

Klasifikasi Kategori Berita Berdasarkan Judul Menggunakan TF-IDF dan *Support Vector Machine*

Ciko Nanda Putra Wardana¹, Mochammad Malikal Mulki²

^{1,2}Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

E-mail: ¹cikonanda08@gmail.com, ²malikal.mulki.x36@gmail.com

Abstrak – Perkembangan teknologi informasi dan komunikasi mendorong peningkatan jumlah berita digital yang dipublikasikan melalui berbagai platform daring. Kondisi ini menuntut adanya sistem pengelompokan berita yang efisien agar proses pengelolaan dan pencarian informasi dapat dilakukan secara cepat dan akurat. Klasifikasi berita secara manual menjadi kurang efektif, terutama pada pengelompokan berbasis judul berita yang bersifat singkat dan memiliki keterbatasan konteks. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem klasifikasi kategori berita otomatis dengan memanfaatkan judul berita sebagai data masukan melalui pendekatan Natural Language Processing (NLP).

Penelitian ini menerapkan tahapan pra-pemrosesan teks yang meliputi case folding, penghapusan tanda baca, stopword removal, dan stemming. Selanjutnya, teks direpresentasikan ke dalam bentuk vektor numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Support Vector Machine (SVM) digunakan sebagai algoritma klasifikasi utama, sedangkan Logistic Regression diterapkan sebagai metode pembandingan. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, precision, recall, F1-score, serta analisis confusion matrix.

Hasil pengujian menunjukkan bahwa SVM memberikan performa terbaik dengan nilai akurasi sebesar 94,41% dan weighted average F1-score sebesar 0,94, sementara Logistic Regression mencapai nilai akurasi sebesar 91,48% dan weighted average F1-score sebesar 0,91. Temuan ini menunjukkan bahwa kombinasi TF-IDF dan SVM efektif dan stabil untuk klasifikasi kategori berita berbasis judul.

Kata Kunci — Klasifikasi Berita, Logistic Regression, Support Vector Machine, TF-IDF, Text Mining

1. PENDAHULUAN

Kemajuan teknologi informasi dan komunikasi mendorong meningkatnya produksi serta penyebaran berita secara digital melalui berbagai platform daring. Kondisi ini menyebabkan jumlah dan ragam berita yang tersedia semakin besar, sehingga dibutuhkan mekanisme pengelompokan informasi yang efisien untuk menunjang pengelolaan serta pencarian informasi secara cepat dan akurat [1] [2]. Di sisi lain, pengelompokan berita secara manual membutuhkan waktu dan sumber daya yang signifikan serta berpotensi menimbulkan ketidakonsistenan dalam penentuan kategori, terutama seiring dengan terus bertambahnya jumlah berita yang dipublikasikan.

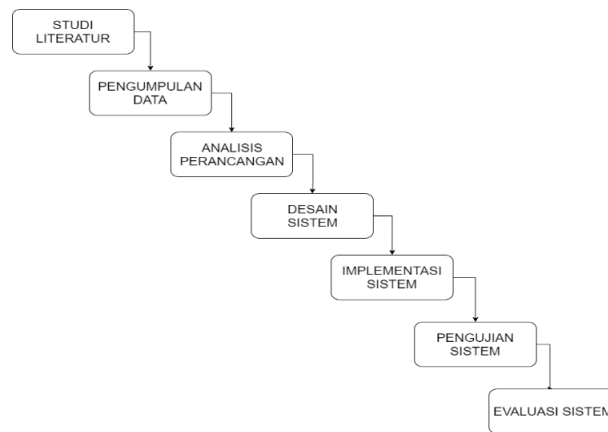
Seiring dengan perkembangan tersebut, pendekatan berbasis Natural Language Processing (NLP) banyak diterapkan untuk melakukan klasifikasi teks secara otomatis [3] [4]. Klasifikasi berita berdasarkan judul menjadi salah satu topik yang menarik karena judul berita berperan sebagai representasi ringkas dari isi berita. Namun demikian, judul berita umumnya bersifat singkat dan memiliki keterbatasan konteks, sehingga proses klasifikasi menjadi lebih menantang dibandingkan klasifikasi menggunakan isi berita secara penuh [5]. Keterbatasan konteks ini menuntut penggunaan metode representasi fitur yang mampu mengekstraksi informasi penting dari teks pendek secara efektif.

