LEMBAR PENGESAHAN

Nomor: 436/LP-LPBK/VIII/2026

Judul	:	KOMPARASI ALGORITMA *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE* (SVM) DALAM PENENTUAN KATEGORI ARTIKEL DI WEBSITE MAJELISTABLIGHPWMJATENG.COM
Nama	:	Agung Dwi Hanggoro

Menerangkan bahwa abstrak dengan judul di atas telah diterjemahkan ke dalam Bahasa Inggris oleh Lembaga Pengembangan Bahasa dan Kerja Sama (LPBK), Universitas Muhammadiyah Pekajangan Pekalongan.

Pekalongan, 19 Agustus 2026

Disahkan oleh,
Kepala Lembaga Pengembangan Bahasa dan Kerja Sama (LPBK)

Aida Rusmariana, S.Kep., Ns., MAN

Program Studi Sarjana Informatika
Fakultas Teknik dan Ilmu Komputer
Universitas Muhammadiyah Pekajangan Pekalongan

ABSTRAK

Agung Dwi Hanggoro[1], Aslam Fatkhudin[2], Fenilinas Adi Artanto[3]

KOMPARASI ALGORITMA *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE (SVM)* DALAM PENENTUAN KATEGORI ARTIKEL DI WEBSITE MAJELISTABLIGHPWMJATENG.COM

Perkembangan publikasi digital mendorong meningkatnya jumlah artikel yang perlu dikelola dan dikelompokkan secara sistematis. Website majelistablighpwmjateng.com menyediakan berbagai artikel dakwah dengan kategori yang beragam. Kondisi tersebut menjadikan proses pengategorian secara manual berpotensi memerlukan waktu yang lebih lama serta menghasilkan pelabelan yang kurang konsisten. Penelitian ini bertujuan untuk membangun model klasifikasi kategori artikel berdasarkan judul dan membandingkan kinerja algoritma *Naive Bayes* dengan *Support Vector Machine* dalam mengklasifikasikan artikel dakwah berbahasa Indonesia. Data penelitian berupa judul artikel beserta kategorinya yang diperoleh dari website majelistablighpwmjateng.com pada periode 10 Juni 2023 hingga 9 Juni 2026. Dataset awal berjumlah 1.369 data yang terbagi dalam 19 kategori. Kategori Khutbah Nikah yang hanya memiliki satu data dihapus agar proses pembagian *dataset* dapat dilakukan secara optimal. Setelah penghapusan tersebut, *dataset* akhir terdiri atas 1.368 data dalam 18 kategori. Tahapan penelitian mencakup pengumpulan data, *preprocessing* yang meliputi *cleaning*, *case folding*, *tokenisasi*, dan *stopword removal*, pembobotan fitur dengan metode *Term Frequency-Inverse Document Frequency*, pelatihan model, serta evaluasi menggunakan *metrik accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian menunjukkan bahwa algoritma *SVM* menghasilkan kinerja yang lebih baik daripada *Naive Bayes*. Model SVM mencapai *accuracy* sebesar 79,56%, *precision* sebesar 72,11%, *recall* sebesar 79,56%, dan *F1-score* sebesar 74,36%. Adapun model *Naive Bayes* memperoleh *accuracy* sebesar 73,36%, *precision* sebesar 64,73%, *recall* sebesar 73,36%, *dan F1-score* sebesar 66,86%. Perbedaan tersebut mengindikasikan bahwa SVM lebih efektif dalam menangani data teks berdimensi tinggi dan pola kategori yang kompleks. Meskipun demikian, ketidakseimbangan jumlah data pada setiap kategori masih memengaruhi kemampuan model dalam mengidentifikasi kelas minoritas. Berdasarkan hasil penelitian, kombinasi TF-IDF dan SVM lebih sesuai digunakan untuk mengklasifikasikan kategori judul artikel pada website majelistablighpwmjateng.com dibandingkan kombinasi TF-IDF dan *Naive Bayes*. Penelitian ini diharapkan dapat mendukung proses pengelompokan artikel secara lebih cepat, konsisten, dan efisien, serta menjadi referensi bagi pengembangan sistem klasifikasi teks berbahasa Indonesia dalam domain dakwah.

