KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 6, Juni 2024, Hal 2902-2908
DOI 10.30865/klik.v4i6.1944
https://djournals.com/klik

Klasifikasi Sentimen SVM Dengan Dataset yang Kecil Pada Kasus Kaesang Sebagai Ketua Umum PSI

Yoga El Saputra, Surya Agustian*, Yusra, Siti Ramadhani

Sains Dan Teknologi, Teknik Informatika, Universitas Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12050116789@students.uin-suska.ac.id, [2]surya.agustian@uin-suska.ac.id, [3]yusra@uin-suska.ac.id, [4]siti.rahmadani@uin-suska.ac.id
Email Penulis Korespondensi: surya.agustian@uin-suska.ac.id

Abstrak−Media sosial telah menjadi platform utama bagi masyarakat untuk menyampaikan pandangan dan opini terhadap berbagai peristiwa, termasuk pengangkatan Kaesang Pangarep sebagai Ketua Umum Partai Solidaritas Indonesia (PSI). Penelitian ini bertujuan untuk mengklasifikasikan sentimen masyarakat terhadap pengangkatan tersebut menggunakan metode Support Vector Machine (SVM) dengan pendekatan Term Frequency-Inverse Document Frequency (TF-IDF). Data dikumpulkan dari Twitter menggunakan kata kunci "Kaesang PSI" serta data eksternal dengan topik terkait Covid-19. Pada data kaeasang diambil 300 data dengan masing-masing label (positif, netral, negatif) mendapatkan 100 tweet serta menambahkan data eksternal sebanyak 900 data dengan masing label (positif, netral, negatif) mendapatkan 300 tweet. Setelah proses text preprocessing yang meliputi case folding, stopword removal, dan stemming. Model diuji menggunakan confusion matrix untuk mengevaluasi performa berdasarkan metrik akurasi, presisi, recall, dan F1 Score. Hasil menunjukkan bahwa model SVM dengan TF-IDF memiliki F1 Score sebesar 0.53, akurasi 0.62, presisi 0.52, dan recall 0.57. Penambahan data ekstenal terkait Covid-19 pada fitur TF-IDF terbukti meningkatkan kinerja model secara signifikan. Kesimpulannya, metode SVM dengan TF-IDF efektif digunakan dalam analisis sentimen di media sosial meskipun dengan dataset yang kecil.

Kata Kunci: SVM; TF-IDF; Klasifikasi Sentiment; Machine learning; PSI

Abstract−Social media has become the main platform for the public to express views and opinions on various events, including the appointment of Kaesang Pangarep as General Chair of the Indonesian Solidarity Party (PSI). This research aims to classify public sentiment towards the appointment using the Support Vector Machine (SVM) method with the Term Frequency-Inverse Document Frequency (TF-IDF) approach. Data was collected from Twitter using the keyword "Kaesang PSI" as well as external data on topics related to Covid-19. In the kaeasang data, 300 data were taken with each label (positive, neutral, negative) to get 100 tweets and added external data of 900 data with each label (positive, neutral, negative) to get 300 tweets. After the text preprocessing process which includes case folding, stopword removal, and stemming. The model was tested using a confusion matrix to evaluate performance based on accuracy, precision, recall and F1 Score metrics. The results show that the SVM model with TF-IDF has an F1 Score of 0.53, accuracy of 0.62, precision of 0.52, and recall of 0.57. Adding external data related to Covid-19 to the TF-IDF feature has been proven to significantly improve model performance. In conclusion, the SVM method with TF-IDF is effective in analyzing sentiment on social media even with small datasets.

Keywords: SVM; TF-IDF; Klasifikasi Sentiment; Machine learning; PSI

1. PENDAHULUAN

Dalam era digital saat ini, platform media sosial telah menjadi saluran utama bagi masyarakat untuk berbagi pandangan, opini, dan sentimen mereka terhadap berbagai peristiwa dan tokoh publik[1]. Peningkatan kemajuan teknologi informasi dan komunikasi telah memungkinkan masyarakat untuk secara aktif berpartisipasi dalam diskusi dan pertukaran pendapat melalui platform daring. Twitter atau X sebagai salah satu platform media sosial terkemuka, menjadi wadah utama di mana pengguna dapat mengekspresikan pendapat mereka secara langsung dan terbuka. Dalam konteks ini, analisis sentimen di media sosial telah menjadi subjek penting dalam memahami opini dan persepsi masyarakat terhadap berbagai topik.Salah satu topik yang trending adalah pengangkatan kaesang sebagai ketua umum PSI. Pengangkatannya dilakukan dalam acara Kopi Darat Nasional (Kopdarnas) PSI di Djakarta Theater, Jakarta Pusat. Kaesang bergabung dengan PSI hanya dua hari sebelumnya, pada 23 September 2023[2]. Sebagai putra presiden, pengangkatan Kaesang bisa dianggap sebagai bentuk nepotisme. Kritikus mungkin menilai bahwa keputusan ini lebih didorong oleh hubungan keluarga daripada kapasitas dan kualifikasi politik. Ini dapat mempengaruhi citra PSI sebagai partai yang memperjuangkan transparansi dan meritokrasi[3]. Penunjukan Kaesang dapat dianggap sebagai langkah untuk memperbarui kepemimpinan di PSI dengan menghadirkan sosok muda dan dinamis. Ini sesuai dengan citra PSI sebagai partai yang mendukung kaum muda dan inovasi, yang dapat menarik pemilih dari kalangan milenial dan Gen Z[4]

