JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume XIV No 2, Desember 2025
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

PERBANDINGAN ALGORITMA SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK ANALISIS SENTIMEN ISU IJAZAH PALSU JOKO WIDODO DI MEDIA SOSIAL X

Deni[1)], Roki Fatih Musthofa[2)], Hanum Surya Herfiana[3)], Betha Nurina Sari[4)]
Program Studi Sistem Informasi, Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang
Karawang, Indonesia
email: 2210631250008@student.unsika.ac.id[1)], 2210631250029@student.unsika.ac.id[2)],
2210631250014@student.unsika.ac.id[3)] , betha.nurina@staff.unsika.ac.id[4)]

Abstrak
Media sosial, khususnya X, telah menjadi ruang diskusi publik yang aktif dalam membahas berbagai isu sosial dan politik. Salah satu isu yang menimbulkan banyak perdebatan adalah dugaan ijazah palsu milik Presiden Joko Widodo. Penelitian ini bertujuan untuk menganalisis sentimen pengguna X terhadap isu tersebut serta membandingkan performa dua algoritma klasifikasi, yaitu *Support Vector Machine* (SVM) dan *Random Forest*. Proses analisis diawali dengan pengumpulan data menggunakan teknik *scraping*, diikuti tahap pra-pemrosesan, pelabelan data secara manual, ekstraksi fitur menggunakan TF-IDF, serta penyeimbangan data dengan SMOTE untuk mengatasi ketidakseimbangan label. Dari total 1.783 komentar yang terkumpul, ditemukan 1.661 komentar negatif dan 122 komentar positif. Setelah diterapkan SMOTE, distribusi data menjadi seimbang dengan total 3.322 data. Hasil pengujian pada beberapa skenario menunjukkan bahwa algoritma SVM mencapai akurasi tertinggi sebesar 100%, sementara Random Forest juga memberikan performa sangat baik dengan akurasi mencapai 99,24%. Temuan ini menunjukkan bahwa SVM lebih unggul dalam mengklasifikasikan sentimen teks pada isu sensitif di media sosial, khususnya ketika data telah melalui proses penyeimbangan menggunakan SMOTE.
Kata Kunci : Algoritma, Random Forest, Support Vector Machine, Analisis Sentimen, X

1. Pendahuluan

Media sosial menandai era media baru yang mendasar, mengubah cara masyarakat modern berinteraksi satu sama lain. Melalui media sosial, individu memiliki wadah untuk menampilkan identitas diri, berkomunikasi, berkolaborasi, berbagi informasi, dan membangun hubungan secara virtual dengan pengguna lain [1]. Platform seperti X (sebelumnya Twitter) sebagai contoh mikroblogging bersifat sangat cepat dan real-time, memberi ruang bagi pengguna untuk mengekspresikan pendapat, menyebarkan informasi, serta menanggapi isu secara terbuka dan luas [2].

Data yang tersedia di media sosial seperti X sangat besar dan beragam, merefleksikan berbagai persepsi, emosi, serta reaksi publik terhadap peristiwa atau tokoh tertentu. Namun, karena data ini biasanya tidak terstruktur, bersifat singkat, dan penuh dengan ciri khas media sosial seperti emotikon, singkatan, sarkasme, dan bahasa informal, dibutuhkan pendekatan analisis yang lebih cermat dan canggih. Khusus dalam isu kontroversial berupa dugaan ijazah palsu terkait mantan Presiden Joko Widodo, data yang terkumpul tidak hanya mengandung sikap individu terhadap sosok publik tersebut, tetapi juga mencerminkan tingkat kepercayaan masyarakat terhadap institusi pendidikan dan legitimasi pemerintah. Oleh sebab itu, analisis sentimen menjadi metode strategis untuk mengungkap tren persepsi publik secara objektif dan berdasarkan data nyata. Selain itu, perbandingan antara algoritma klasifikasi seperti *Support Vector Machine* dan *Random Forest* perlu dilakukan guna mengetahui metode mana yang lebih efektif dalam mengolah opini publik di media sosial, terutama dalam kasus dengan tingkat sensitivitas politik tinggi. Penelitian ini diharapkan memberikan kontribusi baik secara teoretis maupun praktis dalam mengembangkan analisis sentimen pada era digital.

Isu ijazah palsu yang menyeret nama mantan Presiden Joko Widodo menjadi sorotan besar di masyarakat dan media sosial. Meski terdapat penelitian akademik yang menyatakan bahwa tuduhan pemalsuan tersebut secara hukum tidak berdasar, dengan analisis pidana dan administratif yang mendukung keaslian ijazahnya [3], isu tetap meluas di ruang digital dan memengaruhi kepercayaan publik. Penelitian lain juga memperlihatkan bagaimana gerakan sosial digital dan aktivisme massa merespons isu ini sebagai bagian dari dinamika legitimasi politik, menandakan bahwa kontroversi ini berpotensi mengguncang legitimasi institusi pendidikan dan kepercayaan terhadap sistem politik [4]. Lebih lanjut, kajian arsip hukum menggarisbawahi pentingnya arsip digital dan transparansi dokumen pendidikan publik dalam menjaga kredibilitas figur publik dan institusi negara [5].

