# Perbandingan Labelling Menggunakan USR Terhadap Labelling Manual Untuk Analisis Sentimen Media Sosial Pada Kemacetan Lalu Lintas Di Kota Palembang

1st Elifaza Alvarez Patra
Jurusan Sistem Komputer
*Fakultas Ilmu Komputer UNSRI*
Palembang, Indonesia
09011382025162@student.unsri.ac.id

2nd Yoppy Sazaki
*Jurusan Sistem Komputer*
*Fakultas Ilmu Komputer UNSRI*
Palembang, Indonesia
yoppysazaki@unsri.ac.id

3rd Ahmad Fali Oklilas
*Jurusan Sistem Komputer*
*Fakultas Ilmu Komputer UNSRI*
Palembang, Indonesia
fali@ilkom.unsri.ac.id

***Abstract*—This study aims to analyze public sentiment toward traffic conditions in Palembang City using comments from Instagram. The data were collected through web scraping from traffic information accounts, then cleaned and labelled using two approaches: keyword-based automatic labelling and manual labelling. The results show that most comments express negative sentiments, reflecting public dissatisfaction with traffic conditions. The consistency between automatic and manual labelling reached 62.02%, indicating that data balancing labelling is fairly effective but still needs contextual adjustments. In conclusion, social media–based sentiment analysis can serve as an alternative source of information for understanding public perceptions of urban traffic conditions.**

**Keywords—sentiment analysis, traffic, data labelling, social media**

## I. PENDAHULUAN

Analisis sentimen adalah prosedur untuk mengkategorikan pandangan yang diungkapkan atas suatu objek tertentu. Dengan munculnya berbagai perangkat teknologi, menjadi penting untuk mengetahui pandangan massa dalam bisnis, produk, atau dalam hal-hal yang disukai dan tidak disukai. Melacak emosi di balik unggahan di media sosial dapat membantu menghubungkan konteks di mana pengguna akan bereaksi dan berkembang [1].

Media sosial merupakan salah satu platform utama yang memfasilitasi interaksi dan ekspresi publik [2]. Salah satu media sosial yang digunakan untuk penelitian ini adalah *Instagram*. *Instagram* dipilih sebagai platform untuk analisis sentimen terutama karena platform ini mengutamakan arsip individual berisi kiriman dan gambar yang dapat mengartikulasikan percakapan, salah satunya yaitu percakapan mengenai kemacetan lalu lintas di Kota Palembang [3].

Kemacetan lalu lintas menjadi salah satu masalah yang terus berkembang di berbagai kota besar di seluruh dunia termasuk salah satunya ialah di Kota Palembang. Kemacetan tersebut tidak hanya menyebabkan gangguan mobilitas harian, tetapi juga memiliki dampak negatif terhadap lingkungan, kesehatan masyarakat, dan perekonomian. Peningkatan jumlah kendaraan bermotor, pembangunan infrastruktur yang tidak sebanding, serta kurangnya pengelolaan lalu lintas yang efektif menjadi penyebab utama kemacetan lalu lintas.

Berdasarkan laporan *TomTom Traffic Index* 2024, Kota Palembang termasuk salah satu kota termacet di Asia Tenggara dengan rata-rata waktu tempuh mencapai 27 menit 55 detik per 10 kilometer [4]. Kondisi ini menggambarkan tingginya kepadatan lalu lintas, terutama pada jam sibuk di titik strategis kota. Meski demikian, pihak kepolisian setempat menegaskan bahwa kondisi lalu lintas telah mengalami perbaikan berkat rekayasa lalu lintas dan pembangunan infrastruktur, sehingga data *TomTom* tidak sepenuhnya mencerminkan situasi aktual [5]. Sebagai informasi, *TomTom Traffic Index* sendiri merupakan indikator

global yang membandingkan waktu tempuh ideal dalam kondisi lancar dengan waktu tempuh aktual pada jam padat, menggunakan data GPS *(floating car data)* dari miliaran perjalanan kendaraan setiap tahun [6].

Oleh karena itu, hasil dari penelitian ini diharapkan mampu memberikan gambaran yang lebih menyeluruh mengenai faktor-faktor penyebab kemacetan, respons masyarakat, serta tingkat kepercayaan terhadap informasi yang berkaitan dengan kondisi lalu lintas, khususnya di jalanjalan protokol di Kota Palembang. Informasi ini diperoleh melalui analisis data dari media sosial. Temuan ini ditujukan untuk menjadi masukan yang bermanfaat bagi pihak-pihak terkait, seperti pemerintah daerah, instansi pengelola transportasi, serta masyarakat umum, terutama warga Kota Palembang.

## II. LANDASAN TEORI

### *A. Analisis Sentimen*

Analisis Sentimen sering dikenal sebagai *opinion mining* yang digunakan untuk menganalisis atau mengklasifikasikan pengguna dari kata, kalimat atau dokumen [7]. Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks. Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau pendapat. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif [8].

### *B. Media Sosial*

Media sosial merupakan platform daring yang memungkinkan pengguna internet di seluruh wilayah geografis untuk terhubung, berkomunikasi, dan berbagi konten multimedia secara publik termasuk perspektif, opini, pengalaman, foto, dan video. Selain pesan langsung, media sosial mendistribusikan konten buatan pengguna ke khalayak yang lebih luas melalui kiriman, berbagi, tagar, dan algoritma yang bertujuan untuk menyebarkan materi populer secara viral [9].

