalam mendeteksi kalimat sindiran dan sarkasme menjadi penyebab utama masih tingginya persentase komentar yang salah dikelompokkan ke dalam kelas netral [15].

Terlepas dari keberhasilannya mengoreksi *typo*, algoritma Jaro-Winkler memiliki keterbatasan karena hanya menilai kemiripan susunan huruf tanpa mampu memahami makna kata. Hal ini berisiko menimbulkan kesalahan deteksi (false positive) pada dua kata bermakna berlawanan yang ejaannya mirip, seperti kata "suap" (negatif) dan "siap" (positif), atau "balik" dan "baik". Oleh karena itu, penerapan ambang batas kemiripan tinggi ($\geq 0{,}90$) dan batas panjang kata (≥ 4) pada penelitian ini terbukti menjadi langkah krusial untuk meminimalisir risiko kesalahan pemetaan kata tersebut.

5. KESIMPULAN

Berdasarkan serangkaian tahapan penelitian, mulai dari pengumpulan data, *preprocessing* hingga evaluasi komparatif perhitungan sentimen terhadap komentar publik di Instagram terkait isu dugaan ijazah palsu, maka dapat ditarik kesimpulan dalam bentuk poin-poin sebagai berikut:

1. Integrasi algoritma Jaro-Winkler Distance dengan parameter ambang batas kemiripan ≥ 0,90 dan panjang kata ≥ 4 karakter terbukti sangat efektif dalam mengatasi kelemahan pencocokan kata kaku pada metode leksikon murni. Dari 5.485 data bersih yang dievaluasi, sistem berhasil mengoreksi kesalahan ejaan (*typo*) pada kata-kata pendek maupun panjang, sehingga mampu menyelamatkan 570 komentar bermuatan sentimen yang sebelumnya salah dikelompokkan sebagai opini Netral.
2. Hasil klasifikasi akhir memperlihatkan bahwa opini warganet didominasi oleh sentimen Negatif sebesar 43,8% (2.405 komentar), disusul oleh kelas Netral sebesar 36,9% (2.025 komentar), dan Positif sebesar 19,2% (1.055 komentar). Dominasi sentimen negatif ini menjadi bukti empiris yang merepresentasikan tingginya keraguan publik, kritik tajam, serta tuntutan masyarakat terhadap keterbukaan informasi dan penegakan hukum yang transparan.
3. Meskipun turun drastis, sentimen Netral masih menempati urutan kedua (36,9%). Hal ini disebabkan oleh perilaku warganet yang kerap menyampaikan pendapatnya melalui kalimat sindiran, candaan, atau pertanyaan tanpa kata kunci negatif yang langsung terlihat. Karena metode leksikon hanya menilai kata satu per satu, kalimat sindiran tersebut tidak dikenali polaritasnya dan otomatis dihitung sebagai kelas netral.
4. Penelitian selanjutnya disarankan untuk menerapkan pendekatan hibrida yang memadukan metode Leksikon dan Jaro-Winkler dengan pemodelan semantik seperti Word Embedding (Word2Vec/FastText) atau Deep Learning (IndoBERT), serta menambahkan modul pendeteksi sarkasme, guna memvalidasi kedekatan makna kata, memahami konteks kalimat, dan meminimalisir kesalahan klasifikasi pada kelas Netral.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] R. Sri, R. A. Kartodinoto, D. Purnomo, A. W. Utomo, U. Kristen, and S. Wacana, "Analisis Instagram Sebagai Ruang Publik Terhadap Respon Generasi-Z Dalam Kontestasi Pilpres 2024 (Studi Pada Generasi-Z Di Kota Salatiga)," vol. 17, no. 1, pp. 20–29, 2024.

[2] M. Resa, A. Yudianto, A. Rahim, P. Sukmasetya, and R. A. Hasani, "Perbandingan Metode Support Vector Machine Dengan Metode Lexicon Dalam Analisis Sentimen Bahasa Indonesia," *Jurnal Teknologi Informasi)*, vol. 6, no. 1, [Online]. Available: https://github.com/fajri91/InSet.