Isu politik seperti dugaan pemalsuan ijazah tidak hanya menimbulkan kegaduhan di ruang publik digital, tetapi juga dapat memengaruhi tingkat kepercayaan masyarakat terhadap kredibilitas figur publik maupun institusi pemerintahan [6]. Dalam konteks ini, analisis sentimen sangat penting untuk memetakan pola persepsi masyarakat secara objektif berdasarkan data nyata dari media sosial. Karena data Twitter bersifat tidak terstruktur dan dinamis,

diperlukan algoritma yang efektif dan efisien untuk klasifikasinya. Oleh karena itu, sangat krusial mengevaluasi dan memilih algoritma paling relevan guna menganalisis sentimen masyarakat terkait isu sensitif seperti ini. Analisis sentimen adalah bagian dari *text mining* yang fokus pada pemrosesan teks untuk mengekstraksi emosi, opini, dan persepsi seseorang terhadap suatu subjek menggunakan teknik pemrosesan bahasa alami [7]. Analisis sentimen di media sosial mampu memberikan informasi berharga bagi objek opini, dengan mengklasifikasikan pernyataan pengguna ke dalam kategori positif, negatif, atau netral [8].

Teknik dan algoritma data mining dapat digunakan untuk mengidentifikasi pola, mengklasifikasikan data, dan membuat prediksi yang mendukung analisis sentimen [9]. Dua algoritma yang banyak digunakan dalam analisis sentimen adalah *Support Vector Machine* (SVM) dan *Random Forest*. SVM dikenal efektif dalam memisahkan kelas data melalui pemetaan ke ruang berdimensi tinggi, sehingga mampu menghasilkan akurasi yang baik pada data teks yang kompleks [10]. Sementara itu, *Random Forest* bekerja dengan membangun banyak pohon keputusan dan melakukan voting untuk menentukan hasil klasifikasi sehingga lebih robust terhadap data beragam, termasuk untuk tugas analisis sentimen di media sosial [11]. Kedua algoritma ini sering menjadi pilihan utama dalam pemodelan sentimen karena performanya yang stabil pada data teks yang besar dan tidak terstruktur.

Support Vector Machine merupakan algoritma klasifikasi yang sangat efektif untuk pemrosesan bahasa alami dan pada penelitian ini digunakan untuk mengkategorikan komentar pengguna aplikasi Twitter ke dalam kategori komentar positif dan negatif [4]. Meskipun SVM cenderung membutuhkan waktu komputasi lebih tinggi dibandingkan beberapa algoritma lain, algoritma tersebut sering menunjukkan performa yang unggul dalam tugas klasifikasi yang kompleks. Dalam konteks analisis sentimen, beberapa studi membandingkan SVM dengan RF dan menemukan bahwa SVM mencatat akurasi yang kompetitif, tetapi RF juga memberikan hasil yang sangat kuat dan lebih tahan terhadap variasi data [12].

Penelitian sebelumnya yang dilakukan oleh Muasaroh et al. menganalisis sentimen komentar YouTube terhadap isu ijazah Presiden Joko Widodo menggunakan algoritma *Support Vector Machine* dan *Random Forest*. Hasil penelitian tersebut menunjukkan bahwa algoritma Support Vector Machine memperoleh akurasi terbaik, yaitu sebesar 97%, lebih tinggi dibandingkan *Random Forest* yang memiliki akurasi 95%. [13] Hasil penelitian lain yang dilakukan oleh Wulandari et al. juga menegaskan keunggulan algoritma *Support Vector Machine* dalam menganalisis sentimen. Pada studi tersebut, yang membandingkan kinerja Naïve Bayes dan SVM dalam klasifikasi sentimen Twitter mengenai isu dugaan ijazah palsu Jokowi, algoritma SVM menghasilkan tingkat akurasi 69,23%, lebih tinggi dibandingkan Naïve Bayes yang hanya mencapai 65%. [14]

Tujuan penelitian ini adalah membandingkan algoritma klasifikasi *Support Vector Machine* dan *Random Forest* dalam analisis sentimen data X terkait isu ijazah palsu Presiden Joko Widodo, untuk mengetahui algoritma mana yang lebih unggul dalam mengolah data media sosial yang tidak terstruktur dan dinamis. Dari penelitian ini, diharapkan akan membantu peneliti lain dalam menentukan teknik analisis sentimen yang lebih efektif dan akurat untuk mengolah data media sosial yang tidak terstruktur.

2. Landasan Teori

X

X atau twiter merupakan salah satu platform microblogging yang banyak digunakan untuk membagikan pemikiran dan membahas berbagai isu yang sedang berkembang. Sebagai media sosial yang memungkinkan pengguna menyampaikan pendapat secara *real-time*, Twitter menjadi sumber data yang relevan untuk analisis sentimen. Penelitian sentimen pada platform ini umumnya bertujuan menggali dan mengelompokkan opini pengguna ke dalam kategori positif, negatif, atau netral. Di Indonesia, Twitter sering dimanfaatkan dalam berbagai penelitian untuk memahami pandangan publik terhadap beragam topik, mulai dari kebijakan pemerintah, isu sosial, hingga peluncuran teknologi baru [15].

Analisis Sentimen

Analisis sentimen merupakan salah satu teknik dalam *Natural Language Processing (NLP)* yang bertujuan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau emosi yang terkandung dalam sebuah teks ke dalam kategori seperti positif, negatif, atau netral [16]. Proses ini dilakukan melalui beberapa tahapan, mulai dari pengumpulan data, pembersihan teks (preprocessing), ekstraksi fitur menggunakan metode seperti TF-IDF, hingga klasifikasi sentimen menggunakan algoritma *machine learning*, misalnya *Support Vector Machine (SVM)*, *Random Forest*, atau metode lainnya. Analisis sentimen menjadi penting karena mampu memberikan gambaran mengenai persepsi publik terhadap suatu isu, produk, kebijakan, atau peristiwa secara cepat dan *real-time*, terutama ketika diterapkan pada data berjumlah besar seperti tweet di media sosial [17].

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency (TF-IDF) adalah metode untuk mengubah kata menjadi nilai numerik dalam bentuk vektor guna menentukan bobot atau tingkat kepentingan suatu kata dalam sebuah dokumen maupun keseluruhan korpus. Bobot tersebut membantu mengukur seberapa signifikan kata tersebut dalam konteks dokumen. Secara umum, perhitungan TF-IDF terdiri dari dua komponen utama, yaitu *Term Frequency (TF)* dan *Inverse Document Frequency (IDF)*, yang masing-masing memiliki cara perhitungan berbeda dan digabungkan pada tahap akhir untuk menghasilkan nilai bobot akhir [18].

1. *Term Frequency (TF)*

Perhitungan TF dilakukan dengan menghitung seberapa sering sebuah kata muncul dalam suatu dokumen. Karena setiap dokumen memiliki panjang yang berbeda, nilai TF dinormalisasi dengan membaginya terhadap total jumlah kata pada dokumen tersebut.

$$tf_{t.d} = \frac{n_{t,d}}{(Total\ number\ of\ term\ in\ document)} \quad (2.1)$$

Keterangan
Tf = frekuensi kemunculan kata pada sebuah dokumen

2. *Inverse Document Frequency*

Setelah nilai TF diperoleh, langkah berikutnya adalah menghitung IDF, yaitu ukuran yang menunjukkan tingkat kepentingan suatu kata dalam seluruh korpus. Semakin kecil nilai IDF, maka kata tersebut dianggap kurang penting karena muncul di banyak dokumen, dan sebaliknya.

$$idf_d = log\ \frac{Number\ of\ document}{(Total\ number\ of\ term\ in\ document)} \quad (2.2)$$

Keterangan
Idf = mengukur penting/tidak sebuah kata dalam dokumen

3. *Term Frequency-Inverse Document Frequency*

Setelah kedua nilai tersebut dihitung, bobot TF-IDF diperoleh dengan mengalikan nilai TF dan IDF sehingga menghasilkan ukuran akhir yang menunjukkan relevansi kata dalam dokumen.

$$tfidf_{t,d} = tf_{t,d} \ X\ idf_d \quad (2.3)$$

Keterangan
TF-IDF = hasil penggabungan antara TF dan IDF

SMOTE

Synthetic Minority Oversampling Technique (SMOTE) adalah teknik *oversampling* yang digunakan untuk mengatasi masalah data tidak seimbang dalam klasifikasi. SMOTE menghasilkan sampel sintetis dari kelas minoritas dengan interpolasi antara contoh yang ada, sehingga meningkatkan representasi kelas minoritas dan membantu model menghindari bias terhadap kelas mayoritas. Penerapan SMOTE dalam penelitian analisis sentimen dan klasifikasi terbukti dapat meningkatkan performa model dalam hal akurasi, *recall*, dan *F1-score* [19].

Support Vector Machine

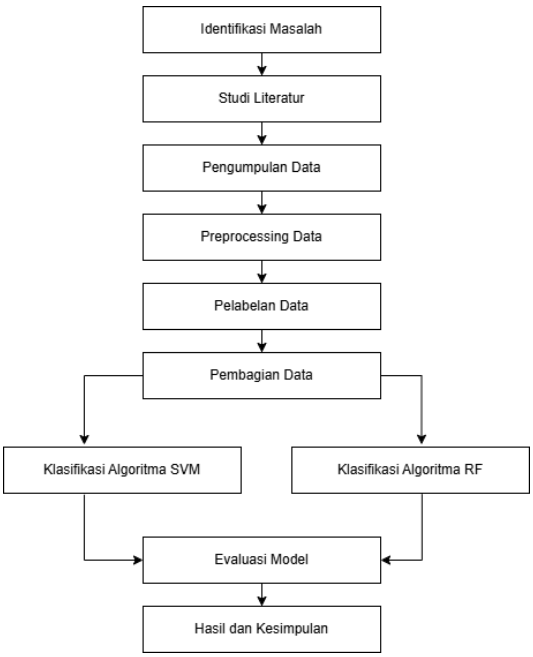
Support Vector Machine (SVM) adalah algoritma klasifikasi yang berupaya menemukan *hyperplane* optimal yang dapat memisahkan kelas-kelas data secara maksimal. Algoritma ini efektif untuk data berdimensi tinggi dan sangat populer dalam klasifikasi teks dan analisis sentimen. Penggunaan SVM dengan data yang sudah diolah menggunakan TF-IDF dan SMOTE memungkinkan pengklasifikasian sentimen dengan ketepatan yang baik, terutama pada dataset berimbalance [20].

Random Forest

Random Forest (RF) adalah metode *ensemble* yang menggabungkan banyak pohon keputusan untuk meningkatkan kemampuan klasifikasi dan mengurangi risiko *overfitting*. Algoritma ini populer karena robust, mudah digunakan, dan memberikan hasil akurasi yang tinggi pada berbagai aplikasi klasifikasi dan prediksi. *Random Forest* dapat diaplikasikan pada data hasil ekstraksi fitur TF-IDF untuk melakukan klasifikasi sentimen atau prediksi lain dengan performa yang konsisten [21].

3. Metode Penelitian

Gambar 1 menyajikan diagram alir yang menguraikan tahapan penelitian, memungkinkan peneliti untuk melaksanakan studi secara komprehensif dari awal hingga akhir serta menjamin keterpaduan dan keselarasan setiap elemen penelitian dengan rencana yang telah ditetapkan.



Gambar 1. Bagan Alur Penelitian

Identifikasi Masalah

Isu kontroversial mengenai dugaan ijazah palsu mantan Presiden Joko Widodo telah memicu diskusi intensif di berbagai saluran media sosial, khususnya Twitter. Akibat dari penyebaran beragam komentar dan opini, proses pengelompokan sentimen publik menjadi kompleks. Penelitian ini berfokus pada evaluasi komparatif performa algoritma *Support Vector Machine* dan *Random Forest* dalam mengklasifikasikan sentimen khalayak terhadap polemik tersebut.

Studi Literatur

Studi literatur dilakukan untuk mengumpulkan teori dan temuan penelitian sebelumnya yang berkaitan dengan text mining, analisis sentimen, preprocessing teks, TF-IDF, serta kinerja algoritma *Support Vector Machine* dan *Random Forest*. Tahap ini memberikan dasar ilmiah dalam menyusun metode yang tepat untuk penelitian.

Pengumpulan Data

Data diperoleh melalui *scraping* tweet dari *platform* X menggunakan pustaka *snscrape*. Tweet dikumpulkan berdasarkan kata kunci relevan, kemudian digabungkan, dibersihkan dari duplikasi, dan dipilih hanya kolom yang diperlukan. Dataset ini menjadi bahan utama dalam analisis sentimen.

Preprocessing Data

Preprocessing dilakukan untuk membersihkan dan menyiapkan teks sebelum pemodelan. Prosesnya meliputi pembersihan teks dari simbol atau URL, tokenisasi, penghapusan *stopwords*, serta *stemming* untuk mengembalikan kata ke bentuk dasar. Tahap ini memastikan data siap dipakai pada proses ekstraksi fitur.

Pelabelan Data

Pada tahap ini, setiap tweet diberi label sentimen sesuai kategori positif dan negatif. Pelabelan dilakukan secara manual berdasarkan penilaian peneliti. Proses ini memastikan bahwa data memiliki target label yang dibutuhkan untuk pelatihan model *supervised learning*. Label yang akurat akan membantu model belajar pola sentimen dengan lebih baik sehingga prediksi menjadi lebih valid.

Pembagian Data

Setelah data dibersihkan dan diberi label, dataset dibagi menjadi dua bagian, yaitu data latih *(training set)* dan data uji *(testing set)* dengan beberapa skenario seperti 90:10, 80:20, 70:30, dan 60:40. Pembagian ini bertujuan untuk melatih model menggunakan sebagian besar data, sementara sisanya digunakan untuk menguji performanya. Teknik penyeimbangan seperti SMOTE dapat diterapkan sebelumnya jika dataset memiliki distribusi sentimen yang tidak seimbang.

Klasifikasi Algoritma Support Vector Machine

Pada tahap ini, algoritma *Support Vector Machine* dilatih menggunakan data latih yang telah melalui *preprocessing* dan transformasi TF-IDF. SVM bekerja dengan mencari *hyperplane* terbaik yang memisahkan kelas sentimen berdasarkan fitur teks. Model kemudian digunakan untuk memprediksi sentimen pada data uji. Proses ini menghasilkan metrik performa seperti akurasi, *precision*, *recall*, dan *F1-score* untuk evaluasi.

Klasifikasi Algoritma Random Forest

Pada tahap ini dilakukan pelatihan menggunakan algoritma *Random Forest*. Algoritma ini membangun sejumlah pohon keputusan *(decision trees)* dan menghasilkan prediksi berdasarkan voting mayoritas. *Random Forest* cocok untuk menangani dataset dengan pola kompleks dan memiliki toleransi tinggi terhadap noise. Setelah dilatih, model digunakan untuk memprediksi data uji dan menghasilkan metrik performa yang sama seperti SVM untuk tujuan komparasi.

Evaluasi Model

Evaluasi dilakukan untuk membandingkan performa kedua model pada berbagai skenario pembagian data. Parameter yang digunakan antara lain akurasi, *precision, recall, F1-score*, serta *confusion matrix*. Evaluasi ini bertujuan untuk menentukan model mana yang lebih optimal dalam mengklasifikasikan sentimen terkait isu ijazah Presiden Joko Widodo. Hasil evaluasi menjadi dasar untuk menarik kesimpulan objektif mengenai model terbaik.

4. Hasil Penelitian

a. Pengumpulan Data

Proses pengumpulan data dilakukan dengan memanfaatkan teknik *scraping* pada platform media sosial X (Twitter). Teknik ini memungkinkan peneliti untuk mengunduh sejumlah besar data berdasarkan kata kunci spesifik yang relevan dengan isu yang sedang diteliti, seperti "ijazah palsu Jokowi", "ijazah Joko Widodo", dan "ijazah asli Jokowi" dengan periode waktu 1 April - 1 Juni 2025, data yang berhasil diambil sebanyak 1783 cuitan. Pemilihan kata kunci tersebut dilakukan agar data yang dikumpulkan dapat merepresentasikan opini masyarakat secara komprehensif, baik yang mendukung maupun yang meragukan keaslian ijazah tersebut. Tabel 1 merupakan contoh data komentar yang diperoleh dari X.

Tabel 1. Contoh Komentar Pengguna Twitter

No	Teks Komentar
1.	UUD ITE lagi menyasar Roy Suryo cs... IJAZAH
2.	@mihan_sandy @DokterTifa @UGMYogyakarta Pihak pelapor mendatangkan saksi ahli dari ugm
3.	Selesaikan masalah ijazah yok!!!
4.	tunjukkan wujud esemka dan ijazah aslinya

b. *Preprocessing Data*

Preprocessing merupakan tahapan krusial untuk membersihkan dan mengubah data mentah menjadi format yang lebih terstruktur agar dapat diolah oleh algoritma. Tahapan *preprocessing* dalam penelitian ini meliputi *cleaning text, tokenisasi, stopword removal,* dan *stemming*.

Cleaning Text, tahap ini bertujuan membersihkan noise dari data teks, seperti menghapus URL, *mention (@), hashtag (#)*, angka, dan tanda baca, serta mengubah seluruh huruf menjadi huruf kecil *(lowercase)*. Contoh hasil dari *cleaning text* dapat dilihat pada Tabel 3.

Tabel 3. Hasil *Cleaning Text*

Data Mentah	Hasil *Cleaning*
UUD ITE lagi menyasar Roy Suryo cs... IJAZAH	uud ite lagi menyasar roy suryo cs ijazah palsu
Selesaikan masalah ijazah yok!!!	selesaikan masalah ijazah yok
@PrakosoTim1574 @tukangpermales @DokterTifa….	saksi ahli dari ugm datang menyatakan ijazah…

Tokenisasi, setelah data dibersihkan, dilakukan proses tokenisasi untuk memecah kalimat menjadi potongan kata atau token yang berdiri sendiri. Hasil dari tokenisasi dapat dilihat pada Tabel 4.

Tabel 4. Hasil Tokenisasi

Hasil *Cleaning*	Hasil Tokenisasi

uud ite lagi menyasar roy suryo cs ijazah palsu	['uud', 'ite', 'lagi', 'menyasar', 'roy', 'suryo', 'cs', 'ijazah', 'palsu']
selesaikan masalah ijazah yok	['selesaikan', 'masalah', 'ijazah', 'yok']
tunjukkan wujud esemka dan ijazah aslinya	['tunjukkan', 'wujud', 'esemka', 'dan', 'ijazah', 'aslinya']

Stopword Removal, proses ini menghilangkan kata-kata umum yang tidak memiliki makna sentimen yang signifikan (seperti "dan", "yang", "di") menggunakan pustaka NLTK bahasa Indonesia dan daftar *stopword* tambahan. Hasil dari proses *stopword removal* dapat dilihat pada Tabel 5.

Tabel 5. Hasil *Stopword Removal*

Hasil Tokenisasi	Hasil *Stopword Removal*
['uud', 'ite', 'lagi', 'menyasar', 'roy', 'suryo', 'cs', 'ijazah', 'palsu']	['uud', 'ite', 'menyasar', 'roy', 'suryo', 'cs', 'ijazah', 'palsu']
['selesaikan', 'masalah', 'ijazah', 'yok']	['selesaikan', 'ijazah', 'yok']
['tunjukkan', 'wujud', 'esemka', 'dan', 'ijazah', 'aslinya']	['tunjukkan', 'wujud', 'esemka', 'ijazah', 'aslinya']

Stemming, tahap akhir *preprocessing* adalah *stemming* menggunakan pustaka Sastrawi untuk mengubah kata berimbuhan menjadi kata dasarnya. Hasil dari proses *stemming* dapat dilihat pada Tabel 6.

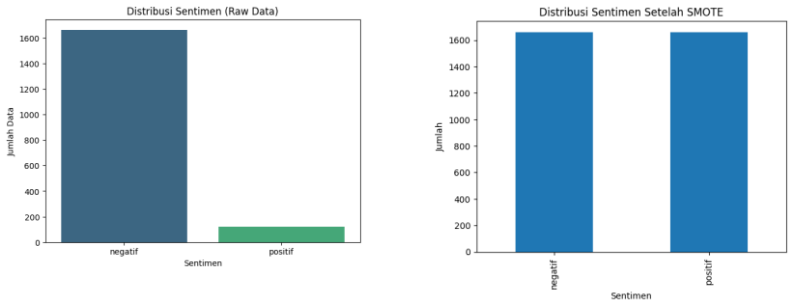
Tabel 6. *Stemming*

Hasil *Stopword Removal*	Hasil *Stemming*
['uud', 'ite', 'menyasar', 'roy', 'suryo', 'cs', 'ijazah', 'palsu']	['uud', 'ite', 'sasar', 'roy', 'suryo', 'cs', 'ijazah', 'palsu']
['selesaikan', 'ijazah', 'yok']	['selesai', 'ijazah', 'yok']
['tunjukkan', 'wujud', 'esemka', 'ijazah', 'aslinya']	['tunjuk', 'wujud', 'esemka', 'ijazah', 'asli']

c. Pelabelan Data

Pelabelan data dilakukan secara manual untuk memastikan ketepatan klasifikasi sentimen pada setiap cuitan. Proses ini melibatkan peninjauan satu per satu terhadap isi cuitan guna menentukan apakah sentimen yang disampaikan bersifat positif atau negatif. Komentar dikategorikan ke dalam dua kelas sentimen: positif dan negatif, berdasarkan kemunculan kata kunci tertentu (misalnya "palsu" untuk negatif, "sah" untuk positif). Distribusi awal data menunjukkan ketidakseimbangan yang signifikan dengan dominasi sentimen negatif. Untuk mengatasi hal ini, dilakukan penyeimbangan data menggunakan teknik SMOTE, sehingga menghasilkan proporsi data yang seimbang untuk pelatihan model. Contoh hasil dari pengimplementasian SMOTE dapat dilihat pada Gambar ?.

Gambar 2. Sebelum dan sesudah SMOTE



Sebelum dilakukan SMOTE, distribusi data sentimen menunjukkan ketidakseimbangan yang sangat mencolok, di mana sentimen negatif jauh lebih banyak yaitu 1663 data dibandingkan sentimen positif yang hanya 122 data. Kondisi ini dapat menimbulkan bias pada model karena algoritma cenderung lebih mengenali pola pada kelas mayoritas. Setelah dilakukan SMOTE, jumlah data pada kelas positif ditingkatkan secara sintetis hingga mencapai proporsi yang seimbang dengan kelas negatif. Dengan demikian, data yang dihasilkan menjadi lebih seimbang dan memungkinkan model untuk belajar secara lebih optimal pada kedua kelas sentimen. Hasil pelabelan dapat dilihat pada Tabel 7.

Tabel 7. Hasil Pelabelan

Kategori	Text Komentar
Negatif	uud ite sasar roy suryo cs ijazah palsu
Negatif	lapor saksi ahli ugm ijazah asli bless masuk
Positif	tunjuk wujud esemka ijazah asli

d. Pembagian Data

Data yang telah melalui tahap preprocessing dan penyeimbangan menggunakan SMOTE kemudian dibagi menjadi data latih (training data) dan data uji (testing data) dengan beberapa skenario proporsi, yaitu 90:10, 80:20, 70:30, dan 60:40. Pembagian ini dilakukan untuk mengevaluasi konsistensi performa model pada berbagai komposisi data latih dan uji. Setelah proses penyeimbangan, total data yang digunakan berjumlah 3.322 sampel, terdiri dari 1.661 data sentimen positif dan 1.661 data sentimen negatif. Melalui pembagian ini, model dilatih untuk mengenali pola sentimen secara optimal dan kemudian diuji pada data yang belum pernah ditemui sebelumnya, sehingga hasil evaluasi lebih objektif dan mencerminkan kemampuan generalisasi model. Berikut ini hasil dari skenario pengujian dapat dilihat pada Tabel 8.

Tabel 8. Hasil Pengujian Model pada Berbagai Skenario Pembagian Data

Skenario	Train Size	Test Size	Model	Akurasi
90:10	2.989	333	SVM (RBF)	1.0000
			Random Forest	0.9909
80:20	2.657	665	SVM (RBF)	1.0000
			Random Forest	0.9924
70:30	2.325	997	SVM (RBF)	0.9979
			Random Forest	0.9909
60:40	1.993	1.329	SVM (RBF)	0.9947
			Random Forest	0.9947

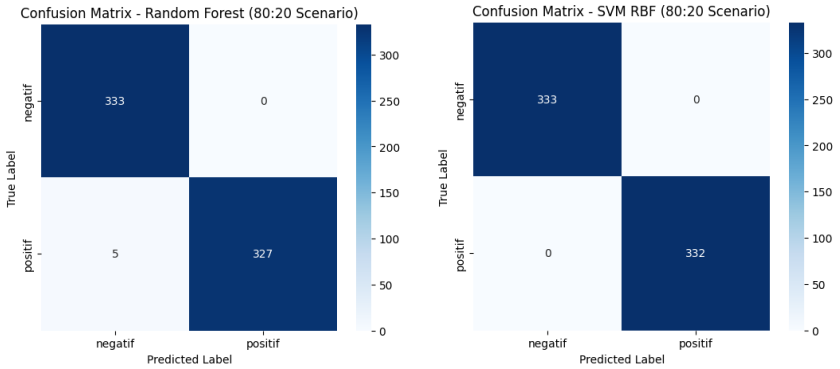
Hasil pengujian pada keempat skenario menunjukkan bahwa seluruh model memiliki performa yang sangat baik dengan akurasi diatas 98%. Perbedaan persentase pembagian data latih dan uji tidak berpengaruh signifikan terhadap penurunan performa, sehingga model terbukti stabil dan mampu mengenali pola sentimen dengan baik setelah data diseimbangkan menggunakan SMOTE.

e. Evaluasi Model

Evaluasi model pada penelitian ini dilakukan menggunakan confusion matrix, yaitu metode evaluasi performa klasifikasi melalui tabel matriks yang membandingkan hasil prediksi model dengan label sebenarnya. Berdasarkan hasil confusion matrix pada skenario pengujian 80:20, model Random Forest berhasil memprediksi 333 data

negatif dan 327 data positif dengan benar dari total keseluruhan data uji. Meskipun demikian, masih terdapat kesalahan prediksi pada 5 data positif yang salah diklasifikasikan sebagai negatif.

Secara keseluruhan, hasil confusion matrix menunjukkan bahwa model Random Forest memiliki kemampuan klasifikasi yang sangat baik, dengan tingkat kesalahan yang sangat rendah dan konsistensi prediksi yang tinggi terhadap kedua kelas sentimen.



Gambar 3. Confusion Matrix

Berikut ini adalah tabel perbandingan dari dua algoritma yang digunakan.

Tabel 9. Perbandingan

Algoritma	Accuracy	Precision	Recall	F1-Score
SVM	100%	100%	100%	100%
Random Forest	99,24%	100%	98,48%	99,23%

f. Visualisasi Word Cloud

Gambar 4. Visualisasi Wordcloud

Pada Word Cloud sentimen positif, kata-kata yang paling dominan seperti “cemar nama”, “bukti”, “asli”, dan “dukung” menunjukkan bahwa pengguna cenderung menyuarakan pembelaan terhadap isu ijazah tersebut. Kata-kata ini menggambarkan kecenderungan opini yang menekankan pentingnya klarifikasi, pembuktian, serta dukungan terhadap keaslian dokumen.

Sebaliknya, pada Word Cloud sentimen negatif, kemunculan kata-kata seperti “palsu”, “polisi”, “hukum”, dan “bohong” mencerminkan ekspresi keraguan, tuduhan, dan dorongan masyarakat agar isu ini diproses secara hukum. Dominasi kata-kata tersebut memperlihatkan bahwa kelompok sentimen negatif cenderung mengasosiasikan isu ini dengan tindakan manipulatif atau pelanggaran yang memerlukan penegakan aturan.

5. Kesimpulan

Penelitian ini menganalisis sentimen masyarakat terhadap isu dugaan ijazah palsu Presiden Joko Widodo di media sosial X serta membandingkan performa algoritma *Support Vector Machine* (SVM) dan *Random Forest*. Berdasarkan hasil pengolahan data melalui tahapan *preprocessing*, pelabelan manual, ekstraksi fitur menggunakan TF-IDF, serta penyeimbangan data menggunakan SMOTE, diperoleh bahwa kedua algoritma memberikan performa klasifikasi yang sangat baik. Algoritma SVM menunjukkan hasil paling optimal dengan akurasi mencapai 100% pada beberapa skenario pengujian, sedangkan Random Forest tetap memberikan akurasi tinggi, yaitu 99,24%. Temuan ini menunjukkan bahwa SVM lebih unggul dalam mengklasifikasikan sentimen teks pada

isu sensitif di media sosial X. Selain itu, visualisasi word cloud memperlihatkan pola perbedaan penggunaan kata pada sentimen positif dan negatif, yang mencerminkan kecenderungan opini publik terhadap isu tersebut. Secara keseluruhan, penelitian ini menegaskan bahwa algoritma SVM merupakan metode paling efektif digunakan untuk analisis sentimen pada data teks yang telah diseimbangkan menggunakan SMOTE.

Penelitian selanjutnya disarankan untuk menggunakan jumlah data yang lebih besar agar hasil analisis semakin representatif. Selain itu, dapat dipertimbangkan penggunaan algoritma lain seperti metode *deep learning* untuk melihat kemungkinan peningkatan performa. Peningkatan kualitas *preprocessing*, terutama pada tahap pelabelan dan pengayaan kamus sentimen, juga diperlukan agar hasil klasifikasi lebih akurat dan konsisten.

6. Daftar Pustaka

[1] A. A. Salsabila dan H. Nur, "Representasi Diri di Sosial Media: Antara Identitas Nyata dan Identitas Virtual," *PESHUM: Jurnal Pendidikan, Sosial dan Humaniora*, vol. 4, no. 4, hlm. 5601–5620, 2025.

[2] N. W. Putri, "Komunikasi Masa Kini: Transformasi Melalui Media Sosial," *Digihub*, 30 Okt. 2025.

[3] G. Gunawan dan R. B. Aji, "Kontroversi Ijazah Joko Widodo: Antara Tuduhan Palsu dan Fakta Hukum yang Terverifikasi," Law and Humanity, vol. 3, no. 2, hlm. 139-152, Sept. 2025.

[4] R. Fahmi dan P. Aswirna, "The Alleged Fake Degree Certificate of Joko Widodo (Jokowi): Social Movement, Democracy, and Accountability," Universitas Islam Negeri Imam Bonjol, diakses via ResearchGate, Agust. 2025.

[5] N. S. Meilani, "Arsip Sebagai Bukti Hukum Dalam Isu Kasus Dugaan Pemalsuan Ijazah Mantan Presiden Joko Widodo," Jurnal Ilmu Komunikasi, vol. 4, no. 2, hlm. 132–146, 2025.

[6] T. Mufhimah, S. A. Yumnatusta, dan N. A. Rakhmawati, "Analisis Reaksi Masyarakat Indonesia atas Isu Ijazah Jokowi di Media Sosial X Menggunakan K-Means Clustering," *Etika Teknologi Informasi*, vol. 1, no. 2, 2024/2025.

[7] D. Alita, "Implementasi Metode SVM pada Sentimen Analisis terhadap Opini Politik di Twitter seputar Pemilu 2024," *J. Informatika Poltek Harber*, vol. ?, no. ?, 2024.

[8] M. F. Kono, I. N. Fajri, dan Y. Pristyanto, "Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM," *J. Appl. Inform. Comput.*, vol. 9, no. 5, 2025.

[9] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," *J. Ilm. Inform. dan Ilmu Komput.*, vol. 3, no. 1, pp. 46–56, 2024.

[10] P. Cahyani dan L. Abdillah, "Perbandingan Performa Algoritma Support Vector Machine dan Random Forest Studi Kasus Analisis Sentimen Pengguna Sosial Media X," *Kalbiscientia: Jurnal Sains dan Teknologi*, vol. 11, no. 2, 2024, doi: 10.53008/kalbiscientia.v11i02.3624.

[11] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Random Forest dan Support Vector Machine," *J. Transformatika*, vol. 18, no. 1, 2020, doi: 10.26623/transformatika.v18i1.2317.

[12] M. Samantri dan A. Afiyati, "Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terhadap Kebijakan Pemerintah Terkait Kenaikan Harga BBM 2022," *J. Teknol. Inf. dan Komunikasi (JTIK)*, vol. 8, no. 1, 2024, doi: 10.35870/jtik.v8i1.1202.

[13] Y. I. Muasaroh, Z. Fatah, and A. Baijuri, "Analisis Sentimen Komentar YouTube Terhadap Isu Ijazah Presiden Jokowi Menggunakan Support Vector Machine dan Random Forest," in *Prosiding Seminar Nasional (Semnas) 2025 Sekolah Tinggi Teknologi Dumai*, Dumai, 24 Juni 2025, vol. 1, no. 2, pp. xx–xx, ISSN: 2581-267X.

[14] O. M. Wulandari, I. Maulana, F. Syamsudin, and R. Waluyo, "Perbandingan Algoritma Naive Bayes dan SVM dalam Analisis Sentimen Twitter terhadap Isu Ijazah Jokowi Palsu," *JUMISTIK*, vol. 4, no. 1, pp. 392–400, Jun. 2025, doi: 10.70247/jumistik.v4i1.145.

[15] N. P. G. Naraswati, D. C. Rosmilda, D. Desinta, F. Khairi, R. Damaiyanti, dan R. Nooraeni, "Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification," *J. Sistem Informasi*, vol. 10, no. 1, pp. 222-238, Jan. 2021.

[16] A. Safira and F. N. Hasan, "ANALISIS SENTIMEN MASYARAKAT TERHADAP PAYLATER MENGGUNAKAN METODE NAIVE BAYES," ZONAsi (Jurnal Sistem Informasi), vol. 5, no. 1, pp. 69-70, 2023.

[17] B. Ramadhani and R. R. Suryono, "Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 8, no. 2, pp. 714-725, 2024.

[18] K. T. Putra, M. A. Hariyadi and C. Crysdian, "PERBANDINGAN FEATURE EXTRACTION TF-IDF DAN BOW UNTUK ANALISIS SENTIMEN BERBASIS SVM," Jurnal Cahaya MANDALIKA, vol. 3, no. 2, pp. 1449-1463, 2023.

[19] Ridwan, E. H. Hermaliani and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," Computer Science (CO-SCIENCE), vol. 4, no. 1, pp. 80-88, 2024.

[20] T. R. Salsabilla and . N. Pratiwi, "Penerapan Support Vector Machine Untuk Analisis Sentimen pada X (Twitter) Mengenai Obat Penyebab Gagal Ginjal Akut pada Anak," Jurnal Teknik Informatika dan Komputer, vol. 3, no. 2, pp. 67-74, 2024.

[21] S. Amaliah, M. Nusrang and Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng," VARIANSI: Journal of Statistics and Its Application on Teaching and Research, vol. 4, no. 2, pp. 121-127, 2022