Kata Kunci: klasifikasi teks, judul artikel, TF-IDF, *Naive Bayes*, *Support Vector Machine*, website dakwah, majelistablighpwmjateng.com.

Undergraduate Program in Informatics
Faculty of Engineering and Computer Science
University of Muhammadiyah Pekajangan Pekalongan

ABSTRACT

Agung Dwi Hanggoro[1], Aslam Fatkhudin[2], Fenilinas Adi Artanto[3]

THE COMPARISON OF *NAIVE BAYES ALGORITHMS* AND *SUPPORT VECTOR MACHINE (SVM)* IN DETERMINING ARTICLE CATEGORIES ON THE MAJELISTABLIGHPWMJATENG.COM WEBSITE

The development of digital publications has led to an increase in the number of articles that need to be managed and grouped systematically. majelistablighpwmjateng.com website provides a variety of da'wah articles in various categories. This condition makes the manual categorization process potentially take longer and result in less consistent labeling. This study aims to build a model for the classification of article categories based on titles and compare the performance of *the Naive Bayes* algorithm with *the Support Vector Machine* in classifying da'wah articles in Indonesian. The research data is in the form of article titles and their categories obtained from the majelistablighpwmjateng.com website in the period from June 10, 2023 to June 9, 2026. The initial dataset totaled 1,369 data points divided into 19 categories. The category of Marriage Sermon, that only has one data is deleted so that the dataset sharing process could be carried out optimally. After the deletion, the final *dataset* consisted of 1,368 data points in 18 categories. The research stages include data collection, *preprocessing,* which includes *cleaning*, *case folding*, *tokenization*, and *stopword removal*, feature weighting with *the Term Frequency-Inverse Document Frequency* method, model training, and evaluation using *accuracy*, *precision*, and *recall metrics*, then *F1 score*. The test results showed that *the SVM* algorithm produced better performance than *Naive Bayes*. The SVM model achieved *an accuracy* of 79.56%, *precision* of 72.11%, *recall* of 79.56%, and *an F1-score* of 74.36%. The *Naive Bayes* model obtained *an accuracy* of 73.36%, *precision* of 64.73%, *recall* of 73.36%, *and an F1-score* of 66.86%. These differences indicate that SVM is more effective at handling high-dimensional text data and complex category patterns. However, the imbalance in the amount of data in each category still affects the model's ability to identify minority classes. Based on the results of the study, the combination of TF-IDF and SVM is more suitable to be used to classify the category of article titles on the majelistablighpwmjateng.com website than the combination of TF-IDF and *Naive Bayes*. This research is expected to support the process of grouping articles more quickly, consistently, and efficiently, as well as become a reference for the development of a classification system for Indonesian texts in the domain of da'wah.

Keywords: text classification, article title, TF-IDF, *Naive Bayes*, *Support Vector Machine*, da'wah website, majelistablighpwmjateng.com.

KOMPARASI ALGORITMA *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE (SVM)* DALAM PENENTUAN KATEGORI ARTIKEL DI WEBSITE MAJELISTABLIGHPWMJATENG.COM

COMPARISON OF NAIVE BAYES AND SUPPORT VECTOR MACHINE (SVM) ALGORITHMS IN DETERMINING ARTICLE CATEGORIES ON THE MAJELISTABLIGHPWMJATENG.COM WEBSITE

Agung Dwi Hanggoro[1], Aslam Fatkhudin[2], Fenilinas Adi Artanto[3]
Program Studi Informatika, Fakultas Teknik dan Ilmu Komputer,
Universitas Muhammadiyah Pekajangan Pekalongan
Email: agungdh@umpp.ac.id , aslamfatkhudin@umpp.ac.id, fenilinas@umpp.ac.id