Agar data teks yang berupa judul berita dapat diproses oleh mesin, diperlukan tahapan pra-pemrosesan dan ekstraksi fitur. Salah satu metode yang banyak digunakan adalah Term Frequency–Inverse Document Frequency (TF-IDF), yang mengonversi teks menjadi vektor numerik dengan memperhitungkan frekuensi kemunculan kata dalam suatu dokumen relatif terhadap keseluruhan kumpulan dokumen [6]. Pada tahap klasifikasi, Support Vector Machine (SVM) dikenal efektif dalam menangani data berdimensi tinggi seperti teks dengan membentuk hyperplane optimal untuk memisahkan kelas data. Untuk menilai kinerja SVM secara objektif dalam klasifikasi judul berita, Logistic Regression digunakan sebagai metode pembandingan karena memiliki struktur sederhana, efisiensi komputasi yang baik, serta kemampuan dalam menangani klasifikasi biner dan multikelas [7].

Berdasarkan uraian latar belakang tersebut, penelitian ini bertujuan mengembangkan sistem klasifikasi kategori berita secara otomatis dengan memanfaatkan judul berita sebagai data masukan utama. Proses ekstraksi fitur dilakukan menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF), sementara algoritma Support Vector Machine (SVM) diterapkan sebagai model klasifikasi utama. Untuk menilai efektivitas pendekatan yang diusulkan dalam mengklasifikasikan teks pendek berupa judul berita, kinerja SVM selanjutnya dievaluasi dan dibandingkan dengan algoritma Logistic Regression sebagai metode pembandingan.

2. METODE PENELITIAN

Penelitian ini dilaksanakan melalui tahapan yang tersusun secara sistematis, dimulai dari studi literatur dan pengumpulan data, dilanjutkan dengan analisis dan perancangan sistem, serta desain dan implementasi model klasifikasi. Tahap akhir penelitian meliputi pengujian dan evaluasi sistem untuk menilai kinerja metode yang diusulkan dalam mengklasifikasikan kategori berita berdasarkan judul. Alur penelitian ini telah dicantumkan sebagai berikut:



Gambar 1. Diagram Waterfall Model

2.1 Studi Literatur

Studi literatur dilakukan untuk memperoleh landasan teoritis terkait konsep dan metode yang digunakan dalam klasifikasi teks berita berbasis Natural Language Processing. Kajian ini difokuskan pada penelitian-penelitian sebelumnya yang membahas klasifikasi berita menggunakan judul sebagai data masukan, termasuk penerapan tahapan pra-pemrosesan dan teknik ekstraksi fitur teks. Selain itu, studi literatur mencakup pembahasan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai representasi fitur, serta algoritma *Support Vector Machine* (SVM) dan *Logistic Regression* yang banyak digunakan dalam klasifikasi teks. Berdasarkan hasil kajian tersebut, SVM dipilih sebagai metode klasifikasi utama, sementara *Logistic Regression* digunakan sebagai model pembandingan untuk mengevaluasi kinerja pendekatan yang diusulkan.

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa kumpulan judul berita berbahasa Indonesia yang diperoleh dari basis data berita digital. Dataset awal mencakup 1878 judul berita yang telah diberi label kategori, seperti ekonomi, politik, olahraga, teknologi, dan kategori lainnya. Seluruh data disimpan dalam basis data untuk memudahkan proses pengelolaan serta memungkinkan pembaruan apabila terdapat penambahan judul berita baru.

Pada tahap pengumpulan data, dilakukan proses seleksi dengan mengeluarkan kategori berita yang memiliki jumlah data sangat terbatas. Langkah ini bertujuan untuk meminimalkan masalah ketidakseimbangan data (*imbalanced data*) yang berpotensi menurunkan kinerja model klasifikasi.

Dataset tersebut kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan teknik *stratified sampling* untuk menjaga proporsi setiap kategori. Data latih dimanfaatkan dalam proses pembentukan model, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan judul berita yang belum pernah dipelajari.

2.3 Analisis Perancangan

Tahap analisis dan perancangan sistem dilakukan untuk mengidentifikasi kebutuhan fungsional, alur pemrosesan, serta komponen utama dalam pengembangan sistem klasifikasi kategori berita. Sistem dirancang untuk menerima judul berita sebagai data masukan dan menghasilkan kategori berita sebagai keluaran. Mengingat judul berita merupakan teks yang relatif singkat, padat, dan memiliki keterbatasan konteks, diperlukan perancangan alur pemrosesan yang tepat agar informasi penting dapat diekstraksi secara optimal tanpa menimbulkan gangguan (*noise*) dalam proses klasifikasi.

Berdasarkan hasil analisis kebutuhan, sistem ditetapkan harus mampu melakukan tahapan pra-pemrosesan teks sebelum proses pemodelan. Tahapan ini meliputi *case folding* untuk menyeragamkan bentuk huruf, penghapusan tanda baca dan karakter non-alfanumerik, serta *stopword removal* guna menghilangkan kata-kata umum yang tidak memiliki kontribusi signifikan dalam penentuan kategori berita [8]. Selain itu, diterapkan proses stemming untuk mengonversi kata berimbuhan ke bentuk dasar, sehingga kata-kata dengan makna serupa dapat direpresentasikan secara konsisten dalam sistem.