[3] M. Hamka and D. Ratna Sari, "Analisis Sentimen Dan Information Extraction Pembelajaran Daring Menggunakan Pendekatan Lexicon," 2022.

[4] S. A. Nugraha, "Penerapan Lexicon Based Untuk Analisis Sentimen Masyarakat Indonesia Terhadap Danantara," 2025.

[5] Fitria Wulandari, Elisabet Simanjuntak, Piladeplia Trifosa Ginting, Raisha Mei Nabel Pakpahan, and Anggia Puteri, "Kesalahan Berbahasa dalam Penulisan di Media Sosial Kajian Sintaksis dan Semantik," *Sintaksis : Publikasi Para ahli Bahasa dan Sastra Inggris*, vol. 3, no. 1, pp. 01–09, Dec. 2024, doi: 10.61132/sintaksis.v3i1.1238.

[6] Darmanto, Kharisma, Muhammad Ar-Razy, Pradasari Novi Indah, and Wahyudi Eka, "Perbandingan Kinerja Algoritma Jaro-Winkler dan Levenshtein Distance untuk Deteksi Kesalahan Penulisan Bahasa Indonesia pada Karya Ilmiah Mahasiswa," *JURTI Jurnal Rekayasa Teknologi Informasi*, vol. 9, pp. 112–121, Jun. 2025.

[7] K. Khotimah, M. Martanto, A. R. Dikananda, and A. Rifa'i, "Analisis Sentimen Ulasan Aplikasi Pintu Di Google Play Store Menggunakan Algoritma Naive Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5789.

[8] D. Chrisinta and J. E. Simarmata, "Eksplorasi Teknik Web Scraping pada Data Mining: Pendekatan Pencarian Data Berbasis Python," *Faktor Exacta*, vol. 17, no. 1, May 2024, doi: 10.30998/faktorexacta.v17i1.22393.

[9] S. Ratnaswari, N. C. Wibowo, and D. S. Y. Kartika, "Analisis Sentimen Menggunakan Metode Lexicon-Based Dan Support Vector Machine Pada Presiden Dan Wakil Presiden Indonesia Periode 2024–2029," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5604.

[10] M. Bagas, D. Putra, and E. Setiawan, "Metode Lexicon Based Untuk Analisis Sentimen Pengguna Twitter Terhadap Kinerja Isp (Studi Kasus : Indihome, Biznet, Myrepublic)," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 6, pp. 12079–12087, 2024.

[11] M. Syahrullah, F. H. Rachman, and I. O. Suzanti, "Deteksi Kemiripan Dokumen Abstrak Skripsi menggunakan Metode Jaro-Winkler Distance dan Synonym Recognition," *Sains Data Jurnal Studi Matematika dan Teknologi*, vol. 2, no. 2, pp. 68–79, Dec. 2024, doi: 10.52620/sainsdata.v2i2.136.

[12] D. Pramana, M. Afdal, M. Mustakim, and I. Permana, "Analisis Sentimen Terhadap Pemindahan Ibu Kota Negara Menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neightbors," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, p. 1306, Jul. 2023, doi: 10.30865/mib.v7i3.6523.

[13] M. F. Y. Herjanto and C. Carudin, "Analisis Sentimen Ulasan Pengguna Aplikasi Sirekap Pada Play Store Menggunakan Algoritma Random Forest Classifer," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4192.

[14] M. I. Abidin and E. W. Pamungkas, "Analisis Sentimen Terhadap Timnas Indonesia Di Piala Asia 2023 Dengan Model Transformer Berbahasa Indonesia," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, pp. 482–496, Jul. 2025, doi: 10.36341/rabit.v10i2.6142.

[15] Eko Arip Winanto, S. M. Z. Ali Difyah, Pareza Alam Jusia, and Sharipuddin, "Analisis Sentimen Terhadap Tagar Kabur Aja Dulu di Twitter Menggunakan Metode Lexicon-Based," *Jurnal PROCESSOR*, vol. 20, no. 2, Oct. 2025, doi: 10.33998/processor.2025.20.2.2542.