### *C. Instagram*

*Instagram* adalah platform media sosial yang memungkinkan pengguna untuk berbagi foto, video, dan cerita dalam format visual. Diluncurkan pada tahun 2010, *Instagram* awalnya fokus pada berbagi foto dengan filter artistik untuk meningkatkan penampilan gambar. Namun, seiring berjalannya waktu, platform ini berkembang menjadi salah satu media sosial terbesar di dunia dengan lebih dari satu miliar pengguna aktif bulanan pada tahun 2023 [10].

*Instagram* menjadi salah satu media sosial yang sangat populer bagi pengguna internet terutama di kalangan remaja dan orang dewasa. Instagram menawarkan beragam fitur menarik seperti unggahan foto dan video, fitur cerita *(stories)*, penggunaan *hashtag* untuk mencari dan mengikuti topik-topik tertentu [11].

### *D. Kemacetan Lalu Lintas*

Kemacetan merujuk pada keadaan di mana suatu hal dalam konteks ini jalan raya tidak mampu beroperasi sebagaimana mestinya, sehingga fungsi utamanya sebagai jalur transportasi terganggu. Kemacetan terjadi pada daerah dengan aktivitas atau kegiatan masyarakat, penggunaan lahan, serta tingkat kepadatan penduduknya yang tinggi. Kemacetan pada lalu lintas terjadi oleh adanya ketidakseimbangan antara populasi penduduk, dan jumlah kendaraan yang meningkat tiap tahunnya dengan kapasitas suatu jalan yang dapat dilintasi. Kegiatan berlalu lintas tersebut seharusnya diimbangi dengan adanya penambahan kapasitas jalan seperti pelebaran jalan atau penggunaan jalan alternatif sehingga aktivitas lalu lintas tidak terpaku pada suatu ruas jalan saja dan dapat dialihkan dengan penggunaan jalan lainnya [12].

### *E. Kota Palembang*

Kota Palembang adalah salah satu Kota di Provinsi Sumatra Selatan yang merupakan Ibukota Provinsi dari Provinsi Sumatera Selatan. Kota Palembang memiliki luas wilayah ± 400,61 km2 yang secara administrasi terbagi atas 18 kecamatan dan 107 kelurahan. Kota Palembang merupakan ibukota Provinsi Sumatera Selatan dengan batas wilayah yaitu di sebelah utara, timur dan barat dengan Kabupaten Banyuasin, sedangkan sebelah selatan berbatasan dengan Kabupaten Muara Enim. Dengan keadaan geografis dan jumlah penduduk yang semakin lama semakin meningkat untuk mewujudkan pelayanan masyarakat yang cepat dan efektif, transparan dan akuntabel diperlukan sistem pelayanan yang berbasis elektronik di Kota Palembang [13].

### *F. Crawling Data*

*Crawling data* adalah teknik pengambilan data secara otomatis dari situs web menggunakan program komputer. Metode ini memungkinkan pengambilan data dari berbagai sumber secara efisien dan terstruktur, sehingga dapat digunakan untuk berbagai keperluan seperti penelitian, analisis, dan pengembangan aplikasi [14]. Dalam penelitian ini, metode *crawling* digunakan untuk memperoleh data opini masyarakat mengenai kemacetan lalu lintas di Kota Palembang melalui media sosial, khususnya *Instagram*.

Salah satu pendekatan yang digunakan dalam *crawling data* adalah *web scraping*, yaitu metode untuk memperoleh data dalam bentuk dokumen dari halaman situs web tertentu. Proses kerja *web scraper* dimulai dengan memasukkan satu atau lebih URL yang akan di *scrape*. Setelah itu, *scraper* akan mengekstrak data dari halaman web yang dipilih sesuai dengan URL yang telah ditentukan [15].

### *G. Pre-processing Data*

*Pre-processing data* adalah proses persiapan data sebelum digunakan pada tahap pembagian data. Proses ini terdiri dari proses penghapusan atau mentransformasi data yang kurang relevan agar sistem dapat memproses data tersebut dengan mudah. Tahapan ini merupakan tahapan krusial, terutama khusus karena data di media sosial memiliki banyak item teks unik sehingga perlu dihilangkan data teks umum dan informal

untuk mengurangi ruang fitur [16]. Berikut terdapat beberapa tahapan-tahapan *preprocessing data* :

*1) Cleaning*

Tahapan cleaning merupakan suatu proses pembersihan dokumen dari kata atau karakter yang dihilangkan pada komentar adalah karakter simbol, angka, link, url, hastag (#), *mention (@username)*, tanda koma, titik dan tanda baca lainnya [17].

*2) Case Folding*

Tahapan *case folding* mengkonversikan keseluruhan teks dari huruf capital menjadi huruf kecil. Hanya huruf "a" sampai "z" yang akan diterima dalam proses ini selain huruf tersebut semuanya dianggap tidak diterima [17].

*3) Stemming*

Tahapan stemming merupakan proses membentuk suatu kata menjadi kata dasar. Proses ini menghilangkan semua imbuhan [17].

*4) Tokenization*

Tahapan *tokenization/tokenizing* merupakan pemilahan kalimat menjadi perkata. Suatu proses pemotongan seluruh karakter dan sebuah set dokumen yang diberikan menjadi satu potongan kata. Potongan tersebut dikenal dengan istilah token [17].

*5) Stopword Removal*

Tahapan *stopword removal* merupakan kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Seperti kata konjungsi "yang", "dan", "dari", dll. Atau sebagai penghapus kata penghubung dan imbuhan [17].