Salah satu komponen penting dalam analisis data dan kecerdasan buatan adalah klasifikasi atau yang lebih dikenal dengan analisis sentimen. Klasifikasi sentimen adalah metode untuk menguji pendapat individu atau kelompok tentang topik, produk, layanan, atau kelompok tertentu [5]. Machine learning adalah sistem yang memungkinkan manusia belajar mengambil keputusan sendiri tanpa pemrograman berulang-ulang, dan memungkinkan komputer menjadi lebih pintar dengan belajar dari data yang dimilikinya [6]. Machine learning mempunyai metode seperti Support Vector Machine (SVM), Regresi Linear (Linear Regression), Regresi Logistik (Logistic Regression), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree, Random Forest, K-Means Clustering, Neural Networks dan Deep Learning, Gradient Boosting Machines (GBM).Algoritme pembelajaran mesin telah diterapkan untuk melatih dan mengenali fitur-fitur utama, serta untuk mengklasifikasikan kelompok. Metode *machine learning* memiliki kemampuan mengidentifikasi pola yang sulit dilihat dalam kumpulan data yang besar, kacau, atau kompleks. Kemampuan ini sangat bermanfaat ketika diterapkan pada data genom yang rumit[7].

Metode SVM lebih akurat sebagai metode pengelompokkan untuk proses analisis sentimen opini masyarakat berbahasa Indonesia pada Twitter dibandingan Naïve Bayes dengan akurasi SVM 83% sedangkan Naïve Bayes 74,6% dari 5000 tweet[8]. Hasil penelitian menunjukkan bahwa algoritma SVM mempunyai kinerja yang lebih baik dengan akurasi 73% dibandingkan algoritma KNN yang akurasinya 60% [9]. Minimnya jumlah data training yang tersedia adalah masalah yang sering muncul dalam implementasi nyata[10]. Padahal biasanya hanya ada data latih dalam jumlah terbatas, pengumpulan dan pelabelan data dalam jumlah besar secara manual membutuhkan banyak sumber daya. SVM juga merupakan salah satu metode yang tepat untuk digunakan dalam pemecahan masalah berdimensi tinggi namun jumlah sampel data terbatas dan SVM mudah untuk diimplementasikan pada data yang telah memiliki library[11].Menggunakan fitur TF-IDF pada SVM memiliki akurasi yang lebih baik dibandingkan dengan fitur ekstraksi yang lain[12]. Jika data latih tidak mencukupi, performa SVM biasanya buruk. Ketika data training terbatas, SVM mungkin kesulitan menemukan batasan keputusan yang optimal, yang mengakibatkan model kurang mampu menggeneralisasi pola[13]. Hal ini dapat menyebabkan model rentan terhadap overfitting atau underfitting, yang berdampak pada keakuratannya dalam mengklasifikasikan data baru. Jumlah data training yang sedikit sangat mempengaruhi kinerja model karena dapat menyebabkan overfitting, di mana model terlalu menyesuaikan diri dengan data pelatihan dan gagal melakukan generalisasi pada data baru. Data yang sedikit juga tidak mampu merepresentasikan distribusi data secara keseluruhan, mengakibatkan kesalahan generalisasi yang lebih tinggi dan performa yang buruk pada data yang belum pernah dilihat. Selain itu, kekurangan data membuat sulit untuk melakukan validasi dan pengujian yang akurat, dan biasanya kurang bervariasi, sehingga model tidak terpapar pada berbagai skenario nyata. Untuk mengatasi masalah ini, teknik seperti augmentasi data, transfer learning, dan cross-validation sering digunakan.

