Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

Klasifikasi Sentimen pada Dataset Terbatas Menggunakan Random Forest dan Word2Vec

Dina Deswara Fitri, Surya Agustian*, Pizaini, Suwanto Sanjaya

Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12050120494@students.uin-suska.ac.id, [2,*]surya.agustian@uin-suska.ac.id, [3]pizaini@uin-suska.ac.id, [4]suwanto.sanjaya@uin-suska.ac.id
Email Penulis Korespondensi: surya.agustian@uin-suska.ac.id
Submitted: 13/11/2024; Accepted: 26/11/2024; Published: 26/11/2024

Abstrak−Pengukuran sentimen terhadap opini publik di media sosial penting untuk memahami pandangan masyarakat terhadap berbagai isu, termasuk tokoh publik dan peristiwa politik. Penelitian ini mengeksplorasi efektivitas algoritma Random Forest dengan representasi kata berbasis Word2Vec untuk klasifikasi sentimen pada dataset terbatas. Studi kasus melibatkan tweet mengenai Kaesang Pangarep sebagai Ketua Umum PSI, ditambah data eksternal terkait Covid-19 dan topik umum. Dataset diproses menggunakan teknik cleaning, case folding, stopword removal, stemming, dan tokenisasi. Kata-kata dalam dataset direpresentasikan menggunakan model Word2Vec dengan arsitektur Continuous Bag of Words (CBOW) dan dimensi vektor 500.Random Forest digunakan untuk mengklasifikasikan sentimen ke dalam kategori positif, negatif, atau netral. Pada tahap awal, model dilatih menggunakan 300 sampel data per label, tetapi hasilnya menunjukkan performa yang kurang memuaskan dengan F1-Score 49,00% dan akurasi 50,00%. Untuk meningkatkan kinerja, dataset diperluas dengan menambahkan 900 sampel dari Kaesang dan 1.080 sampel dari topik eksternal. Hasil akhir menunjukkan peningkatan dengan F1-Score 49,89%, akurasi 58,29%, presisi 49,16%, dan recall 56,47%. Penelitian ini mengonfirmasi bahwa penggunaan Random Forest dengan representasi kata dari Word2Vec dapat meningkatkan performa klasifikasi sentimen meskipun dengan dataset terbatas, serta memberikan kontribusi dalam pengembangan teknik analisis sentimen di bidang pembelajaran mesin.

Kata Kunci: Klasifikasi sentimen; Random Forest; Word2vec ; Dataset Terbatas; Sosial Media

Abstract−Sentiment measurement of public opinion on social media is essential for understanding societal views on various issues, including public figures and political events. This research explores the effectiveness of the Random Forest algorithm with Word2Vec-based word representation for sentiment classification on a limited dataset. The case study involves tweets regarding Kaesang Pangarep as the Chairman of PSI, supplemented by external data related to Covid-19 and general topics. The dataset was processed using cleaning techniques, case folding, stopword removal, stemming, and tokenization. Words in the dataset were represented using the Word2Vec model with a Continuous Bag of Words (CBOW) architecture and a vector dimension of 500. Random Forest was employed to classify sentiment into positive, negative, or neutral categories. In the initial phase, the model was trained using 300 samples per label; however, the results showed unsatisfactory performance with an F1-Score of 49.00% and an accuracy of 50.00%. To improve performance, the dataset was expanded by adding 900 samples from Kaesang and 1,080 samples from external topics. The final results indicated an improvement with an F1-Score of 49.89%, an accuracy of 58.29%, precision of 49.16%, and recall of 56.47%. This research confirms that the use of Random Forest with word representation from Word2Vec can enhance sentiment classification performance, even with a limited dataset, and contributes to the development of sentiment analysis techniques in the field of machine learning.

Keywords: Sentiment Classification; Random Forest; Word2Vec; Limited Dataset; Media Sosial

1. PENDAHULUAN

Di Indonesia, media sosial telah menjadi platform yang sangat populer dan sering digunakan oleh masyarakat. Platform berbasis online ini, baik dalam bentuk situs web maupun aplikasi, memungkinkan interaksi pengguna tanpa harus bertemu secara langsung[1]. Twitter, sebagai salah satu platform media sosial, telah menjadi media utama untuk menyuarakan pendapat publik. Platform ini populer karena memungkinkan pengguna untuk berkomentar dan mengungkapkan opini secara bebas[2]. Salah satu peristiwa yang menjadi perhatian publik adalah pengangkatan Kaesang Pangarep sebagai Ketua Umum Partai Solidaritas Indonesia (PSI) pada 25 September 2023[3]. Keputusan ini memicu berbagai reaksi, baik positif maupun negatif, yang tersebar luas di platform Twitter. Dengan keragaman opini yang ada, menganalisis sentimen masyarakat terhadap topik-topik semacam ini menjadi tantangan tersendiri, terlebih lagi ketika jumlah data yang tersedia terbatas.

Analisis sentimen telah menjadi alat yang sangat penting dalam memahami opini publik di era digital. Analisis ini bertujuan untuk menilai polaritas opini, apakah bersifat positif, negatif, atau netral, terhadap suatu topik atau entitas tertentu[4]. Salah satu teknik yang umum digunakan dalam analisis sentimen adalah dengan menggunakan machine learning. Machine learning adalah sistem yang menggunakan data yang telah dipelajari untuk melakukan klasifikasi sentimen dengan berbagai algoritma[5]. Beberapa algoritma yang umum digunakan dalam analisis sentimen meliputi Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest.

Random Forest menjadi salah satu metode yang populer karena kemampuannya dalam menangani data berdimensi tinggi, mengurangi overfitting, dan memberikan hasil yang andal dibandingkan dengan algoritma lain[6]. Keunggulan Random Forest terletak pada kemampuannya untuk menggabungkan keputusan dari

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

beberapa decision tree, sehingga menghasilkan model yang lebih stabil dan efektif, bahkan ketika jumlah data yang tersedia terbatas. Hal ini menjadikan Random Forest sebagai pilihan yang efektif untuk klasifikasi sentimen, meskipun keterbatasan data sering kali menjadi masalah utama dalam penerapan pembelajaran mesin di dunia nyata[7]. Jumlah data yang terbatas dapat menyebabkan model rentan terhadap overfitting atau underfitting, yang dapat menurunkan akurasi prediksi. Sebagian besar penelitian sebelumnya lebih fokus pada penerapan pembelajaran mesin pada dataset besar, namun jarang yang membahas penerapan teknik seperti Random pada dataset kecil.