ABSTRAK

Perkembangan publikasi digital mendorong meningkatnya jumlah artikel yang perlu dikelola dan dikelompokkan secara sistematis. Website majelistablighpwmjateng.com menyediakan berbagai artikel dakwah dengan kategori yang beragam, sehingga proses pengategorian secara manual berpotensi memerlukan waktu lebih lama dan menghasilkan pelabelan yang kurang konsisten. Penelitian ini bertujuan membangun model klasifikasi kategori artikel berdasarkan judul serta membandingkan kinerja algoritma *Naive Bayes* dengan *Support Vector Machine* (*SVM*) dalam mengklasifikasikan artikel dakwah berbahasa Indonesia. Data penelitian berupa judul artikel beserta kategorinya yang diperoleh dari website majelistablighpwmjateng.com pada periode 10 Juni 2023 hingga 9 Juni 2026, dengan dataset awal berjumlah 1.369 data dalam 19 kategori. Kategori Khutbah Nikah yang hanya memiliki satu data dihapus sehingga dataset akhir terdiri atas 1.368 data dalam 18 kategori. Tahapan penelitian meliputi pengumpulan data, preprocessing (cleaning, case folding, tokenisasi, dan stopword removal), pembobotan fitur dengan *Term Frequency-Inverse Document Frequency (TF-IDF)*, pelatihan model, serta evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian menunjukkan bahwa algoritma *SVM* menghasilkan kinerja lebih baik daripada *Naive Bayes*, dengan *accuracy* 79,56%, *precision* 72,11%, *recall* 79,56%, dan *F1-score* 74,36%, sedangkan *Naive Bayes* memperoleh *accuracy* 73,36%, *precision* 64,73%, *recall* 73,36%, dan *F1-score* 66,86%. Perbedaan tersebut mengindikasikan bahwa *SVM* lebih efektif dalam menangani data teks berdimensi tinggi dan pola kategori yang kompleks, meskipun ketidakseimbangan jumlah data pada tiap kategori masih memengaruhi kemampuan model dalam mengidentifikasi kelas minoritas. Berdasarkan hasil penelitian, kombinasi *TF-IDF* dan *SVM* lebih sesuai digunakan untuk mengklasifikasikan kategori judul artikel pada website majelistablighpwmjateng.com dibandingkan kombinasi *TF-IDF* dan *Naive Bayes*.

Kata kunci: klasifikasi teks; judul artikel; *TF-IDF*; *Naive Bayes*; *Support Vector Machine*; website dakwah.

1. PENDAHULUAN

Perkembangan teknologi informasi yang pesat telah mengubah cara manusia memproduksi, mendistribusikan, dan mengelola data tekstual. Berbagai platform digital seperti portal berita, media sosial, dan website organisasi menghasilkan volume teks yang sangat besar setiap hari, sehingga menimbulkan tantangan baru dalam pengorganisasian dan pengelompokan dokumen secara manual (Alfando & Hayami, 2023). Kebutuhan sistem klasifikasi otomatis untuk dokumen berbahasa Indonesia, termasuk konten keagamaan pada website dakwah, masih memerlukan pengelompokan kategori yang konsisten dan efisien (Aziza et al., 2025)

Website majelistablighpwmjateng.com merupakan situs resmi Majelis Tabligh Pimpinan Wilayah Muhammadiyah Jawa Tengah yang memuat artikel dakwah dengan kategori beragam, antara lain Artikel, Dinamika Persyarikatan, Dinamika Tabligh, Khutbah Jum'at, dan Tokoh. Berdasarkan data yang dihimpun, terdapat 1.369 artikel yang terdistribusi ke dalam 19 kategori berbeda. Banyaknya kategori tersebut menunjukkan bahwa proses pelabelan manual memiliki kelemahan, baik dari sisi waktu maupun konsistensi hasil, karena sangat bergantung pada perspektif individu (Rani et al., 2025).

Klasifikasi teks merupakan salah satu cabang *text mining* yang digunakan untuk mengelompokkan dokumen ke dalam kelas tertentu berdasarkan karakteristik isi teks. Judul artikel dipilih sebagai sumber data karena merepresentasikan inti pembahasan secara ringkas dan membutuhkan waktu komputasi yang lebih singkat dibandingkan pengolahan seluruh isi dokumen (Dewi et al., 2024; Ramadhan, 2025). Setelah melalui tahap *preprocessing* yang lazim digunakan pada klasifikasi teks berbahasa Indonesia (Salsabila et al., 2022), teks diekstraksi menjadi fitur numerik menggunakan metode *Term Frequency-Inverse Document Frequency (TF-IDF)* karena mampu merepresentasikan tingkat kepentingan suatu kata dalam dokumen berdasarkan frekuensi kemunculan dan persebarannya pada seluruh korpus (Efrizoni et al., 2022).