Setelah tahapan pra-pemrosesan selesai, langkah berikutnya adalah merancang metode representasi fitur yang sesuai untuk karakteristik teks pendek. Metode *Term Frequency–Inverse Document Frequency* (TF-IDF) dipilih karena mampu memberikan bobot lebih besar pada kata-kata yang memiliki tingkat kepentingan tinggi, yaitu kata yang jarang muncul secara global tetapi sering muncul dalam dokumen tertentu[9]. Melalui TF-IDF, data teks diubah menjadi bentuk vektor numerik sehingga dapat digunakan sebagai masukan bagi algoritma pembelajaran mesin. Secara matematis, nilai TF-IDF diperoleh dari hasil perkalian antara *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF), yang dirumuskan sebagai berikut:

$$TF(t, d) = f_{t,d} \quad (1)$$

$$IDF(t) = \log \frac{N}{df_t} \quad (2)$$

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (3)$$

di mana $f_{t,d}$ menyatakan frekuensi kemunculan term t dalam dokumen d , N merupakan jumlah keseluruhan dokumen, dan df_t menunjukkan jumlah dokumen yang mengandung term t [10]. Skema pembobotan ini memberikan nilai yang lebih besar pada term yang sering muncul dalam dokumen tertentu namun jarang ditemukan pada keseluruhan dokumen, sehingga mampu merepresentasikan kata-kata yang bersifat informatif. Dengan representasi tersebut, sistem diharapkan dapat mengekstraksi karakteristik utama dari judul berita meskipun teks yang digunakan relatif singkat [11].

Pada tahap perancangan model klasifikasi, algoritma *Support Vector Machine* (SVM) ditetapkan sebagai metode utama karena kemampuannya dalam menangani data dengan dimensi tinggi serta membentuk batas pemisah (*hyperplane*) yang optimal antar kelas [12]. SVM bekerja dengan prinsip memaksimalkan margin pemisahan antar kelas, sehingga dapat meningkatkan kemampuan generalisasi model terhadap data baru. Secara matematis, tujuan SVM adalah meminimalkan fungsi objektif sebagai berikut:

$$\min_{w,b,\xi} \frac{1}{2} |w|^2 + C \sum_{i=1}^n \xi_i \quad (4)$$

$$y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \quad (5)$$

dengan w sebagai vektor bobot, C sebagai parameter regularisasi, dan ξ_i sebagai variabel kesalahan (*slack variable*) [13].

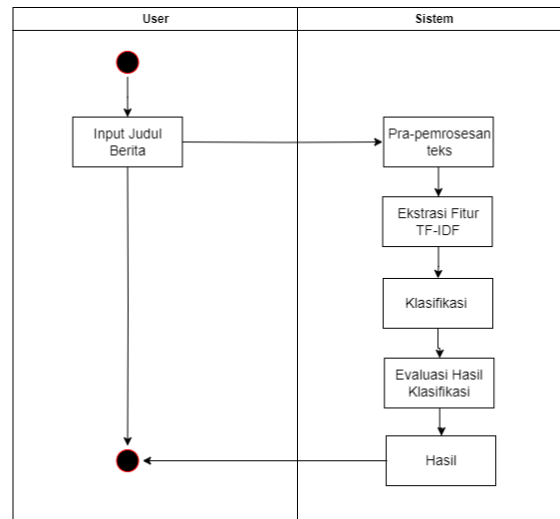
Sebagai pembanding kinerja, *Logistic Regression* diterapkan untuk menilai efektivitas model SVM. Algoritma ini memodelkan probabilitas suatu data termasuk ke dalam kelas tertentu melalui fungsi sigmoid. *Logistic Regression* dipilih karena memiliki struktur model yang sederhana, efisiensi komputasi yang tinggi, serta umum digunakan sebagai model dasar (*baseline*) dalam berbagai tugas klasifikasi teks [14]. Fungsi sigmoid pada *Logistic Regression* dirumuskan sebagai berikut:

$$P(y = 1 | x) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (6)$$

di mana x merepresentasikan vektor fitur hasil ekstraksi TF-IDF, w merupakan bobot yang dipelajari oleh model, dan b menyatakan nilai bias [15].

Tahap analisis dan perancangan ini menghasilkan alur sistem klasifikasi yang tersusun secara sistematis, mulai dari proses pra-pemrosesan teks, ekstraksi fitur menggunakan TF-IDF, penerapan algoritma klasifikasi SVM dan *Logistic Regression*, hingga evaluasi kinerja model. Dengan rancangan tersebut, setiap komponen sistem diharapkan dapat terintegrasi secara optimal dan mendukung pencapaian tujuan penelitian dalam melakukan klasifikasi kategori berita otomatis berbasis judul.