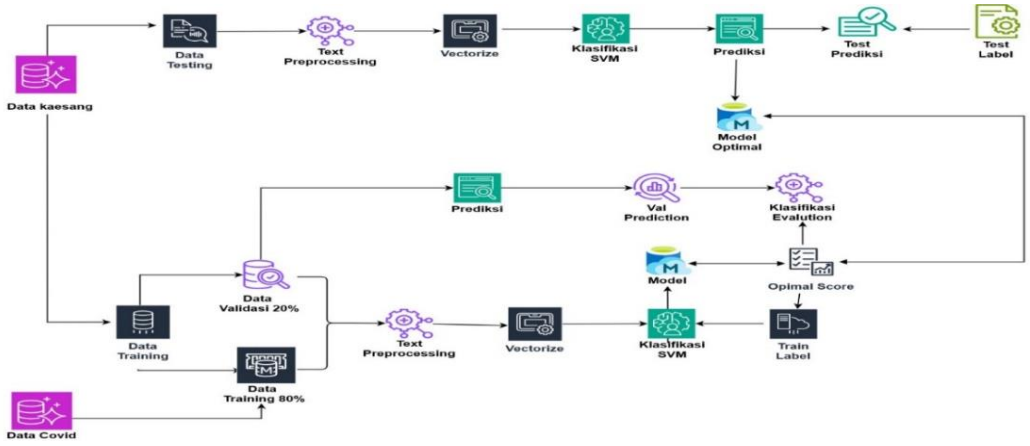
Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka pada penelitian ini akan dilakukan penelitian Peningkatan Performa Klasifikasi Dari SVM Pada Kasus Data Training Yang Kecil (Studi Kasus Sentimen Masyarakat Terhadap Pengangkatan Kaesang Sebagai Ketum PSI Di Twitter). Dengan tujuan memberikan kontribusi penelitian melalui penerapan algoritma machine learning pada SVM dengan hasil penelitian berupa tingkat akurasi dan konvusion matriks pada analisisi SVM.

2. METODOLOGI PENELITIAN

Dalam penelitian ini, metodologi yang diterapkan mencakup beberapa tahapan yang terstruktur untuk memastikan keakuratan dan keandalan hasil yang diperoleh. Tahapan-tahapan ini meliputi tahapan penelitian, dataset, text preprocessing , penerapan metode TF-IDF, penggunaan algoritma SVM, dan pengujian model. Setiap tahapan dirancang dengan hati-hati untuk menangani tantangan yang terkait dengan analisis sentimen pada dataset yang kecil dan memastikan bahwa model yang dikembangkan dapat memberikan performa yang optimal.

2.1 Tahapan Penelitian

Penelitian ini dimulai dengan pengumpulan data dari dua sumber utama: Data Kaesang dan Data Covid, yang kemudian dibagi menjadi data training (80%) dan data validasi (20%). Selanjutnya, dilakukan preprocessing teks pada kedua set data ini, yang mencakup pembersihan, normalisasi, tokenisasi, dan penghapusan stopwords. Setelah itu, teks yang telah diproses diubah menjadi bentuk numerik melalui metode vectorisasi. Model kemudian dilatih menggunakan Klasifikasi SVM (Support Vector Machine) pada data yang sudah di-vectorisasi. Evaluasi model dilakukan menggunakan Data Validasi untuk memprediksi hasil dan mengukur performa model dengan metrik tertentu, serta melakukan evaluasi klasifikasi untuk mendapatkan *Optimal Score*. Berdasarkan evaluasi ini, model dioptimalkan dan disimpan untuk prediksi lebih lanjut. Model optimal tersebut digunakan untuk memprediksi hasil pada data testing, yang kemudian dibandingkan dengan label sebenarnya (*test label*) untuk mengevaluasi performa model. Hasil akhirnya adalah prediksi yang dievaluasi secara keseluruhan untuk memastikan model memiliki performa yang baik dan dapat diaplikasikan pada data nyata. Berikut adalah gambar 1 dari metodologi dari penetian ini :



Gambar 1 Metodologi Penelitian

2.2 Dataset

Pada awalnya, tim data yang bertindak sebagai pihak primer melaksanakan proses pengumpulan data melalui teknik crawling. Mereka berhasil mengumpulkan sebanyak 2033 tweet dengan menggunakan kata kunci "Kaesang PSI" selama periode pengambilan data dari tanggal 25 September 2022 hingga 03 Oktober 2022. Selanjutnya, data yang telah terkumpul diberikan label positif, netral, atau negatif melalui metode crowd sourcing. Setiap tweet minimal dilabel oleh tiga orang anotator untuk memastikan konsistensi dan keandalan label. Label akhir kemudian ditentukan berdasarkan mayoritas suara dari anotator-anotator tersebut. Dari total 2033 tweet yang terkumpul hanya 1.674 data yang digunakan.

2.3 Text Processing

Salah satu tujuan utama text preprosesing adalah untuk membakukan bentuk kata indeks. Di sini, indeks adalah gambaran isi dokumen yang dapat dicari. Karena banyak komputer yang kesulitan membedakan huruf besar dan huruf kecil, tanda baca, dan fitur linguistik lainnya, Pemrosesan Teks menjadi sangat penting [14].

a. Case Folding

Case folding adalah mengubah karakter huruf besar menjadi huruf kecil, dan menghapus kata-kata yang tidak perlu untuk mengurangi noise[15].

b. Word Normalization

Dengan menggunakan aturan yang ditetapkan oleh Kamus Besar Bahasa Indonesia (KBBI), normalisasi kata mengubah kata-kata yang tidak beraturan menjadi kata-kata yang sesuai dengan standar. Dalam penelitian ini, normalisasi kata digunakan untuk mengubah bahasa gaul, akronim, dan kata-kata yang salah eja menjadi bahasa yang lebih konvensional dan digunakan sehari-hari. Karena ada begitu banyak istilah non-standar dalam tweet dan komentar Instagram yang ditulis dalam bahasa Indonesia, normalisasi kata merupakan langkah penting sebelum algoritma dapat mulai mengklasifikasikan konten [16].