*6) Normalization*

Tahapan Normalisasi adalah teknik yang sering diterapkan sebagai bagian dari penyiapan data untuk pembelajaran mesin. Normalisasi juga diperlukan untuk beberapa algoritma untuk pembuatan model [17]. *H. Penilaian Kecenderungan Media Sosial*

Saat ini terdapat dua pendekatan utama dalam menilai kecenderungan pada analisis sentimen media sosial, yaitu *lexicon-based classificaiton* dan klasifikasi berbasis *machine learning*. Pendekatan *lexicon-based* menggunakan kamus untuk menentukan polaritas teks untuk mengklasifikasikannya dengan tepat. Pendekatan *machine learning* berbeda dari pendekatan *lexicon-based*, karena mereka mengubah data teks menjadi data numerik setelah beberapa operasi *pre-processing* [18].

Dalam penelitian ini, pendekatan yang digunakan penulis adalah pendekatan berbasis *machine learning*. Pendekatan ini dipilih karena kemampuannya dalam menangani data dalam jumlah besar.

*I. Sentimen Positif, Negatif, dan Netral*

Analisis sentimen dapat diklasifikasikan kedalam kelas sentimen bersifat positif, negatif dan netral [19]:

*1) Sentimen Positif*

Menurut Kamus Besar Bahasa Indonesia (KBBI) sentimen positif merupakan reaksi atau sikap yang meningkatkan nilai seseorang atau sesuatu. Sentimen positif menunjukkan kondisi lancar, yaitu arus kendaraan berjalan normal tanpa hambatan berarti.

*2) Sentimen Negatif*

Menurut Kamus Besar Bahasa Indonesia (KBBI) sentimen negatif merupakan reaksi atau sikap yang menurunkan nilai seseorang atau sesuatu, jadi kalimat bersentimen negatif akan menyebabkan penyurutan nilai pandang terhadap sesuatu, sehingga membentuk *tren down*. Umumnya kalimat bersentimen negatif ditandai dengan penggunaan kata negasi. Negasi merupakan sesuatu yang dikenal dalam semua bahasa dan biasanya negasi digunakan untuk mengubah polaritas dari suatu pernyataan. Sentimen negatif menunjukkan kondisi macet, yaitu situasi lalu lintas yang padat hingga menyebabkan kendaraan terhambat atau berhenti.

*3) Sentimen Netral*

Kata netral sendiri dalam Kamus Besar Bahasa Indonesia (KBBI) berarti tidak berpihak, kalimat bersentimen netral merupakan ekspresi kalimat yang tidak bersifat positif maupun negatif. Sentimen netral menggambarkan kondisi ramai lancar, yaitu situasi lalu lintas yang mulai padat namun kendaraan masih dapat bergerak dengan baik.

Detail dari analisis sentimen dalam mengklasifikasikan sentimen positif, negatif, dan netral dapat dilihat pada Tabel I [20].

*Tabel I. Kategori Sentimen*

| No | Sentimen | Keterangan |
|---|---|---|
| 1. | Positif | Sentimen positif merupakan respon ataupun opini sesorang yang menandakan dirinya setuju akan suatu keadaan. |
| 2. | Negatif | Sentimen negatif merupakan respon ataupun opini sesorang yang menandakan dirinya bertentangan, marah, tidak setuju dan sebagainya, biasanya sentimen negatif ditandai dengan negasi. |
| 3. | Netral | Sentimen netral merupakan respon atau opini yang tidak bersifat negatif ataupun positif. |

*J. TomTom Traffic Index*

*TomTom Traffic Index* adalah indikator yang dikembangkan oleh *TomTom* untuk mengukur dan membandingkan tingkat kemacetan di kota-kota dunia. Indeks ini dibentuk dengan membandingkan waktu tempuh ideal dalam kondisi lalu lintas lancar dengan waktu tempuh

aktual pada jam puncak, berdasarkan data *floating car* (GPS/perangkat navigasi) [6]. Palembang menduduki posisi ketiga dengan rerata waktu tempuh 27 menit 55 detik per 10 km dan tingkat kemacetan 41%. Kota ini berada di peringkat ke-53 dunia dan peningkatan waktu perjalanan terjadi sebesar 30 detik pada tahun 2024 [21].

Berikut adalah tampilan grafik kemacetan lalu lintas di Kota Palembang yang disediakan oleh *TomTom Traffic Index* pada tanggal 27 Juli 2025 pukul 08.00 WIB, 12.00 WIB, 16.00 WIB, dan 21.00 WIB, terlihat pada gambar 1 sampai gambar 4.



*Gambar 1. Grafik Kemacetan Lalu Lintas Kota Palembang Pagi Hari*



*Gambar 2. Grafik Kemacetan Lalu Lintas Kota Palembang Siang Hari*



*Gambar 3. Grafik Kemacetan Lalu Lintas Kota Palembang Sore Hari*



*Gambar 4. Grafik Kemacetan Lalu Lintas Kota Palembang Malam Hari*

## III. METODOLOGI PENELITIAN

Penelitian ini menggunakan data dari media sosial Instagram. Tujuan utama dari penelitian ini adalah untuk membangun dataset yang siap untuk diolah model klasifikasi kondisi kemacetan lalu lintas berdasarkan data dari media sosial.

Topik penelitian ini berfokus pada analisis sentimen terhadap kondisi kemacetan lalu lintas di Kota Palembang dengan memanfaatkan data dari media sosial, khususnya *Instagram*. media sosial menawarkan informasi aktual yang tersebar luas dan diperbarui secara spontan oleh pengguna.

Sebelum melalui tahapan utama, penulis melakukan pengkajian terhadap berbagai literatur yang berkaitan dengan topik penelitian, baik dari jurnal ilmiah, publikasi akademik, maupun sumber daring yang kredibel. Tujuan dari studi literatur ini adalah untuk membentuk landasan teori yang kokoh, memastikan bahwa metode yang digunakan telah teruji dalam penelitian terdahulu, serta menemukan ruang kosong atau kekurangan dalam penelitian sebelumnya yang dapat diisi melalui studi ini.