2.4 Desain Sistem



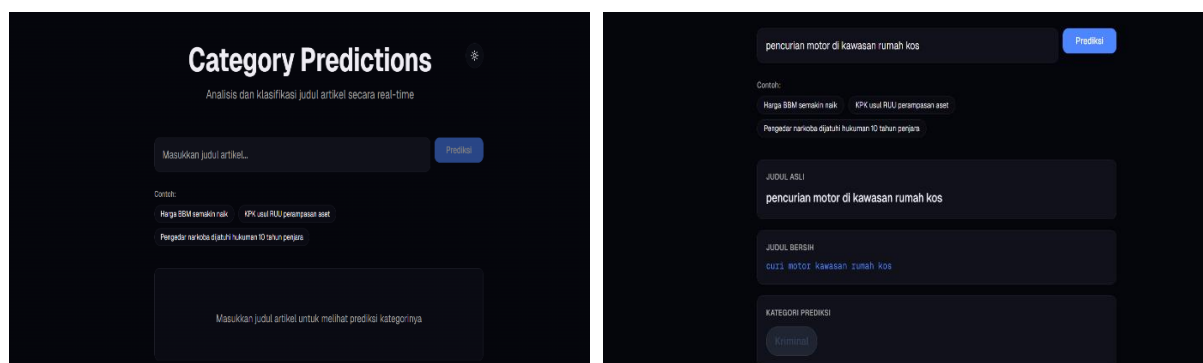
Gambar 2. Activity Diagram

Pada tahap input judul berita, selanjutnya data tersebut diproses melalui tahap pra-pemrosesan teks untuk menghasilkan data yang bersih dan representatif. Tahapan ini meliputi case folding untuk menyeragamkan huruf, penghapusan tanda baca dan karakter non-alfanumerik, tokenisasi untuk memecah teks menjadi unit kata, stopword removal untuk menghilangkan kata-kata umum yang kurang informatif, serta stemming untuk mengonversi kata berimbuhan ke bentuk dasar.

Setelah pra-pemrosesan, teks yang telah dibersihkan diubah ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Metode ini memberikan bobot pada setiap term berdasarkan tingkat kepentingannya dalam dokumen dan distribusinya pada keseluruhan dataset. Hasil dari proses ini berupa vektor fitur numerik yang digunakan sebagai masukan pada tahap klasifikasi.

Pada tahap klasifikasi, vektor fitur TF-IDF diproses menggunakan algoritma *Support Vector Machine* (SVM) sebagai model utama. SVM dipilih karena kemampuannya dalam menangani data berdimensi tinggi serta membentuk *hyperplane* pemisah antar kelas secara optimal. Sebagai pembanding, *Logistic Regression* digunakan untuk mengevaluasi kinerja SVM, mengingat kemampuannya dalam memodelkan probabilitas kelas dan perannya sebagai baseline dalam klasifikasi teks.

Dengan rancangan tersebut, setiap tahapan sistem dirancang saling terintegrasi guna mendukung proses klasifikasi kategori berita secara otomatis berdasarkan judul berita, mulai dari pengolahan data teks hingga evaluasi hasil klasifikasi.



Gambar 3. Desain Antarmuka Prediksi Judul Berita

Gambar menunjukkan proses kerja sistem klasifikasi kategori berita setelah pengguna memasukkan judul artikel dan menekan tombol prediksi. Sistem menampilkan judul asli yang dimasukkan pengguna, kemudian hasil pra-pemrosesan teks berupa judul bersih yang telah melalui tahapan *case folding*, penghapusan kata tidak relevan, dan normalisasi kata. Selanjutnya, judul bersih diproses oleh model klasifikasi untuk menghasilkan kategori prediksi, yang pada contoh ini diklasifikasikan sebagai kategori Kriminal.

2.5 Implementasi Sistem

Tahap implementasi sistem merupakan proses penerjemahan rancangan metode klasifikasi ke dalam bentuk aplikasi yang dapat dijalankan untuk mengelompokkan kategori berita berdasarkan judul. Pada tahap ini, metode *Term Frequency–Inverse Document Frequency* (TF-IDF) digunakan untuk mengubah teks judul berita menjadi representasi vektor numerik, sedangkan algoritma *Support Vector Machine* (SVM) diterapkan sebagai model klasifikasi utama. Selain itu, *Logistic Regression* diimplementasikan sebagai metode pembandingan guna mengevaluasi kinerja sistem secara objektif.