c. Stopword Removal

Kata atau konjungsi yang tidak relevan, seperti tetapi, dengan, untuk, yang, dan sejenisnya, akan ditiadakan [17].

d. Stemming

Proses stemming kuncinya adalah menghilangkan imbuhan di awal terlebih dahulu, lalu menghilangkan sufiksnya [18]. Proses menghilangkan prefiks atau sufiks pada kata yang mengandung konjungsi, preposisi, dan kata ganti sehingga menjadi kata dasar sesuai KBBI[15].

2.4 TF-IDF

Untuk menentukan signifikansi dari setiap kata yang telah diambil, digunakan teknik yang dikenal dengan perhitungan Term Frequency-Inverse Document Frequency (TF-IDF)[19]. Penggunaan strategi ini sering dilakukan dalam proses pencarian informasi untuk menghitung istilah yang sering muncul. Salah satu model pembelajaran mesin yang paling efektif adalah support vector machine (SVM) yang memanfaatkan TF-IDF[20]. Term Frequency (tf(w,d)) dianggap memiliki proporsi kepentingan sesuai dengan total kemunculannya dalam teks atau dokumen. Semakin sering sebuah kata muncul dalam teks, semakin tinggi nilai Term Frequency-nya, yang menunjukkan bahwa kata tersebut memiliki relevansi yang lebih besar dalam konteks dokumen tersebut. Sebaliknya, Inverse Document Frequency (IDF) adalah metode pembobotan token yang berfungsi untuk memonitor kemunculan token dalam himpunan teks. IDF mengukur seberapa sering sebuah kata muncul di berbagai dokumen dalam kumpulan teks. Semakin jarang sebuah kata muncul dalam keseluruhan dokumen, semakin tinggi nilai IDF-nya. Ini berarti bahwa kata tersebut memiliki kekhususan yang lebih besar dan mungkin memberikan informasi yang lebih signifikan. Dengan demikian, penggabungan TF dan IDF dalam metode TF-IDF membantu mengidentifikasi kata-kata yang penting dan informatif dalam analisis teks, sehingga meningkatkan keakuratan dan efektivitas proses text mining.

2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah pengklasifikasi pembelajaran mesin yang diawasi yang umum digunakan dan terkenal dengan keefektifannya dalam tugas-tugas seperti klasifikasi gambar, pengenalan pola, dan diagnosis penyakit. Namun demikian, kompleksitas komputasi yang signifikan menimbulkan tantangan yang signifikan untuk masalah klasifikasi dalam skala yang luas[21]. Pada tahun 1990-an, Vapnik dan timnya menemukan teori SVM. Ide di balik SVM berasal dari jaringan saraf, atau bisa dikatakan bahwa SVM adalah perluasan matematis dari jaringan saraf. SVM dapat menyortir data linier dan nonlinier. Hal ini dilakukan dengan mengubah data pelatihan menjadi ruang multidimensi dan membuat hyper-plane dalam dimensi yang lebih tinggi. SVM adalah cara terbaik untuk belajar matematika menggunakan hyper-planes[22].

2.6 Pengujian

Pada tahap ini, model SVM akan diuji dengan confusion matrix setelah memperoleh kombinasi fitur dan parameter terbaik dari setiap pencarian optimal. Confusion matrix memiliki peran penting dalam menampilkan hasil klasifikasi dengan memberikan representasi visual dari data yang diklasifikasikan dengan benar atau salah. Meskipun confusion matrix

umumnya digunakan untuk menghitung akurasi, presisi, dan recall, dalam penelitian ini hanya fungsi akurasi yang digunakan.

3. HASIL DAN PEMBAHASAN

3.1 Pengolahan Dataset

Proses selanjutnya adalah pengelompokan data menjadi data training dan data testing. Dari 1.674 data hanya 300 data yang diambil sebagai data training. Pembagian ini dilakukan dengan masing-masing label (positif, netral, negatif) mendapatkan 100 tweet. Meskipun jumlah data training terbilang sedikit, hal ini mungkin dipilih dengan pertimbangan tertentu, seperti keterbatasan sumber daya atau untuk menghindari overfitting pada model yang akan dibuat. Pada penelitian ini juga menggunaka dataset eksternal yaitu dataset sentimet covid-19 sebanyak 900 data dengan masing label (positif, netral, negatif) mendapatkan 300 tweet. Sisa tweet, yang tidak digunakan pada twet kaesang sebagai data training, dialokasikan sebagai data testing. Proses ini bertujuan untuk memastikan bahwa model yang akan dikembangkan memiliki akurasi dan generalisasi yang baik saat diuji pada data yang belum pernah dilihat sebelumnya.

Tabel 1. Data Kaesang

No	Kalimat	Label
1	@HusinShihab @kaesangp @psi_id Masuknya kaesang akan membawa anak muda milenial (gen Y &amp Z) ikut pro aktif dlm kancah politik nasional dan saya yakin psi bakal mendulang suara hanya utk DPR sementara pilpres tetap om rambut putih ðŸ˜€ðŸ˜€	Positif
2	@republikaonline Kasihan sang Kaesang hanya di jadikan alat .. Seolah olah dengan kehadiran dia PSI bisa tambah lebih maju. Padahal hanya ingin sekedar MENJILAT sang presiden yg notabene nya kaesang itu anak nya sendiri. Cara cepat naikan elektoral parpol nya. Mumpung masih 10 bulan lagiðŸ¤£	Negatif
3	Data_Train_Kaesang_FINAL@AliAsro60643529 @saidiman @psi_id @RajaJuliAntoni @Andy_Budiman_ @grace_nat banyak org yg spt kaesang. Tp karena beliau bukan anak walikota, anak gubernur dan anak presiden maka gak ada istimewanya. Mensholati istri mantan presiden atau cium tangan sama mantan gubernur menurut kacamata saya biasa2. Kalaulah dia membina umkm dgn jaringannya baru hebat.	Netral
4	@Metro_TV @psi_id Menurut saya partai PSI krisis kepemimpinan makanya utk mendongkrak suara PSI dipasang mas Kaesang utk mendongkrak suara, harapan kita sih kedepan PSI menjadi partai anak muda yg idealisme nya tinggi sesuai visi dan misi perjuangan anak2 muda ketika zaman kemerdekaan.	Positif
5	@Dennysiregar7 @jokowi Alaahhh loe mah bacod ajah.. Privilege anak Presiden y mmg melekat di diri mrk Melek sikiitt!! Mana ada anak Presiden yg ga mndapatkan privilege baik di minta maupun tidak! Kaesang jg KeTum @psi_id jg kan itu hak-nya para pengurus..Ngapain jg elu yg bkn pengurus jd refott?@Dennysiregar7 @jokowi Alaahhh loe mah bacod ajah.. Privilege anak Presiden y mmg melekat di diri mrk Melek sikiitt!! Mana ada anak Presiden yg ga mndapatkan privilege baik di minta maupun tidak! Kaesang jg KeTum @psi_id jg kan itu hak-nya para pengurus..Ngapain jg elu yg bkn pengurus jd refott	Negatif
6	@Dennysiregar7 @jokowi Konteknya berbeda, pemilihan Kpl daerah langsung oleh rakyat dan yang menentukan calonnya adalah ketum partai.. Kaesang jd ketum PSI pun mrpkn pilihan partainya dan tentunya kl ada irisan ttg latar blkng keluarga itu hal yg wajar Sah-sah sj.Tapi mengingatkan adlh hal yg bagus..	Netral

Pada tabel 1 menggambarkan data tweet yang telah dikategorikan berdasarkan sentimen untuk digunakan dalam penelitian klasifikasi sentimen. Sentimen positif menunjukkan pandangan mendukung atau optimis terhadap Kaesang sebagai Ketua Umum PSI, sentimen negatif menunjukkan pandangan kritis atau tidak mendukung, dan sentimen netral menunjukkan pandangan yang netral atau tidak condong ke positif maupun negatif.

Tabel 2. Data Covid

No	Kalimat	Label
1	Banyak yang belum paham bahwa vaksin tidak mencegah terjadinya penularan Covid-19, tapi vaksin menurunkan risiko seâ€¦ https://t.co/JCteJHXON8 Banyak yang belum paham bahwa vaksin tidak mencegah terjadinya penularan Covid-19, tapi vaksin menurunkan risiko seâ€¦ https://t.co/JCteJHXON8	Positif
2	alasan subsidi kampus dipotong untuk bansos covid dan pembelian vaksin covid, lha ngentot bansos covid aja dikorupsâ€¦ https://t.co/ldqRw0Oz80	Negatif
3	bingung dapet vaksin gratis, like emang gue siapaaapas dikasih tau karena kerja di industri pendidikan oh iya lupa ehe .-.	Netral
4	Alhamdulillah selesai 2nd dose vaksin. Semoga dipermudahkan segala urusan olehNya ðŸ¤²ðŸ_x008f_¼â˜ï¸_x008f_ðŸŽ‰ #sayanakbalikraya @ Sunwayâ€¦ https://t.co/gPIdv6FI5L	Positif

5	Apakah para nakes ga sadar juga melihat diskriminasi kpd sesama manusia gegara vaksin.....??Apa mereka akan membeâ€¦ https://t.co/dHDaVUAI8H	Negatif
6	Giat Pengawalan Vaksin Covid-19 Dari Dinkes Provinsi Sumatera Selatan ke Dinkes Kota Pagaralam Oleh Personil Polresâ€¦ https://t.co/RrzfKkufuL	Netral

Pada tabel 2 menggambarkan data tweet yang telah dikategorikan berdasarkan sentimen untuk digunakan dalam penelitian klasifikasi sentimen. Sentimen positif menunjukkan pandangan optimis atau mendukung terkait vaksin Covid-19, sentimen negatif menunjukkan pandangan kritis atau tidak mendukung, dan sentimen netral menunjukkan pandangan yang tidak condong ke positif maupun negatif.

3.2 Text Preprocessing

Pada tahap ini, dilakukan pembersihan teks pada dataset tweet agar siap untuk proses berikutnya. Proses ini mencakup beberapa langkah preprocessing teks seperti menghapus URL, mengurangi spasi berlebihan, menghilangkan hashtag, dan menghapus username. Setelah data dibersihkan, langkah selanjutnya adalah mengombinasikan berbagai fitur preprocessing untuk meningkatkan akurasi model machine learning. Beberapa fitur preprocessing yang digunakan meliputi case folding, tokenisasi, penghapusan stopword, normalisasi, dan stemming. Hasil dari proses pembersihan dan kombinasi fitur ini dapat dilihat dalam Tabel 3, yang menunjukkan hasil dari preprocessing teks.

Tabel 3. Text Preprocessing

No	Proses	Sebelum	Sesudah
1	Case Folding	@@republikaonline Pan pks demokrat nasdem pkb golkar sdh pasti ambruk dg hadirnya kaesang ketum psi, justru psi yg akan meroket di urutan ke 2 nanti	republikaonline pan pks demokrat nasdem pkb golkar sdh pasti ambruk dg hadirnya kaesang ketum psi justru psi yg akan meroket di urutan ke nanti
2	Stopword Removal	@@republikaonline Pan pks demokrat nasdem pkb golkar sdh pasti ambruk dg hadirnya kaesang ketum psi, justru psi yg akan meroket di urutan ke 2 nanti	republikaonline pan pks demokrat nasdem pkb golkar sdh ambruk dg hadirnya kaesang ketum psi psi yg meroket urutan
3	Stemming	@@republikaonline Pan pks demokrat nasdem pkb golkar sdh pasti ambruk dg hadirnya kaesang ketum psi, justru psi yg akan meroket di urutan ke 2 nanti	republikaonline pan pks demokrat nasdem pkb golkar sdh ambruk dg hadir kaesang tum psi psi yg roket urut
4	Word normalization	@@republikaonline Pan pks demokrat nasdem pkb golkar sdh pasti ambruk dg hadirnya kaesang ketum psi, justru psi yg akan meroket di urutan ke 2 nanti	@republikaonline pan pks demokrat nasdem pkb golkar sudah pasti ambruk dengan hadirnya kaesang ketum psi, justru pssi yang akan meroket di urutan ke 2 nanti

Pada tabel 3 menggambarkan bagaimana teks diubah melalui beberapa tahap text prprocessing seperti case folding, stopword removal, stemming, dam word normalization. Proses-proses ini sangat penting untuk meningkatkan akurasi dan kinerja model klasifikasi sentimen yang digunakan dalam penelitian ini.

3.3 Pengujian Text Preprocessing

Pada Tahapan text preprocessing ini bertujuan untuk menemukan model SVM yang optimal dengan kinerja terbaik. Pada tahap ini, beberapa eksperimen dilakukan untuk mengevaluasi dampak dari berbagai langkah preprocessing teks, seperti case folding, stemming, normalisasi, dan penghapusan stopword. Komposisi langkah-langkah preprocessing ini dijelaskan dalam 5abel 4 berikut:

Tabel 4. Pengujian Text Preprocessing

Eksperimen	Stopword Normalizatiom	Stemming	Word Normalization
E1	Iya	Iya	Iya
E2	Iya	Iya	Tidak
E3	Iya	Tidak	Iya
E4	Tidak	Iya	Iya
E5	Tidak	Tidak	Tidak
E6	Tidak	Iya	Tidak
E7	Tidak	Tidak	Iya
E8	Iya	Tidak	Tidak

Pada tabel 4 Pengujian Text Preprocessing, dilakukan proses penerapan dan non-penerapan beberapa langkah preprocessing. Tanda "iya" menunjukkan bahwa langkah tersebut diterapkan, sedangkan tanda "tidak" menunjukkan bahwa langkah tersebut tidak diterapkan dalam tahap text preprocessing..

3.4 Pengujian Data Uji

Eksperimen ini bertujuan untuk menemukan model SVM yang optimal dengan kinerja terbaik. Pada tahap ini, beberapa eksperimen dilakukan untuk mengevaluasi dampak dari berbagai langkah pada data balencing . Komposisi langkah-langkah data balencing ini dijelaskan dalam tabel 5.

Tabel 5. Pengujian data val

Eksperimen	F1 Score	Acc	Prec	Recall
E1	0.59	0.60	0.61	0.60
E2	0.57	0.58	0.58	0.58
E3	0.55	0.57	0.58	0.56
E4	0.53	0.55	0.55	0.55
E5	0.53	0.55	0.56	0.55
E6	0.52	0.53	0.53	0.53
E7	0.56	0.57	0.57	0.56
E8	0.59	0.60	0.60	0.60

Pada Tabel 5 Pengujian data val menunjukkan hasil pengujian beberapa eksperimen model terhadap data validasi berdasarkan metrik F1 Score, Accuracy (Acc), Precision (Prec), dan Recall. Eksperimen diberi label dari E1 hingga E8, dengan nilai F1 Score berkisar antara 0.52 hingga 0.59, nilai akurasi antara 0.53 hingga 0.60, nilai presisi antara 0.53 hingga 0.61, dan nilai recall antara 0.53 hingga 0.60. Eksperimen E1 dan E8 menonjol dengan nilai F1 Score tertinggi sebesar 0.59 serta nilai accuracy dan recall sebesar 0.60, menunjukkan bahwa model pada kedua eksperimen ini memiliki kinerja terbaik dibandingkan dengan eksperimen lainnya dalam tabel.

Tabel 6. Perbandingan pengujian data test

Metode	F1 Score	Acc	Prec	Recall
SVM TF-IDF (Kaesang dan Covid)	0.51	0.61	0.52	0.59
SVM TF-IDF (kaesang)	0.43	0.50	0.47	0.53

Dari pengujian pada tabel 6 di atas membandingkan hasil pengujian dua metode SVM dengan fitur TF-IDF pada data uji berdasarkan metrik F1 Score, Accuracy (Acc), Precision (Prec), dan Recall. Metode pertama, SVM TF-IDF (Kaesang dan Covid), memiliki F1 Score 0.51, akurasi 0.61, presisi 0.52, dan recall 0.59, menunjukkan kinerja yang lebih baik dibandingkan dengan metode kedua, SVM TF-IDF (Kaesang), yang memiliki F1 Score 0.43, akurasi 0.50, presisi 0.47, dan recall 0.53. Hal ini mengindikasikan bahwa dengan menambahkan data terkait Covid pada fitur TF-IDF, model SVM menunjukkan peningkatan kinerja yang signifikan dalam hal akurasi, presisi, dan F1 Score dibandingkan dengan model yang hanya menggunakan data terkait Kaesang.

Tabel 7. Hasil Leaderboard

Nama	Metode	F1 Score	Acc	Prec	Recall
Yoga El Saputra	SVM TF-IDF	0.51	0.61	0.52	0.59
Reza Mahendra	SGD TF-IDF	0.43	0.50	0.47	0.52
admin	Baseline	0.40	0.45	0.49	0.48

Dari tabel 7 menampilkan hasil leaderboard dari tiga peserta yang menggunakan metode berbeda untuk mengolah data dengan teknik TF-IDF. Yoga El Saputra menggunakan metode SVM TF-IDF dan berhasil memperoleh F1 Score sebesar 0.51, akurasi 0.61, presisi 0.52, dan recall 0.59. Menggunakan metode SGD TF-IDF dan mencatat F1 Score sebesar 0.43, akurasi 0.50, presisi 0.47, dan recall 0.52. Admin, yang menggunakan metode baseline, memperoleh F1 Score sebesar 0.40, akurasi 0.45, presisi 0.49, dan recall 0.48. Dari hasil ini, metode SVM TF-IDF menunjukkan kinerja terbaik dalam hal F1 Score, akurasi, presisi, dan recall dibandingkan dengan Reza Mahendra dan metode baseline yang digunakan oleh admin.

4. KESIMPULAN

Penelitian ini berhasil menerapkan metode Support Vector Machine (SVM) dengan pendekatan Term Frequency-Inverse Document Frequency (TF-IDF) untuk melakukan klasifikasi sentimen terkait pengangkatan Kaesang Pangarep sebagai Ketua Umum Partai Solidaritas Indonesia (PSI). Hasil penelitian menunjukkan bahwa model SVM yang digunakan memiliki kinerja yang cukup baik dengan nilai F1 Score sebesar 0.51, akurasi 0.60, presisi 0.52, dan recall 0.59. Nilai-nilai ini menunjukkan bahwa metode SVM dengan TF-IDF mampu mengklasifikasikan sentimen dengan cukup efektif dalam konteks dataset yang digunakan.Namun, penelitian ini juga mengidentifikasi beberapa keterbatasan yang perlu diperhatikan. Salah satu keterbatasan utama adalah ukuran dataset yang relatif kecil, hanya terdiri dari 300 data training dan 924 data testing. Ukuran dataset yang kecil dapat mempengaruhi kemampuan model untuk menggeneralisasi hasilnya

pada data yang lebih luas. Selain itu, proses anotasi data yang dilakukan melalui metode crowd sourcing juga memiliki potensi bias, meskipun sudah dilakukan validasi oleh tiga anotator untuk memastikan konsistensi dan keandalan label. Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan lebih beragam untuk meningkatkan akurasi dan kemampuan generalisasi model. Selain itu, penerapan teknik-teknik pemrosesan teks tambahan, seperti penggunaan word embeddings atau deep learning, dapat dipertimbangkan untuk meningkatkan kinerja model. Penelitian ini memberikan dasar yang kuat untuk analisis sentimen menggunakan SVM dan TF-IDF, namun masih ada ruang untuk pengembangan lebih lanjut agar hasil yang diperoleh dapat lebih akurat dan andal dalam berbagai konteks aplikasi.

REFERENCES

[1] M. Apriliansyah, "Dinamika Partisipatif Media Dan Jaringan Sosial: Analisis Kasus Isu Prabowo Dalam Pemilu 2024," *J. Sekr. Adm.*, vol. 21, no. 2, hal. 81–95, 2023.

[2] "Profil PSI yang Angkat Kaesang Pangarep Jadi Ketua Umum Halaman all - Kompas.com." https://www.kompas.com/tren/read/2023/09/27/103000465/profil-psi-yang-angkat-kaesang-pangarep-jadi-ketua-umum-?page=all (diakses Mei 21, 2024).

[3] "Kaesang Jadi Ketum PSI Dinilai Semakin Merusak Kaderisasi dan Tak Beri Teladan Halaman all - Kompas.com." https://nasional.kompas.com/read/2023/09/26/14552721/kaesang-jadi-ketum-psi-dinilai-semakin-merusak-kaderisasi-dan-tak-beri?page=all (diakses Mei 21, 2024).

[4] "Suara Anak Muda dan Pengaruh Politik Gembira ala PSI - Partai Solidaritas Indonesia." https://psi.id/suara-anak-muda-dan-pengaruh-politik-gembira-ala-psi/ (diakses Mei 21, 2024).

[5] S. Suryono, E. Utami, dan E. T. Luthfi, "Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes Classifier," *Angkasa J. Ilm. Bid. Teknol.*, vol. 10, no. 1, hal. 89, 2018, doi: 10.28989/angkasa.v10i1.218.

[6] E. Retnoningsih dan R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insa. Ict J.*, vol. 7, no. 2, hal. 156, 2020, doi: 10.51211/biict.v7i2.1422.

[7] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, dan X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, hal. 41–51, 2018, doi: 10.21873/cgp.20063.

[8] M. I. Petiwi, A. Triayudi, dan I. D. Sholihati, "Analisis Sentimen Gofood Berdasarkan Twitter Menggunakan Metode Naïve Bayes dan Support Vector Machine," *J. Media Inform. Budidarma*, vol. 6, no. 1, hal. 542, 2022, doi: 10.30865/mib.v6i1.3530.

[9] A. Baita, Y. Pristyanto, dan N. Cahyono, "Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN)," *Inf. Syst. J.*, vol. 4, no. 2, hal. 42–46, 2021.

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

[11] R. Wahyudi dan G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *J. Inform.*, vol. 8, no. 2, hal. 200–207, 2021, doi: 10.31294/ji.v8i2.9681.

[12] I. Muslim Karo Karo *et al.*, "Analisis Sentimen Ulasan Aplikasi Info BMKG di Google Play Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Syst. Res.*, vol. 4, no. 4, hal. 1423–1430, 2023, doi: 10.47065/josh.v4i4.3943.

[13] S. Khairunnisa, A. Adiwijaya, dan S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, hal. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[14] N. Satya Marga, A. Rahman Isnain, dan D. Alita, "Sentimen Analisis Tentang Kebijakan Pemerintah Terhadap Kasus Corona Menggunakan Metode Naive Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 453, no. 4, hal. 453–463, 2021.

[15] D. Darwis, E. S. Pratiwi, dan A. F. O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, hal. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.

[16] N. Charibaldi, A. Harfiani, dan O. S. Simanjuntak, "Bayes Classifier dan K-Nearest Neighbor untuk Analisis Sentimen," vol. 9, no. 1, 2024.

[17] H. Syah dan A. Witanti, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (Svm)," *J. Sist. Inf. dan Inform.*, vol. 5, no. 1, hal. 59–67, 2022, doi: 10.47080/simika.v5i1.1411.

[18] A. N. Ulfah dan M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, hal. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[19] W. Athira Luqyana, I. Cholissodin, dan R. S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, hal. 4704–4713, 2018.

[20] A. Bhattacharjee *et al.*, "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022 - Findings*, 2022, hal. 1318–1327, doi: 10.18653/v1/2022.findings-naacl.98.

[21] H. Wang, W. Zhou, dan Y. Shao, "A new fast ADMM for kernelless SVM classifier with truncated fraction loss," *Knowledge-Based Syst.*, vol. 283, hal. 111214, Jan 2023, doi: 10.1016/j.knosys.2023.111214.

[22] M. Arya dan C. S. S. Bedi, "Survei tentang SVM dan aplikasinya dalam klasifikasi citra," 2018.