JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i1.5604

ANALISIS SENTIMEN MENGGUNAKAN METODE LEXICON-BASED DAN SUPPORT VECTOR MACHINE PADA PRESIDEN DAN WAKIL PRESIDEN INDONESIA PERIODE 2024–2029

Syahrani Ratnaswari [1*], Nur Cahyo Wibowo [2], Dhian Satria Yudha Kartika [3]

[1,2,3] Universitas Pembangunan Nasional Veteran Jawa Timur; Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya; 60294

Received: 4 Desember 2024
Accepted: 14 Januari 2025
Published: 20 Januari 2025

Keywords:
lexicon;
president;
sentiment;
twitter;
SVM.

Corespondent Email:
sratnaswari@gmail.com

Abstrak. Twitter merupakan salah satu platform media sosial yang memiliki pengaruh besar di Indonesia. Banyak pengguna Twitter di Indonesia sering membagikan pendapat atau perasaan mereka tentang berbagai isu, termasuk tentang Presiden dan Wakil Presiden terpilih periode 2024–2029. Tujuan utama penelitian ini adalah penerapan algoritma *machine learning* analisis sentimen berbasis leksikon (*Lexicon-Based*). Algoritma yang digunakan untuk pembuatan model adalah *Support Vector Machine* (SVM). Proses uji coba melibatkan beberapa langkah, mulai dari pengambilan data dari Twitter, kemudian proses pembersihannya (*preprocessing*) sampai *train* model. Dalam pembuatan model SVM, berbagai konfigurasi diuji untuk menemukan model terbaik agar hasil analisis menjadi lebih akurat. Hasil uji coba menunjukkan bahwa metode SVM memiliki performa yang baik, dengan tingkat akurasi sekitar 93%. Laporan klasifikasi menunjukkan kinerja yang sangat baik di semua kelas. Kelas "negatif" memiliki *F1-score* sebesar 0,93, sedangkan kelas "netral" mencapai *F1-score* 0,95, dan kelas "positif" memperoleh *F1-score* 0,94. Presisi dan *recall* juga sangat tinggi di setiap kelas, dengan kelas "netral" mencatat presisi tertinggi sebesar 0,96 dan *recall* sebesar 0,94.

Abstract. *Twitter holds substantial influence in Indonesia, with users frequently expressing their views on various topics, including the upcoming Presidential and Vice Presidential elections for 2024–2029. This study aims to implement a lexicon-based sentiment analysis using the Support Vector Machine (SVM) algorithm. The process began with* data *collection from Twitter, followed by preprocessing and testing various configurations to build an optimal SVM model for accurate analysis. The results demonstrated that the SVM method performed exceptionally well, achieving an accuracy rate of approximately 93%. The classification report highlighted excellent performance across all classes, with F1-scores of 0.93 for the "negative" class, 0.95 for the "neutral" class, and 0.94 for the "positive" class. Precision and recall were also notably high, particularly for the "neutral" class, which showed a precision of 0.96 and recall of 0.94*

1. PENDAHULUAN

Indonesia merupakan negara demokrasi, sesuai pembukaan UUD 1945 yang menatakan bahwa kedaulatan berada di tangan rakyat (UUD 1945). Salah satu implementasi negara demokrasi adalah pelaksanaan pemilu [1]. Belakangan ini, topik mengenai pemilu, terlebih pada topik presiden dan wakil presiden Indonesia terpilih untuk periode 2024–2029 menjadi pembahasan yang sering muncul di berbagai media sosial. Tak terkecuali media sosial X (Twitter). Melaui media sosial X (twitter) ini, semua masyarakat memiliki kesempatan yang sama untuk berbagi opini mereka, termasuk mengenai pasangan presiden dan wakil presiden Indonesia terpilih periode 2024–2029 yang akhir-akhir ini marak dibicarakan.

Analisis sentimen atau *opinion m*ining adalah cabang penelitian yang mengkaji evaluasi, pendapat, sikap, penilaian, sentimen, dan emosi yang diungkapkan oleh masyarakat [2]. Analisis sentimen ini dilakukan pada sejumlah data besar, yaitu data media sosial X (Twitter) yang diperoleh melalui media sosial X (Twitter) API untuk mengidentifikasi arah atau kecenderungan masyarakat Indonesia terhadap presiden dan wakil presiden Indonesia terpilih periode 2024–2029 apakah cenderung ke arah pro (positif), kontra (negatif), atau netral.