Implementasi sistem dilakukan secara bertahap dan sistematis, diawali dengan pengambilan data judul berita dari basis data. Data tersebut kemudian melalui tahap pra-pemrosesan teks yang meliputi *case folding*, penghapusan karakter *non-alfanumerik*, *stopword removal*, serta *stemming*. Tahapan ini bertujuan menghasilkan data teks yang bersih, seragam, dan siap digunakan dalam proses pemodelan.

Setelah pra-pemrosesan selesai, sistem melakukan ekstraksi fitur menggunakan TF-IDF untuk memperoleh bobot kata yang mencerminkan tingkat kepentingannya dalam kumpulan dokumen. Vektor fitur yang dihasilkan digunakan sebagai masukan dalam proses pelatihan dan pengujian model klasifikasi. Pada tahap ini, model SVM dan Logistic Regression dilatih menggunakan data latih dan diuji pada data uji untuk menghasilkan prediksi kategori berita.

Keluaran sistem berupa hasil klasifikasi kategori berita serta nilai evaluasi kinerja, yang mencakup akurasi, classification report dengan metrik precision, recall, dan F1-score, serta confusion matrix sebagai alat analisis kesalahan klasifikasi antar kategori. Hasil tersebut menjadi dasar dalam menganalisis performa sistem pada tahap pengujian dan evaluasi.

2.6 Pengujian Sistem

Pengujian sistem dilakukan untuk menilai kinerja model klasifikasi kategori berita yang dikembangkan menggunakan algoritma Support Vector Machine (SVM) sebagai metode utama, dengan Logistic Regression sebagai metode pembandingan. Tahap ini bertujuan untuk mengukur kemampuan masing-masing model dalam mengklasifikasikan judul berita ke dalam kategori yang tepat, sekaligus membandingkan efektivitas kedua algoritma dalam menangani karakteristik teks pendek.

Data yang digunakan dalam pengujian merupakan data yang telah melewati tahap pra-pemrosesan dan ekstraksi fitur menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80% untuk pelatihan dan 20% untuk pengujian menggunakan teknik *train-test split*. Pembagian data dilakukan secara stratified untuk menjaga proporsi setiap kategori berita, dengan rasio 80% sebagai data latih dan 20% sebagai data uji, sehingga diperoleh sebanyak 376 judul berita sebagai data uji.

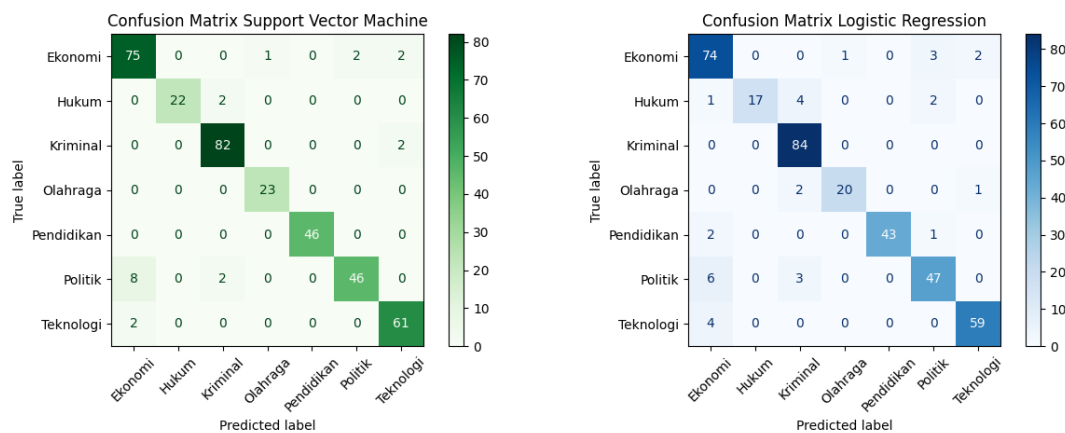
Model SVM dan Logistic Regression dilatih menggunakan data latih, kemudian dievaluasi menggunakan data uji yang tidak terlibat dalam proses pelatihan. Evaluasi performa model dilakukan dengan mengacu pada metrik akurasi, classification report yang mencakup nilai precision, recall, dan F1-score, serta confusion matrix untuk menggambarkan pola kesalahan klasifikasi antar kategori. Penggunaan beberapa metrik evaluasi ini dimaksudkan untuk memberikan gambaran kinerja model secara komprehensif, terutama pada kondisi distribusi data yang tidak seimbang antar kelas.

2.7 Evaluasi Sistem

Hasil evaluasi menunjukkan bahwa algoritma Support Vector Machine (SVM) dan Logistic Regression sama-sama mampu mengklasifikasikan kategori berita berdasarkan judul dengan tingkat kinerja yang baik. Penilaian performa dilakukan menggunakan metrik akurasi, *classification report*, dan *confusion matrix* guna memperoleh gambaran kinerja model secara komprehensif.