### *A. Pengumpulan Data*

Penelitian ini memanfaatkan data teks dari media sosial dan. Data media sosial diperoleh melalui proses *scraping*, yang bertujuan mengumpulkan informasi mengenai situasi lalu lintas dari unggahan pengguna yang berkaitan dengan jalan-jalan protokol di Kota Palembang.

Pengumpulan data dari media sosial dilakukan melalui teknik *scraping* dengan memanfaatkan platform *Apify* yang dikombinasikan dengan ekstensi *IG Comment Exporter*. Data yang dikumpulkan berupa komentar pengguna dari unggahan di *Instagram* yang berkaitan dengan kemacetan lalu lintas di Kota Palembang. Komentar-komentar ini dipilih berdasarkan relevansi dengan topik penelitian, seperti kondisi jalan, kepadatan kendaraan, dan keluhan pengguna terhadap lalu lintas di wilayah tersebut.

### *B. Pre-processing Data*

Pra-pemrosesan data merupakan langkah penting dalam analisis data guna memastikan bahwa data yang digunakan dalam pembangunan model telah berada dalam kondisi yang bersih dan layak olah. Dalam penelitian ini, tahapan prapemrosesan diterapkan pada data komentar yang diperoleh dari Instagram. Proses ini mencakup beberapa tahapan utama, yaitu *cleaning*, *case folding*, *tokenization*, *normalization*, *stopwords removal*, dan *stemming*.

### *C. Labelling Data*

Pelabelan data dilakukan setelah proses pembersihan komentar, sehingga hanya tersisa komentar yang relevan dengan kondisi lalu lintas di Kota Palembang. Dua metode digunakan, yaitu pelabelan manual dan pelabelan otomatis menggunakan teknik *balancing data*.

Pelabelan manual dilakukan dengan membaca komentar satu per satu untuk memahami konteks dan emosi yang terkandung. Cara ini dianggap lebih akurat karena

memperhatikan makna keseluruhan kalimat, dan hasilnya dijadikan acuan utama *(ground truth)* dalam penelitian.

Pelabelan otomatis dilakukan dengan memberikan label sentimen awal pada setiap komentar. Selanjutnya, pada tahap pengolahan data diterapkan pendekatan USR *(Under Sampling, SMOTE,* dan *Resampling)* sebagai bagian dari strategi penyesuaian jumlah data antar kategori sentimen. USR digunakan untuk menjaga agar setiap kelas sentimen memiliki representasi yang memadai dengan cara mengurangi data pada kelas yang terlalu banyak *(under sampling)* serta menambah data pada kelas yang jumlahnya lebih sedikit melalui teknik SMOTE dan metode *resampling* lainnya.

*D. Analisa*

Pada tahap ini, dilakukan analisis terhadap hasil akhir berupa persentase tingkat kemiripan antara data dari media sosial yang telah diberi label berdasarkan kata kunci dan label manual.

*E. Kesimpulan*

Tahap ini mencakup penarikan kesimpulan berdasarkan hasil pengolahan data dan analisis yang telah dilakukan sebelumnya.

## IV. HASIL DAN PEMBAHASAN

Setelah perancangan selesai dikerjakan, langkah selanjutnya adalah mengumpulkan data teks dari media sosial *Instagram* melalui teknik scraping menggunakan ekstensi dari *Google Chrome* bernama *IG Comment Exporter*.
Pengumpulan data melalui platform *Instagram* terkait kondisi lalu lintas di Kota Palembang dalam rentang Juni 2024 sampai Juni 2025.

*A. Crawling Data*

Pada tahap ini, dilakukan proses pengumpulan data berupa postingan dan komentar dari *Instagram*. Untuk data postingan berupa link diporses menggunakan *Apify*, sedangkan data komentar berupa teks diambil menggunakan ekstensi.

*1) Data Postingan*

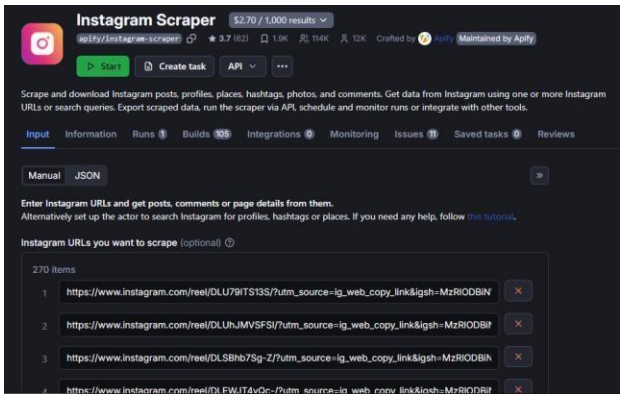
Pengumpulan URL postingan terkait kondisi lalu lintas ini akan digunakan untuk proses *scraping data*. URL postingan *Instagram* ditemukan sebanyak 270 URL terlihat pada gambar 5.

| | URL |
|---|---|
| 0 | https://www.instagram.com/reel/DLU79ITS13S/?ut... |
| 1 | https://www.instagram.com/reel/DLUhJMVSFSI/?ut... |
| 2 | https://www.instagram.com/reel/DLSBhb7Sg-Z/?ut... |
| 3 | https://www.instagram.com/reel/DLEWJT4yQc-/?ut... |
| 4 | https://www.instagram.com/reel/DLAPCezSWXT/?ut... |
| ... | ... |
| 265 | https://www.instagram.com/reel/C_uxjXxSFdJ/?ut... |
| 266 | https://www.instagram.com/p/C_VAOXxy-Fy/?utm_s... |
| 267 | https://www.instagram.com/reel/C-zqNlnSbz0/?ut... |
| 268 | https://www.instagram.com/reel/C9tufPgy7ch/?ut... |
| 269 | https://www.instagram.com/reel/C9hL33gSSJM/?ut... |

