Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1846~1851
e-ISSN: 2548-6861

A Sentiment Analysis of Public Perception Toward Pets in Public Spaces Using Logistic Regression and Word Embedding

Dennita Noor Febianty[1]*, Majid Rahardi[2]*
* Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
dennita_nf@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-07-15
Revised 2025-07-29
Accepted 2025-07-30

Keyword:

Logistic Regression,
Pet in Public Space,
Sentiment Analysis,
Word Embeddings.

ABSTRACT

Addressing the complex social debate over pets in public areas, this study assesses public sentiment by analyzing a dataset of YouTube comments. We employed a machine learning pipeline beginning with data collection via the YouTube API, followed by rigorous text preprocessing and SMOTE-based class balancing for the training data. For classification, a Logistic Regression model was trained on contextual features generated by Word Embeddings (Word2Vec) and optimized through hyperparameter tuning. The final model proved highly effective, yielding a test accuracy of 92.74% with F1-scores of 0.84 for the negative class and 0.95 for the positive class. Ultimately, this research establishes an effective approach to measuring public opinion on social issues in Indonesia, providing actionable insights for public space administrators and policy makers.

I. PENDAHULUAN

Keberadaan hewan peliharaan di ruang publik merupakan isu sosial yang menimbulkan berbagai tanggapan dari masyarakat [1]. Sebagian pihak mendukung keberadaan hewan peliharaan karena dianggap memberikan manfaat emosional dan menjadi bagian dari keluarga. Namun, tidak sedikit juga yang menganggap kehadiran hewan peliharaan di ruang publik menimbulkan gangguan, seperti suara bising, kotoran, hingga potensi penyakit. Fenomena ini menarik untuk dikaji lebih dalam guna mengetahui bagaimana persepsi publik sesungguhnya terhadap isu tersebut.

YouTube, sebagai salah satu platform berbagi video terbesar di dunia, menyediakan banyak data dalam bentuk komentar pengguna yang dapat dijadikan sumber opini publik [2]. Komentar-komentar ini mencerminkan reaksi spontan dan jujur dari masyarakat terhadap berbagai topik, termasuk keberadaan hewan peliharaan di ruang publik.

Penelitian ini menggunakan metode *Logistic Regression* sebagai algoritma klasifikasi, dan Word2Vec untuk representasi fitur berbasis teks. *Logistic Regression* dipilih karena kemampuannya dalam klasifikasi biner yang sederhana namun efektif [3]. Word2Vec memungkinkan representasi kata dalam bentuk vektor numerik yang mempertahankan makna semantik [4]. Untuk memperbaiki distribusi kelas yang tidak seimbang, digunakan metode *Synthetic Minority Oversampling Technique* (SMOTE) agar model dapat belajar secara adil terhadap kedua kelas sentimen [5].

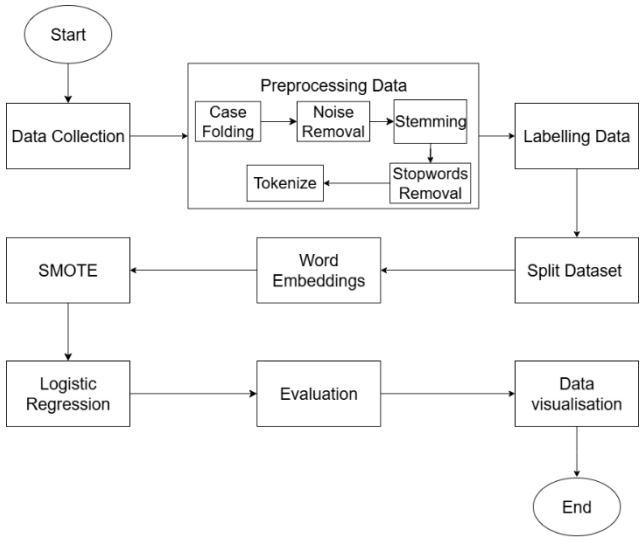
Melalui analisis sentimen terhadap komentar tersebut, penelitian ini dibuat untuk mengidentifikasi pola opini dan menyimpulkan sentimen dominan yang muncul. Penelitian ini memiliki tujuan untuk membangun model klasifikasi sentimen yang mampu mengenali opini publik terhadap keberadaan hewan peliharaan di ruang publik berdasarkan data komentar YouTube [6].