Secara keseluruhan, SVM menunjukkan performa yang lebih unggul dibandingkan Logistic Regression. Hal ini tercermin dari nilai akurasi SVM sebesar 0,94 yang lebih tinggi dibandingkan Logistic Regression sebesar 0,91. Selain itu, nilai *macro average* dan *weighted average F1-score* pada SVM juga lebih besar, yang mengindikasikan kemampuan generalisasi model yang lebih baik terhadap seluruh kategori berita.

Berdasarkan *classification report*, kedua model memperlihatkan performa yang cukup baik pada beberapa kategori tertentu dengan nilai *precision* dan *recall* yang tinggi. Namun, perbedaan kinerja mulai terlihat pada kategori yang memiliki kemiripan topik dan kosakata. Logistic Regression menunjukkan nilai *recall* yang lebih rendah pada beberapa kategori, yang menandakan masih adanya data aktual yang belum berhasil terklasifikasi secara optimal. Sebaliknya, SVM menghasilkan nilai *precision* dan *recall* yang lebih seimbang sehingga memberikan nilai *F1-score* yang lebih stabil antar kategori.



Gambar 4. Perbandingan Confusion Matrix pada Model Support Vector Machine dan Logistic Regression

Berdasarkan Gambar 4, confusion matrix pada kedua model, terlihat bahwa Support Vector Machine (SVM) dan Logistic Regression sama-sama mampu mengklasifikasikan sebagian besar data uji dengan benar, yang ditunjukkan oleh dominasi nilai pada diagonal utama. Namun, terdapat perbedaan yang cukup jelas pada pola kesalahan klasifikasi yang dihasilkan oleh masing-masing model.

Pada model SVM, tingkat klasifikasi yang sangat baik terlihat pada kategori Pendidikan dan Olahraga dengan jumlah prediksi benar yang tinggi dan kesalahan yang relatif minimal. Kesalahan klasifikasi pada SVM umumnya terkonsentrasi pada kategori Ekonomi dan Politik, yang saling tertukar akibat kemiripan kosakata dan konteks pemberitaan. Meskipun demikian, kesalahan tersebut bersifat terbatas dan tidak menyebar luas ke kategori lain, yang menunjukkan bahwa SVM mampu membentuk batas pemisah antar kelas secara lebih optimal pada ruang fitur TF-IDF berdimensi tinggi.

Sementara itu, Logistic Regression menunjukkan pola kesalahan yang lebih menyebar antar kategori. Kesalahan klasifikasi lebih sering terjadi pada kategori seperti Ekonomi, Politik, dan Hukum, di mana judul berita dari kategori tersebut cenderung diprediksi sebagai kategori lain yang memiliki kosakata serupa. Selain itu, pada kategori dengan jumlah data relatif lebih sedikit, Logistic Regression cenderung mengklasifikasikan data ke kategori mayoritas. Hal ini mengindikasikan bahwa Logistic Regression lebih sensitif terhadap ketidakseimbangan distribusi data serta keterbatasan dalam memodelkan hubungan linier antar fitur TF-IDF.

Berdasarkan hasil evaluasi tersebut, dapat disimpulkan bahwa SVM memiliki performa yang lebih stabil dan unggul dalam klasifikasi kategori berita berbasis judul. Meskipun demikian, Logistic Regression tetap menunjukkan kinerja yang baik serta efisien sebagai metode pembandingan dalam sistem klasifikasi yang dikembangkan.

3. HASIL DAN PEMBAHASAN

Hasil pengujian menunjukkan bahwa algoritma Logistic Regression dan Support Vector Machine (SVM) sama-sama mampu mengklasifikasikan kategori berita berdasarkan judul dengan tingkat kinerja yang tinggi. Evaluasi performa dilakukan menggunakan metrik akurasi, precision, recall, F1-score, serta analisis confusion matrix untuk memperoleh gambaran kinerja model secara kuantitatif dan kualitatif.

Secara umum, SVM menunjukkan performa yang lebih baik dibandingkan Logistic Regression. Model SVM mencapai tingkat akurasi sebesar 94,41% dengan nilai weighted average F1-score sebesar 0,94, sementara Logistic Regression memperoleh akurasi sebesar 91,48% dan weighted average F1-score sebesar 0,91. Hasil ini mengindikasikan bahwa SVM memiliki kemampuan generalisasi yang lebih baik dalam menangani data teks berdimensi tinggi yang direpresentasikan menggunakan TF-IDF, khususnya pada permasalahan klasifikasi multikelas.

Berdasarkan classification report, kedua model memberikan performa yang sangat baik pada kategori Kriminal, Pendidikan, dan Teknologi, yang ditunjukkan oleh nilai precision dan recall yang tinggi. Hal ini menunjukkan bahwa kategori-kategori tersebut memiliki karakteristik kosakata yang lebih spesifik dan konsisten, sehingga lebih mudah dikenali oleh model. Sebaliknya, kategori Ekonomi, Politik, dan Hukum memperlihatkan nilai precision dan recall yang berubah-ubah akibat adanya kemiripan kosakata dan keterkaitan konteks antar kategori.

Analisis confusion matrix memperlihatkan bahwa kesalahan klasifikasi pada kedua model umumnya terjadi pada kategori dengan kedekatan topik, terutama antara Ekonomi dan Politik. Pada Logistic Regression, kesalahan klasifikasi cenderung lebih menyebar, khususnya pada kategori dengan jumlah data yang relatif lebih sedikit, yang

sering diprediksi sebagai kelas mayoritas. Hal ini menunjukkan bahwa Logistic Regression lebih sensitif terhadap ketidakseimbangan data dan hubungan linier antar fitur TF-IDF.

Sebaliknya, confusion matrix pada SVM menunjukkan dominasi prediksi pada diagonal utama dengan jumlah kesalahan yang lebih rendah dan terkontrol. Kondisi ini menunjukkan bahwa SVM lebih efektif dalam membentuk batas pemisah antar kelas meskipun terdapat tumpang tindih fitur dalam ruang vektor TF-IDF. Temuan ini sejalan dengan nilai F1-score SVM yang lebih stabil pada hampir seluruh kategori.

Secara keseluruhan, hasil evaluasi menegaskan bahwa pendekatan klasifikasi berbasis TF-IDF dengan algoritma SVM dan Logistic Regression efektif untuk pengelompokan judul berita secara otomatis. Namun, berdasarkan hasil pengujian, SVM terbukti lebih unggul dan stabil dibandingkan Logistic Regression dalam konteks klasifikasi teks pendek multikelas. Hal ini menegaskan pentingnya pemilihan algoritma dan evaluasi yang komprehensif dalam meningkatkan keandalan sistem klasifikasi berita.

4. SIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan tahapan pra-pemrosesan teks yang dikombinasikan dengan metode Term Frequency–Inverse Document Frequency (TF-IDF) mampu merepresentasikan judul berita secara efektif ke dalam bentuk vektor numerik. Pendekatan ini memungkinkan sistem menangkap karakteristik kata yang relevan meskipun judul berita memiliki keterbatasan konteks sebagai teks pendek.

Hasil evaluasi menunjukkan bahwa algoritma Support Vector Machine (SVM) dan Logistic Regression sama-sama mampu mengklasifikasikan kategori berita dengan tingkat akurasi yang tinggi. Namun, berdasarkan perbandingan nilai akurasi, macro dan weighted average F1-score, serta analisis confusion matrix, SVM terbukti memiliki kinerja yang lebih unggul dan stabil. Keunggulan ini terlihat dari kemampuannya membentuk batas pemisah antar kelas secara lebih optimal pada ruang fitur berdimensi tinggi hasil representasi TF-IDF.

Meskipun Logistic Regression tetap menunjukkan performa yang kompetitif dan efisien sebagai metode pembanding, kinerjanya lebih sensitif terhadap ketidakseimbangan data dan tumpang tindih fitur antar kategori. Secara keseluruhan, penelitian ini menunjukkan bahwa pendekatan klasifikasi judul berita berbasis TF-IDF dengan algoritma SVM merupakan solusi yang efektif dan andal untuk pengelompokan teks pendek secara otomatis.

5. SARAN

Berdasarkan hasil penelitian serta keterbatasan yang ditemukan, terdapat beberapa rekomendasi yang dapat dipertimbangkan untuk pengembangan penelitian selanjutnya. Penelitian mendatang disarankan untuk menggunakan metode representasi fitur berbasis embedding semantik, seperti Word2Vec, atau BERT, agar konteks makna kata pada judul berita dapat direpresentasikan dengan lebih baik dan mengurangi tumpang tindih fitur antar kategori.

Selain itu, penerapan teknik penanganan data tidak seimbang, seperti metode resampling atau pemberian bobot kelas, dapat dilakukan untuk meningkatkan kinerja model pada kategori dengan jumlah data yang relatif sedikit. Penelitian algoritma klasifikasi lain, seperti Random Forest, maupun model berbasis deep learning, juga dapat dijadikan alternatif untuk memperoleh performa klasifikasi yang lebih optimal.

Dari aspek data, disarankan untuk memperluas jumlah dan variasi dataset, baik dari segi kategori maupun sumber berita, guna meningkatkan kemampuan generalisasi model. Penggunaan teks berita yang lebih panjang, seperti ringkasan atau isi berita, juga dapat dipertimbangkan agar sistem memperoleh konteks yang lebih kaya dalam proses klasifikasi.

DAFTAR PUSTAKA

- [1] P. Rama Bena Putra and R. Setya Perdana, “Klasifikasi Judul Berita Online menggunakan Metode Support Vector Machine (SVM) dengan Seleksi Fitur Chi-square,” 2023. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [2] M. F. Ramadhan, “Klasifikasi Topik dan Sentimen Judul Berita dengan Augmentasi dan TF-IDF,” *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 6732–6741, Jul. 2025, doi: 10.31004/riggs.v4i2.1692.
- [3] M. Fazri and A. Voutama, “ANALISIS SENTIMEN PUBLIK TERHADAP DANANTARA DI MEDIA SOSIAL X MENGGUNAKAN NLP DAN PEMBELAJARAN MESIN,” *JOISIE (Journal Of Information Systems And Informatics Engineering)*, vol. 9, no. 1, p. 197, Jul. 2025, doi: 10.35145/joisie.v9i1.4924.

-
- [4] R. Rakhmat Sani, Y. Ayu Pratiwi, S. Winarno, E. Devi Udayanti, and dan Farrikh Al Zami, “Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Hoax pada Berita Online Indonesia,” 2022.
- [5] O. N. Cahyani and F. Budiman, “Performa Logistic Regression dan Naive Bayes dalam Klasifikasi Berita Hoax di Indonesia,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 60–68, Apr. 2025, doi: 10.29408/edumatic.v9i1.28987.
- [6] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, “Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity,” 2018. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [7] N. Maulana, M. R. Handayani, W. D. Yuniarti, and K. Umam, “PENERAPAN METODE SUPPORT VECTOR MACHINE DALAM MENGLASIFIKASIKAN BERITA POLITIK : FAKTA VS HOAKS,” *JIPi (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 3, pp. 2679–2688, Sep. 2025, doi: 10.29100/jipi.v10i3.8077.
- [8] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 406, Apr. 2021, doi: 10.30865/mib.v5i2.2835.
- [9] D. Septiani and I. Isabela, “ANALISIS TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DALAM TEMU KEMBALI INFORMASI PADA DOKUMEN TEKS”.
- [10] A. M. Putri, Rosita Yanuarti, Moh Dasuki, and Agus Milu Susetyo, “Analisis Sentimen Terhadap Identitas Kependudukan Digital Menggunakan Algoritma Multinomial Naïve Bayes,” *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, vol. 10, no. 2, pp. 85–95, Aug. 2025, doi: 10.32528/justindo.v10i2.2777.
- [11] A. Pranata and N. Azmi Verdikha, “KLASIFIKASI TEKS QUICK COUNT PEMILIHAN PRESIDEN 2024 PADA TWITTER MENGGUNAKAN METODE TF-IDF DAN NAIVE BAYES,” *Jurnal Informatika Terpadu*, vol. 10, no. 2, pp. 93–100, 2024, [Online]. Available: <https://journal.nurulfikri.ac.id/index.php/JIT>
- [12] I. Komang Dharmendra *et al.*, “IMPLEMENTASI TEXT MINING UNTUK KLASIFIKASI OPINI ALUMNI PADA PERGURUAN TINGGI,” *Jurnal Teknologi Informasi dan Komputer*, vol. 9, no. 3, Jul. 2023, doi: 10.36002/JUTIK.V9I3.2504.
- [13] M. I. Adrian and E. A. Laksana, “Optimalisasi Parameter Support Vector Machine dengan Algoritma PSO untuk Tugas Klasifikasi Sentimen Ulasan IMDb,” *Jurnal Algoritma*, vol. 22, no. 1, pp. 288–299, Jun. 2025, doi: 10.33364/algoritma/v.22-1.2306.
- [14] A. Mu, amar Wahid, K. Adi Nugroho, T. Safitri, and F. Setyo Utomo, “Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis TF-IDF,” *Jurnal Pendidikan dan Teknologi Indonesia (JPTI)*, vol. 4, no. 8, pp. 381–392, 2024, doi: 10.52436/1.jpti.602.
- [15] A. Firizkiansah, A. Muhammad, and I. R. Maulana, “Optimasi Klasifikasi Data Teks Menggunakan Algoritma Logistic Regression dengan TF-IDF dan SMOTE,” *JIKOMTI : Jurnal Ilmiah Ilmu Komputer dan Teknologi Informasi*, vol. 2, no. 1, pp. 29–36, May 2025, Accessed: Jan. 04, 2026. [Online]. Available: <https://ojs.sains.ac.id/index.php/Jikomti/article/view/97>