Penelitian ini memilih menggunakan metode *Lexicon-Based* dan *Support Vector Machine* (SVM). *Lexicon-Based* digunakan untuk mengidentifikasi dan menganalisis sentimen atau emosi dalam teks dengan cara mencocokkan kata-kata dalam teks dengan kamus kata-kata yang telah dikategorikan berdasarkan sentimen atau emosi tertentu [3]. Sedangkan SVM digunakan untuk membuat model klasifikasi sehingga dapat digunakan untuk memprediksi sentimen teks secara lebih akurat berdasarkan pola yang dipelajari dari data *train* [4].

2. TINJAUAN PUSTAKA

2.1 Demokrasi di Indonesia

Demokrasi dan prinsip negara hukum adalah elemen yang saling terkait dalam sistem pemerintahan demokratis. Demokrasi memerlukan kerangka hukum untuk memfasilitasi prosesnya dan keteraturan dari implementasi hukum agar demokratisasi efektif. Prinsip hukum juga memerlukan demokrasi sebagai landasan agar hukum responsif terhadap kepentingan masyarakat. Tanpa demokrasi, risiko otoriterisme dalam pengaturan hukum meningkat. Berdasarkan hal tersebut, dapat dipahami bahwa salah satu prinsip utama dalam menjalankan kedaulatan rakyat adalah melalui keterlibatan langsung rakyat dalam proses pemerintahan. Salah satu bentuk keterlibatan yang penting adalah melalui proses pemilihan dan pencalonan dalam transisi kekuasaan, di mana rakyat memiliki hak untuk memilih serta dipilih sebagai wujud nyata dari kedaulatan yang mereka pegang. Hubungan erat antara demokrasi dan prinsip negara hukum esensial untuk memastikan pemerintahan yang adil, transparan, dan berkeadilan bagi seluruh rakyat [1].

2.2 Analisis Sentimen

Analisis sentimen dalam NLP mengevaluasi opini dalam teks dengan metode berbasis leksikon dan *machine learning* untuk mengukur sentimen positif, netral, dan negatif. Penggunaannya meluas ke sektor seperti produk konsumen, layanan, perawatan kesehatan, keuangan, serta acara sosial dan politik. Pentingnya analisis ini memotivasi banyak penelitian yang berfokus pada metodologinya [2].

2.3 Media Sosial X (Twitter)

Platform media sosial menjadi alat penting bagi individu untuk berinteraksi dan bertukar informasi dengan cepat. Studi menunjukkan bahwa Media Sosial X (Twitter) memiliki fitur-fitur yang meningkatkan efisiensi komunikasi dan berbagi informasi dibandingkan dengan platform lain, menjadikannya destinasi populer untuk interaksi sosial sehari-hari [5]. Penelitian lain oleh Hangbo Fang menemukan bahwa media sosial X memiliki volume diskusi yang jauh lebih tinggi daripada platform lain seperti Reddit dan HackerNews, dengan jumlah kiriman setidaknya sepuluh kali lebih banyak. Perbedaan serupa juga terlihat dalam jumlah pengguna yang mengirimkan konten [5].

2.4 *Text Mining*

Text mining adalah bagian dari data mining yang berfokus pada penggalian dan analisis data dalam bentuk teks. *Text mining* biasanya digunakan untuk *classification, clustering,*

information extraction, dan information retrival. Melalui *text mining,* informasi berharga dapat diambil dari data teks, misalnya dengan melakukan analisis sentimen terhadap ulasan pelanggan di media *online*. Hal ini memungkinkan perusahaan untuk lebih memahami perspektif, opini, atau perasaan pelanggan terhadap produk atau layanan mereka dengan lebih akurat dan mendalam [6].

2.5 *Lexicon Based*

Lexicon Based adalah pendekatan dalam klasifikasi data *mining* yang menggunakan kamus kata-kata dengan penanda polaritas atau orientasi sentimen. Dengan metode ini, kamus kata-kata digunakan untuk secara otomatis mengevaluasi sentimen dalam teks, seperti positif, negatif, atau netral. Metode *Lexicon* sering digunakan dalam analisis sentimen di berbagai bidang, seperti analisis media sosial, penilaian produk, dan analisis opini [3].