Penelitian ini menggunakan pendekatan Word2Vec dan metode berbasis leksikon (*lexicon-based*), mengenali kedua pendekatan tersebut telah umum digunakan dalam penelitian terdahulu dan menunjukkan kecenderungan hasil yang lebih baik dalam klasifikasi sentimen [7][8]. Dengan menggabungkan Word2Vec dan *Logistic Regression* serta teknik penyeimbangan data SMOTE, diharapkan model mampu memberikan klasifikasi sentimen yang memiliki akurat tinggi dan bermanfaat dalam pemetaan opini publik [9].

II. METODE

Metodologi penelitian dirancang secara sistematis, dimulai dari pengumpulan data hingga evaluasi model. Pendekatan ini bertujuan untuk menganalisis berbagai faktor yang memengaruhi persepsi dan tingkat kepuasan konsumen,

sekaligus menentukan model prediksi yang paling sesuai [10]. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

A. Data Collection

Data yang digunakan dalam penelitian ini di *scrapping* menggunakan API Youtube. Pengumpulan data dilakukan dalam periode waktu Juli 2025 dan berhasil mengumpulkan sekitar 9.100 komentar. Data dikumpulkan dari sebelas video YouTube yang membahas keberadaan hewan peliharaan di ruang publik; kategori ini termasuk video berita, video opini masyarakat, video edukasi, dan video vlog sehari-hari. Kumpulan data teks komentar inilah yang menjadi dasar untuk analisis sentimen dalam penelitian ini.

B. Prepocessing Data

Tahap pra-pemrosesan data dilakukan untuk membersihkan dan menstandarisasi data teks mentah agar model dapat mengolahnya secara efektif dan akurat [11]. Proses ini terdiri dari beberapa tahapan fundamental sebagai berikut:

1) *Case Folding*: Tahap ini mengubah seluruh teks komentar menjadi format huruf kecil. Tujuannya adalah untuk menyeragamkan teks, sehingga kata-kata dengan makna yang sama namun memiliki kapitalisasi berbeda akan dianggap sebagai satu token yang identik oleh model [12].
2) *Noise Removal*: Ini adalah tahap untuk menghapus berbagai karakter dan elemen yang tidak relevan yang dapat mengganggu proses analisis sentimen. Dalam penelitian ini, elemen yang dihapus mencakup alamat situs web, tanda pagar, dan karakter-karakter non-ASCII.
3) *Stemming*: Proses *stemming* dilakukan untuk mengubah setiap kata menjadi bentuk dasarnya [13]. Proses ini sangat penting untuk mengurangi variasi kata dan memfokuskan analisis pada makna inti. Pada penelitian ini, proses stemming untuk Bahasa Indonesia dilakukan menggunakan library Sastrawi.
4) *Stopwords Removal*: Tahap ini bertujuan untuk menghilangkan kata-kata yang sering muncul namun kurang mengandung informasi penting dalam analisis sentimen, seperti kata sambung dan kata depan. Penghapusan kata-kata tersebut membantu model agar lebih fokus pada kata-kata yang berkontribusi terhadap penentuan sentimen [14].
5) *Tokenize*: Proses memecah teks menjadi kata individual [15]. Pada penelitian ini, proses *tokenize* dilakukan menggunakan *library* dari NLTK dengan fungsi word_tokenize.

C. Labelling Data

Setelah melewati tahap pra-pemrosesan, setiap komentar yang sudah bersih serta terstruktur dilabeli sentimen sebelum dapat digunakan untuk melatih model untuk klasifikasi. Penelitian ini menggunakan daftar kata kunci sentimen yang disusun secara manual untuk melakukan labelan sentimen secara otomatis. Penelitian awal terhadap korpus komentar serta referensi dari kamus sentimen Bahasa Indonesia digunakan untuk menyusun kata kunci positif dan negatif. Untuk menyamakan bentuk dasar kata, proses stemming digunakan pada semua kata dalam daftar tersebut. Proses pelabelan dilakukan dengan mencocokkan token dalam setiap komentar terhadap daftar kata kunci. Komentar diberi label 1 jika memiliki kecocokan dengan kata positif, 0 jika memiliki kecocokan dengan kata negatif, dan 2 jika tidak ada kecocokan sama sekali. Untuk menilai tingkat akurasi dan konsistensi pelabelan otomatis yang digunakan, verifikasi manual terhadap beberapa komentar secara acak direncanakan sebagai langkah evaluasi lanjutan. Hal ini dilakukan meskipun metode ini dinilai efektif dan sesuai dengan karakteristik data yang digunakan [16]. Dalam konteks analisis sentimen berbahasa Indonesia, evaluasi ini sangat penting untuk menilai kredibilitas metode berbasis leksikon.