270 rows × 1 columns

*Gambar 5. URL Instagram*

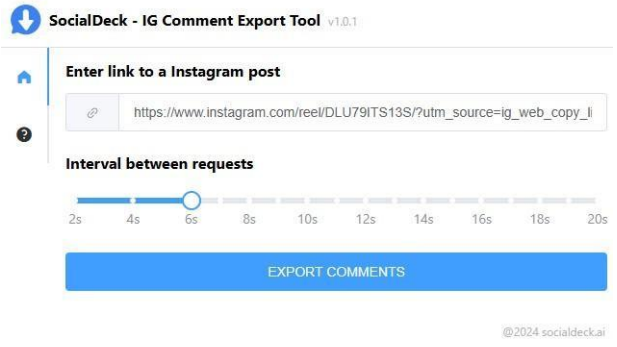
Setelah URL dari *Instagram* terkumpul, *scraping* dilakukan menggunakan tool *Instagram Scraper* di *Apify*. URL dimasukkan ke tool tersebut dan *scraper* akan otomatis menarik data dari setiap postingan pengguna yang berkaitan dengan kondisi lalu lintas terlihat pada gambar 6.




*Gambar 6. Pengambilan Data Instagram pada Apify*

*2) Data Komentar*

Untuk scraping dengan ekstensi, URL yang sudah dikumpulkan disalin satu per satu ke dalam ekstensi *IG Comment Export Tool* dari *SocialDeck*. Ekstensi ini akan melakukan *scraping* komentar dari postingan instagram secara otomatis dan hasilnya dapat diunduh dalam bentuk file csv terlihat pada gambar 7.




*Gambar 7. Pengambilan Data Komentar Instagram dengan*

*Ekstensi*

Data *Instagram* yang dikumpulkan dari *Apify* dan ekstensi didapatkan dari akun @oypalembang, @palembanginfo, @infopalembangterbaru, @palembang.update dan @palembangkulukilir. Dataset dapat dilihat seperti pada gambar 8.



*Gambar 8. Dataset Komentar Instagram*

Dataset yang didapatkan berjumlah 18963 baris dengan 8 kolom yaitu *User ID, User Name, Profile Url, Profile Pic Url, Comment ID, Comment Text, Account Name,* dan *Comment Time*.

*B. Pre-processing Data*

Data yang diperoleh melalui *web scraper Apify* dan ekstensi *IG Comment Export Tool* merupakan data mentah yang masih mengandung banyak *noise*. Oleh karena itu, perlu dilakukan tahapan *preprocessing* teks agar data lebih bersih dan dapat diproses. Tahapan *preprocessing* dilakukan melalui enam langkah, yaitu *cleaning, case folding, stemming, tokenization*, *stopword removal,* dan *normalization*.

*1) Cleaning*

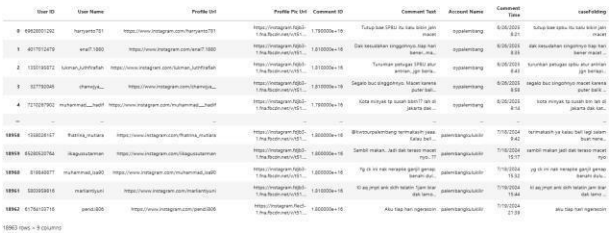
*Cleaning* merupakan tahap pertama dalam proses *preprocessing*, yang bertujuan untuk menghilangkan karakter atau simbol yang tidak diperlukan dalam data teks, seperti *hashtag*, *mention*, tautan (URL), simbolsimbol, angka, tanda baca, dan elemen lainnya yang dianggap tidak relevan terlihat pada gambar 9.



*Gambar 9. Cleaning Data Teks*

*2) Case Folding*

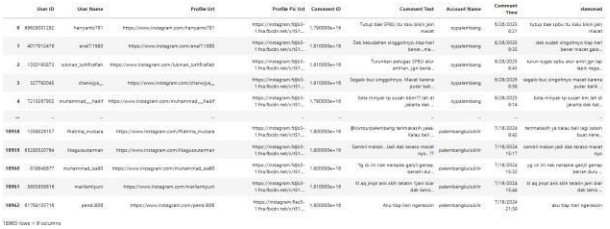
*Case folding* adalah proses mengubah seluruh karakter dalam teks menjadi huruf kecil *(lowercase)*. Tujuan dari langkah ini adalah untuk menyamakan bentuk kata, sehingga perbedaan antara huruf kapital dan huruf kecil tidak memengaruhi analisis, khususnya dalam perhitungan frekuensi kata terlihat pada gambar 10.



*Gambar 10. Case Folding Data Teks*

*3) Stemming*

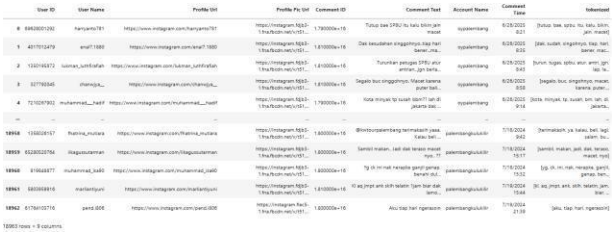
*Stemming* merupakan proses mengubah kata menjadi bentuk dasarnya dengan cara menghapus imbuhan atau akhiran, tanpa mempertimbangkan struktur tata bahasa. *Stemming* bertujuan untuk menyatukan berbagai turunan dari suatu kata agar dapat diproses secara lebih efisien. Contohnya, kata-kata seperti "bertemu", "pertemuan", dan "menemui" akan dikembalikan ke bentuk dasar "temu". Tahapan *stemming* dilakukan dengan memanfaatkan library Sastrawi terlihat pada gambar 11.



*Gambar 11. Stemming Data Teks*

*4) Tokenization*

Pada tahap ini, teks dibagi menjadi bagian-bagian kecil yang disebut token, yang bisa berupa kata, frasa, atau simbol. *Tokenization* sangat penting dalam analisis sentimen karena memungkinkan setiap kata dalam sebuah kalimat atau teks dipisahkan, sehingga model dapat mengevaluasi frekuensi kemunculan serta pola dari masing-masing token. Proses tokenisasi dilakukan dengan memanfaatkan fungsi *word_tokenize* yang diambil dari pustaka NLTK terlihat pada gambar 12.

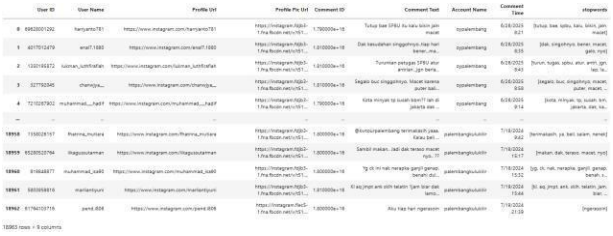


*Gambar 12. Tokenisasi Data Teks*

*5) Stopword Removal*

*Stopword Removal* adalah tahap di mana kata-kata umum yang sering muncul dalam teks, seperti "dan", "itu", atau "adalah", dihapus karena umumnya tidak memiliki kontribusi signifikan terhadap makna. Penghapusan *stopword* bertujuan untuk menyederhanakan data yang dianalisis, sehingga proses analisis sentimen menjadi lebih efisien tanpa mengorbankan informasi yang relevan. Seperti halnya proses

tokenisasi, tahap *stopword removal* juga memanfaatkan pustaka NLTK dengan mengimpor fungsi *stopword* terlihat pada gambar 13.



*Gambar 13. Stopword Data Teks*

*6) Normalization*

Tahap normalisasi bertujuan untuk menyelaraskan teks dengan mengubah berbagai variasi kata menjadi bentuk yang seragam. Proses ini mencakup penulisan ulang singkatan, koreksi ejaan, serta penyamaan format angka dan simbol, sehingga model dapat lebih mudah memproses dan menganalisis informasi yang terkandung dalam data. Proses normalisasi ini memanfaatkan library *Colloquial Indonesian Lexicon* untuk melakukan standarisasi kata-kata tidak baku dalam Bahasa Indonesia terlihat pada gambar 14.

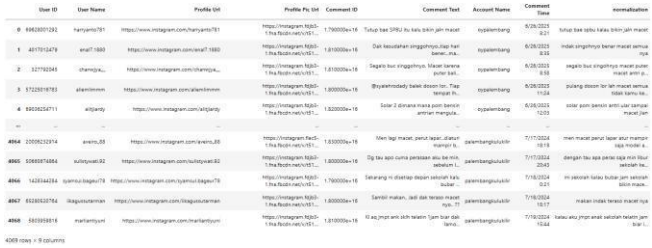


*Gambar 14. Normalisasi Data Teks*

Setelah melalui tahap normalisasi, ditemukan sebanyak 928 entri kosong yang dihasilkan dari tahapan preprocessing. Setelah data NaN tersebut dihapus, jumlah data yang tersisa menjadi 18.035 baris.

*C. Filtering Data*

Pada tahap ini, data yang sudah melewati proses preprocessing disaring kembali untuk memastikan hanya data yang berkaitan dengan kondisi lalu lintas yang digunakan. Hasil dari proses filtering ditampilkan pada Gambar 15.



*Gambar 15. Filtering Data Teks*

Melalui proses *filtering*, diperoleh sebanyak 4.069 baris data yang mengandung kata kunci tertentu yang berasal dari total 18.963 data mentah. Kemudian dari data *filtering*, komentar yang memiliki relevansi terhadap kondisi kepadatan lalu lintas terdapat sebanyak 3594 data yang dapat digunakan sebagai *dataset*. Tujuan dari tahapan ini adalah untuk meningkatkan akurasi model dalam mengenali pola-pola yang berhubungan dengan kondisi lalu lintas. *Filtering* dilakukan dengan menggunakan kata kunci seperti "macet", "ramai", "padat", dan "lancar", serta memastikan bahwa konteksnya benar-benar menggambarkan situasi lalu lintas.

*D. Labelling Data*

Setelah proses *filtering*, dilakukan analisis terhadap data untuk menilai sentimen. Analisis sentimen ini dilakukan melalui proses filter manual untuk menilai sentimen, yaitu dengan membaca secara langsung setiap komentar dan menentukan kecenderungan emosional yang terkandung di dalamnya. Komentar yang bernada keluhan atau ketidakpuasan dimasukkan ke dalam kategori negatif dan diberi label nomor 3, komentar yang berisi dukungan atau penilaian baik dikategorikan sebagai positif dan diberi label nomor 1, sedangkan komentar yang bersifat netral, informatif, atau tidak menunjukkan emosi tertentu dimasukkan ke kategori netral dan diberi label nomor 2 terlihat pada tabel II.

*Tabel II. Penilaian Sentimen Pada Media Sosial Berdasarkan Labelling Manual*

| **Sentimen** | **Negatif** | **Positif** | **Netral** |
|---|---|---|---|
| **Jumlah Sentimen** | 3365 | 138 | 91 |
| **Persentase** | 93,60% | 3,83% | 2,53% |

Hasil *labelling* manual ditampilkan pada Gambar 16.



*Gambar 16. Labelling Manual pada Data Media Sosial*

Setelah dilakukan pelabelan manual terhadap 3.594 komentar hasil penyaringan, terlihat distribusi sentimen yang sangat timpang. Komentar berlabel negatif mencapai 3.365 contoh, menyumbang sekitar 93,63% dari seluruh dataset. Sebaliknya, komentar positif hanya sebanyak 138 data (3,84%) dan komentar netral berjumlah 91 data (2,53%). Kondisi ini memperlihatkan adanya ketidakseimbangan kelas yang cukup ekstrem.

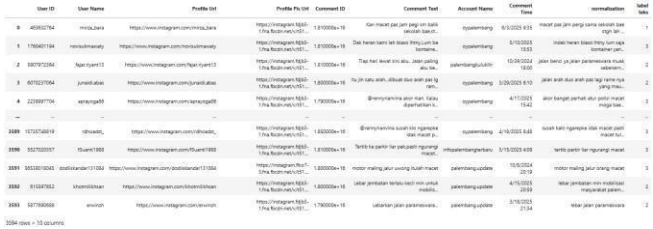
Ketidakseimbangan tersebut berisiko memengaruhi proses pembelajaran mesin selanjutnya. Model cenderung "belajar" dari pola kelas mayoritas sehingga performanya terhadap kelas minoritas (positif dan netral) menjadi buruk, misalnya tingkat *recall* untuk kelas minoritas rendah, atau model sering salah mengklasifikasikan komentar positif/netral sebagai negatif. Oleh karena itu, diperlukan penerapan teknik penyeimbangan data *(balancing)* seperti *oversampling*, *undersampling*, atau *Synthetic Minority Oversampling Technique (SMOTE)* dengan rasio pembagian

data 65% untuk sentimen negatif, 20% untuk sentimen positif, dan 15% untuk sentimen netral terlihat pada tabel III.

*Tabel III. Penilaian Sentimen Pada Media Sosial Berdasarkan Labelling Otomatis*

| **Sentimen** | **Negatif** | **Positif** | **Netral** |
|---|---|---|---|
| **Jumlah Sentimen** | 2337 | 718 | 539 |
| **Persentase** | 65,03% | 19,98% | 15% |

Hasil *balancing data* dari *labelling manual* ditampilkan pada Gambar 17.



*Gambar 17. Balancing Data Media Sosial*

Setelah proses *balancing data* dilakukan, diperoleh distribusi sentimen yang lebih proporsional dibandingkan kondisi awal. Jumlah data pada kelas negatif menjadi 2.337 atau sekitar 65,03% dari total *dataset*. Sementara itu, kelas positif meningkat menjadi 718 data (19,98%) dan kelas netral menjadi 539 data (15%).

Setelah dilakukan proses pelabelan manual dan balancing data, langkah selanjutnya adalah melakukan pencocokan antara kedua label manual dengan hasil pelabelan otomatis menggunakan teknik balancing data. Proses pencocokan dilakukan dengan memberikan nilai 1 apabila kedua label memiliki kategori yang sama, serta nilai 0 apabila kedua label berbeda terlihat pada tabel IV.

*Tabel IV. Kecocokan Antar Penilaian Manual dan Otomatis*

| | **Label yang Cocok** | **Label yang Berbeda** |
|---|---|---|
| **Total** | 2229 | 1365 |
| **Akurasi** | 62,02% | 37,98% |

Untuk hasil kecocokan antara penilaian sentimen berdasarkan filter manual dan hasil *balancing data* ditampilkan pada Gambar 18.



*Gambar 18. Kecocokan antar Kedua Penilaian Sentimen*

Setelah dilakukan proses pencocokan antara label hasil pelabelan manual dengan label yang diperoleh dari teknik *balancing data* dari pelabelan manual, ditemukan bahwa sebanyak 2.229 data menunjukkan hasil yang konsisten atau sama antara kedua metode tersebut. Jumlah tersebut merepresentasikan tingkat kecocokan sebesar 62,02% dari total data yang dibandingkan. Sementara itu, terdapat 1.365 data atau sekitar 37,98% yang memiliki perbedaan hasil label antara proses manual dan otomatis.

## V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, proses pengumpulan data menghasilkan total 18.963 komentar Instagram. Setelah melalui tahap preprocessing (cleaning, case folding, stemming, tokenizing, stopword removal, dan normalization), dilakukan proses filtering menggunakan kata kunci terkait kemacetan lalu lintas sehingga diperoleh sebanyak 4.069 data. Dari jumlah tersebut, terdapat 3.594 data relevan yang digunakan sebagai dataset utama penelitian.

Berdasarkan analisis sentimen, hasil labelling manual menunjukkan distribusi komentar dengan sentimen negatif sebesar 93,60%, positif sebesar 3,83%, dan netral sebesar 2,53%. Sementara itu, hasil *labelling* otomatis menggunakan teknik balancing data menghasilkan distribusi komentar dengan sentimen negatif sebesar 65,03%, positif sebesar 19,98%, dan netral sebesar 15%.

Perbandingan antara *hasil labelling* otomatis dan *labelling manual* menunjukkan tingkat kecocokan sebesar 62,02%.

## REFERENCES


[1] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A Survey of Sentiment Analysis from Social Media Data," IEEE Trans. Comput. Soc. Syst., vol. 7, no. 2, pp. 450–464, 2020, doi: 10.1109/TCSS.2019.2956957.

[2] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based Dan Random Forest," J. Ilm. Inform., vol. 11, no. 02, pp. 159–169, 2023, doi: 10.33884/jif.v11i02.7987.

[3] G. S. Cornelio and A. Roig, "Mixed methods on Instagram research: Methodological challenges in data analysis and visualization," Convergence, vol. 26, no. 5–6, pp. 1125–1143, 2020, doi: 10.1177/1354856520941613.

[4] "The Worst Cities in Southeast Asia for Traffic in 2024." [Online]. Available: https://seasia.co/2025/01/31/the-worst-cities-in-southeastasia-for-traffic-in-2024

[5] "Kasatlantas Palembang Bantah Klaim TomTom Traffic Index: 'Kondisi Lalu Lintas Sudah Jauh Lebih Baik.'" [Online]. Available: https://tribratanews.sumsel.polri.go.id/main/detail/6059/

[6] "About | TomTom Traffic Index." [Online]. Available: https://www.tomtom.com/traffic-index/about/

[7] D. T. Hermanto, A. Setyanto, and E. T. Luthfi, "Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online," Creat. Inf. Technol. J., vol. 8, no. 1, p. 64, 2021, doi: 10.24076/citec.2021v8i1.264.

[8] A. Z. Amrullah, A. Sofyan Anas, and M. A. J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," J. BITe, vol. 2, no. 1, pp. 40–44, 2020, doi: 10.30812/bite.v2i1.804.

[9] A. S. George and T. Baskar, "Leveraging Big Data and Sentiment Analysis for Actionable Insights : A Review of Data Mining

Approaches for Social Media Partners Universal International Innovation Journal ( PUIIJ ),” Partners Univers. Int. Innov. J., no. August, pp. 39–59, 2024, doi: 10.5281/zenodo.13623777.

[10] R. Zulfiqri, B. N. Sari, T. N. Padilah, U. S. Karawang, and T. Timur, “ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI MEDIA SOSIAL INSTAGRAM PADA SITUS GOOGLE PLAY STORE MENGGUNAKAN NAÏVE BAYES CLASSIFIER,” JITET (Jurnal Inform. dan Tek. Elektro Ter., vol. 12, no. 3, 2024.

[11] F. Baehaqi and N. Cahyono, “Analisis Sentimen Terhadap Cyberbullying Pada Komentar di Instagram Menggunakan Algoritma Naïve Bayes,” Indones. J. Comput. Sci., vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[12] R. O. Andini, L. N. Arief, and S. F. Hana, “PEMETAAN KEMACETAN LALU LINTAS DI UNIVERSITAS DIPONEGORO (STUDI KASUS: KECAMATAN TEMBALANG DAN KECAMATAN BANYUMANIK, KOTA SEMARANG),” Geod. Undip, vol. 2, no. Sistem Informasi Geografis, pp. 240–252, 2013.

[13] D. Antoni, M. I. Herdiansyah, M. Akbar, and A. Sumitro, “Pengembangan Infrastruktur Jaringan Untuk Meningkatkan Pelayanan Publik di Kota Palembang,” J. Media Inform. Budidarma, vol. 5, no. 4, p. 1652, 2021, doi: 10.30865/mib.v5i4.3318.

[14] I. Syahrohim, S. D. Saputra, and R. W. Saputra, “PERBANDINGAN ANALISIS SENTIMEN SETELAH PILPRES 2024 DI TWITTER MENGGUNAKAN ALGORITMA MACHINE LEARNING,” JITET (Jurnal Inform. dan Tek. Elektro Ter., vol. 12, no. 2, 2024.

[15] J. A. Wardhana and Y. Sibaroni, “Aspect Level Sentiment Analysis on Zoom Cloud Meetings App Review Using LDA,” J. RESTI (Rekayasa Sist. dan Teknol. Informasi), no. 10, pp. 631–638, 2021.

[16] C. L. T. Boro, A. Faisol, and D. Rudhistiar, “ANALISIS SENTIMEN TERHADAP KAMPANYE PENGURANGAN PLASTIK PADA MEDIA SOSIAL MENGGUNAKAN METODE SVM,” J. Inform. Teknol. dan Sains, vol. 6, no. 1, pp. 147–155, 2025.

[17] S. Nanda, D. Mualfah, and D. A. Fitri, “Analisis Sentimen Kepuasan Pengguna Terhadap Layanan Streaming Mola Menggunakan Algoritma Random Forest,” J. Apl. Teknol. Inf. dan Manaj., vol. 3, no. 2, pp. 210– 219, 2022, doi: 10.31102/jatim.v3i2.1592.

[18] O. Rakhmanov, “A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify StudentLecturer Comments,” Procedia Comput. Sci., vol. 178, pp. 194–204, 2020, doi: 10.1016/j.procs.2020.11.021.

[19] L. Ardiani, H. Sujaini, and T. Tursina, “Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak,” J. Sist. dan Teknol. Inf., vol. 8, no. 2, p. 183, 2020, doi: 10.26418/justin.v8i2.36776.

[20] F. Rizqi Irawan, “Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearset Neighbors,” JIKO (Jurnal Inform. dan Komputer), vol. 5, no. 1, pp. 62–68, 2022, doi: 10.33387/jiko.v5i1.4267.

[21] “Daftar Kota Termacet di Indonesia, Nomor 1 Bukan Jakarta.” [Online]. Available: https://data.goodstats.id/statistic/daftar-kota-termacet-diindonesia-nomor-1-bukan-jakarta-UsTgP#:~: