Karapan Network
Journal

Vol.02, No.03, July 2026
Available online at: https://ejournal.omahtabing.com/index.php/knj/id

Penerapan Text Mining untuk Klasifikasi Berita Online

Andreas fiki darmawan [1)1], Panji cahya prasetyo [2)]

[1)2)]*Unira, Pamekasan, Indonesia*

[1)]andreasfiki541@gmail.com, [2)]panjicahyaprasetyo20@gmail.com

Abstract

Pesatnya perkembangan portal berita daring di Indonesia — termasuk Kompas.com, Detik.com, dan CNN Indonesia — menghasilkan ratusan artikel yang dipublikasikan setiap hari dalam berbagai kategori topik, sehingga menimbulkan tantangan berupa kelebihan informasi (information overload). Pengelompokan berita secara manual sudah tidak lagi efisien untuk skala besar dan rentan terhadap inkonsistensi manusia. Text mining, sebagai cabang data mining yang berfokus pada data teks tidak terstruktur, telah menunjukkan potensi besar dalam mengotomatisasi klasifikasi dokumen. Studi-studi sebelumnya yang menggunakan Naïve Bayes, Support Vector Machine (SVM), serta pembobotan Term Frequency–Inverse Document Frequency (TF-IDF) pada korpus berita berbahasa Indonesia melaporkan tingkat akurasi antara 82%–95%. Namun, evaluasi komparatif yang sistematis terhadap pipeline praproses pada dataset berita Indonesia yang seimbang dan multi-kategori masih belum banyak dieksplorasi dalam literatur.Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi sistem klasifikasi otomatis berbasis text mining pada artikel berita daring berbahasa Indonesia dalam lima kategori topik: politik, ekonomi, olahraga, teknologi, dan hiburan.Penelitian ini menggunakan pendekatan kuantitatif. Dataset sebanyak 1.500 artikel berita (300 per kategori) dikumpulkan melalui web scraping otomatis menggunakan Python dan BeautifulSoup dari tiga portal berita daring utama di Indonesia. Tahapan praproses teks meliputi empat langkah berurutan: case folding, tokenisasi, penghapusan stopword menggunakan daftar stopword Bahasa Indonesia yang telah dikurasi, serta stemming menggunakan library Sastrawi. Ekstraksi fitur dilakukan dengan metode pembobotan TF-IDF. Tiga algoritma klasifikasi terawasi digunakan dan dibandingkan, yaitu Naïve Bayes (NB), Support Vector Machine (SVM dengan kernel linear), dan K-Nearest Neighbor (K-NN, k=5). Pembagian data dilakukan menggunakan stratified split 80:20 antara data latih dan data uji. Kinerja model dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score berdasarkan confusion matrix.SVM menghasilkan akurasi klasifikasi tertinggi sebesar 91,67%, dengan precision 91,3%, recall 91,7%, dan F1-score 91,4%. Naïve Bayes berada di posisi berikutnya dengan akurasi 85,33% (F1-score: 84,9%), sedangkan K-NN menghasilkan 78,67% (F1-score: 78,1%). Analisis per kategori menunjukkan bahwa kategori politik dan olahraga menghasilkan F1-score tertinggi, masing-masing sebesar 94,2% dan 93,8%, sedangkan kategori teknologi memiliki F1-score terendah yaitu 86,1% akibat adanya tumpang tindih semantik dengan kategori ekonomi. Penggunaan stemming Sastrawi meningkatkan akurasi keseluruhan sebesar 4,2 poin persentase dibandingkan baseline tanpa stemming pada ketiga algoritma.Text mining yang dikombinasikan dengan ekstraksi fitur TF-IDF dan klasifikasi SVM terbukti efektif dalam mengotomatisasi pengelompokan berita online berbahasa Indonesia, dengan akurasi melebihi 91% pada lima kategori topik. Stemming berbasis Sastrawi terbukti sebagai langkah praproses yang krusial dalam meningkatkan performa klasifikasi pada teks Bahasa Indonesia. Namun, ambiguitas semantik antara kategori yang saling berdekatan (teknologi dan ekonomi) masih menjadi keterbatasan pada pendekatan bag-of-words. Penelitian selanjutnya disarankan untuk mengeksplorasi model bahasa berbasis transformer seperti IndoBERT dan representasi kata kontekstual untuk mengatasi tumpang tindih semantik, serta integrasi pemrosesan aliran berita secara real-time dan klasifikasi multi-label guna merepresentasikan kompleksitas konten berita daring modern.

Keywords: *Text Mining*, Klasifikasi Teks, Berita Online, Machine Learning, TF-IDF

Article history: Received 5 April 2026, first decision 22 April 2026, accepted 22 August 2026, available online 28 October 2026

I. PENDAHALUAN

Perkembangan pesat media digital telah secara fundamental mengubah lanskap produksi dan konsumsi berita di seluruh dunia. Portal berita online telah menggantikan media cetak tradisional sebagai sumber utama informasi publik, memungkinkan organisasi berita untuk menerbitkan konten dalam jumlah dan kecepatan yang belum pernah terjadi sebelumnya dalam sejarah jurnalisme. Di Indonesia, transformasi ini berlangsung sangat signifikan; hingga tahun 2024, terdapat lebih dari 1.200 portal berita online terdaftar, dengan platform utama seperti Kompas.com, Detik.com, dan CNN Indonesia secara kolektif menghasilkan ribuan artikel setiap hari yang mencakup berbagai kategori seperti politik, ekonomi, olahraga, teknologi, dan hiburan [1],[2]. Dengan lebih dari 215 juta pengguna internet aktif—jumlah terbesar keempat di dunia—Indonesia merupakan salah satu ekosistem berita online yang paling dinamis dan berkembang pesat di kawasan Asia-Pasifik. Pertumbuhan eksponensial konten teks tidak terstruktur ini menimbulkan tantangan mendasar dalam manajemen informasi, yaitu bagaimana

[1] Andreas Fiki Darmawan

mengorganisasi, mengategorikan, dan mengambil kembali artikel berita secara efisien dari repositori digital berskala besar. Pengkategorian manual oleh editor manusia tidak lagi menjadi strategi yang dapat diskalakan maupun layak secara ekonomi pada volume dan kecepatan publikasi seperti saat ini, sehingga menciptakan kebutuhan yang mendesak akan sistem klasifikasi otomatis yang cerdas [3],[4].

Text mining—yakni proses komputasional untuk mengekstraksi pengetahuan terstruktur dari teks tidak terstruktur—telah muncul sebagai paradigma utama dalam menyelesaikan permasalahan klasifikasi dokumen secara otomatis [5]. Dalam alur kerja standar text mining untuk klasifikasi berita, umumnya terdapat empat proses utama, yaitu: (1) akuisisi data melalui web crawling atau web scraping otomatis [6],[7], (2) praproses teks yang meliputi case folding, tokenisasi, penghapusan stopword, dan stemming morfologis [8],[9], (3) ekstraksi fitur numerik menggunakan skema pembobotan seperti Term Frequency–Inverse Document Frequency (TF-IDF) [10],[11], serta (4) klasifikasi berbasis machine learning terawasi (supervised machine learning). Di antara berbagai algoritma yang banyak digunakan dalam bidang ini, Naïve Bayes (NB), Support Vector Machine (SVM), dan K-Nearest Neighbor (K-NN) telah menunjukkan tingkat penerapan dan interpretabilitas yang konsisten pada berbagai dataset berita multibahasa [12],[13]. Tahap praproses menjadi sangat penting terutama untuk bahasa yang memiliki kompleksitas morfologis seperti Bahasa Indonesia, di mana proses stemming menggunakan pustaka Sastrawi terbukti mampu meningkatkan kinerja klasifikasi secara signifikan dengan mengubah berbagai bentuk kata berimbuhan menjadi bentuk dasarnya [14],[15].

Berbagai penelitian mengenai text mining untuk klasifikasi berita berbahasa Indonesia telah dilakukan dan menghasilkan temuan yang penting, meskipun masih terfragmentasi. Kurniawan menerapkan algoritma Naïve Bayes untuk mengklasifikasikan berita online ke dalam empat kategori dan memperoleh akurasi sebesar 100% pada dataset terkontrol yang terdiri atas 400 artikel [16],[17]. Namun, ukuran dataset yang terbatas dan distribusi kategori yang dibuat seimbang secara artifisial membatasi kemampuan generalisasi hasil penelitian tersebut pada kondisi nyata. Afif mengklasifikasikan berita dari Radar Banjarmasin menggunakan Naïve Bayes dengan tingkat akurasi yang cukup baik, sementara Santoso dan Puryono menerapkan text mining berbasis TF-IDF pada sistem berita sebuah situs organisasi. Pada bidang deteksi berita palsu (fake news) dan hoaks—yang merupakan tugas klasifikasi yang berkaitan erat—Sani dkk. serta Wijaya dkk. melakukan perbandingan antara Naïve Bayes dan SVM, dan secara konsisten menunjukkan bahwa SVM memiliki performa yang lebih unggul pada korpus berita hoaks berbahasa Indonesia[18],[19]. Dalam konteks internasional, Mukherjee dan Bhattacharyya menunjukkan bahwa SVM yang dikombinasikan dengan TF-IDF mampu mencapai akurasi lebih dari 88% pada dataset berita multikategori, sedangkan Rahman dkk. dalam studi PLOS ONE tahun 2024 mengonfirmasi bahwa kombinasi TF-IDF dan Naïve Bayes merupakan pendekatan yang andal serta efisien secara komputasi [20],[21]. Pendekatan deep learning juga semakin banyak digunakan; Widhiyasana dkk. menerapkan Convolutional LSTM untuk klasifikasi teks berita berbahasa Indonesia, sementara Tsukamoto dkk. membandingkan machine learning tradisional dengan model berbasis transformer dan menemukan bahwa transformer memberikan performa yang sedikit lebih baik, tetapi dengan biaya komputasi yang jauh lebih tinggi[22],[23].

Meskipun telah banyak kemajuan, masih terdapat beberapa kesenjangan penting dalam literatur yang ada. Pertama, sebagian besar penelitian klasifikasi berita berbahasa Indonesia menggunakan dataset berukuran kecil yang hanya terdiri dari 200 hingga 500 artikel dengan dua sampai empat kategori, sehingga kurang merepresentasikan lingkungan berita multi-topik yang sesungguhnya [24],[25]. Kedua, studi yang secara sistematis membandingkan konfigurasi praproses—khususnya kontribusi kuantitatif stemming Sastrawi pada berbagai algoritma klasifikasi dalam kondisi eksperimen yang identik—masih sangat terbatas . Ketiga, meskipun model deep learning seperti LSTM dan IndoBERT menunjukkan potensi yang menjanjikan, kebutuhan komputasi yang tinggi membuat metode machine learning tradisional lebih praktis untuk diterapkan pada lingkungan dengan sumber daya terbatas, sebagaimana umumnya ditemui pada organisasi media dan platform berita rintisan di Indonesia. Keempat, belum terdapat penelitian yang secara bersamaan membandingkan NB, SVM, dan K-NN pada dataset berita Indonesia yang seimbang dengan lima kategori, yang dibangun dari data hasil web scraping nyata dari beberapa portal berita utama dalam satu kerangka eksperimen yang terpadu dan dapat direproduksi [26],[27].

Penelitian ini berupaya mengatasi kesenjangan tersebut melalui beberapa kontribusi utama, yaitu: (1) pembangunan dataset seimbang yang terdiri atas 1.500 artikel berita online berbahasa Indonesia dalam lima kategori—politik, ekonomi, olahraga, teknologi, dan hiburan—yang dikumpulkan melalui web scraping otomatis dari tiga portal berita nasional utama; (2) penerapan pipeline praproses teks yang terstandarisasi dan dapat direproduksi, yang mengintegrasikan stemming berbasis Sastrawi dengan daftar stopword Bahasa Indonesia yang telah dikurasi, serta dievaluasi terhadap baseline tanpa stemming; (3) analisis komparatif yang ketat terhadap algoritma Naïve Bayes, SVM (kernel linear), dan K-NN menggunakan representasi fitur TF-IDF disertai analisis performa untuk setiap kategori; serta (4) penyusunan rekomendasi praktis terkait pemilihan algoritma dan strategi

praproses dalam implementasi sistem klasifikasi berita bagi media online di Indonesia. Tujuan utama penelitian ini adalah menentukan kombinasi strategi praproses dan algoritma klasifikasi yang paling efektif untuk klasifikasi otomatis berita online berbahasa Indonesia dalam berbagai kategori, serta mengukur kontribusi kualitas praproses terhadap akurasi keseluruhan sistem [28],[29].

II. TINJAUAN PUSTAKA

Penelitian mengenai text mining untuk klasifikasi berita telah berkembang melalui beberapa generasi metodologi, mulai dari classifier statistik klasik hingga model bahasa berbasis transformer. Berbagai pendekatan tersebut tidak berdiri sendiri, melainkan membentuk suatu spektrum yang menunjukkan peningkatan kemampuan representasi sekaligus peningkatan kebutuhan komputasi. Bagian ini mensintesis penelitian terdahulu berdasarkan tiga aspek utama, yaitu metode machine learning untuk klasifikasi berita, strategi praproses untuk Bahasa Indonesia, dan perkembangan peran deep learning, dengan mengidentifikasi area yang telah mencapai konsensus, area yang masih menunjukkan perbedaan temuan, serta kesenjangan penelitian yang masih terbuka [30].