D. Split Dataset

Data penelitian ini dibagi dengan rasio 80:20 untuk melatih dan menguji model. Sebesar 80% data (X_train, y_train) digunakan dalam proses pelatihan untuk mempelajari pola, sementara 20% sisanya (X_test, y_test) berfungsi sebagai data independen untuk mengukur seberapa baik kemampuan generalisasi model.

E. Word Embeddings

Penelitian ini memanfaatkan teknik *Word Embedding* dengan mengimplementasikan model Word2Vec melalui *library* Gensim. Model Word2Vec dilatih secara khusus hanya menggunakan data latih (X_train) untuk mempelajari representasi vektor dari setiap kata unik berdasarkan konteks kemunculannya. Setelah model terbentuk, setiap komentar (baik dari data latih maupun data uji) diubah menjadi satu vektor fitur tunggal dengan cara menghitung nilai rata-rata dari seluruh vektor kata yang terdapat di dalamnya. Hasil dari

tahap ini adalah sebuah representasi numerik untuk setiap komentar yang siap digunakan pada tahap klasifikasi model.

F. SMOTE (Synthetic Minority Oversampling Technique)

Untuk mengatasi masalah ketidakseimbangan distribusi kelas pada data latih, penelitian ini menerapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE bekerja dengan cara membuat sampel data sintetis baru untuk kelas minoritas (sentimen negatif), sehingga distribusi data menjadi lebih seimbang tanpa hanya menduplikasi data yang sudah ada [17]. Proses ini dilakukan setelah data diubah menjadi vektor fitur dan sebelum tahap pelatihan model klasifikasi.

Penerapan SMOTE berhasil menyeimbangkan data latih, di mana jumlahnya meningkat dari 3.305 menjadi 5.112 sampel. Hasilnya, kelas positif dan negatif memiliki distribusi yang seimbang, yaitu masing-masing sebanyak 2.556 data. Dengan distribusi yang proporsional ini, model diharapkan dapat mempelajari pola dari kedua kelas sentimen secara lebih akurat dan tidak bias terhadap salah satu kelas.

G. Logistic Regression

Tahap klasifikasi sentimen menggunakan algoritma *Logistic Regression*, yang diimplementasikan melalui *library* Scikit-learn. Model ini dipilih karena kemampuannya yang baik dalam menangani masalah klasifikasi biner serta kemudahan interpretasi hasilnya [18]. Proses pelatihan melibatkan penggunaan vektor fitur dari *Word Embedding* (X_train_vectors) sebagai variabel independen dan label sentimen (y_train) sebagai variabel dependen. Optimasi hyperparameter dilakukan dengan *GridSearchCV* untuk memastikan kinerja yang optimal [19]. Teknik ini secara sistematis menguji berbagai kombinasi parameter (seperti parameter regularisasi C dan jenis solver) melalui validasi silang untuk menemukan konfigurasi model terbaik, yang kemudian digunakan sebagai model final untuk evaluasi.

H. Evaluation

Untuk mengetahui seberapa baik model *Logistic Regression* dapat diterapkan pada data uji baru, evaluasi kinerja dilakukan. Analisis kuantitatif ini dilakukan menggunakan *library* Scikit-learn dengan metrik evaluasi standar. Kinerja model dianalisis secara rinci menggunakan *Confusion Matrix* untuk melihat jumlah prediksi yang benar dan salah, serta metrik kunci seperti accuracy untuk akurasi keseluruhan, ditambah precision, recall, dan F1-score untuk setiap kelas sentimen. Kumpulan metrik ini memberikan pemahaman komprehensif tentang efektivitas model dalam mengklasifikasi sentimen secara tepat dan meminimalkan kesalahan [20]. Terakhir, kinerja pada data uji juga dibandingkan dengan data latih untuk mengidentifikasi kemungkinan adanya overfitting.