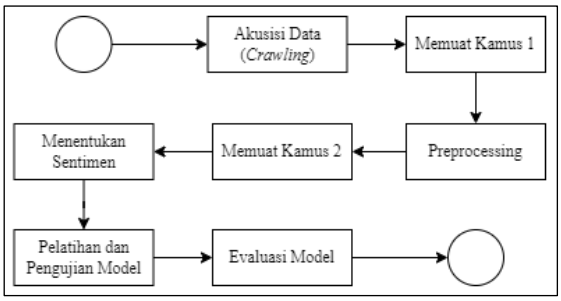
Lexicon-Based digunakan untuk mengidentifikasi dan menganalisis sentimen atau emosi dalam teks dengan cara mencocokkan kata-kata dalam teks dengan kamus kata-kata yang telah dikategorikan berdasarkan sentimen atau emosi tertentu [3].

2.6 *Support Vector Machine*

Support Vector Machine (SVM) adalah teknik yang didasarkan pada teori pembelajaran statistik dan telah terbukti menghasilkan performa yang baik dalam berbagai aplikasi praktis, mulai dari pengenalan digit tulisan tangan hingga klasifikasi teks. SVM juga unggul dalam menangani data berdimensi tinggi dan mampu mengatasi permasalahan yang terkait dengan dimensionalitas [4]. SVM digunakan untuk membuat model klasifikasi sehingga dapat digunakan untuk memprediksi sentimen teks secara lebih akurat berdasarkan pola yang dipelajari dari data *train* [4].

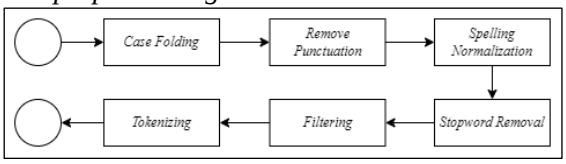
3. METODE PENELITIAN

Dalam penelitian ini, data yang digunakan adalah data primer, yang diperoleh secara langsung melalui proses *crawling* dari media sosial X (Twitter). Gambar 1 menunjukkan diagram alur penelitian.



Gambar 1. Alur Penelitian

Proses akuisisi data (crawling) diambil dari media sosial X (Twitter) menggunakan tweet-harvest. Tweet-harvest merupakan *tools* yang digunakan untuk melakukan *crawling* data pada media sosial X (Twitter) dengan menggunakan *Application Programming Interface* (API) [7]. Memuat kamus 1 atau *lexicon load* 1 bertujuan memastikan bahwa data yang dianalisis hanya mencakup kata yang relevan dan sesuai dengan kata yang ada di KBBI dan KBBA. KBBI adalah Kamus Besar Bahasa Indonesia, sedangkan KBBA adalah Kamus Besar Bahasa Alay [8]. *Preprocessing* bertujuan untuk mengolah teks sehingga dapat diubah menjadi format data yang siap untuk diproses tahap pemrosesan berikutnya [7]. Gambar 2 adalah alur *preprocessing* data.

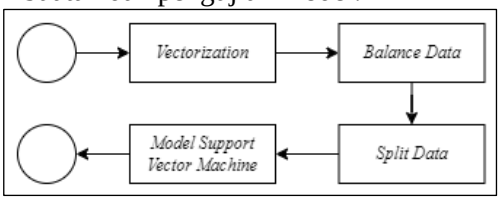


Gambar 2. Alur *Preprocessing* data

Case folding adalah proses mengubah karakter huruf besar (*capital*) menjadi huruf kecil (*lowercase*) [9]. *Remove Punctuation* adalah proses memisahkan tanda baca dan simbol lainnya selain alfabet [8]. *Normalization* adalah proses yang tujuannya adalah untuk mengubah daftar kata slang menjadi kata yang lebih formal sesuai KBBI atau kata yang belum menjadi kata dasar [9]. *Stopword removal* adalah proses eliminasi kata yang dianggap tidak signifikan atau berlebihan dari sebuah dokumen [10]. Proses *filtering* atau penyaringan melibatkan penghapusan data duplikat dan data kosong untuk memastikan bahwa hanya data yang relevan yang digunakan dalam analisis [11]. *Tokenizing* membagi kalimat menjadi token atau bagian kata merupakan langkah yang krusial dalam proses *preprocessing text* [9].

Memuat kamus 2 atau *lexicon load* 2 bertujuan memastikan bahwa data yang dianalisis hanya mencakup kata yang relevan dan diperlukan untuk penelitian, dengan menggunakan kamus yang berisi daftar kata atau frasa penting untuk analisis sentimen, seperti *positive keyword* dan *negative keyword*, yang memungkinkan normalisasi kalimat dan ekstraksi kata kunci [4]. Pembobotan metode *lexicon* menggunakan kamus leksikal yang diambil dari daftar kata-kata opini Liu yang telah dimodifikasi dan diterjemahkan ke dalam bahasa Indonesia [12]. Apabila tweet lebih banyak mengandung kata-kata yang bersifat negatif, maka tweet tersebut akan diklasifikasikan sebagai sentimen negatif. Pun sebaliknya. Namun, dalam situasi lain, jika kedua kata tersebut memiliki nilai yang setara, maka tweet akan dikategorikan sebagai netral [13]. Untuk penentuan batas ambang label positif, negatif, dan netral. Batas ambang untuk label positif adalah jika skornya > 0, label negatif jika skornya < 0, dan netral jika skornya = 0 [13].

Hasil analisis sentimen ini mencakup visualisasi *wordcloud* untuk menampilkan persebaran kata yang sering muncul, *pie chart* yang menunjukkan distribusi label sentimen, *line plot* yang menggambarkan tren sentimen, serta bar chart untuk menampilkan kata yang paling sering muncul. Proses pe*train* an dan pengujian model melibatkan beberapa langkah, termasuk vektorisasi (*text vectorization*), penyeimbangan data (*balance* dataset), pemisahan data (*split* dataset), dan pembuatan model (*build model*). Gambar 3 adalah proses pembuatan dan pengujian model.



Gambar 3. Alur Pembutan dan Pengujian Model

Text Vectorization dilakukan menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*), teknik yang mengubah teks menjadi vektor numerik dengan memberi bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam dokumen serta seberapa unik kata tersebut di seluruh kumpulan dokumen [12]. *Balance* data (penyeimbangan data) dilakukan dengan teknik *oversampling* untuk mengatasi ketidak seimbangan kelas dengan menambah jumlah contoh dari kelas minoritas [14]. *Split data* adalah proses membagi dataset menjadi dua bagian, yaitu data *train* (*train*) dan data *test* (*test*) [13]. Pembagian *dataset* dilakukan dengan rasio 80:20, di mana 80% data digunakan untuk *train* model dan 20% digunakan untuk test model.

Algoritma *Support Vector Machine* (SVM) adalah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, yang bekerja dengan mencari hyperplane optimal yang memisahkan data dari berbagai kelas. Kernel SVM yang digunakan dalam penelitian ini menerapkan kernel polinomial, yang memungkinkan pemisahan data yang tidak linear dengan memproyeksikannya ke ruang dimensi yang lebih tinggi [15]. Evaluasi dalam penelitian ini dilakukan menggunakan confusion matrix dan classification report, dengan mengukur performa melalui metrik akurasi, presisi, recall, dan *F1-score*, yang dihitung berdasarkan *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [16].

4. HASIL DAN PEMBAHASAN

Dataset didapat dengan melakukan crawling pada media sosial X (twitter) menggunakan tools tweet-harvest. Data yang diperlukan dalam penelitian ini adalah opini masyarakat Indonesia di media sosial X (Twitter) mengenai pemilihan presiden dan wakil presiden Indonesia periode 2024–2029. Data yang dikumpulkan berupa tweet yang mengandung kata-kata “Prabowo”, “Gibran”, dan “Paslon 02” selama periode pengambilan data dari 14 Februari 2024 hingga 30 April 2024. Jumlah data yang dikumpulkan pada proses ini adalah sebanyak 19.852 data.

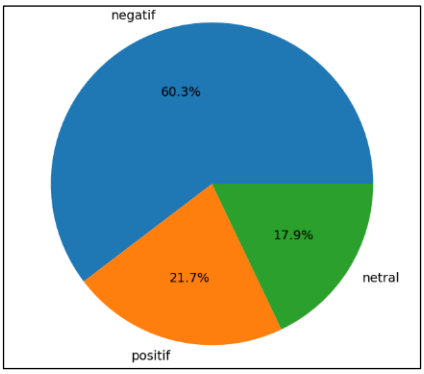
Load dictionary 1 dilakukan dengan menghapus data yang tidak mengandung kata bebahasa Indonesia atau bahasa alay sama sekali. Pada proses ini, baris-baris yang tidak memiliki kata-kata tersebut dihapus dari data frame. Setelah tahap ini dilakukan, jumlah data yang tersisa adalah 12.141 dataset. Pada proses *preprocessing*, hanya kolom created_at, username, dan full_text yang diambil dari file yang telah disaring pada tahap sebelumnya.

Proses *Case Folding* menggunakan fungsi str.lower() dari *library* Pandas dengan tujuan

mengonversi huruf kapital pada atribut ‘full_text’ menjadi huruf kecil. Hal ini dilakukan agar teks dapat dikenali dengan makna yang setara. Tahapan *Remove Punctuation* ini berfungsi untuk membersihkan teks dalam sebuah *dataset* dengan menghilangkan berbagai elemen yang tidak diinginkan, seperti karakter khusus, tanda baca, dan whitespace yang berlebihan. Hasil dari preprocess ini, dimasukkan dalam kolom bernama ‘cleaned_text’. Proses *Normalization* ini mengubah kata-kata yang belum sesuai dengan KBBI menjadi kata-kata yang sesuai dengan KBBI. Hasil dari *preprocessing Normalization* pertama akan disimpan dalam kolom ‘normalized_text’. *Stopword removal* bertujuan menghapus kata-kata umum yang tidak memberikan informasi penting dari teks yang telah dinormalisasi, dengan memanfaatkan daftar *Stopword* bahasa Indonesia yang tersedia di *library* 'nltk'.

Pada proses *filtering* ini dilakukan penyaringan data duplikat dan data kosong pada subset kolom ‘Stopwords’ untuk kemudian dihapuskan. Data yang yang didapat setelah proses ini sebanyak 11.792 dengan tipe data *object*. Pada proses *tokenizing* digunakan untuk melakukan tokenisasi pada teks yang telah diproses sebelumnya untuk dipecah kalimat menjadi token-token individu. Memuat kamus 2 dilakukan dengan menghapus data yang tidak mengandung kata sentimen. Baris-baris yang tidak mengandung kata-kata positif atau negatif (di mana jumlah_positif dan jumlah_negatif keduanya bernilai 0) dihapus. Setelah tahap ini dilakukan, jumlah data yang tersisa adalah 11.430 dataset.

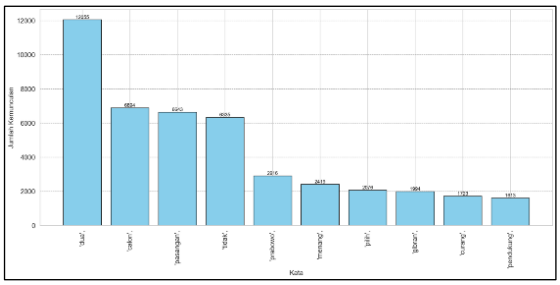
Setiap token yang telah diproses diperiksa dengan kamus leksikon untuk diberi nilai, yang kemudian dijumlahkan guna menentukan skor polaritas dan mengklasifikasikan sentimen tweet sebagai positif, negatif, atau netral. Hasil analisis sentimen menggunakan berbasis leksikon terdapat 2.409 tweet netral, 2.489 tweet positif, dan 6.897 tweet negatif sesuai pada Gambar 4. Berdasarkan hasil ini, mayoritas tweet mengenai Presiden dan Wakil Presiden Indonesia Terpilih periode 2024–2029 memiliki sentimen negatif. Gambar 5 menjelaskan kata yang sering muncul untuk sentimen presiden dan wakil presiden terpilih Indonesia periode 2024–2029.



Gambar 4. Persebaran Label Sentimen

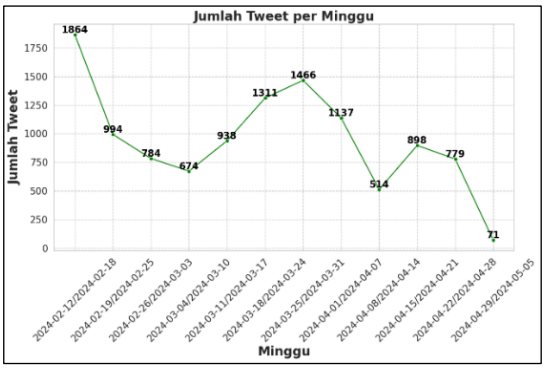
Gambar 5. *Wordcloud* Persebaran Kata

Sedangkan top-10 frekuensi kata yang sering muncul dari mulai yang paling banyak digunakan untuk analisis sentiment ini adalah ‘dua’, ‘calon’, ‘pasangan’, ‘tidak’, ‘Prabowo’, ‘menang’, ‘pilih’, ‘Gibran’, ‘orang’, dan ‘pendukung’. Sesuai yang tertera pada Gambar 6.



Gambar 6. Top-10 Kata yang Sering Muncul

Jumlah tweet mengalami fluktuasi dari minggu ke minggu seperti yang tergambarkan pada gambar 7. Puncak tertinggi berada pada interval pertama yaitu pada minggu terakhir Januari 2024 (22 Januari 2024 hingga 29 Januari 2024) dengan mencapai 1854 cuitan. Sedangkan interval terendah berada pada interval terahir dengan hanya 71 cuitan.



Gambar 7. Frekuensi Tweet Perminggu

ini dilakukan untuk melakukan pengujian model. Proses vektorisasi menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF adalah teknik yang digunakan untuk mengubah teks menjadi vektor numerik berdasarkan pentingnya kata-kata dalam dokumen. Hasil vektorisasi ini disimpan dalam matriks TF-IDF yang berbentuk *array*. *Balance* data dilakukan dengan Random *oversampling*. Distribusi setelah *oversampling* menunjukkan bahwa setiap kelas dalam dataset kini memiliki jumlah contoh yang sama, yaitu 6897 contoh untuk masing-masing kelas: 'negatif', 'positif', dan 'netral'. Dataset yang telah di-*oversampling* dibagi menjadi data *train* dan data *test*, dengan 80% data digunakan untuk data *train* dan 20% untuk data *test*.

Hasil evaluasi model menunjukkan bahwa proses prediksi memerlukan waktu sekitar 45,60 detik dan mencapai akurasi keseluruhan sebesar 93,96%. Laporan klasifikasi menunjukkan performa yang sangat baik di seluruh kelas. Kelas "negatif" memiliki *F1-score* 0,93, kelas "netral" mencapai *F1-score* 0,95, dan kelas "positif" memiliki *F1-score* 0,94. Presisi dan *recall* juga sangat tinggi di setiap kelas, dengan kelas "netral" menunjukkan presisi tertinggi sebesar 0,96 dan *recall* 0,94. Gambar 8 menunjukkan hasil dari model SVM.

```
Time taken: 45.599759340286255
Accuracy: 0.9395989369412902

Classification Report:
              precision    recall  f1-score   support

     negatif       0.90      0.96      0.93      1379
      netral       0.96      0.94      0.95      1380
     positif       0.96      0.93      0.94      1380

    accuracy                           0.94      4139
   macro avg       0.94      0.94      0.94      4139
weighted avg       0.94      0.94      0.94      4139


Confusion Matrix:
 [[1317   34   28]
 [  65 1294   21]
 [  78   24 1278]]
```

Gambar 8. Hasil Model SVM

5. KESIMPULAN

Berdasarkan analisis dan implementasi yang dilakukan, penggunaan metode *lexicon-based* sentimen analisis tentang Presiden dan Wakil Presiden Indonesian Periode 2024–2029 pada media sosisal X (Twitter) dapat disimpulkan bahwa distribusi label yang diperoleh menunjukkan bahwa mayoritas data tergolong dalam kategori sentimen negatif, yaitu sebanyak 6,879. Di sisi lain, data dengan sentimen positif berjumlah 2,482, dan data dengan sentimen netral berjumlah 2,046. Dari hasil ini, terlihat bahwa sentimen negatif mendominasi, menunjukkan bahwa sebagian besar data yang dianalisis cenderung mengandung sentimen negatif dibandingkan dengan sentimen positif atau netral. Hal ini dapat memberikan wawasan penting mengenai persepsi umum atau reaksi pengguna terhadap topik tertentu dalam dataset yang digunakan.