Beberapa penelitian terdahulu menunjukkan bahwa kombinasi TF-IDF dan *Naive Bayes* masih relevan untuk klasifikasi teks berbahasa Indonesia, namun hasilnya sangat bergantung pada karakteristik data (Pranata et al., 2024; Winarti et al., 2021). Di sisi lain, *Support Vector Machine* (*SVM*) secara konsisten menunjukkan performa yang

lebih unggul dibanding *Naive Bayes* pada berbagai penelitian teks berbahasa Indonesia. Alfando dan Hayami (2023) melaporkan *SVM* mencapai akurasi tertinggi 94,24% dalam klasifikasi teks berita, sedangkan Nashir dkk. (2025) melaporkan *SVM* mencapai akurasi 98% dibandingkan *Naive Bayes* yang hanya 88% pada analisis sentimen komentar YouTube. Tuhenay dan Mailoa (2021) juga menunjukkan bahwa *SVM* (0,9634) lebih unggul dibanding *Naive Bayes Classifier* (0,9378) pada identifikasi bahasa daerah, sementara Alkaff dkk. (2021) membuktikan *SVM* bekerja sangat baik pada data teks berdimensi tinggi dan bersifat sparse.

Naive Bayes memiliki keunggulan pada kesederhanaan model, kecepatan komputasi, dan kemampuan bekerja baik pada data latih yang tidak terlalu besar, sedangkan *SVM* bekerja dengan menemukan hyperplane optimal yang memisahkan kelas dengan margin maksimum sehingga lebih stabil pada data teks berdimensi tinggi dan sparse (Nashir et al., 2025). Pada konteks artikel dakwah yang kaya istilah keagamaan dan memiliki distribusi kategori tidak seimbang, pemilihan algoritma yang mampu mempertahankan performa meskipun data bersifat imbalanced menjadi faktor penting (Muhammadin & Sobari, 2021).

Berdasarkan uraian tersebut, penelitian ini bertujuan (1) membangun model klasifikasi kategori artikel secara otomatis berdasarkan judul artikel pada website majelistablighpwmjateng.com menggunakan *TF-IDF* sebagai ekstraksi fitur, dan (2) membandingkan kinerja algoritma *Naive Bayes* dan *Support Vector Machine* untuk menentukan algoritma yang paling sesuai dengan karakteristik data artikel dakwah berbahasa Indonesia. Penelitian ini diharapkan memberikan kontribusi praktis dalam pengelolaan artikel dakwah secara lebih efisien serta kontribusi akademis dalam memperkaya literatur text mining berbahasa Indonesia pada domain dakwah yang masih terbatas eksplorasinya.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komparatif untuk menguji secara objektif efektivitas algoritma *Naive Bayes* dan *Support Vector Machine* dalam mengklasifikasikan artikel pada website majelistablighpwmjateng.com. Penelitian difokuskan pada proses klasifikasi teks dari

judul artikel ke dalam kategori tertentu, tanpa membangun sistem website secara menyeluruh. Tahapan penelitian meliputi pengumpulan data, preprocessing, pembentukan fitur *TF-IDF*, pembagian data latih dan uji, pelatihan model *Naive Bayes* dan *SVM* secara terpisah, serta evaluasi kinerja menggunakan *accuracy, precision, recall, F1-score,* dan *cross validation 5-fold*.

2.1. Sumber dan Objek Data

Data penelitian merupakan data sekunder berupa judul artikel dan kategorinya yang diperoleh dari website majelistablighpwmjateng.com pada rentang waktu 10 Juni 2023 hingga 9 Juni 2026, tersedia dalam format .xlsx. Data yang bersifat duplikat, kosong, atau tidak memiliki label kategori yang jelas dieliminasi untuk menjaga integritas dataset. Dataset awal berjumlah 1.369 data dalam 19 kategori. Karena kategori Khutbah Nikah hanya memiliki satu data sehingga tidak memadai untuk proses stratified train-test split, kategori tersebut dihapus sehingga dataset akhir menjadi 1.368 data dalam 18 kategori.

2.2. *Preprocessing Data*

Tahap *preprocessing* dilakukan pada kolom judul melalui empat proses: (1) *case folding*, mengubah seluruh huruf menjadi huruf kecil; (2) *cleaning*, menghapus tanda baca, angka, simbol, dan karakter yang tidak diperlukan; (3) *tokenisasi*, memecah kalimat menjadi kata-kata tunggal; dan (4) *stopword removal*, menghapus kata-kata umum berbahasa Indonesia yang tidak berkontribusi signifikan terhadap proses klasifikasi. Seluruh tahapan diimplementasikan menggunakan bahasa pemrograman *Python* pada platform *Google Colab*. Selanjutnya label kategori diubah menjadi nilai numerik menggunakan *Label Encoder*, menghasilkan 18 kelas dengan rentang indeks 0 sampai 17.

2.3. *Ekstraksi Fitur TF-IDF* dan Pembagian Data

Teks judul yang telah dibersihkan ditransformasikan ke dalam representasi *numerik* menggunakan *TF-IDF* Vectorizer dengan parameter max_features=5000, ngram_range=(1,2), min_df=2, max_df=0.95, dan sublinear_tf=True. Dataset selanjutnya dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan

train_test_split secara *stratified* (*random_state*=42), menghasilkan 1.094 data latih (80%) dan 274 data uji (20%). Proses *TF-IDF* menghasilkan *matriks* fitur berukuran 1.094 × 1.265 pada data latih dan 274 × 1.265 pada data uji, atau sebanyak 1.265 fitur unik.

2.4. Klasifikasi *Naive Bayes* dan *Support Vector Machine*

Naive Bayes diimplementasikan sebagai algoritma klasifikasi probabilistik yang menentukan label kelas berdasarkan estimasi *probabilitas* fitur *TF-IDF*. Model dilatih menggunakan data latih melalui perhitungan probabilitas kemunculan fitur pada setiap kelas, kemudian diuji pada data uji untuk memprediksi kategori artikel. *Support Vector Machine* (SVM) diimplementasikan dengan *kernel linear,* yang umum digunakan pada klasifikasi teks berdimensi tinggi seperti hasil representasi *TF-IDF*. Model *SVM* dilatih untuk menentukan *hyperplane* pemisah optimal antar kelas, kemudian diuji pada data uji yang sama dengan *Naive Bayes* untuk menjamin objektivitas perbandingan.

2.5. Metode Evaluasi

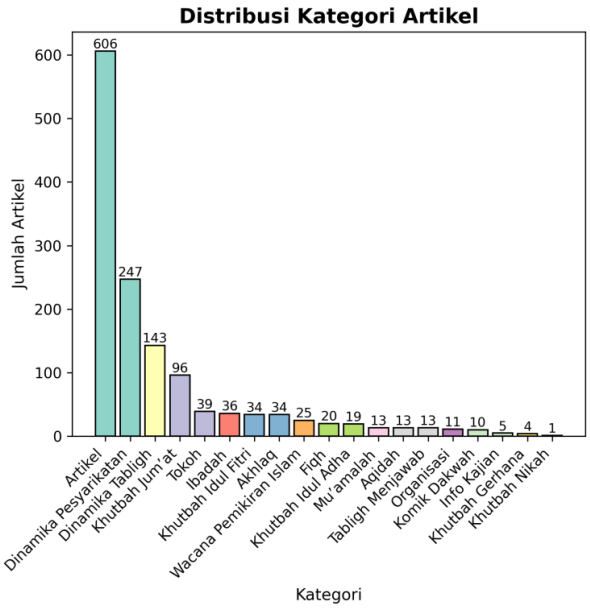
Evaluasi performa kedua model dilakukan menggunakan empat metrik, yaitu *accuracy* (ketepatan prediksi secara keseluruhan), *precision* (ketepatan model dalam memprediksi suatu kelas), *recall* (kemampuan model menemukan seluruh data yang benar-benar termasuk kelas tertentu), dan *F1-score* (rata-rata harmonik *precision* dan *recall*). Evaluasi juga didukung teknik *cross validation* 5-fold untuk mendapatkan hasil yang lebih valid dan reliabel. Proses penelitian dilakukan menggunakan *Python* pada *platform Google Colab*, dengan bantuan pustaka *Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Re, dan NLTK/Sastrawi*.