Beberapa penelitian terdahulu telah menerapkan penggunaan Random Forest dalam analisis sentimen dengan dataset terbatas. penelitian[8] mengembangkan klasifikasi analisis sentimen menggunakan algoritma Random Forest pada data Twitter yang terbatas, menunjukkan bahwa Random Forest mampu mengklasifikasikan sentimen dengan baik pada dataset kecil. Penelitian[9] menggunakan Random Forest dengan SMOTE untuk analisis sentimen Wattpad, dengan akurasi 84,05%, menghadapi tantangan ketidakseimbangan kelas dan overfitting pada dataset kecil. Penelitian[10] membandingkan efektivitas beberapa algoritma, termasuk Random Forest, dalam klasifikasi sentimen pada dataset yang terbatas, menemukan bahwa Random Forest dapat bersaing dengan algoritma lain meskipun dengan jumlah data yang sedikit dengan akurasi memiliki akurasi 90,77%, recall 90,77%. Selain itu penelitian oleh [11] juga mengusulkan penggunaan metode ensemble, termasuk Random Forest, untuk meningkatkan analisis sentimen pada dataset kecil, menunjukkan bahwa pendekatan ini dapat meningkatkan akurasi klasifikasi secara signifikan.

Beberapa penelitian terdahulu telah mengkaji penggunaan Random Forest dalam analisis sentimen dengan dataset terbatas. Meskipun penelitian-penelitian tersebut menunjukkan bahwa Random Forest mampu mengklasifikasikan sentimen dengan baik pada dataset kecil, tidak ada yang secara spesifik mengeksplorasi kombinasi antara Random Forest dan representasi kata menggunakan Word2Vec. Penelitian ini bertujuan untuk mengisi celah tersebut dengan mengeksplorasi efektivitas Random Forest dalam klasifikasi sentimen pada dataset kecil, dengan fitur Word2Vec sebagai representasi vektor kata untuk meningkatkan akurasi klasifikasi

Penelitian ini bertujuan untuk mengisi celah tersebut dengan mengeksplorasi efektivitas Random Forest dalam klasifikasi sentimen pada dataset kecil. Dalam penelitian ini, fitur Word2Vec digunakan sebagai representasi vektor kata untuk meningkatkan akurasi klasifikasi. Word2Vec memungkinkan pengubahan kata-kata dalam teks menjadi representasi numerik yang lebih bermakna, sehingga model dapat menangkap hubungan semantik antar kata dengan lebih baik dibandingkan dengan metode bag of words dan TF-IDF dan mampu dalam meingkatkan akurasi pada algoritma random forest[12]. Pada penelitian [13] membandingkan kinerja algoritma Logistic Regression, SVM, dan Random Forest dengan dan tanpa metode Feature Expansion Word2Vec. Hasilnya menunjukkan bahwa penggunaan Word2Vec pada Random Forest mampu meningkatkan akurasi sebesar 1,46%, mencapai total 89,53%, sehingga efektif untuk klasifikasi hoaks di Twitter.

Dari isu yang telah dijelaskan sebelumnya diatas penulis berencana melakukan Penelitian ini sejalan dengan tantangan yang diuraikan dalam shared task klasifikasi sentimen yang membahas penggunaan dataset terbatas, seperti yang dijelaskan dalam[14], yang menggunakan hanya 300 tweet sebagai data pelatihan. Fokus penelitian ini untuk mengeksplorasi efektivitas kombinasi metode Random Forest dan fitur Word2Vec dalam meningkatkan performa klasifikasi sentimen pada dataset kecil. Berbagai optimasi, seperti pengujian pada preprocessing teks dan penambahan data eksternal, juga dilakukan untuk meningkatkan akurasi model. Dengan menggunakan tweet-tweet terkait Kaesang Pangarep sebagai Ketua Umum PSI. Urgensi penelitian ini terletak pada minimnya kajian terkait penerapan Random Forest untuk analisis sentimen pada dataset kecil, meskipun metode ini terbukti efektif dalam menangani data berdimensi tinggi dan terbatas dengan harapan hasil penelitian ini dapat memberikan kontribusi signifikan dalam bidang analisis sentimen, khususnya dalam konteks pengukuran sentimen terhadap tokoh publik dengan keterbatasan data.

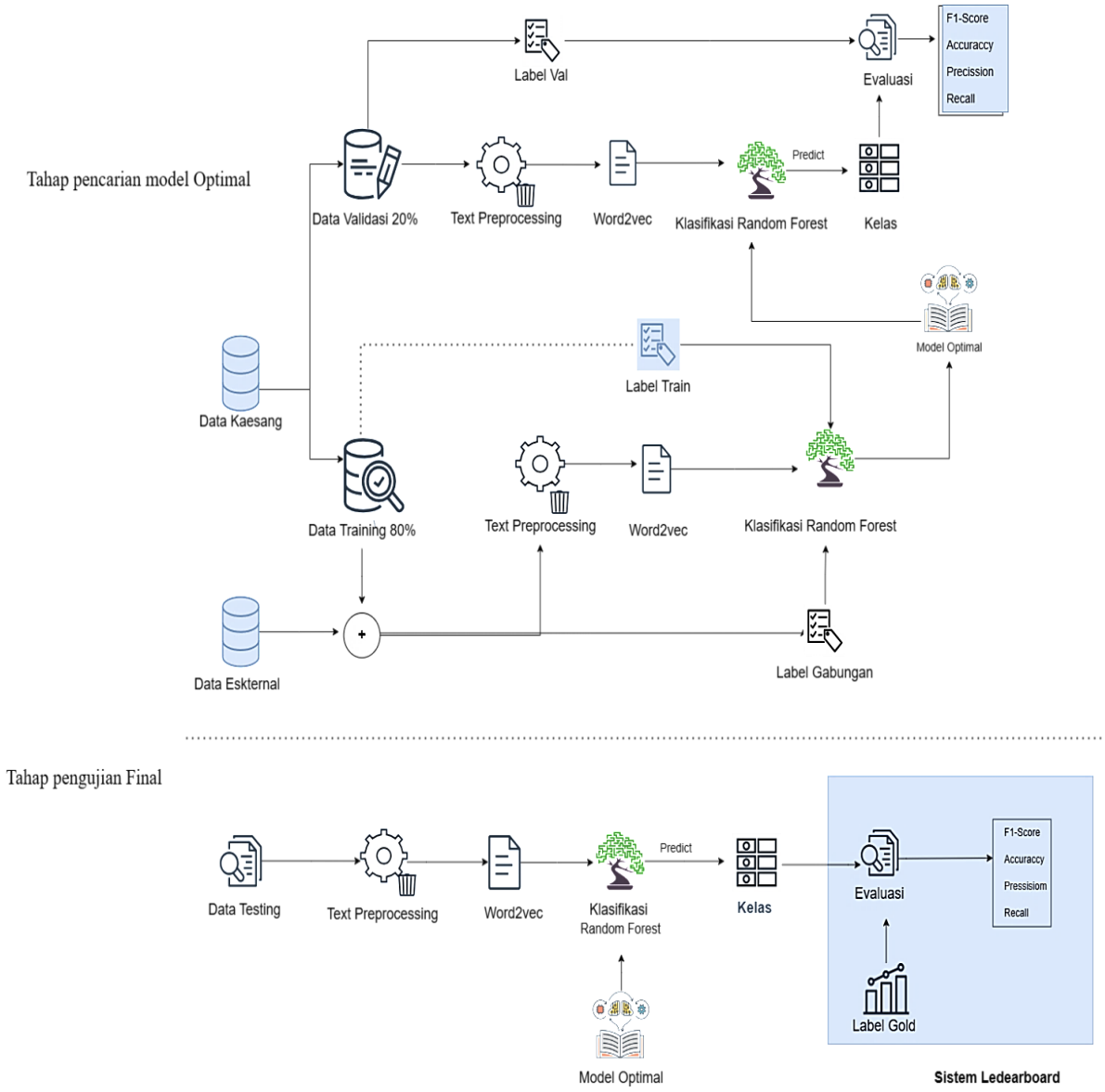
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini diawali dengan proses pengumpulan data terkait Kaesang serta data tambahan dari sumber eksternal. Data yang diperoleh kemudian dibagi menjadi dua bagian, yaitu data pelatihan (80%) dan data validasi (20%). Pra-pemrosesan teks dilakukan pada kedua set data, yang meliputi pembersihan, konversi teks menjadi huruf kecil, penghapusan kata-kata umum, dan tokenisasi. Setelah itu, teks yang telah diproses diubah menjadi bentuk numerik menggunakan metode vektorisasi (word2vec). Selanjutnya, model dilatih dengan algoritma Random Forest menggunakan data yang telah diubah ke dalam bentuk vektor Model dievaluasi dengan menggunakan data validasi untuk memprediksi output dan menilai kinerjanya berdasarkan metrik yang telah ditentukan, sekaligus melaksanakan analisis klasifikasi guna memperoleh skor terbaik. Berdasarkan evaluasi tersebut, model dioptimalkan dan disimpan untuk keperluan prediksi selanjutnya. Model yang sudah dioptimalkan ini digunakan untuk melakukan prediksi pada data uji, yang kemudian hasilnya dibandingkan dengan label sebenarnya (test label) untuk mengevaluasi kinerja model. Akhirnya, hasil prediksi akan disubmit ke sistem leaderboard, yang membandingkan performa model dengan model lain. Pada akhirnya, prediksi

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

dievaluasi secara keseluruhan untuk memastikan bahwa model memiliki kinerja yang baik dan dapat diterapkan pada data nyata. Proses ini dapat dilihat secara bertahap pada Gambar 1 di bawah ini.



Gambar 1. Metodologi Penelitian

2.2 Dataset

Penelitian ini mengumpulkan data terkait dengan pengangkatan Kaesang Pangarep sebagai Ketua Umum Partai Solidaritas Indonesia (PSI), yang menjadi studi kasus dalam penelitian ini. Data ini dikumpulkan pada periode 25 September 2023 hingga 3 Oktober 2023 dengan menggunakan kata kunci 'Kaesang PSI'. Dataset ini dibagi ke dalam tiga label sentimen: positif, negatif, dan netral, yang ditentukan berdasarkan metode majority vote. Selain itu, untuk memperkaya dan meningkatkan performa model dalam mengklasifikasikan sentimen pada dataset terbatas, penelitian ini juga menggunakan data eksternal yang melibatkan topik Covid-19 dan topik-topik umum lainnya. Penambahan data eksternal ini bertujuan untuk memberikan konteks yang lebih luas, namun fokus penelitian tetap pada Kaesang sebagai Ketua Umum PSI. Data eksternal digunakan untuk memperbaiki kinerja model dan tidak mengalihkan perhatian dari studi kasus yang menjadi objek penelitian.

Dataset pelatihan dalam penelitian ini terdiri dari dua kumpulan data, masing-masing berisi 300 tweet dengan label positif, negatif, dan netral. Dataset pelatihan pertama disebut Train Kaesang V1, sedangkan yang kedua dinamakan Train Kaesang V2. Untuk pengujian model, tersedia 924 tweet dengan label gold standard yang disimpan pada server leaderboard. Model terbaik yang dihasilkan dalam penelitian ini akan diuji menggunakan dataset pengujian tersebut, dan hasil prediksinya akan dikirimkan ke sistem leaderboard.Selain itu, data eksternal dengan beragam topik (Open Topic) sebanyak 7.569 tweet dan data Covid sebanyak 8.000 tweet dari penelitian sebelumnya [15][16] juga digunakan sebagai tambahan data pelatihan. Kedua dataset ini telah diberi label positif, negatif, dan netral. Rincian jumlah tweet pada setiap dataset dapat dilihat pada Tabel 1.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

Table 1. Rincian Jumlah Dataset

Data	Jumlah Tweet	Kelas		
		Positif	Netral	Negatif
Train Kaesang V1	300	100	100	100
Train Kaesang V2	300	100	100	100
Data Covid	8000	463	6664	873
Data Open Topic	7569	1505	3408	2656
Data Test Kaesang	924	-	-	-

Pada tabel 1 menunjukkan rincian jumlah dataset yang digunakan dan untuk data test sudah ada di sistem dileaderboard sentimen untuk hasil data uji.

2.3 Teks Preprocessing

Teks Preprocessing adalah langkah untuk membersihkan dan menstandarisasikan data sebelum melanjutkan proses selanjutnya. Beberapa tahap yang dilakukan sebagai berikut.

a. Cleaning : Langkah pembersihan dataset melibatkan menghapus username Twitter, mengganti username dengan token "user", menghilangkan tautan (URLs), menghapus karakter non-alphanumeric kecuali spasi, menghilangkan emoji dan karakter khusus, serta mengonversi teks menjadi huruf kecil.
b. Case Folding : mengubah karakter huruf besar menjadi huruf kecil, dan menghapus kata-kata yang tidak perlu untuk mengurangi noise[17].
c. Stopword Removal :Menghilangkan kata-kata umum yang kurang memberikan kontribusi informasi yang signifikan dan tidak memiliki makna khusus dalam analisis, seperti kata-kata anti, kata depan, dan kata penghubung[18].
d. Stemming : mengubah kata menjadi bentuk awalan[17].
e. Tokenization : Proses membagi teks menjadi unit yang disebut token[18].

2.4 Word2vec

Setelah melakukan pemrosesan teks, langkah berikutnya adalah mengekstrak fitur. Dalam penelitian ini, ekstraksi fitur dilakukan menggunakan Word2Vec. Word2Vec adalah metode yang digunakan untuk mengubah kata- kata menjadi vektor dalam dimensi N[19]. Salah satu parameter yang dimanfaatkan adalah jendela (window) dalam Word2Vec. Dalam penelitian ini model Word2Vec yang digunakan adalah model CBOW dan strategi algoritma seleksi 0 dan dimensi fitur sejumlah 500. model CBOW dipilih karena kemampuannya dalam memprediksi kata target dengan memanfaatkan konteks kata di sekitarnya dalam suatu kalimat[20]. Penulis memilih model ini karena efisiensinya dalam mempelajari representasi vektor kata dalam teks yang tidak terstruktur dengan volume yang besar[21]. Selain itu, CBOW efektif dalam menangani banyak variasi kata, serta tidak terpengaruh oleh urutan kata dalam suatu teks, sehingga mengurangi kemungkinan munculnya kata-kata yang jarang ditemui. Dengan algoritma continuous bag-of-words, urutan dari kalimat di dalam riwayat tidak mempengaruhi proyeks. Pada model word2vec, dimensi fitur sebesar 500 adalah ukuran yang memadai, tergantung pada hasil yang diinginkan. Dimensi fitur dapat divariasikan, mulai dari 200 hingga 600, sesuai dengan kebutuhan dan hasil yang diinginkan.

2.5 Klasifikasi Random Forest

Random Forest adalah teknik pembelajaran mesin yang digunakan untuk klasifikasi data. Pendekatan ini melibatkan pembuatan dan penggabungan beberapa pohon keputusan. Tujuan dari algoritma ini adalah untuk membangun pohon keputusan yang mencakup simpul akar yang mewakili berbagai kemungkinan hasil. Algoritma ini mengintegrasikan prediksi dari berbagai pohon keputusan untuk meningkatkan akurasi keseluruhan[22]. Pohon keputusan dibangun secara acak menggunakan data dan atribut sesuai dengan persyaratan yang ditetapkan. Penggunaan sampling terpandu dalam pembangunan pohon prediksi, di mana setiap pohon menggunakan prediktor secara acak, adalah salah satu elemen kunci dari Random Forest[23]. Teknik ini digunakan untuk kedua tujuan, yakni klasifikasi dan regresi. Penggunaan Random Forest bertujuan untuk mengurangi masalah overfitting yang biasa terjadi pada decision tree dan meningkatkan akurasi model. Random Forest memiliki beberapa keunggulan, termasuk kemampuannya untuk meningkatkan kinerja saat data terbatas, ketahanannya terhadap outlier, dan efektivitasnya saat digunakan pada dataset yang diperluas[24]. Selain itu, algoritma ini juga mampu menghitung tingkat kepentingan masing- masing fitur dalam data, yang berguna untuk mengidentifikasi fitur-fitur yang paling relevan. Random Forest sangat fleksibel dan dapat diaplikasikan pada berbagai jenis data, termasuk data numerik dan teks[25]. Dalam penelitian ini, sebuah model Random Forest dibangun dengan 200 pohon keputusan. Langkah ini membantu meningkatkan akurasi serta mengurangi risiko overfitting dengan menggabungkan hasil dari berbagai pohon. Dengan menerapkan pendekatan ini, kita dapat menghasilkan model klasifikasi teks yang handal, akurat, dan mampu menangani beragam jenis data teks yang kompleks, sehingga memberikan prediksi yang lebih stabil dan tepat.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

2.6 Evaluasi

Berbagai parameter digunakan untuk menilai performa klasifikasi dalam penelitian ini. Eveluasi dilakukan setelah mendapatkan kombinasi terbaik dari fitur-fitur dan parameter dari pencarian optimal. Evaluasi dilakukan menggunakan metrik Akurasi, Presisi, Recall, dan F1-Score[26]. Penilaian kinerja dilakukan berdasarkan hasil eksperimen. Precision, Recall, dan F1-Score dievaluasi dalam skala 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan performa yang lebih baik dan semakin mendekati angka 1. Berikut rumus untuk perihitungan akurasi, presisi, recall dan F1-Score.

$$F_1 = 2\ x\ \left(\frac{presisi\ x\ recall}{presisi+recall}\right) \quad (1)$$

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, hasil prediksi data uji akan disubmit ke dalam sistem leaderboard berbasis web yang telah disediakan oleh penyelenggara. Leaderboard ini berfungsi sebagai papan skor tempat hasil peringkat peserta ditampilkan berdasarkan nilai metrik evaluasi, seperti F1-score, Accuracy, Precision, dan Recall yang dihitung secara macro-average. Setelah submit, skor peserta akan langsung muncul di leaderboard. Sistem leaderboard terdiri atas dua komponen utama, yaitu organizer dan admin. Organizer adalah pihak penyelenggara yang menetapkan standar nilai F1-score untuk penelitian ini, sedangkan admin merupakan metode baseline. Metode baseline, dalam konteks penelitian klasifikasi teks, adalah model dasar yang digunakan sebagai titik acuan dalam mengevaluasi performa model peserta yang lebih kompleks. Sebanyak 924 data digunakan sebagai benchmark untuk pengujian prediksi peserta.

3.1 Dataset

Dalam penelitian ini, data tweet Kaesang yang diperoleh melalui proses crawling di Twitter telah melalui tahap seleksi. Jumlah total data yang berhasil dikumpulkan mencapai 1.524 tweet. meliputi Data Kaesang dimana Data yang digunakan diberi nama data Train Kaesang V1 dan Data Train Kaesang V2, Data Train Kaesang V1 berjumlah 300 tweet dengan komposisi 100 positif, 100 negatif dan 100 netral , dan data Tarain Kaesang V2 berjumlah 300 tweet dengan komposisi berjumlah 100 positif, 100 negatif dan 100 netral. Data ini kemudian dibagi menjadi dua bagian, yaitu 300 tweet sebagai data Train dan 924 tweet sebagai data Test. Pembagian ini dilakukan untuk keperluan pelatihan dan pengujian model, di mana data Test tidak pernah digunakan selama proses pelatihan. Untuk menentukan model yang paling optimal, data latih yang berjumlah 300 tweet akan dibagi kembali dengan komposisi 80% untuk pelatihan dan 20% untuk validasi. Contoh dataset yang digunakan dapat dilihat pada tabel 2 dibawah.

Table 2. Contoh Dataset

No	Data	Kelas
1	@psi_id @kaesangp Asli ini re-marketing @psi_id ke ibu- ibu dan wanita bagus banget... Lgsg salfok sama baju imutnya kaesang. Ini kena banget dan politik jd adeeemmmmm banget... Ngga ada kalimat kasar, vulgar,caci maki, dsb.	Positif
2	@bangherwin Banyak x cakap @adearmando61 basiii. Yang jelas selama ente di @psi_id ga bisa derek parpol itu naik. Harus ada Kaesang.	Netral
3	@abdulmukti691 Kaesang itu hanya boneka PSI untuk mendongkrak suara saja.	Negatif

3.2 Text Preprocessing

Tahapan teks preprocessing ini digunakan untuk membersihkan data. Untuk memastikan data berkualitas tinggi, ulasan yang identik atau duplikat akan dihapus. Langkah-langkah preprocessing meliputi tokenisasi, penghapusan simbol atau angka, case folding, dan penghapusan stopword. Tokenisasi memisahkan kalimat menjadi kata-kata individu, sedangkan case folding mengkonversi semua huruf besar menjadi huruf kecil[20]. Penghapusan stopword bertujuan untuk menghilangkan kata-kata yang tidak relevan atau penting dari dokumen. Tabel 3 menyajikan hasil dari proses pembersihan dan penggabungan fitur yang diperoleh setelah tahap pra-pemrosesan data.

Table 3. Text Preprocessing

No	Proses	Sebelum	Sesudah
1	Cleaning	@CNNIndonesia PSI setelah dimasukikaesang be lyke: https://t.co/ZGgpn7uFhs	Psi setelah dimasuki kaesang be lyke
2	Case Folding	Psi setelah dimasuki kaesang be lyke.	psi setelah dimasuki kaesang be lyke

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

No	Proses	Sebelum	Sesudah
3	Stopword Removal	psi setelah dimasuki kaesang be lyke.	psi dimasuki kaesang be lyke
4	Stemming	psi dimasuki kaesang be lyke	psi masuk kaesang be lyke
5	Tokenizing	psi masuk kaesang be lyke	[psi', 'masuk', 'kaesang',' be', 'lyke']

3.3Pencarian Model Optimal

Langkah selanjutnya adalah memeriksa teks hasil preprocessing melalui eksperimen untuk menemukan model Random Forest yang optimal dan akurat. Tujuan dari eksperimen ini adalah untuk menentukan model Random Forest yang memberikan kinerja terbaik. Proses ini melibatkan langkah-langkah preprocessing, yaitu Cleaning (Cl), Stopword Removal (SR), dan Stemming (ST). Pengujian text preprocessing dilakukan dengan menerapkan dan tidak menerapkan beberapa langkah preprocessing. Tanda "iya" menandakan bahwa langkah tersebut telah diterapkan, sedangkan tanda "tidak" menunjukkan langkah tersebut tidak diterapkan. Optimalisasi dataset dilakukan dengan menambahkan data Kaesang (Ks) serta data eksternal, yaitu data Covid (Cv) dan data dari Open Topik (OT), untuk memperluas data pelatihan. Model yang optimal ditunjukkan pada Tabel 4 di bawah ini.

Tabel 4. Pencarian Model Optimal

ID	Teks Preprocessing			Data Eksternal			F1-Score	Accuraccy
	Cl	SR	ST	KS	CV	OT		
P1	Iya	Tidak	Tidak	300	-	-	49.00%	50.00%
P2	Iya	Iya	Tidak	300	900	900	56.84%	58.33%
P3	Iya	Tidak	Iya	900	900	900	60.61%	61.67%
P4	Iya	Iya	Iya	900	2.400	2.400	63.33 %	61.65 %
P5	Iya	Iya	Iya	900	1.080	-	66.82%	66.67%

Berdasarkan Tabel 4, terlihat bahwa penambahan data eksternal berperan penting dalam peningkatan F1-Score. Nilai tertinggi tercatat pada ID P5, yang mencapai F1-Score sebesar 66.82% dan akurasi 66.67%. Pada konfigurasi ini, semua tahap preprocessing seperti pembersihan teks, penghapusan kata-kata umum (stopwords), dan stemming telah diterapkan, dengan tambahan data eksternal yang mencakup 900 data dari Kaesang dan 1.080 data dari Covid. Pada ID P4, model memperoleh F1-Score sebesar 63.33% dan akurasi 61.65%, dengan seluruh langkah preprocessing diterapkan. Konfigurasi ini melibatkan penambahan 900 data dari Kaesang, serta masing-masing 2.400 data dari Covid dan Open Topic. ID P3 menunjukkan hasil F1-Score sebesar 60.61% dan akurasi 61.67% dengan menerapkan pembersihan teks dan stemming. Data eksternal yang digunakan mencakup 900 data dari Kaesang serta masing-masing 900 data dari Covid dan Open Topic Model final yang diperoleh diterapkan pada data uji yang belum pernah digunakan sebelumnya dalam proses pelatihan. Penggunaan model ini melibatkan algoritma Random Forest untuk klasifikasi serta Word2Vec untuk memilih fitur terbaik dalam pembentukan model bahasa. Model optimal tersebut kemudian diterapkan pada data uji yang terdiri dari 924 data, dan hasilnya disubmit ke sistem leaderboard, sebuah platform berbasis web yang memungkinkan para peneliti mengunggah hasil prediksi mereka. Hasil prediksi label data uji berdasarkan eksperimen menggunakan tiga model terbaik juga dilaporkan, yang dapat dilihat pada Tabel 5.

3.4Pengujian Terhadap Data Uji

Model optimal diterapkan pada data uji yang belum pernah digunakan selama proses pelatihan klasifikasi menggunakan Random Forest . Prediksi kelas positif, negatif, dan netral pada data uji yang berisi 924 tweet tersebut ditampilkan dalam bentuk tabel 5 berikut.

Tabel 5. Hasil Pengujian terhadap data uji (score dari ledearbroad)

ID	Nama	Metode	F1-Score	Accuracy	Precision	Recall
P3	Run 1	Random Forest KS+CV+OT	47.83%	56.01%	49.08%	56.22%
P4	Run 2	Random Forest KS+CV+OT	49.09%	57.47%	49.13%	56.26%
P5	Run 3	Random Forest KS+CV	49.89%	58.29%	49.16%	56.74%

Berdasarkan pengujian pada Tabel 5 di atas, hasil pengujian menunjukkan kinerja tiga model Random Forest dengan fitur Word2Vec pada data uji berdasarkan metrik F1 Score, Accuracy (Acc), Precision (Prec), dan Recall. Model pertama, Random Forest dengan data Kaesang (Run 1), memperoleh F1-Score sebesar 47.83% dan akurasi 56.01%. Model kedua (Run 2), yang menggunakan data Kaesang, Covid, dan Open Topic, menunjukkan peningkatan dengan F1-Score sebesar 49.09% dan akurasi 57.47%. Model ketiga (Run 3) dengan data Kaesang dan Covid, mencapai F1-Score tertinggi sebesar 49.89% dan akurasi 58.29%. Hal ini mengindikasikan bahwa dengan penambahan data terkait Covid, model Random Forest menunjukkan peningkatan kinerja yang signifikan dibandingkan dengan model yang hanya menggunakan data Kaesang. Peringkat model terbaik (Run 3) menunjukkan bahwa kombinasi data Kaesang dan Covid memberikan hasil yang lebih baik dalam hal akurasi, presisi, dan F1 Score.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

3.5 Perbandingan Pengujian

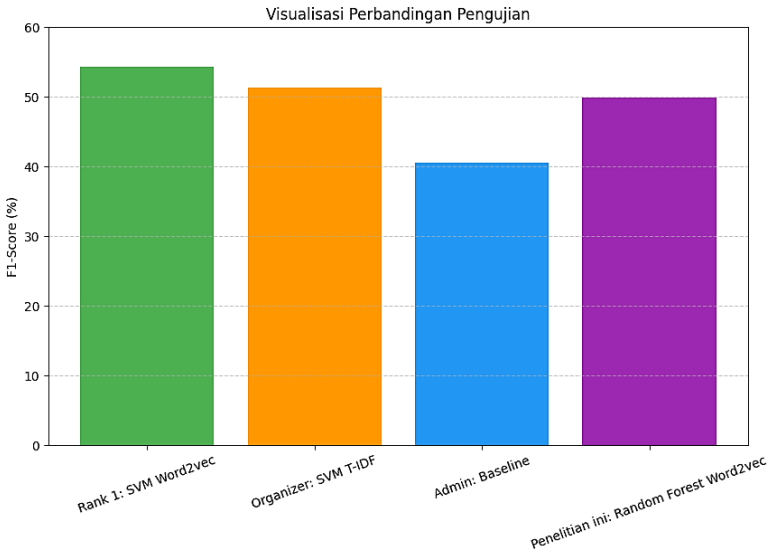
Berdasarkan Tabel 6, menunjukkan bahwa metode SVM dengan Word2Vec mencatatkan F1-Score tertinggi sebesar 54.23% dan akurasi 63.81%. Hal ini mengindikasikan kemampuan SVM dalam menangkap lebih banyak fitur relevan untuk analisis sentimen, memberikan keseimbangan antara presisi dan recall. Sebaliknya, Random Forest dengan Word2Vec mencatatkan F1-Score 49.89% dan akurasi 58.29%. Meskipun hasil ini lebih rendah dibandingkan SVM, Random Forest tetap menunjukkan performa yang lebih baik dibandingkan metode baseline (F1-Score 40.43%). Keunggulan Random Forest terletak pada kemampuannya menangani data dengan kompleksitas tinggi dan fitur beragam tanpa risiko overfitting yang besar. Namun, pada dataset kecil, algoritma ini cenderung kurang optimal dibandingkan SVM.

Hal ini kemungkinan disebabkan oleh sensitivitas Random Forest terhadap volume data yang terbatas, yang mengurangi kemampuan ensemble learning-nya. Sebaliknya, SVM lebih stabil dalam kondisi dataset kecil dan berdimensi tinggi, sehingga memberikan hasil yang lebih baik. Penggunaan Word2Vec sebagai representasi fitur juga berkontribusi pada peningkatan kinerja kedua algoritma dibandingkan baseline, tetapi efektivitasnya dapat terpengaruh oleh jumlah data pelatihan yang terbatas. Representasi numerik dari Word2Vec memungkinkan algoritma menangkap hubungan semantik antar kata, namun pada dataset kecil, fitur ini belum dapat dimanfaatkan secara maksimal oleh Random Forest

Tabel 6. Hasil Terbaik Dari Eksperimen

Nama	Metode	F1-Score	Accuracy	Precision	Recall
Rank 1	SVM Word2vec	54.23%	63.81%	54.72%	59.70%
Organizer	SVM T-IDF	51.28 %	61.21%	52.89%	59.95%
Admin	Baseline	40.43%	40.45%	49.53%	48.80%
Penelitian ini	Random Forest Word2vec	49.89%	58.29%	49.16%	56.74%

Hasil penelitian ini mengindikasikan bahwa, meskipun Random Forest dengan Word2Vec belum mampu melampaui SVM, algoritma ini tetap memiliki potensi besar untuk analisis sentimen. Peningkatan performa dapat dicapai melalui optimalisasi parameter seperti jumlah trees dan kedalaman maksimum, penambahan data pelatihan, atau integrasi metode lain seperti SVM. Dengan langkah-langkah tersebut, Random Forest dapat memberikan kinerja yang lebih kompetitif pada dataset kecil. Untuk pengembangan lebih lanjut, penelitian ini dapat memanfaatkan kombinasi metode, penambahan lebih banyak data pelatihan, atau eksplorasi lebih lanjut terhadap fitur-fitur yang dapat menggali sentimen lebih dalam. Berikut visualisasi perbandingan penguji dapat dilihat pada gambar 2 dibawah ini.



Gambar 2. Visualisasi Perbandingan Pengujian

Gambar 2 menampilkan visualisasi grafik dari seluruh pengujian dalam penelitian ini. Grafik ini memperlihatkan perbandingan kinerja antara beberapa tim yang berpartisipasi dalam tugas shared task ini. Hasil analisis menunjukkan bahwa dalam penelitian ini Metode SVM dengan Word2Vec memperoleh F1-Score tertinggi sebesar 54.23%, diikuti oleh SVM dengan T-IDF (51.28%) yang digunakan oleh organizer. Random Forest dengan Word2Vec yang digunakan dalam penelitian ini berada di posisi ketiga dengan F1-Score 49.89%, menunjukkan kinerja yang lebih baik dibandingkan dengan metode baseline (40.43%). Meskipun hasilnya belum optimal, penggunaan Random Forest dengan Word2Vec masih memberikan kontribusi positif dalam analisis sentimen, dan memperlihatkan potensi untuk perbaikan di masa mendatang.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

4. KESIMPULAN

Penelitian ini mengkaji klasifikasi sentimen pada dataset terbatas menggunakan algoritma Random Forest dengan representasi kata berbasis Word2Vec. Studi kasus ini menggunakan data tweet yang membahas Kaesang Pangarep sebagai Ketua Umum PSI, serta data eksternal untuk memperluas dataset. Hasil awal menunjukkan bahwa performa model masih kurang optimal, dengan F1-Score sebesar 49,00% dan akurasi 50,00%. Setelah dilakukan optimalisasi melalui penambahan data eksternal dan penerapan teknik preprocessing teks, performa model meningkat menjadi F1-Score 49,89% dan akurasi 58,29%.Meskipun Random Forest belum mampu melampaui performa SVM dalam klasifikasi sentimen, kombinasi Random Forest dan Word2Vec tetap menunjukkan potensi besar untuk diterapkan pada dataset kecil. Penelitian ini juga menegaskan bahwa penambahan data eksternal yang relevan dapat meningkatkan akurasi model. Untuk pengembangan lebih lanjut, disarankan untuk mengeksplorasi teknik augmentasi data, algoritma alternatif, atau kombinasi model lain untuk mendapatkan hasil yang lebih optimal. Penelitian ini memberikan kontribusi dalam pengembangan metode analisis sentimen dengan keterbatasan data, terutama dalam analisis opini publik terhadap tokoh dan isu tertentu di media sosial.

REFERENCES

[1] H. Naufal, M. F., Arifin, T., & Wirjawan, "Analisis Perbandingan Tingkat Performa Algoritma SVM , Random Forest , dan Naïve Bayes untuk Klasifikasi Cyberbullying pada Media Sosial," *J. Ris. Sist. Inf. Dan Tek. Inform.*, vol. 8, no. 1, pp. 82–90, 2023, doi: 10.30645/jurasik.v8i1.544.

[2] A. Wandani, "Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN , Random Forest , dan Naive Bayes," *J. Sains Komput. Inform. Vol.*, vol. 5, no. 2, pp. 651–665, 2021, doi: 10.30645/j-sakti.v5i2.365.

[3] A. Nasrudin Yahya, "Pro dan Kontra Kaesang Pangerep Jadi Ketum PSI," Kompas.com. [Online]. Available: https://nasional.kompas.com/read/2023/09/26/16000031/pro-dan-kontra-kaesang-pangarep-jadi-ketum-psi?page=all

[4] R. M. Nailar, "Sistem Deteksi Berita Hoax Menggunakan Algoritma Navie Bayes Dan Random Forest Pada Machine Learning," *Pondasi J. Appl. Sci. Eng.*, vol. 1, no. 2, pp. 43–57, 2024.

[5] D. A. Agustina, S. Subanti, E. Zukhronah, P. S. Statistika, and U. S. Maret, "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine," *Indones. J. Appl. Stat.*, vol. 3, no. 2, pp. 109–122, 2020, doi: https://doi.org/10.13057/ijas.v3i2.44337.

[6] F. W. Kurniawan and W. Maharani, "Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec," *e-Proceeding Eng.*, vol. 7, no. 2, pp. 7821–7829, 2020.

[7] R. Di, K., Tentang, Y., Afdhal, I., Kurniawan, R., Iskandar, I., & Salambue, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 122–130, 2022, doi: 10.32672/jnkti.v5i1.4004.

[8] D. I. A. Susanto.Aji, "Analisis Sentimen Data Twitter Topik Ekonomi Dan Industri Dengan Metode Naive Bayes Dan Random Forest," *J. Ilm. Wahana Pendidik.*, vol. 9, no. 20, pp. 59–65, 2023, doi: 10.5281/zenodo.8398895.

[9] & R. Adhan, S. N., Wibawa, G. N. A., Arisona, D. C., Yahya, I., Agusrawati, "Analisis sentimen ulasan aplikasi wattpad di google play store dengan metode random forest," *AnoaTIK J. Teknol. Inf. dan Komput.*, vol. 2, no. 1, pp. 6–15, 2024, doi: 10.33772/anoatik.v2i1.32.

[10] S. K. Delimasari, "Komparasi Algoritma Machine Learning Untuk Menganalisis Sentimen Ulasan Pada Aplikasi Digital Korlantas Polri," *G-Tech J. Teknol. Terap.*, vol. 8, no. 4, pp. 2411–2419, 2024, doi: 10.70609/gtech.v8i4.5089.

[11] N. B. S. N. R. R. N. S. Fatonah, "PENGGUNAAN METODE SVM DAN RANDOM FOREST UNTUK ANALISIS SENTIMEN ULASAN PENGGUNA TERHADAP KAI ACCESS DI GOOGLE PLAYSTORE," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 3, pp. 1901–1906, 2023, doi: 10.36040/jati.v7i3.6899.

[12] I. R. Hendrawan, E. Utami, and A. D. Hartanto, "Analisis Perbandingan Metode Tf-Idf dan Word2vec pada Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal di Indonesia," *Smart Comp*, vol. 11, no. 3, pp. 497–503, 2022, doi: 10.30591/smartcomp.v11i3.3902.

[13] F. Ismayanti and E. B. Setiawan, "Deteksi Konten Hoax Berbahasa Indonesia di Twitter Menggunakan Fitur Ekspansi dengan Word2Vec," *eProceedings Eng.*, vol. 8, no. 5, pp. 10288–10300, 2021.

[14] S. Agustian, M. I. Syah, and R. Abdillah, "Arah Baru Penelitian Klasifikasi Teks: Memaksimalkan Kinerja Klasifikasi Sentimen dari Data Terbatas," *MALCOM (Indonesia J. Mach. Learn. Comput. Sci.*, vol. 4, no. 8, pp. 1–10, 2024.

[15] M. Sahbuddin and S. Agustian, "Support Vector Machine Method with Word2vec for Covid-19 Vaccine Sentiment Classification on Twitter," *J. Informatics Telecommun. Eng.*, vol. 6, no. 1, pp. 288–297, 2022, doi: 10.31289/jite.v6i1.7534.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 6, No. 1, November 2024, Page 214-222
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v6i1.6246

[16] M. Ihsan *et al.*, "LSTM (Long Short Term Memory) for Sentiment COVID-19 Vaccine Classification on Twitter 1,2,3," *Digit. Zo.*, vol. 13, no. 1, pp. 79–89, 2022, doi: 10.31849/digitalzone.v13i1.9950.
[17] S. Agustian and A. Nazir, "Klasifikasi Sentimen Terhadap Pengangkatan Kaesang Sebagai Ketua Umum Partai PSI Menggunakan Metode Support Vector Machine," *Build. Informatics, Technol. Sci.*, vol. 6, no. 1, pp. 216–225, 2024, doi: 10.47065/bits.v6i1.5340.
[18] M. Dimas Lutfiyanto, E. B. Setiawan, and S. Si, "Expansion Feature dengan Word2Vec untuk Analisis Sentimen pada Opini Politik di Twitter dengan Klasifikasi Support Vector Machine, Naïve Bayes, dan Random Forest," *eProceedings Eng.*, vol. 8, no. 5, pp. 10399–10410, 2021.
[19] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," *J. Media Inform. Budidarma*, vol. 5, no. 3, pp. 1018–1026, 2021, doi: 10.30865/mib.v5i3.3111.
[20] Y. A. Pradana, I. Cholissodin, and D. Kurnianingtyas, "Analisis Sentimen Pemindahan Ibu Kota Indonesia pada Media Sosial Twitter menggunakan Metode LSTM dan Word2Vec," *J. Pengemb. Teknol. dan Ilmu Komput.*, vol. 7, no. 5, pp. 2389–2397, 2023.
[21] T. A. A. D. Ananey-obiri, "Word2vec neural model-based technique to generate protein vectors for combating COVID-19 : a machine learning approach," *Int. J. Inf. Technol.*, vol. 14, no. 7, pp. 3291–3299, 2022, doi: 10.1007/s41870-022-00949-2.
[22] M. S. Efendi and A. K. Zyen, "Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk," *RESOLUSI Rekayasa Tek. Inform. dan Inf.*, vol. 5, no. 1, pp. 12–20, 2024, doi: 10.30865/resolusi.v5i1.2149.
[23] R. F. Amir and I. A. Sobari, "Penerapan PSO Over Sampling Dan Adaboost Random Forest Untuk Memprediksi Cacat Software," *Indones. J. Softw. Eng.*, vol. 6, no. 2, pp. 230–239, 2020, doi: 10.31294/ijse.v6i2.9258.
[24] M. R. Adrian and M. P. Putra, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," *J. Inform. UPGRIS*, vol. 7, no. 1, pp. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099.
[25] E. Elgeldawi, A. Sayed, and A. R. Galal, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, doi: 10.3390/informatics8040079.
[26] H. M. Falah, M. R. Jamil, A. Taufik, and M. Botha, "Analysis Sentiment Terhadap Ginjal Akut pada Twitter Menggunakan Algoritma Random Forest," *Jurnla Ilmu Komput. dan Inform.*, vol. 3, no. 2, pp. 99–106, 2023, doi: 10.54082/jiki.65.