Dalam melakukan *lexicon-based* sentimen analisis, metode SVM mampu mencapai tingkat akurasi pengujian yang baik. Hasil akurasi pengujiannya yaitu sekitar 93,96%. Laporan klasifikasi menunjukkan performa yang sangat baik di seluruh kelas. Kelas "negatif" memiliki *F1-score* 0,93, kelas "netral" mencapai *F1-score* 0,95, dan kelas "positif" memiliki *F1-score* 0,94. Presisi dan *recall* juga sangat tinggi di setiap kelas, dengan kelas "netral" menunjukkan presisi tertinggi sebesar 0,96 dan *recall* 0,94.

UCAPAN TERIMA KASIH

Penelitian ini tidak lepas dari bantuan berbagai pihak. Ucapan terima kasih diucapkan kepada kedua orang tua penulis yang telah mendukung penulis dalam hal materi dan moral. Berkat doa, motivasi, dan dukungan penuh yang mereka berikan, penulis mampu menjalani proses penelitian ini dengan lebih

tenang dan percaya diri. Selain itu, penulis ingin menyampaikan penghargaan dan ucapan terima kasih yang tulus kepada dosen pembimbing. Terima kasih atas kesabaran dan wawasan yang diberikan oleh dosen pembimbing sangat membantu penulis dalam menghadapi tantangan-tantangan yang ada dalam penelitian ini. Penulis menyadari bahwa penelitian ini masih memiliki kekurangan, namun besar harapan penulis agar penelitian ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan.

DAFTAR PUSTAKA

[1] Suhartini, S. (2019). Demokrasi Dan Negara Hukum. Jurnal de jure, 11(1).

[2] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.

[3] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[4] Ariadi, D., & Fithriasari, K. (2016). Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer. Jurnal Sains dan Seni ITS, 4(2).

[5] Fang, H., Vasilescu, B., & Herbsleb, J. (2023, May). Understanding information diffusion about open-source projects on Twitter, HackerNews, and Reddit. In 2023 IEEE/ACM 16th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE) (pp. 56-67).

[6] Huda, K., Pohan, S. D., & Herlina, Y. (2024). PENERAPAN PEMBOBOTAN *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* DAN ALGORITMA K-NEAREST NEIGHBOR UNTUK ANALISIS ULASAN HOTEL DI SITUS TRIPADVISOR. Jurnal Informatika dan Teknik Elektro Terapan, 12(3).

[7] Steven A. P., Andri W., 2023. Analisis Sentimen Artifical Intelligence (AI) pada Media Sosial Twitter Menggunakan Metode Lexicon Based. Jurnal Sistem & Teknologi Informasi Komunikasi, 7(1), 21-28.

[8] Matulatuwa, F. M. (2017). Text mining dengan metode lexicon based untuk sentiment analysis pelayanan PT. Pos Indonesia melalui media sosial Twitter (Doctoral dissertation, Magister Sistem Informasi Program Pascasarjana FTI-UKSW).

[9] Undap, M., Rantung, V. P., & Rompas, P. T. (2021). Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based. JOINTER: Journal of Informatics Engineering, 2(02), 39-46.

[10] Chanda, S., & Pal, S. (2023). The effect of *Stopword removal* on information retrieval for code-mixed data obtained via social media. SN Computer Science, 4(5), 494.

[11] Sulianta, F. (2023). Basic Data Mining from A to Z. Feri Sulianta.

[12] Wahid, D. H., & Azhari, S. N. (2016). Peringkasan sentimen esktraktif di twitter menggunakan hybrid TF-IDF dan cosine similarity. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 10(2), 207-218.

[13] Syakur, A. (2022). Implementasi Metode Lexicon Base Untuk Analisis Sentimen Kebijakan Pemerintah Dalam Pencegahan Penyebaran Virus Corona Covid-19 Pada Twitter. Jurnal Ilmiah Informatika Komputer, 26(3), 247-260.

[14] Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi algoritma synthetic minority over-sampling technique untuk menangani ketidakseimbangan kelas pada dataset klasifikasi. SISTEMASI: Jurnal Sistem Informasi, 10(2), 445-459.

[15] Wahyuni, R. T., Prastiyanto, D., & Supraptono, E. (2017). Penerapan algoritma cosine similarity dan pembobotan tf-idf pada sistem klasifikasi dokumen skripsi. Jurnal Teknik Elektro, 9(1), 18-23.

[16] Rahmad, F., Suryanto, Y., & Ramli, K. (2020, July). Performance comparison of anti-spam technology using confusion matrix classification. In IOP Conference Series: Materials Science and Engineering (Vol. 879, No. 1, p. 012076). IOP Publishing.