I. Data Vizualization

Untuk mendapatkan pemahaman kualitatif yang lebih mendalam dari kumpulan data komentar, penelitian ini memanfaatkan teknik visualisasi data. Secara spesifik, Word Cloud dibuat secara khusus menggunakan *library* wordcloud di Python, dengan dukungan matplotlib untuk tampilan visual. Visualisasi ini dirancang untuk menonjolkan dan menampilkan kata-kata yang memiliki frekuensi kemunculan tertinggi atau paling dominan dalam keseluruhan data komentar yang telah melalui tahap pra-pemrosesan [21]. Dengan demikian, Word Cloud memberikan ringkasan tentang istilah-istilah penting atau topik utama yang paling sering didiskusikan pengguna YouTube seputar isu keberadaan hewan peliharaan di ruang publik, mempermudah identifikasi cepat.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data penelitian diperoleh melalui proses web scraping menggunakan API YouTube yang diimplementasikan dalam bahasa pemrograman Python. Pengumpulan data dilakukan selama bulan Juli 2025. Sebanyak 9.100 komentar mentah berhasil dikumpulkan dari 11 video bertema keberadaan hewan peliharaan di ruang publik. Data ini menjadi dasar utama proses analisis sentimen yang dilakukan dalam penelitian ini.

TABEL I
CONTOH DATA MENTAH

Username	Komentar
@ekafatmawati5336	Opini yang sangat bagus bang Saya sangat se...
@NuyBran	Kami sukaaa bikin trus ya❤ □

B. Hasil Prepocessing Data

Komentar yang dikumpulkan selanjutnya melalui tahap pra-pemrosesan untuk menghilangkan noise seperti URL, tanda baca, angka, dan stopwords. Tahapan ini juga mengembalikan kata ke bentuk dasar (stemming). Berikut contoh perbandingan komentar sebelum dan sesudah pra-pemrosesan pada tabel 2.

TABEL II
CONTOH HASIL DATA SETELAH PRA-PEMROSESAN

Komentar Asli	Setelah Pra-pemrosesan
Padahal baru liat judul 😱	padahal baru liat judul
<a href="https://www.youtube..	cie mimin sekarang udah boss mantap min

C. Labeling Data Sentimen

Setelah proses labeling dilakukan, diperoleh total 4.132 komentar yang siap digunakan untuk pelatihan model. Distribusi sentimen menunjukkan bahwa terdapat 3.196 komentar positif dan 936 komentar negatif. Hal ini menunjukkan proporsi data yang cenderung lebih banyak memuat opini positif terhadap keberadaan hewan peliharaan di ruang publik.

D. Word Embeddings

Model Word2Vec, yang dilatih khusus menggunakan korpus data latih, digunakan untuk membentuk representasi fitur penelitian ini menggunakan pendekatan Embedding Word. Pelatihan model dilakukan dengan library Gensim, dengan parameter vector_size=200, window=5, min_count=1, sg=1 (menggunakan arsitektur skip-gram), dan jumlah iterasi 30 epoch. Replikasi kata yang dihasilkan bersifat kontekstual terhadap domain dan gaya bahasa dalam dataset karena model dilatih dari awal (train from scratch) menggunakan 3.305 komentar berbahasa Indonesia yang telah melalui tahapan pra-pemrosesan.

Korpus menggambarkan setiap kata menjadi vektor 200 dimensi yang menunjukkan makna semantiknya sesuai dengan konteks di mana mereka muncul. Teknik agregasi mean pooling, yang menghitung rata-rata vektor dari setiap kata yang termasuk dalam satu komentar, digunakan untuk membentuk representasi kalimat atau komentar secara utuh. Akibatnya, setiap komentar digambarkan sebagai satu vektor numerik berdimensi tetap yang dapat digunakan langsung pada tahap pelatihan model klasifikasi.

Sebagai gambaran, komentar seperti "kucingnya lucu banget" akan di-tokenisasi menjadi ['kucing', 'lucu', 'banget']. Masing-masing token kemudian dipetakan ke vektor berdimensi 200 berdasarkan hasil pelatihan Word2Vec, dan selanjutnya dirata-rata secara elemen untuk menghasilkan vektor representasi kalimat.

Meskipun pelatihan model Word2Vec secara mandiri memberikan fleksibilitas dalam menyesuaikan representasi terhadap domain data, keterbatasan jumlah data dapat memengaruhi cakupan semantik dan generalisasi [22]. Oleh karena itu, pada penelitian selanjutnya disarankan untuk mempertimbangkan pemanfaatan model embedding pra-latih (pre-trained) seperti IndoNLU atau IndoBERT, yang telah dilatih pada korpus Bahasa Indonesia berskala besar. Untuk meningkatkan representasi fitur dan akurasi klasifikasi, integrasi embedding dapat bermanfaat, terutama untuk menangani keragaman linguistik yang lebih kompleks di media sosial.