3. HASIL DAN PEMBAHASAN

3.1. Distribusi Data

Dataset penelitian terdiri atas 1.369 artikel yang terbagi ke dalam 19 kategori. Data menunjukkan adanya ketimpangan yang cukup signifikan (*class imbalance*): kategori Artikel mendominasi dengan 660 data (44,3%), diikuti Dinamika Pesyarikatan 247 data (18,0%) dan Dinamika Tabligh 143 data (10,4%), sedangkan beberapa kategori

seperti Info Kajian, Khutbah Gerhana, dan Khutbah Nikah hanya memiliki 5, 4, dan 1 data. Sebaran distribusi kategori artikel dapat dilihat pada **Gambar 3. 1** di bawah ini :



Gambar 3. 1 Sebaran Kategori

Ketimpangan distribusi ini berpotensi menyebabkan model lebih optimal mempelajari pola kelas mayoritas namun kurang sensitif terhadap kelas minoritas (Sihombing & Yuliati, 2021). Kategori Khutbah Nikah dihapus karena jumlah datanya tidak memadai untuk *stratified split*, sehingga *dataset* akhir menjadi 1.368 data dalam 18 kategori.

3.2. Hasil Evaluasi Model

Model *Naive Bayes* dan *SVM* diuji menggunakan 274 data uji setelah dilatih pada 1.094 data latih dengan representasi fitur *TF-IDF* sebanyak 1.265 fitur. Hasil evaluasi kedua model disajikan pada Tabel 3. 1 di bawah ini:

Tabel 3. 1 Perbandingan Kinerja *Naive Bayes* dan *SVM*

Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***CV Mean***	***CV Std***
Naive Bayes	*0,7336 (73,36%)*	*0,6473 (64,73%)*	*0,7336 (73,36%)*	*0,6686 (66,86%)*	*0,6846*	*0,0231*
Support Vector Machine (SVM)	*0,7956 (79,56%)*	*0,7211 (72,11%)*	*0,7956 (79,56%)*	*0,7436 (74,36%)*	*0,7349*	*0,0105*

Berdasarkan data Tabel 3. 1, algoritma *Support Vector Machine* menunjukkan performa yang lebih baik dibandingkan *Naive Bayes* pada seluruh metrik evaluasi. *SVM* memperoleh *accuracy* 79,56%, *precision* 72,11%, *recall* 79,56%, dan *F1-score* 74,36%, sedangkan *Naive Bayes* memperoleh *accuracy* 73,36%, *precision* 64,73%, *recall* 73,36%, dan *F1-score* 66,86%. Selisih *accuracy* sebesar 6,2 poin persentase menunjukkan bahwa proporsi prediksi benar yang dihasilkan *SVM* lebih besar dibandingkan *Naive Bayes*. Keunggulan *SVM* juga diperkuat oleh hasil *cross validation 5-fold*, di mana *SVM* menghasilkan rata-rata 0,7349 dengan standar deviasi 0,0105, sedangkan *Naive Bayes* menghasilkan rata-rata 0,6846 dengan standar deviasi 0,0231. Nilai rata-rata yang lebih tinggi dan standar deviasi yang lebih rendah pada *SVM* menunjukkan performa yang lebih baik sekaligus lebih stabil pada setiap lipatan pengujian.

3.3. *Analisis Confusion Matrix*

Confusion matrix Naive Bayes menunjukkan bahwa prediksi benar paling banyak terkonsentrasi pada kelas-kelas besar seperti kategori Artikel, sedangkan kelas kecil lebih sering salah diprediksi ke kelas lain, yang mengindikasikan ketidakseimbangan distribusi data cukup memengaruhi hasil klasifikasi. *Confusion matrix SVM* memperlihatkan diagonal utama yang lebih kuat dan merata pada beberapa kelas penting dibandingkan *Naive Bayes*, menandakan bahwa *SVM* memiliki kemampuan pemisahan kelas yang lebih baik dalam mengenali kategori judul artikel, meskipun beberapa kelas minoritas masih memiliki precision dan recall rendah pada kedua model.

3.4. Pembahasan

Hasil penelitian menunjukkan bahwa *SVM* memiliki kinerja lebih baik dibandingkan *Naive Bayes* pada seluruh metrik evaluasi. Secara teoretis, *SVM* dikenal memiliki kemampuan yang baik dalam menangani data dengan jumlah fitur besar seperti hasil pembobotan *TF-IDF*. Pada penelitian ini, jumlah fitur yang dihasilkan mencapai 1.265 fitur sehingga ruang representasi data menjadi sangat luas, dan dalam kondisi tersebut *SVM* mampu mencari hyperplane pemisah antar kelas secara lebih optimal dibandingkan *Naive Bayes*.

Performa *Naive Bayes* yang berada di bawah *SVM* dapat dipahami dari asumsi dasar algoritma tersebut yang menganggap setiap fitur bersifat independen. Pada data teks, asumsi ini tidak sepenuhnya sesuai karena kata-kata dalam judul artikel sering memiliki keterkaitan satu sama lain, sehingga model kurang mampu menangkap pola hubungan antar kata secara menyeluruh, terutama pada kategori dengan jumlah data kecil.

Selain karakteristik algoritma, hasil penelitian ini juga sangat dipengaruhi oleh distribusi data yang tidak seimbang, di mana kategori Artikel mendominasi jumlah data dibandingkan kategori lainnya. Ketidakseimbangan tersebut membuat model lebih banyak mempelajari pola dari kelas mayoritas, sementara kelas minoritas memiliki representasi yang jauh lebih terbatas, terlihat dari nilai precision dan recall yang rendah bahkan bernilai nol pada beberapa kategori. Kondisi ini menunjukkan bahwa performa klasifikasi tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh keseimbangan dan kecukupan data latih pada setiap kelas.

Temuan penelitian ini sejalan dengan sejumlah penelitian terdahulu yang menunjukkan keunggulan *SVM* dalam tugas klasifikasi teks berbahasa Indonesia, seperti Alfando dan Hayami (2023), Nashir dkk. (2025), Tuhenay dan Mailoa (2021), serta Alkaff dkk. (2021), yang seluruhnya melaporkan *SVM* lebih unggul dibandingkan *Naive Bayes* atau metode probabilistik lain pada data teks berdimensi tinggi. *Naive Bayes* tetap memiliki kelebihan dari sisi kesederhanaan dan kecepatan komputasi, namun dalam kasus artikel dakwah pada penelitian ini keunggulan tersebut belum mampu mengimbangi performa *SVM*. Dengan demikian, kombinasi *TF-IDF* dan *SVM* merupakan pendekatan yang paling sesuai untuk dataset judul artikel yang digunakan, meskipun untuk meningkatkan performa pada kelas minoritas diperlukan teknik penyeimbangan data pada penelitian lanjutan.

4. KESIMPULAN

Berdasarkan hasil penelitian mengenai klasifikasi kategori artikel menggunakan metode *TF-IDF* dan algoritma *Naive Bayes* serta *Support Vector Machine*, dapat disimpulkan bahwa proses pengolahan data berhasil menghasilkan dataset yang layak digunakan untuk klasifikasi, yaitu 1.368 data dalam 18 kategori setelah pembersihan

dari 1.369 data awal dalam 19 kategori. Hasil pengujian menunjukkan bahwa algoritma *Support Vector Machine* memiliki kinerja lebih baik dibandingkan *Naive Bayes* pada seluruh metrik evaluasi, dengan accuracy 79,56%, precision 72,11%, recall 79,56%, dan F1-score 74,36%, sedangkan *Naive Bayes* memperoleh accuracy 73,36%, precision 64,73%, recall 73,36%, dan F1-score 66,86%. Perbedaan performa tersebut dipengaruhi oleh karakteristik data teks yang berdimensi tinggi serta ketidakseimbangan kelas, di mana *SVM* lebih mampu membentuk batas pemisah antar kelas secara optimal sedangkan *Naive Bayes* lebih dipengaruhi oleh dominasi kelas mayoritas. Dengan demikian, kombinasi *TF-IDF* dan *SVM* dinilai lebih sesuai digunakan untuk klasifikasi kategori judul artikel pada website majelistablighpwmjateng.com dibandingkan kombinasi *TF-IDF* dan *Naive Bayes*. Penelitian selanjutnya disarankan menambah jumlah data pada kategori minoritas, menerapkan teknik penyeimbangan data seperti oversampling atau undersampling, serta membandingkan algoritma lain seperti Random Forest, Logistic Regression, atau metode berbasis deep learning untuk memperoleh gambaran yang lebih luas mengenai metode terbaik untuk klasifikasi teks dakwah berbahasa Indonesia.