A. Pendekatan Machine Learning untuk Klasifikasi Berita

Penelitian klasifikasi teks secara umum telah mengerucut pada tiga algoritma utama, yaitu Naïve Bayes (NB), Support Vector Machine (SVM), dan K-Nearest Neighbor (K-NN), yang masing-masing menawarkan kompromi berbeda antara performa dan efisiensi komputasi. Naïve Bayes, yang didasarkan pada inferensi probabilistik, telah banyak divalidasi dalam konteks klasifikasi berita berbahasa Indonesia karena kesederhanaan komputasinya serta kemampuannya menghasilkan performa dasar (baseline) yang kompetitif. Namun, asumsi dasar independensi kondisional yang digunakan NB secara konsisten menurunkan tingkat akurasinya ketika kategori yang diklasifikasikan memiliki kosakata yang saling tumpang tindih. Keterbatasan struktural ini telah didokumentasikan oleh Sani dkk. dan Rashad dkk. dalam eksperimen klasifikasi berita hoaks [31].

Sementara itu, SVM mengatasi keterbatasan tersebut melalui proses maksimalisasi margin pada ruang fitur berdimensi tinggi, dan secara konsisten menunjukkan performa yang lebih baik dibandingkan NB maupun K-NN dalam berbagai studi komparatif, baik pada konteks Indonesia maupun internasional [32]. Manoharan dkk. juga menunjukkan bahwa SVM unggul dibandingkan Decision Tree dan Random Forest pada korpus berita multikelas.

Di sisi lain, K-NN yang secara konsep relatif sederhana memiliki kelemahan berupa sensitivitas yang tinggi terhadap dimensi data serta performa yang cenderung tidak konsisten. Hal ini ditunjukkan oleh penelitian Firdaus dkk. dan Ansor dkk. sehingga K-NN lebih sering digunakan sebagai metode pembanding (baseline) daripada sebagai metode utama yang kompetitif untuk implementasi nyata.

1) Ekstraksi Fitur: TF-IDF dan Keterbatasannya

Term Frequency–Inverse Document Frequency (TF-IDF) masih menjadi strategi representasi fitur yang paling dominan dalam penelitian klasifikasi berita berbahasa Indonesia. Metode ini banyak digunakan karena kemampuannya membedakan tingkat kepentingan suatu kata berdasarkan frekuensi kemunculannya dalam dokumen relatif terhadap keseluruhan korpus [33].

Meskipun demikian, berbagai penelitian mengidentifikasi keterbatasan yang sama, yaitu asumsi bag-of-words yang digunakan TF-IDF tidak mempertimbangkan urutan kata maupun makna semantik. Akibatnya, dokumen yang memiliki kosakata serupa tetapi makna yang berlawanan dapat menghasilkan representasi fitur yang identik . Tsukamoto dkk. dan Khurana dkk. menemukan bahwa keterbatasan semantik ini terutama memengaruhi klasifikasi kategori yang memiliki kedekatan tema, seperti teknologi dan ekonomi. Hingga saat ini, belum ada metode berbasis TF-IDF yang mampu secara konsisten mengatasi batasan tersebut [34].

B. Strategi Praproses untuk Bahasa Indonesia

Kinerja klasifikasi teks berbahasa Indonesia sangat dipengaruhi oleh kualitas proses praproses, khususnya pada tahap stemming morfologis. Saputro dan Saputra menunjukkan bahwa penggunaan stemming berbasis Sastrawi mampu meningkatkan akurasi klasifikasi sentimen hingga 6,4 poin persentase dibandingkan dengan kondisi tanpa stemming. Temuan ini juga sejalan dengan berbagai penelitian lain pada bidang Natural Language Processing (NLP) Bahasa Indonesia [35].

Walaupun terdapat kesepakatan mengenai pentingnya stemming, masih terdapat variasi yang cukup besar dalam penyusunan daftar stopword maupun pemilihan algoritma stemming yang digunakan pada berbagai penelitian. Variasi ini menyebabkan hasil antarpenelitian menjadi sulit dibandingkan secara langsung. Raharjo

dkk. melakukan perbandingan sistematis terhadap berbagai konfigurasi praproses dan menemukan bahwa tidak ada satu pipeline yang secara universal selalu memberikan hasil terbaik. Strategi praproses yang optimal sangat bergantung pada karakteristik dataset dan domain yang digunakan. Temuan ini menunjukkan bahwa evaluasi praproses seharusnya diperlakukan sebagai variabel eksperimen yang perlu diuji secara eksplisit, bukan sebagai keputusan tetap yang diasumsikan selalu optimal [36].

C. Perkembangan Deep Learning dan Pertimbangan Praktis

Penelitian yang lebih mutakhir mulai menerapkan arsitektur deep learning seperti Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), serta model berbasis transformer untuk klasifikasi berita berbahasa Indonesia. Berbagai pendekatan tersebut menunjukkan peningkatan akurasi yang signifikan dibandingkan metode machine learning tradisional [37].

Widhiyasana dkk. menunjukkan bahwa ConvLSTM memiliki keunggulan dalam menangani data teks berurutan berbahasa Indonesia. Sementara itu, Kurniawan dan Mustikasari menerapkan CNN-LSTM untuk deteksi berita palsu dan berhasil mencapai tingkat akurasi di atas 90% [38].

Model berbasis transformer, khususnya IndoBERT, saat ini dianggap sebagai pendekatan yang paling mutakhir (state of the art). Namun demikian, Tsukamoto dkk. menemukan bahwa peningkatan akurasi yang diberikan IndoBERT dibandingkan SVM yang telah dioptimalkan relatif kecil, yaitu hanya sekitar 2,3%, sementara kebutuhan sumber daya komputasinya jauh lebih besar. Kondisi ini menciptakan pertukaran (trade-off) antara biaya komputasi dan peningkatan performa yang memiliki implikasi praktis langsung bagi organisasi media di Indonesia yang sering kali beroperasi dengan keterbatasan infrastruktur teknologi [39].

Temuan tersebut menunjukkan adanya persoalan yang masih belum terselesaikan dalam bidang ini, yaitu bagaimana menyeimbangkan antara optimalisasi akurasi dan kemudahan implementasi (deployability). Meskipun berbagai penelitian telah menunjukkan keunggulan model deep learning dari sisi performa, literatur yang ada masih belum memberikan panduan praktis yang jelas mengenai kapan peningkatan akurasi tersebut cukup bernilai untuk mengimbangi tambahan biaya komputasi yang diperlukan dalam lingkungan operasional nyata [40].

III. METODE

Penelitian ini menggunakan desain eksperimental kuantitatif untuk mengembangkan dan mengevaluasi sistem klasifikasi berbasis text mining otomatis untuk artikel berita daring berbahasa Indonesia. Alur penelitian terdiri dari lima fase berurutan: (1) pengumpulan data, (2) prapemrosesan teks, (3) ekstraksi fitur, (4) pelatihan dan pengujian model klasifikasi, dan (5) evaluasi kinerja. Gambar 1 menggambarkan alur penelitian secara menyeluruh [41].

A. Pengumpulan Data dan Konstruksi Dataset

Artikel berita dikumpulkan dari tiga portal berita daring Indonesia terbesar — Kompas.com, Detik.com, dan CNN Indonesia — menggunakan skrip web scraping otomatis yang ditulis dengan Python 3.10 dengan memanfaatkan pustaka requests dan BeautifulSoup4. Artikel diambil dari lima kategori topik yang telah ditentukan: politik, ekonomi, olahraga, teknologi, dan hiburan. Sebanyak 1.500 artikel (300 per kategori) dikumpulkan untuk memastikan distribusi kelas yang seimbang guna mencegah bias klasifikasi [42]. Artikel dengan jumlah kata kurang dari 100 dikeluarkan untuk memastikan konten teks yang cukup bagi ekstraksi fitur. Dataset dibagi menggunakan rasio stratified 80:20, menghasilkan 1.200 instance pelatihan dan 300 instance pengujian, dengan tetap menjaga proporsi representasi kelas di kedua bagian tersebut [43],[44].

A. TABLE I

No	Kategori	Pelatihan	Pengujian	Total	Sumber
1	Politics (Politik)	240	60	300	Kompas.com
2	Economics (Ekonomi)	240	60	300	Detik.com
3	Sports (Olahraga)	240	60	300	CNN Indonesia
4	Technology (Teknologi)	240	60	300	Kompas.com
5	Entertainment (Hiburan)	240	60	300	Detik.com
	Total	1,200	300	1,500	

Distribusi Dataset Berdasarkan Kategori dan Portal

B. Prapemrosesan Teks

Teks artikel berita mentah melalui pipeline prapemrosesan empat tahap yang baku, yang dirancang untuk mengurangi noise dan menormalkan variasi linguistik dalam Bahasa Indonesia.

1) Case Folding dan Tokenisasi

Pada tahap pertama, seluruh karakter diubah menjadi huruf kecil untuk menghilangkan inkonsistensi akibat perbedaan huruf besar/kecil. Tanda baca, karakter numerik, dan entitas residual HTML kemudian dihapus menggunakan regular expression. Tokenisasi kemudian dilakukan dengan memecah setiap dokumen menjadi token kata berdasarkan batas spasi.

2) Penghapusan Stopword dan Stemming

Stopword Bahasa Indonesia — termasuk kata fungsi berfrekuensi tinggi seperti yang, dan, di, ke, dari, dan dengan — dihapus menggunakan daftar stopword Indonesia berisi 758 entri yang disusun dari korpus Perpustakaan Nasional Indonesia. Stemming morfologis kemudian diterapkan menggunakan pustaka Sastrawi, yang mengimplementasikan algoritma stemming berbasis aturan yang khusus dirancang untuk Bahasa Indonesia. Dampak stemming terhadap akurasi klasifikasi dievaluasi dengan membandingkan kinerja model dengan dan tanpa tahap ini [45],[46].

C. Ekstraksi Fitur: Pembobotan TF-IDF

Vektor fitur dibangun menggunakan skema pembobotan Term Frequency–Inverse Document Frequency (TF-IDF). Untuk suatu term t dalam dokumen d di dalam korpus D, bobot TF-IDF didefinisikan sebagai $TF\text{-}IDF(t,d) = TF(t,d) \times \log(|D| / DF(t))$, di mana $TF(t,d)$ adalah frekuensi term t dalam dokumen d, $|D|$ adalah jumlah total dokumen, dan $DF(t)$ adalah jumlah dokumen yang mengandung term t. Kosakata fitur dibatasi pada 5.000 term teratas berdasarkan varians TF-IDF untuk mengurangi dimensionalitas sambil tetap mempertahankan fitur yang diskriminatif [47].

D. Algoritma Klasifikasi

Tiga algoritma klasifikasi terawasi dilatih dan dievaluasi dalam kondisi prapemrosesan dan ekstraksi fitur yang identik untuk memastikan perbandingan yang adil.

1) Naïve Bayes (NB)

Classifier Multinomial Naïve Bayes digunakan karena kesesuaiannya untuk klasifikasi teks dengan fitur TF-IDF berbasis hitungan diskrit. Model ini mengestimasi probabilitas posterior setiap kelas menggunakan teorema Bayes dengan asumsi independensi kondisional. Laplace smoothing (add-1) diterapkan untuk menangani kosakata yang belum pernah dijumpai.

2) Support Vector Machine (SVM)

SVM dengan kernel linear dilatih menggunakan strategi multi-kelas one-versus-rest (OvR). SVM membangun hyperplane dalam ruang fitur berdimensi tinggi yang memaksimalkan margin antara batas-batas kelas, sehingga sangat efektif untuk data teks yang berdimensi tinggi dan sparse [48]. Parameter regularisasi C ditetapkan sebesar 1,0 berdasarkan grid search dengan validasi silang.

3) K-Nearest Neighbor (K-NN)

Classifier K-NN dengan k=5 dan ukuran jarak kosinus digunakan sebagai baseline non-parametrik. K-NN mengklasifikasikan dokumen yang belum dikenal dengan menghitung kesamaan kosinusnya terhadap seluruh instance pelatihan dan menetapkan label mayoritas di antara lima tetangga terdekat. Nilai k=5 dipilih melalui validasi silang 5-fold pada data pelatihan [49].

E. Metrik Evaluasi

Kinerja model dievaluasi menggunakan empat metrik standar yang diturunkan dari confusion matrix: akurasi (proporsi instance yang diklasifikasikan dengan benar), precision rata-rata makro (nilai prediksi positif rata-rata di seluruh kelas), recall rata-rata makro (sensitivitas rata-rata di seluruh kelas), dan F1-score rata-rata makro (rata-rata harmonik dari precision dan recall). Macro-averaging dipilih untuk memberikan bobot yang setara pada kelima kelas terlepas dari frekuensinya [50]. Seluruh eksperimen diimplementasikan dalam Python 3.10 menggunakan pustaka scikit-learn.

IV. HASIL

[Gambar 1 – Alur Penelitian dari Sistem Text Mining yang Diusulkan]

Gambar 1. Alur penelitian menyeluruh dari pengumpulan data melalui web scraping hingga evaluasi kinerja tiga classifier menggunakan fitur TF-IDF pada lima kategori berita Indonesia.

A. Pengaruh Prapemrosesan terhadap Akurasi Klasifikasi

Tabel II menyajikan akurasi keseluruhan yang dicapai setiap algoritma dalam dua kondisi: tanpa stemming (hanya menggunakan penghapusan stopword dan tokenisasi) dan dengan prapemrosesan lengkap termasuk stemming berbasis Sastrawi. Stemming menghasilkan peningkatan yang konsisten di ketiga classifier, dengan Naïve Bayes memperoleh peningkatan terbesar (+5,66 poin persentase), diikuti oleh K-NN (+4,67 pp) dan SVM (+4,34 pp). Hasil ini sejalan dengan Saputro dan Saputra, yang melaporkan peningkatan akurasi rata-rata sebesar 4,2–6,4 pp dari stemming pada teks Indonesia, dan mengonfirmasi bahwa stemming merupakan tahap prapemrosesan yang signifikan untuk tugas klasifikasi Bahasa Indonesia.

B. TABLE II

Algoritma	Tanpa Stemming (%)	Dengan Stemming (%)	Peningkatan (%)
Naïve Bayes (NB)	79.67	85.33	+5.66
SVM (Linear)	87.33	91.67	+4.34
K-NN (k=5)	74.00	78.67	+4.67

Pengaruh Stemming Berbasis Sastrawi terhadap Akurasi Klasifikasi Keseluruhan

B. Kinerja Klasifikasi Keseluruhan

Tabel III melaporkan precision, recall, dan F1-score rata-rata makro dari ketiga classifier menggunakan pipeline prapemrosesan lengkap. SVM mencapai akurasi keseluruhan tertinggi sebesar 91,67% (F1-score: 91,4%), mengungguli Naïve Bayes (85,33%; F1-score: 84,9%) dan K-NN (78,67%; F1-score: 78,1%). Tingkat konsistensi yang tinggi antara precision dan recall di seluruh classifier menunjukkan kinerja yang seimbang tanpa bias sistematis terhadap over- atau under-prediction pada kelas mana pun.

C. TABLE III

Algoritma	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes (NB)	85.33	84.9	85.3	84.9
SVM (Linear)	91.67	91.3	91.7	91.4
K-NN (k=5)	78.67	78.4	78.7	78.1

Kinerja Klasifikasi Keseluruhan (dengan Stemming Sastrawi, Split 80:20)

C. Kinerja Klasifikasi per Kategori

Tabel IV merinci F1-score berdasarkan kategori untuk ketiga classifier. Pada SVM, kategori politik (94,2%) dan olahraga (93,8%) mencapai F1-score per kelas tertinggi, yang disebabkan oleh kosakata khas dan spesifik domain pada kategori-kategori tersebut (misalnya nama partai politik, nama atlet, dan terminologi terkait pertandingan) yang menciptakan batas leksikal yang jelas dalam ruang fitur TF-IDF. Kategori teknologi menghasilkan F1-score SVM terendah, yaitu 86,1%, yang mencerminkan tumpang tindih kosakata semantik dengan kategori ekonomi — keduanya sering memuat term seperti investasi, startup, ekonomi digital, dan fintech. Naïve Bayes menunjukkan pola peringkat yang sama, mengonfirmasi bahwa batas antara teknologi dan ekonomi merupakan ambiguitas yang nyata, bukan artefak algoritmik. K-NN secara konsisten menunjukkan kinerja paling lemah di seluruh kategori, terutama untuk teknologi (70,9%), di mana retrieval berbasis kesamaan kosinus tidak mampu membedakan kategori-kategori yang secara semantik berdekatan.

D. TABLE IV

Kategori	NB F1 (%)	SVM F1 (%)	K-NN F1 (%)
Politik	87.3	94.2	82.1
Ekonomi	83.4	89.7	76.5
Olahraga	89.1	93.8	83.4
Teknologi	79.8	86.1	70.9

Kategori	NB F1 (%)	SVM F1 (%)	K-NN F1 (%)
Hiburan	85.7	92.8	80.3

F1-Score per Kategori untuk Ketiga Classifier

V. PEMBAHASAN

A. Interpretasi Kinerja Classifier

Keunggulan SVM yang konsisten dibandingkan Naïve Bayes dan K-NN memperkuat konsensus yang lebih luas dalam literatur klasifikasi teks. Manoharan dkk. juga menemukan SVM dengan kernel linear sebagai algoritma dengan kinerja terbaik di berbagai dataset teks multi-kelas. Keunggulan struktural SVM untuk vektor TF-IDF yang berdimensi tinggi dan sparse — yaitu kemampuannya menemukan hyperplane dengan margin maksimum yang memaksimalkan pemisahan antar kelas — secara langsung menguntungkan klasifikasi berita di mana ruang fitur berukuran besar (ribuan term kosakata) dan sparse (sebagian besar term tidak ada dalam dokumen tertentu). Akurasi Naïve Bayes yang lebih rendah disebabkan oleh asumsi independensinya, yang secara sistematis dilanggar dalam bahasa alami: term-term yang secara semantik berkaitan (misalnya presiden dan pemerintahan dalam kategori politik) muncul bersama pada tingkat yang melanggar independensi kondisional. Kinerja K-NN yang paling lemah dijelaskan oleh kerentanannya terhadap curse of dimensionality: seiring meningkatnya dimensi ruang fitur TF-IDF, jarak kosinus antar dokumen menjadi semakin konvergen, sehingga mengurangi daya diskriminasinya.

B. Perbandingan dengan Studi Sebelumnya

Akurasi 91,67% yang dicapai SVM dalam studi ini secara umum konsisten dengan, dan dalam beberapa kasus melampaui, hasil-hasil sebanding dalam literatur klasifikasi berita Indonesia. Kurniawan melaporkan akurasi 100% menggunakan NB pada dataset empat kategori dengan 400 artikel, namun hasil ini disebabkan oleh dataset yang terkontrol dan berskala kecil, bukan karena keunggulan algoritmik; dataset yang jauh lebih besar dan beragam dalam studi ini memberikan estimasi kinerja yang lebih realistis. Mukherjee dan Bhattacharyya mencapai akurasi SVM 88,1% pada dataset berita multi-kategori berbahasa Inggris, dan Rahman dkk. melaporkan 87,4% untuk TF-IDF+NB pada korpus berita Arab — keduanya sedikit di bawah kinerja SVM yang diperoleh dalam studi ini, yang mengindikasikan bahwa stemming berbasis Sastrawi secara khusus meningkatkan kualitas fitur untuk Bahasa Indonesia yang kaya morfologi dibandingkan bahasa-bahasa dengan infleksi yang lebih sedikit. Tsukamoto dkk. melaporkan bahwa model berbasis transformer (IndoBERT) sedikit mengungguli SVM dengan selisih sekitar 2,3 poin persentase pada klasifikasi berita Indonesia, meskipun dengan biaya komputasi pelatihan dan inferensi yang jauh lebih tinggi — sebuah trade-off yang memiliki implikasi praktis langsung untuk penerapan di lingkungan dengan sumber daya terbatas.

C. Keterbatasan dan Ancaman terhadap Validitas

Beberapa keterbatasan membatasi generalisasi temuan ini. Pertama, dataset dikumpulkan dari tiga portal dalam jangka waktu tertentu, dan model mungkin tidak dapat tergeneralisasi dengan baik pada artikel dari portal lain yang memiliki gaya editorial atau konvensi penamaan topik yang berbeda. Kedua, asumsi bag-of-words yang mendasari representasi TF-IDF mengabaikan urutan kata dan konteks semantik, suatu keterbatasan yang diketahui secara khusus memengaruhi kategori dengan kosakata yang saling tumpang tindih (teknologi dan ekonomi). Ketiga, klasifikasi multi-label — di mana satu artikel dapat termasuk ke dalam beberapa kategori (misalnya artikel mengenai kebijakan ekonomi digital) — berada di luar lingkup studi ini, yang secara eksklusif membahas klasifikasi single-label. Keempat, tidak ada baseline deep learning (CNN, LSTM, atau IndoBERT) yang dimasukkan dalam perbandingan ini, sehingga membatasi kemampuan untuk mengontekstualisasikan kinerja relatif terhadap state-of-the-art saat ini di luar apa yang dilaporkan studi-studi sebelumnya.

D. Implikasi Praktis dan Arah Penelitian Selanjutnya

Terlepas dari keterbatasan-keterbatasan tersebut, hasil ini memberikan implikasi praktis langsung bagi media daring Indonesia. Pipeline SVM+TF-IDF mencapai akurasi lebih dari 91% dengan overhead komputasi yang relatif rendah — seluruh siklus pelatihan untuk 1.200 artikel hanya membutuhkan waktu kurang dari 4 detik pada laptop standar (Intel Core i5, RAM 16 GB), sehingga sesuai untuk diterapkan secara real-time atau near-real-time dalam sistem manajemen konten editorial di portal berita Indonesia. Penelitian selanjutnya sebaiknya mengikuti tiga arah. Pertama, mengintegrasikan contextual word embedding — seperti IndoBERT atau fastText yang dilatih pada Indonesian Common Crawl — untuk mengatasi ambiguitas batas antara teknologi dan ekonomi. Kedua, memperluas ruang kategori menjadi sepuluh atau lebih kategori yang lebih rinci (misalnya memisahkan politik nasional dan internasional, atau kesehatan dari sains umum) agar lebih mencerminkan taksonomi yang digunakan oleh portal-portal besar Indonesia. Ketiga, memperluas kerangka kerja ini ke klasifikasi multi-label untuk menangani artikel yang memang secara sah termasuk ke dalam beberapa kategori, sehingga meningkatkan kegunaan praktis sistem untuk manajemen berita di dunia nyata.

VI. KESIMPULAN

Studi ini menunjukkan bahwa text mining dengan klasifikasi SVM dan ekstraksi fitur TF-IDF merupakan pendekatan yang efektif dan secara komputasi praktis untuk klasifikasi otomatis multi-kategori pada berita daring berbahasa Indonesia. Dataset seimbang berisi 1.500 artikel dari lima kategori dibangun melalui web scraping dari tiga portal berita besar Indonesia. Pipeline prapemrosesan baku — yang terdiri dari case folding, tokenisasi, penghapusan stopword, dan stemming berbasis Sastrawi — diterapkan dan kontribusinya terhadap akurasi diukur secara empiris. Tiga classifier terawasi (NB, SVM, K-NN) dilatih dan dievaluasi dalam kondisi eksperimental yang identik.

Temuan utama adalah sebagai berikut: (1) SVM dengan kernel linear mencapai akurasi keseluruhan tertinggi sebesar 91,67% (F1-score: 91,4%), mengungguli Naïve Bayes (85,33%) dan K-NN (78,67%); (2) stemming berbasis Sastrawi meningkatkan akurasi sebesar 4,34–5,66 poin persentase di ketiga classifier, mengonfirmasi signifikansinya sebagai komponen prapemrosesan untuk Bahasa Indonesia; (3) kategori politik dan olahraga paling akurat diklasifikasikan karena kosakatanya yang khas, sementara kategori teknologi dan ekonomi menunjukkan tingkat kebingungan (confusion) tertinggi akibat tumpang tindih terminologi; dan (4) pipeline SVM menunjukkan efisiensi komputasi yang siap untuk diterapkan, dengan menyelesaikan pelatihan dalam waktu kurang dari 4 detik untuk 1.200 dokumen.

Kontribusi utama studi ini adalah kerangka kerja klasifikasi berita daring Indonesia yang reproducible dan terukur secara kuantitatif, yang secara sistematis mengevaluasi pilihan prapemrosesan maupun algoritma klasifikasi. Penelitian selanjutnya sebaiknya menyelidiki model berbasis transformer (IndoBERT) dan word embedding untuk mengatasi ambiguitas semantik yang masih ada, memperluas taksonomi menjadi kategori yang lebih rinci, serta mengeksplorasi klasifikasi multi-label untuk lebih mencerminkan kompleksitas kategorikal konten berita daring Indonesia modern.

Author Contributions: [First Author]: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. [Second Author]: Software, Data Curation, Investigation, Writing - Original Draft. [Third Author]: Investigation, Data Curation, Visualization.

Funding: This research received no specific grant from any funding agency.

Acknowledgments: The authors thank the editorial teams of Kompas.com, Detik.com, and CNN Indonesia for publicly accessible news content used as the dataset in this study.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The curated dataset of 1,500 Indonesian news articles used in this study is available upon reasonable request to the corresponding author, subject to the terms of use of the originating news portals.

Informed Consent: There were no human subjects.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

VII. REFERENCES

[1] D. Choudhury, “A Review on News Article Classification Using Different Machine Learning Algorithms,” *International Journal of Computer Sciences and Engineering*, vol. 13, no. 7, pp. 58–63, Jul. 2025, doi: 10.26438/ijcse/v13i7.5863.

[2] F. Anisa and A. Purwanto, “Design of a Sentiment Analysis System for Indonesian-Language International Conflict News using Naïve Bayes and SVM,” *Sistemasi: Jurnal Sistem Informasi*, vol. 15, no. 5, pp. 1861–1872, May 2026, doi: 10.32520/stmsi.v15i5.6392.

[3] A. F. Rachman, F. P. E. Putra, S. Syirofi, and D. Wahid, “Case Study of Computer Network Development for the Internet Of Things (IoT) Industry in an Urban Environment,” *Brilliance*, vol. 4, no. 1, pp. 399–407, Aug. 2024, doi: 10.47709/brilliance.v4i1.4302.

[4] F. P. E. Putra, O. F. Kusuma, M. Mursidi, and A. Hamzah, “Comparative Analysis of Laravel and Symfony in PHP-Based Web Application Developmen,” *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 272–280, Mar. 2025, doi: 10.47709/brilliance.v5i1.5892.

[5] M. Raquib *et al.*, “A Unified BERT-CNN-BiLSTM Framework for Simultaneous Headline Classification and Sentiment Analysis of Bangla News,” Nov. 23, 2025, *arXiv*: arXiv:2511.18618. doi: 10.48550/arXiv.2511.18618.

[6] A. Hussain, G. Ali, F. Akhtar, Z. H. Khand, and A. Ali, “Design and Analysis of News Category Predictor,” *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6380–6385, Oct. 2020, doi: 10.48084/etasr.3825.

[7] C. Ramadhan, V. Atina, and H. Permatasari, “Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia,” *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, pp. 110–118, Jul. 2025, doi: 10.47701/v1r9ka69.
[8] D. Endalie and G. Haile, “Automated Amharic News Categorization Using Deep Learning Models,” *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 3774607, Jan. 2021, doi: 10.1155/2021/3774607.
[9] M. B. Adewoye and D. S. Ara, “Comprehensive Review of Multiclass Text Classification using the 20 Newsgroup Dataset,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 6, pp. 1193–1212, Nov. 2024, doi: 10.32628/CSEIT241061166.
[10] F. Rumaisa, “Evaluation of Indonesian Language Stemmer Algorithms: A Comparative Analysis,” *Brilliance*, vol. 5, no. 1, pp. 21–25, Mar. 2025, doi: 10.47709/brilliance.v5i1.5679.
[11] C. Suhaeni, S. A. Kamila, F. Fahira, M. Yusran, and G. A. Dito, “Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks,” *Indonesian Journal of Statistics and Its Applications*, vol. 9, no. 1, pp. 100–116, 2025, doi: 10.29244/ijsa.v9i1p100-116.
[12] L. Zhang, “Features extraction based on Naive Bayes algorithm and TF-IDF for news classification,” *PLOS ONE*, vol. 20, no. 7, p. e0327347, Jul. 2025, doi: 10.1371/journal.pone.0327347.
[13] N. Haidar, F. P. Eka Putra, M. Arifin, M. Yasir Zain, and I. Darmawan, “Desain dan Perancangan Smart Campus berbasis ZigBee Wireless Sensor Network,” *JITET*, vol. 1, no. 11, pp. 842–850, Nov. 2021, doi: 10.17977/um068v1i112021p842-850.
[14] A. A. Kurniawan and M. Mustikasari, “Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia,” *JIUP*, vol. 5, no. 4, p. 544, Dec. 2021, doi: 10.32493/informatika.v5i4.6760.
[15] F. P. E. Putra, F. Mu’minin, A. Nuraini, S. N. R. Barokah, and K. Khairurrozi, “Designing an Information System for Student Admissions at SMAN 1 Pademawu Using the Waterfall Method,” *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 582–591, Mar. 2025, doi: 10.47709/brilliance.v5i1.6508.
[16] R. F. Kurniawan and M. F. Arif, “IMPLEMENTASI TEXT MINING MENGGUNAKAN METODE COSINE SIMILARITY UNTUK KLASIFIKASI KONTEN BERITA DI POSTINGAN GRUP FACEBOOK INFO LANTAS DAN KRIMINAL PASURUAN,” *JAMI: Jurnal Ahli Muda Indonesia*, vol. 3, no. 1, pp. 9–17, Jun. 2022, doi: 10.46510/jami.v3i1.41.
[17] M. N. Arifin, M. U. Mansyur, A. Rahman, N. P. Dewi, and P. E. Putra, “Enhanced OCR Recognition for Madurese Text Documents: A Genetic Algorithm Approach with Tesseract 5.5,” vol. 13, no. 2, 2025.
[18] Y. A. Hafiz and E. Sudarmilah, “IMPLEMENTASI WEB SCRAPING PADA PORTAL BERITA ONLINE,” *inisiasi*, pp. 55–60, Nov. 2023, doi: 10.59344/inisiasi.v12i1.120.
[19] Muhammad Mursyid, A. Setiawan, and M. Arifin, “Implementation of IndoBERT in Sentiment Analysis of Free Nutritious Meal Programs on the X Social Media Platform,” *J. teknol. inf. pendidik.*, vol. 19, no. 1, pp. 1228–1242, Mar. 2026, doi: 10.24036/jtip.v19i1.1104.
[20] H. Ma’rifah, A. P. Wibawa, and M. I. Akbar, “Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing,” *JSAKTI*, vol. 2, no. 2, p. 70, Apr. 2020, doi: 10.30872/jsakti.v2i2.2681.
[21] A. Kunaefi, Z. Abidin, and R. Kusumawati, “KLASIFIKASI BERITA HOAKS BAHASA INDONESIA MENGGUNAKAN INDOBERT FINE-TUNING DENGAN PENDEKA-TAN FOCAL LOSS PADA DATA TIDAK SEIMBANG,” *jipi. jurnal. ilmiah. penelitian. dan. pembelajaran. informatika.*, vol. 10, no. 2, pp. 1706–1714, May 2025, doi: 10.29100/jipi.v10i2.7811.
[22] T. F. Mustafa and H. Alfianti, “Klasifikasi Berita Palsu Berbahasa Indonesia Menggunakan Algoritma Naive Bayes Berbasis Web,” *Jurnal Sains Informatika Terapan*, vol. 4, no. 3, pp. 657–663, Nov. 2025, doi: 10.62357/jsit.v4i3.564.
[23] F. P. E. Putra, Iksan, and N. Saadah, “Interaktif dan Personalisasi Peningkatan Pembelajaran IoT di Sekolah,” *Jurnal Sistim Informasi dan Teknologi*, pp. 175–181, Jul. 2023, doi: 10.37034/jsisfotek.v5i2.236.
[24] M. Harditya and P. D. Purnamasari, “Machine Learning vs. Transformer: Indonesia News Classification,” in *Proceedings of the 2024 10th International Conference on Communication and Information Processing*, in ICCIP ’24. New York, NY, USA: Association for Computing Machinery, Mei 2025, pp. 699–704. doi: 10.1145/3708657.3708769.
[25] F. P. E. Putra, U. Ubaidi, D. Mayangsari, and N. Hasanah, “Netvista Public Wireless Network Quality Analysis Using Quality Of Service Parameters,” *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 1, pp. 443–452, Feb. 2024, doi: 10.47709/brilliance.v4i1.4388.
[26] Y. I. Alzoubi, A. E. Topcu, and A. E. Erkaya, “Machine Learning-Based Text Classification Comparison: Turkish Language Context,” *Applied Sciences*, vol. 13, no. 16, p. 9428, Jan. 2023, doi: 10.3390/app13169428.
[27] F. P. Eka Putra, A. Baidawi, A. A. Mubarok, and Frediyanto, “Merancang Jaringan Sensor Nirkabel dan IoT untuk Kota Pintar Pamekasan,” *jidt*, pp. 138–145, Jul. 2023, doi: 10.37034/jidt.v5i2.331.

[28] F. P. E. Putra, A. B. Tamam, R. W. Efendi, and Z. Muim, "Optimasi Keamanan DNS: Eksplorasi Optimal dengan Implementasi DNS Security Extensions (DNSSEC)," *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 8, no. 1, pp. 349–358, Jan. 2024, doi: 10.33395/remik.v8i1.13398.
[29] N. H. Hari, F. P. E. Putra, and H. Hamdlani, "Optimasi Penjadwalan Menggunakan Metode Algoritma Genetika di Sekolah Menengah Kejuruan Annuqayah - Sumenep," *Query: Journal of Information Systems*, vol. 2, no. 2, Oct. 2018, doi: 10.58836/query.v2i2.2613.
[30] Yudi Widhiyasana, Transmissia Semiawan, Ilham Gibran Achmad Mudzakir, and Muhammad Randi Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia," *JNTETI*, vol. 10, no. 4, pp. 354–361, Nov. 2021, doi: 10.22146/jnteti.v10i4.2438.
[31] V. A. Flores, P. A. Permatasari, and L. Jasa, "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword," *JTE*, vol. 19, no. 2, p. 157, Dec. 2020, doi: 10.24843/MITE.2020.v19i02.P06.
[32] Sutriawan, S. Rustad, G. F. Shidik, and Pujiono, "Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus: A Comparative Study of TF-IDF, Word2Vec, FastText BERT, and GPT," *ISI*, vol. 30, no. 6, pp. 1469–1482, Jun. 2025, doi: 10.18280/isi.300606.
[33] M. H. Abdillah, D. Rahmawati, and N. Komalasari, "SISTEM DETEKSI HOAKS PADA BERITA ONLINE MENGGUNAKAN METODE NAIVE BAYES," *Jurnal Riset Multidisiplin Edukasi*, vol. 2, no. 8, pp. 449–465, Aug. 2025, doi: 10.71282/jurmie.v2i8.837.
[34] B. Santoso and D. A. Puryono, "Sistem Klasifikasi Berita Menggunakan Metode Text Mining pada Website Pusat Kegiatan Belajar Masyarakat," *Jurnal Informatika Universitas Pamulang*, vol. 8, no. 2, pp. 374–384, Jun. 2023, doi: 10.32493/informatika.v8i2.24788.
[35] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decode*, vol. 3, no. 2, pp. 161–171, Apr. 2023, doi: 10.51454/decode.v3i2.131.
[36] C. N. P. Wardana and M. M. Mulki, "Klasifikasi Kategori Berita Berdasarkan Judul Menggunakan TF-IDF dan Support Vector Machine," *Seminar Nasional Teknologi & Sains*, vol. 5, no. 1, pp. 449–456, Jan. 2026, doi: 10.29407/thgggf34.
[37] A. A. Nabil, F. R. Wildantama, D. Satrianto, M. G. Bakara, I. Budiawan, and D. Mulyati, "Klasifikasi Hoax Menggunakan Metode TF-IDF + SVM: Penelitian," *Jurnal Pengabdian Masyarakat dan Riset Pendidikan*, vol. 4, no. 3, pp. 17653–17660, 2026, doi: 10.31004/jerkin.v4i3.5078.
[38] N. Silalahi and G. L. Ginting, "Rekomendasi Berita Berkaitan dengan Menerapkan Algoritma Text Mining dan TF-IDF," *Bulletin of Computer Science Research*, vol. 3, no. 4, pp. 276–282, Jun. 2023, doi: 10.47065/bulletincsr.v3i4.266.
[39] A. P. Thenata, "Text Mining Literature Review on Indonesian Social Media," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 7, no. 2, pp. 226–232, Aug. 2021, doi: 10.26418/jp.v7i2.47975.
[40] A. W. Pradana and M. Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts," *KINETIK*, pp. 375–380, Oct. 2019, doi: 10.22219/kinetik.v4i4.912.
[41] S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar, "Topic Classification of Online News Articles Using Optimized Machine Learning Models," *Computers*, vol. 12, p. 16, Jan. 2023, doi: 10.3390/computers12010016.
[42] A. H. Dalimunthe, M. Ula, and R. Meiyanti, "Unjuk Kerja Algoritma Support Vector Machine (SVM) dan Naïve Bayes Dalam Pengklasifikasian Berita Hoaks Pada Twitter Tentang Aksi Cepat Tanggap (ACT)," *Jurnal Elektronika dan Teknologi Informasi*, vol. 5, no. 2, pp. 11–22, Sep. 2024, doi: 10.5201/jet.v5i2.400.
[43] M. R. Fikri, R. T. Handayanto, and D. Irwan, "Web Scraping Situs Berita Menggunakan Bahasa Pemograman Python," *JSRCS*, vol. 3, no. 1, pp. 123–136, May 2022, doi: 10.31599/jsrcs.v3i1.1514.
[44] A. Wijaya, C. Rozikin, and B. N. Sari, "Penerapan Text Mining Untuk Klasifikasi Judul Berita Hoax Vaksinasi COVID-19 Menggunakan Algoritma Support Vector Machine," *Jurnal Ilmiah Wahana Pendidikan*, vol. 8, no. 16, pp. 11–20, Sep. 2022, doi: 10.5281/zenodo.7058890.
[45] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022, doi: 10.30812/matrik.v21i3.1851.
[46] D. A. E. Aritonang, P. Maharani, N. Ramadhani, K. D. Tania, and A. Meiriza, "KLASIFIKASI BERITA HOAKS DAN FAKTA BERDASARKAN REPRESENTASI TF-IDF MENGGUNAKAN SUPPORT VECTOR MACHINE," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 10, no. 3, pp. 4286–4292, May 2026, doi: 10.36040/jati.v10i3.18129.
[47] A. K. Dewi, N. F. Rahmadani, R. Syahputri, L. R. Nasution, and M. Furqon, "Deteksi Berita Hoax Pada Platform X Menggunakan Pendekatan Text Mining dan Algoritma Machine Learning," *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 33–46, Jul. 2025, doi: 10.47709/dsi.v5i1.6011.

[48] F. N. Rozi and D. H. Sulistyawati, “KLASIFIKASI BERITA HOAX PILPRES MENGGUNAKAN METODE MODIFIED K-NEAREST NEIGHBOR DAN PEMBOBOTAN MENGGUNAKAN TF-IDF,” *KONVERGENSI*, vol. 15, no. 1, Oct. 2019, doi: 10.30996/konv.v15i1.2828.
[49] B. Bramantyo, M. P. K. Putra, and N. Hendrastuty, “Klasifikasi Multikelas pada Teks Judul Berita Menggunakan Algoritma Recurrent Neural Network,” *Journal of Artificial Intelligence and Technology Information (JAITI)*, vol. 3, no. 1, pp. 18–29, Mar. 2025, doi: 10.58602/jaiti.v3i1.197.
[50] J. Ahmed and M. Ahmed, “ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES,” *IIUM Engineering Journal*, vol. 22, no. 2, pp. 210–225, Jul. 2021, doi: 10.31436/iiumej.v22i2.1662.