E. Menyeimbangkan Data Menggunakan SMOTE

Distribusi awal data latih menunjukkan adanya ketidakseimbangan kelas yang cukup besar, yaitu sebanyak 2.556 data berlabel positif dan hanya 749 data berlabel negatif, dari total 3.305 data. Ketidakseimbangan ini berpotensi menimbulkan bias dalam proses pelatihan model, karena model cenderung lebih mudah mempelajari pola dari kelas mayoritas. Untuk mencegah hal tersebut, teknik penyeimbangan data Synthetic Minority Over-sampling Technique (SMOTE), yang diterapkan khusus pada data latih digunakan.

Setelah proses SMOTE, jumlah data pada kelas negatif meningkat menjadi 2.556, menghasilkan keseimbangan antara kelas positif dan negatif. Dengan demikian, total data latih setelah penyeimbangan bertambah menjadi 5.112 sampel, sebagaimana ditunjukkan pada Tabel 3. Meskipun penyeimbangan ini memberikan keuntungan dalam meningkatkan representasi kelas minoritas, penggunaan data sintetis juga memiliki potensi risiko overfitting, terutama jika sampel yang dihasilkan tidak cukup mewakili variasi alami dari data sebenarnya.

Untuk memitigasi risiko tersebut sekaligus meningkatkan reliabilitas evaluasi, proses pemodelan dan optimasi hyperparameter dilakukan menggunakan GridSearchCV dengan 5-Fold Cross Validation. GridSearchCV secara standar menerapkan K-Fold Stratified karena jenis pemodelan yang digunakan adalah klasifikasi biner. Ini menjaga distribusi kelas proporsional pada setiap fold. Strategi ini memungkinkan hasil evaluasi yang lebih stabil dan mencerminkan performa model secara umum terhadap data tak terlihat.

TABEL III
CONTOH HASIL DATA SETELAH PRA-PEMROSESAN

Label	**Jumlah Sebelum SMOTE**	**Jumlah Sesudah SMOTE**
Positif	2.556	2.556
Negatif	749	2.556
Total	**3.305**	**5.112**

F. Evaluasi Model

Evaluasi Evaluasi dilakukan untuk mengukur performa model *Logistic Regression* dalam mengklasifikasikan sentimen komentar YouTube. Pemilihan algoritma *Logistic Regression* dilatarbelakangi oleh sifatnya yang sederhana, cepat, dan mudah diinterpretasikan, serta performanya yang cukup baik pada kasus klasifikasi biner dengan representasi fitur berdimensi tinggi seperti *Word Embedding*. Selain itu, *Logistic Regression* juga memberikan transparansi dalam interpretasi bobot fitur, yang penting dalam studi berbasis opini publik.

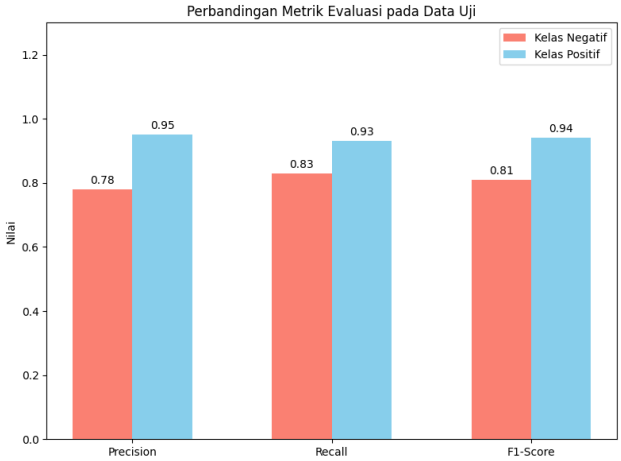
Model dikembangkan melalui pipeline dengan optimasi hyperparameter menggunakan GridSearchCV; Stratified 5-Fold Cross Validation digunakan untuk validasi. Validasi silang ini memastikan bahwa distribusi kelas tetap proporsional dalam setiap fold, sehingga hasil evaluasi lebih stabil dan mengurangi risiko bias akibat pembagian data yang tidak representatif.

Berdasarkan hasil pelatihan pada data latih yang telah diseimbangkan menggunakan SMOTE, model mencapai akurasi sebesar 96,53%, dengan nilai precision dan recall yang tinggi untuk kedua kelas. Precision pada kelas negatif mencapai 0,96 dan recall 0,98, sedangkan pada kelas positif masing-masing 0,98 dan 0,95. Ini menunjukkan kemampuan model untuk mempelajari pola sentimen setelah penyeimbangan kelas.

Selain itu, analisis yang dilakukan pada data uji yang tidak terpengaruh oleh proses SMOTE menunjukkan akurasi sebesar 92,74%. Pada data, kelas berlabel negatif memiliki nilai F1-score 0,84, sedangkan kelas berlabel positif memiliki nilai F1-score 0,95. Meskipun terdapat sedikit penurunan performa dibandingkan data latih, selisih yang

relatif kecil menunjukkan bahwa model tidak mengalami overfitting secara signifikan, dan mampu mempertahankan performa yang baik terhadap data baru.

Meskipun hasil yang diperoleh cukup tinggi, penelitian ini menyadari pentingnya pembanding model. Oleh karena itu, pada penelitian selanjutnya disarankan untuk melakukan evaluasi komparatif terhadap algoritma lain yang lebih kompleks, seperti *Random Forest*, *Support Vector Machine* (SVM), atau model *transformer-based* seperti IndoBERT, guna memastikan keunggulan pendekatan yang digunakan dalam konteks representasi semantik Bahasa Indonesia.



Gambar 1. Jumlah perbandingan hasil evaluasi metrik pada data uji.

G. *Visualisasi dengan WordCloud*

Untuk memperoleh pemahaman yang lebih komprehensif terhadap persepsi masyarakat, dilakukan analisis visual terhadap kata-kata yang paling sering muncul dalam komentar positif dan negatif. Hasilnya divisualisasikan dalam bentuk word cloud, yang memberikan gambaran umum tentang fokus emosi, opini, dan argumen yang terkandung dalam setiap kategori sentimen.

Gambar 2 menunjukkan word cloud dari komentar positif. Kata-kata yang paling dominan di antaranya adalah “suka”, “mau”, “baik”, “sayang”, “adopsi”, “bantu”, “steril”, “rezeki”, dan “alhamdulillah”. Kemunculan kata-kata seperti adopsi, bantu, kasihan, dan makhluk hidup mencerminkan empati publik terhadap hewan, serta dorongan untuk merawat dan melindungi mereka di ruang publik. Kehadiran kata-kata bernuansa religius seperti rezeki, berkah, dan subhanallah juga menunjukkan adanya justifikasi moral dan spiritual atas keberadaan hewan di ruang publik, khususnya dalam konteks budaya Indonesia. Visualisasi ini mengindikasikan bahwa sebagian besar komentar positif bersifat suportif dan penuh kepedulian terhadap kesejahteraan hewan.

Sebaliknya, Gambar 3 menunjukkan word cloud dari komentar negatif. Kata-kata yang paling menonjol adalah “sakit”, “hama”, “buang”, “kotor”, “gigit”, “virus”, “overpopulasi”, dan “ganggu”. Pola ini mencerminkan kekhawatiran masyarakat terhadap risiko kesehatan, kebersihan, dan ketidaknyamanan yang disebabkan oleh hewan di ruang publik. Munculnya istilah seperti berisik, bising, bau, serta kata-kata bernuansa ekstrem seperti musnah, tembak, dan gantung, memperlihatkan adanya kelompok opini yang cenderung menolak atau bahkan menganggap hewan peliharaan sebagai ancaman sosial.

Analisis ini memperkuat hasil kuantitatif bahwa mayoritas masyarakat dalam studi ini cenderung pro terhadap keberadaan hewan peliharaan di ruang publik, sebagaimana ditunjukkan oleh dominasi jumlah komentar positif (77%). Namun demikian, masih terdapat segmen masyarakat yang menyuarakan penolakan, terutama dengan narasi berbasis kesehatan, kenyamanan, dan kebersihan lingkungan.

Dengan demikian, studi ini tidak hanya menyajikan model klasifikasi sentimen yang akurat, tetapi juga membuka wawasan sosial mengenai polarisasi opini masyarakat yang dapat menjadi rujukan bagi pengambil kebijakan dalam merancang tata kelola ruang publik yang inklusif terhadap manusia dan hewan.

Gambar 2. Kata positif yang sering muncul.

Gambar 3. Kata negatif yang sering muncul.

H. *Pembahasan*

Model *Logistic Regression* yang dilatih dengan representasi Word2Vec mendeteksi sentimen dengan sangat baik. Teknik SMOTE berhasil mengatasi ketidakseimbangan data, terbukti dari akurasi data uji yang mencapai lebih dari 92%. Namun, recall pada kelas negatif sedikit lebih rendah dibanding kelas positif, yang dapat disebabkan variasi ekspresi negatif yang lebih beragam dan kontekstual. Secara umum, kombinasi preprocessing, Word2Vec, SMOTE, dan *Logistic Regression* terbukti efektif dalam memetakan persepsi publik terkait keberadaan hewan peliharaan di ruang publik.

IV. KESIMPULAN

Penelitian ini berhasil mencapai tujuannya untuk mengklasifikasikan sentimen masyarakat terhadap keberadaan hewan peliharaan di ruang publik menggunakan data komentar YouTube. Berdasarkan analisis terhadap 4.132 komentar, ditemukan bahwa sentimen publik secara dominan bersifat positif (77%). Penelitian ini telah menunjukkan bahwa model klasifikasi yang menggabungkan teknik penyeimbangan data SMOTE, Word2Vec, dan regresi logistik sangat efektif. Hal ini ditunjukkan dengan hasil evaluasi pada data uji yang mencapai akurasi 92,74%, dengan nilai F1-score 0,84 untuk kelas negatif dan 0,95 untuk kelas positif, menegaskan bahwa model ini andal dalam memetakan persepsi publik secara otomatis.

Untuk pengembangan di masa mendatang, disarankan beberapa perbaikan. Pertama, penelitian selanjutnya dapat memanfaatkan sumber data yang lebih beragam seperti Twitter dan Instagram untuk mendapatkan cakupan opini yang lebih komprehensif. Kedua, agar lebih akurat secara kontekstual, proses pelabelan data dapat diperkuat dengan menggunakan pendekatan semi-supervised atau crowdsourcing. Terakhir, kualitas pra-pemrosesan dapat ditingkatkan dengan mengembangkan kamus kata informal dan pemrosesan emotikon yang lebih spesifik untuk menangani ragam bahasa di media sosial.

DAFTAR PUSTAKA

[1] A. A. B. B. Anggawirya, C. I. A. C. Utari, and I. P. A. J. Wiguna, "Ketidaktaatan Terhadap Penggunaan Ruang Publik Ditinjau dari Theory of Planned Behavior (ToPB) di Kota Denpasar. Studi Kasus : Taman Kota Lumintang," *J. Arsit. Lansek.*, vol. 10, no. 1, p. 133, 2024, doi: 10.24843/jal.2024.v10.i01.p15.

[2] C. Andrew, A. Budiyantara, and I. Lewenusa, "Analisis Sentimen Komentar Pengguna YouTube Untuk Ponsel Di Indonesua Berbasis Web," *Sci. J. Ilm. Sains dan Teknol.*, vol. 3 No. 2, pp. 175–188, 2024, [Online]. Available: https://jurnal.kolibi.org/index.php/scientica/article/view/4212

[3] F. Aprilia, R. A. Anggraini, and Y. D. Putri, "Prediksi Kelulusan Siswa dengan Algoritma Pembelajaran Mesin: Aplikasi Regresi Linear dan Logistik pada Faktor-Faktor Pendidikan," *ROUTERS J. Sist. dan Teknol. Inf.*, vol. 3, no. 1, pp. 55–64, 2025, doi: 10.25181/rt.v3i1.3897.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.

[5] M. F. Yulianto, F. M. Hana, and A. Prihandono, "Penerapan Algoritma Naïve Bayes dalam Analisis Sentimen terhadap Mobil Listrik," vol. 22, no. 1, pp. 109–115, 2025, doi: 10.47065/bit.v5i2.2029.

[6] T. Muhayat, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Komentar Video Youtube Menggunakan Support Vector Machines," *Progresif J. Ilm. Komput.*, vol. 19, no. 1, p. 231, Feb. 2023, doi: 10.35889/progresif.v19i1.1060.

[7] I. Yanti and E. Utami, "Sentiment Analysis Of Indonesia ' S Capital Relocation Using Word2vec And Long Short-Term Memory Method Analisis Sentimen Pemindahan Ibu Kota Indonesia Menggunakan Word Embedding Word2vec Dan Metode Long Short-Term," *J. Tek. Inform.*, vol. 6, no. 1, pp. 149–157, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.1.2712.

[8] S. Al-Saqqa and A. Awajan, "The Use of Word2vec Model in Sentiment Analysis," in *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control*, New York, NY, USA: ACM, Dec. 2019, pp. 39–43. doi: 10.1145/3388218.3388229.

[9] Badriyah, T. Chamidy, and Suhartono, "Application of SMOTE in Sentiment Analysis of MyXL User Reviews on Google Play Store," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, no. 1, pp. 74–86, Jan. 2025, doi: 10.14421/jiska.2025.10.1.74-86.

[10] F. M. Syah Putra, S. Rakasiwi, and N. Ariyanto, "Twitter Sentiment Classification towards Telecommunication Provider Users in Indonesia," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 314–321, 2025, doi: 10.30871/jaic.v9i2.9143.

[11] A. A. Syam, G. Hardy M, A. Salim, D. F. Surianto, and M. Fajar B, "Analisis Teknik Preprocessing Pada Sentimen Masyarakat Terkait Konflik Israel-Palestina Menggunakan Support Vector Machine," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, pp. 1464–1472, 2024, doi: 10.29100/jipi.v9i3.5527.

[12] M. U. Albab, Y. K. P., and M. N. Fawaiq, "Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, pp. 1–12, 2023, doi: 10.26623/transformatika.v20i2.5374.

[13] S. Firman Sodiq, W. Desena, and A. Wibowo, "Penerapan Algoritma Stemming Nazief & Adriani Pada Proses Klasterisasi Berita Berdasarkan Tematik Pada Laman (Web) Direktorat Jenderal HAM Menggunakan Rapidminer," *Syntax J. Inform.*, vol. 11, no. 02, pp. 10–21, 2022, doi: 10.35706/syji.v11i02.7192.

[14] S. J. Angelina, A. Bijaksana, P. Negara, and H. Muhardi, "Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia," *JUARA (Jurnal Apl. dan Ris. Inform.*, vol. 02, no. 1, pp. 165–173, 2023, doi: 10.26418/juara.v2i1.69680.

[15] 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.

[16] M. Z. P. Anugrah, M. F. A. , Muhammad Lutfi, and A. P. S. Mahri, "Sistem Analisis Sentimen Ulasan Produk Berbahasa Indonesia Menggunakan Metode Lexicon Dengan Visualisasi Interaktif," *Kohesi J. Multidisiplin Saintek*, vol. 8, no. No 12, 2025.

[17] I. K. Dharmendra, I. M. Agus, W. Putra, and Y. P. Atmojo, "Evaluasi Efektivitas SMOTE dan Random Under Sampling pada Klasifikasi Emosi Tweet," *Informatics Educ. Prof. J. Informatics*, vol. 9, no. 2, pp. 192–193, 2024.

[18] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3216.

[19] J. Informatika, F. Sains, U. Jenderal, and A. Yani, "Klasifikasi Cuaca Jawa Barat menggunakan Ensemble Learning pada Data Meteorologi Weather Classification in West Java using Ensemble Learning on," vol. 14, pp. 2028–2044, 2025.

[20] R. L. Atimi and Enda Esyudha Pratama, "Implementasi Model Klasifikasi Sentimen Pada Review Produk Lazada Indonesia," *J. Sains dan Inform.*, vol. 8, no. 1, pp. 88–96, 2022, doi: 10.34128/jsi.v8i1.419.

[21] A. I. Kamil, O. N. Pratiwi, and D. Witarsyah, "Analisis Sentimen dan Pemodelan Topik terhadap Aplikasi Pembelajaran Online pada Platform Google Play," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 2, pp. 836–849, 2025, doi: 10.29100/jipi.v10i2.6023.

[22] A. A. Aliero, B. S. Adebayo, H. O. Aliyu, A. G. Tafida, B. U. Kangiwa, and N. M. Dankolo, "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words," *Int. J. Comput. Appl.*, vol. 185, no. 33, pp. 44–55, 2023, doi: 10.5120/ijca2023923106.