DAFTAR PUSTAKA

Alfando, A., & Hayami, R. (2023). Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning dan Deep Learning: Studi Literatur. JATI (Jurnal Mahasiswa Teknik Informatika), 7(1), 681–686. https://doi.org/10.36040/jati.v7i1.6486

Alkaff, M., Baskara, A. R., & Maulani, I. (2021). Klasifikasi Laporan Keluhan Pelayanan Publik Berdasarkan Instansi Menggunakan Metode LDA-*SVM*. Jurnal Teknologi Informasi dan Ilmu Komputer, 8(6), 1265–1276. https://doi.org/10.25126/jtiik.2021863768

Aziza, A., Hasanah, R., Juairiah, J., & Munsyi, M. (2025). Analyzing radicalism sentiments in Indonesian da'wah content on website da'wah through text mining techniques. International Journal of Informatics and Communication Technology (IJ-ICT), 14(2), 575. https://doi.org/10.11591/ijict.v14i2.pp575-585

Dewi, N. L. P. R., Wijaya, I. N. S. W., Purnamawan, I. K., & Marti, N. W. (2024). Model Classifier Judul Berita Pariwisata Indonesia Berdasarkan Sentimen. Jurnal Teknologi Informasi dan Ilmu Komputer, 11(1), 117–124. https://doi.org/10.25126/jtiik.20241117617

Efrizoni, L., Defit, S., Tajuddin, M., & Anggrawan, A. (2022). Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21(3), 653–666. https://doi.org/10.30812/matrik.v21i3.1851

Iqrom, R. A., & Kurniawan, T. B. (2023). Multi-Label Text Classification for Indonesian IT Journal Documents. (Studi kasus klasifikasi dokumen jurnal berbahasa Indonesia).

Muhammadin, F., & Sobari, T. (2021). Analisis Sentimen Ulasan Aplikasi Kredivo Menggunakan Algoritma *Support Vector Machine* dan *Naive Bayes*.

Nashir, M., Kurnia, D., Wijaya, A., et al. (2025). Analisis Sentimen Komentar YouTube Terkait Demonstrasi DPR Menggunakan Algoritma *SVM* dan *Naive Bayes*.

Pranata, R., Rudiman, R., & Verdikha, N. A. (2024). Klasifikasi Teks Quick Count Pemilihan Presiden 2024 Menggunakan Kombinasi *TF-IDF* dan *Naive Bayes*.

Ramadhan, F. (2025). Klasifikasi Ganda Topik dan Sentimen Judul Berita Berbahasa Indonesia Menggunakan *TF-IDF* dan *Support Vector Machine*.

Rani, D., Aslamiah, A., Amrullah, A., et al. (2025). Potensi Otomatisasi Pengelolaan Konten Website Dakwah Menggunakan Algoritma Klasifikasi Teks.

Salsabila, A. D., Helen, A., & Yuliawati, D. (2022). Pengaruh Tahap Preprocessing terhadap Kinerja Klasifikasi Teks Berbahasa Indonesia.

Sihombing, P. R., & Yuliati, A. (2021). Penanganan Ketidakseimbangan Kelas (Class Imbalance) pada Model Klasifikasi Data.

Tuhenay, D. J., & Mailoa, E. (2021). Identifikasi Bahasa Daerah Menggunakan *Naive Bayes* Classifier dan *Support Vector Machine*. Jurnal Ilmiah, 0,9634 vs 0,9378 akurasi *SVM* dan NBC.

Wati, R., Ernawati, S., & Rachmi, A. (2023). Implementasi Multinomial *Naive Bayes* dengan *TF-IDF* untuk Klasifikasi Sentimen Teks Berbahasa Indonesia.

Winarti, T., Indriyawati, H., Vydia, V., et al. (2021). Perbandingan *Naive Bayes* dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia.