

RANCANG BANGUN SISTEM KLASIFIKASI JUDUL BERITA ONLINE BAHASA INDONESIA BERBASIS WEB MENGGUNAKAN ALGORITMA NAIVE BAYES

Ana Mardiana¹; Willy Prihartono²

Program Studi Manajemen Informatika¹
Program Studi Komputerisasi Akuntansi²

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
amardiyana302@gmail.com, willyprihartono@gmail.com

(*) Corresponding Author : amardiyana302@gmail.com
Published : 31 Desember 2025

Abstract— The increasing consumption of digital news in Indonesia has led to a growing number of online news headlines published every day. News headlines are generally short and ambiguous, making manual categorization based on topic difficult. However, effective automated classification systems for short Indonesian-language news headlines remain limited, particularly for media institutions and users who require fast classification. This study aims to design and develop a web-based system for classifying Indonesian online news headlines using the Term Frequency-Inverse Document Frequency (TF-IDF) method for feature weighting and the Naïve Bayes algorithm for classification. The dataset consists of 1,737 news headlines collected from several national news portals. In addition to category classification, this study applies sentiment analysis as a data balancing approach to address class distribution imbalance. The system is developed using Python for model implementation and PHP-MySQL for web deployment. Experimental results show that the model achieves high accuracy on training data (93–98%) but experiences a performance decline on testing data (54–57%) due to a tendency toward overfitting. Nevertheless, the system is able to classify news headlines automatically and consistently, and can be used as an initial solution for news grouping to support media institutions in managing news archives and to serve as a foundation for further development of Indonesian text classification systems.

Keywords: Text classification, Naïve Bayes, Online news, Sentiment analysis.

Abstrak—Peningkatan konsumsi berita digital di Indonesia menyebabkan bertambahnya jumlah judul berita daring setiap hari. Judul berita umumnya bersifat singkat dan ambigu, sehingga menyulitkan proses pengelompokan manual berdasarkan kategori topik. Namun, hingga saat ini masih terbatas sistem klasifikasi otomatis yang efektif untuk judul berita pendek berbahasa Indonesia, khususnya bagi institusi media dan pengguna yang membutuhkan klasifikasi cepat. Penelitian ini bertujuan merancang dan membangun sistem klasifikasi judul berita online berbahasa Indonesia berbasis web menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagai pembobotan fitur dan algoritma *Naïve Bayes* sebagai metode klasifikasi. Dataset yang digunakan terdiri dari 1.737 judul berita yang dikumpulkan dari beberapa portal berita nasional. Selain klasifikasi kategori, penelitian ini menerapkan analisis sentimen sebagai pendekatan penyeimbangan data untuk mengatasi ketidakseimbangan distribusi kelas. Sistem dikembangkan menggunakan Python untuk pemodelan dan PHP-MySQL untuk implementasi web. Hasil pengujian menunjukkan bahwa model mencapai akurasi tinggi pada data pelatihan (93–98%), namun mengalami penurunan pada data pengujian (54–57%) akibat kecenderungan overfitting. Meskipun demikian, sistem mampu melakukan klasifikasi judul berita secara otomatis dan konsisten, serta dapat digunakan sebagai solusi awal pengelompokan berita sehingga dapat membantu institusi media dalam pengelolaan arsip berita serta menjadi dasar pengembangan sistem klasifikasi teks bahasa Indonesia yang lebih lanjut.

Kata kunci: Klasifikasi teks, Naïve Bayes, Berita online, Analisis sentimen.

INTRODUCTION

Perkembangan teknologi informasi dan komunikasi telah mendorong pertumbuhan pesat media daring di Indonesia. Portal berita nasional seperti Kompas.com, Detik.com, dan Liputan6.com secara rutin menerbitkan ratusan hingga ribuan berita setiap hari. Tingginya volume informasi ini menimbulkan tantangan dalam pengelolaan, pengelompokan, dan pencarian berita secara efisien, terutama apabila masih dilakukan secara manual.

Dalam konteks tersebut, sistem klasifikasi berita otomatis menjadi kebutuhan penting untuk membantu proses pengorganisasian informasi digital. Klasifikasi teks memungkinkan pengelompokan dokumen ke dalam kategori tertentu berdasarkan karakteristik linguistiknya, sehingga pengguna dapat memperoleh informasi yang relevan secara lebih cepat dan terstruktur (Fauzi, 2017).

Dalam berita daring, judul berita sering kali berfungsi sebagai representasi ringkas dari isi berita. Judul bersifat padat, informatif, dan langsung menyoroti inti peristiwa, sehingga memiliki potensi besar untuk digunakan sebagai dasar klasifikasi tanpa harus memproses keseluruhan isi artikel. Berbagai penelitian sebelumnya telah membuktikan efektivitas TF-IDF dan Naïve Bayes dalam pengolahan teks berbahasa Indonesia, termasuk pada klasifikasi berita hoaks, berita teknologi, serta pengelompokan berita daring berskala besar (Rahman et al., 2017; Santoso, 2020; Sudrajat et al., 2022; Ionendri et al., 2025; Nisa et al., 2025).

Namun demikian, tantangan utama dalam klasifikasi berita online adalah sifat judul yang singkat dan ambigu, serta distribusi kategori berita yang tidak seimbang. Beberapa kategori seperti politik dan ekonomi cenderung mendominasi, sementara kategori lain memiliki jumlah data yang lebih sedikit. Kondisi ini menyebabkan model klasifikasi berpotensi bias terhadap kelas mayoritas dan menurunkan akurasi pada kategori minoritas.

Meskipun penelitian terkait klasifikasi berita menggunakan TF-IDF dan Naïve Bayes telah banyak dilakukan, terdapat beberapa kesenjangan penelitian yang masih perlu dikaji lebih lanjut. Pertama, sebagian besar penelitian terdahulu berfokus pada klasifikasi berdasarkan isi berita lengkap, bukan pada judul berita yang bersifat teks pendek dan lebih menantang untuk dianalisis.

Kedua, permasalahan ketidakseimbangan data kategori sering kali hanya ditangani melalui teknik oversampling atau undersampling konvensional, tanpa mempertimbangkan karakteristik semantik teks. Pemanfaatan analisis

sentimen sebagai tahap awal penyeimbangan data sebelum klasifikasi kategori masih relatif jarang diterapkan, khususnya pada judul berita berbahasa Indonesia.

Ketiga, masih terbatas penelitian yang mengintegrasikan model klasifikasi berbasis TF-IDF dan Naïve Bayes ke dalam sistem berbasis web yang siap digunakan secara praktis, terutama dengan fokus pada evaluasi pengaruh balancing data terhadap kinerja klasifikasi kategori berita.

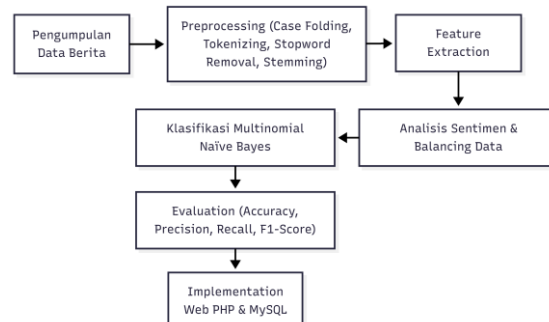
Berdasarkan kesenjangan tersebut, penelitian ini memberikan beberapa kontribusi utama. Pertama, penelitian ini mengusulkan pendekatan klasifikasi otomatis berbasis judul berita online berbahasa Indonesia, sehingga mampu mengolah teks pendek yang padat dan ambigu secara efisien.

Kedua, penelitian ini mengintegrasikan analisis sentimen berbasis leksikon sebagai pendekatan penyeimbangan data sebelum proses klasifikasi kategori dilakukan, guna mengurangi bias model terhadap kelas mayoritas.

Ketiga, penelitian ini merancang dan mengimplementasikan sistem klasifikasi berita berbasis web yang mengombinasikan TF-IDF dan algoritma Naïve Bayes, serta mengevaluasi kinerjanya menggunakan metrik akurasi, presisi, recall, dan F1-score. Kontribusi ini diharapkan dapat memperkaya kajian di bidang data mining dan Natural Language Processing (NLP) berbahasa Indonesia, sekaligus memberikan solusi praktis bagi pengelolaan berita daring.

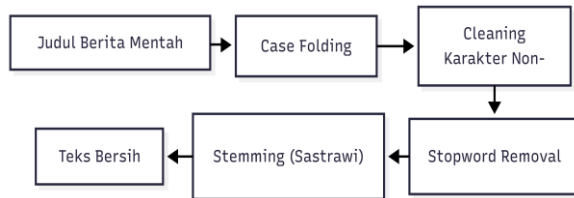
MATERIALS AND METHODS

Penelitian ini menggunakan data berita daring berbahasa Indonesia yang diperoleh melalui proses web scraping dari portal berita nasional seperti Kompas.com, Detik.com, dan Liputan6.com pada periode Oktober–November 2025. Dataset terdiri dari 1.737 judul berita yang telah dilabeli kategori dan sentimen. Setiap judul diperlakukan sebagai teks pendek yang merepresentasikan isi berita, sehingga efisien untuk pemrosesan otomatis.



Gambar 1 KDD Pengolahan Data

Tahapan pengolahan data mengikuti kerangka Knowledge Discovery in Databases (KDD), yang meliputi seleksi data, preprocessing, transformasi fitur, data mining, dan evaluasi. Dataset diperoleh melalui proses scraping dari beberapa portal berita dan digabungkan menjadi satu dataset, sedangkan alur lengkap tahapan penelitian ditampilkan pada Gambar 1



Gambar 2 Tahap Preprocessing Teks

Berdasarkan Gambar 2 pada tahap preprocessing teks judul berita melalui tahapan proses case folding, pembersihan karakter non-alfanumerik, penghapusan stopwords, serta stemming menggunakan pustaka Sastrawi untuk bahasa Indonesia. Tahap ini bertujuan untuk menstandarkan teks dan mengurangi noise sebelum ekstraksi fitur dilakukan. Namun, pada tahap prediksi di aplikasi web, stemming dibatasi untuk menjaga konteks semantik kata pada judul yang bersifat sangat singkat.

	Bencana	Ekonomi	Internasi...	Kriminal	Lainnya	Nasional	Pendidikan	Sosial
aam	-10.4646...	-10.4726...	-10.0278...	-10.4786...	-10.2719...	-10.4771...	-10.4399...	-10.4486...
aam gus	-10.4646...	-10.4726...	-10.4632...	-10.4786...	-10.2590...	-10.4771...	-10.4399...	-10.4486...
aam nilai	-10.4646...	-10.4726...	-10.2872...	-10.4786...	-10.5248...	-10.4771...	-10.4399...	-10.4486...
aam pecat	-10.4646...	-10.4726...	-10.2872...	-10.4786...	-10.5248...	-10.4771...	-10.4399...	-10.4486...
aam urus	-10.4646...	-10.4726...	-10.2872...	-10.4786...	-10.5248...	-10.4771...	-10.4399...	-10.4486...
ab	-10.4646...	-10.4726...	-10.4632...	-10.4786...	-10.5248...	-10.2888...	-10.4399...	-10.4486...
ab yordania	-10.4646...	-10.4726...	-10.4632...	-10.4786...	-10.5248...	-10.2888...	-10.4399...	-10.4486...
abad	-10.2133...	-10.4726...	-10.2358...	-10.4786...	-10.3369...	-10.4771...	-10.4399...	-10.4486...

Gambar 3 Ekstraksi Fitur TF-IDF

Ekstraksi fitur dilakukan menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) dengan konfigurasi unigram dan bigram. TF-IDF digunakan untuk mengubah teks menjadi representasi numerik berdimensi tinggi yang dapat diproses oleh algoritma klasifikasi. Representasi matriksi TF-IDF hasil transformasi ditunjukkan pada Gambar 3.

=== RINGKASAN HASIL KLASIFIKASI ===

Total Data : 1733
Data Training : 1213
Data Testing : 520

=== RINGKASAN HASIL KLASIFIKASI ===

Total Data : 1733
Data Training (Bal) : 2576 (Kategori), 1317 (Sentimen)
Data Testing : 520

Gambar 4 Hasil Train (70%) dan Test (30%)

Berdasarkan gambar 4, Dataset dibagi menjadi data training (70%) dan data testing (30%). Rasio ini dipilih untuk memberikan keseimbangan antara pembelajaran model dan evaluasi performa. Selain dataset asli, dibentuk pula dataset seimbang melalui pendekatan balancing berbasis analisis sentimen, dengan tujuan mengurangi bias akibat ketidakseimbangan distribusi kategori.

label_encoder_category	18/12/2025 19:50	JOBLIB File	1 KB
label_encoder_sentiment	19/12/2025 22:23	JOBLIB File	1 KB
model_category	19/12/2025 22:23	JOBLIB File	3.474 KB
model_sentiment	19/12/2025 22:23	JOBLIB File	1.095 KB
tfidf_kategori	19/12/2025 22:23	JOBLIB File	903 KB
tfidf_sentimen	19/12/2025 22:23	JOBLIB File	757 KB

Gambar 5 Penyimpanan Model Joblib

Klasifikasi dilakukan menggunakan algoritma Multinomial Naïve Bayes untuk dua model terpisah, yaitu model klasifikasi kategori berita dan model klasifikasi sentimen. Model dilatih menggunakan data training dan diuji menggunakan data testing. Proses pelatihan dan penyimpanan model dalam format .joblib ditunjukkan pada Gambar 5.



Gambar 6 Alur Sistem web dan DFD Level 0

Sistem kemudian diimplementasikan dalam bentuk aplikasi berbasis web menggunakan PHP dan MySQL, dengan integrasi model Python melalui mekanisme pemanggilan skrip. Arsitektur sistem dan alur pemrosesan pengguna divisualisasikan dalam Data Flow Diagram (DFD) Level 0 yang menggambarkan interaksi antara pengguna dan sistem klasifikasi seperti pada Gambar 6.

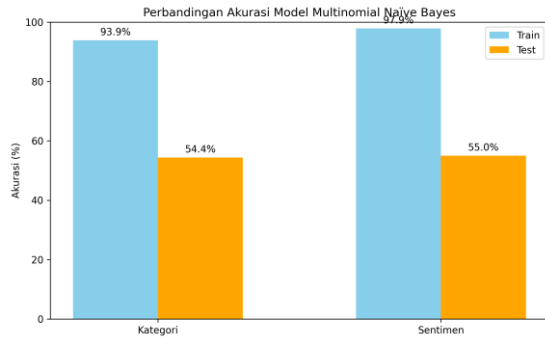
RESULTS AND DISCUSSION

A ^B Judul Berita	A ^B I...	A ^B Tanggal	A ^B Kategori	A ^B Sentim...	A ^B Link sumber
Update Korban Banji...	Jakarta - ...	30/11/2025 01:01...	Bencana	Negatif	https://news.detik.com,
Menteri Agus ke Kelo...	Jakarta - ...	30/11/2025 00:40...	Nasional	Positif	https://news.detik.com,
IKN Bangun PLTS 50 ...	PLTS IKN...	30/11/2025 00:05...	Ekonomi	Netral	https://www.liputan6.c
Jokowi Disebut Akan ...	Dengan ...	30/11/2025 00:05...	Lainnya	Positif	https://www.liputan6.c
PM Australia Anthon...	Jodie Ha...	30/11/2025 00:03...	Ekonomi	Netral	https://www.liputan6.c

Gambar 7 Dataset Hasil Scraping dan Integrasi

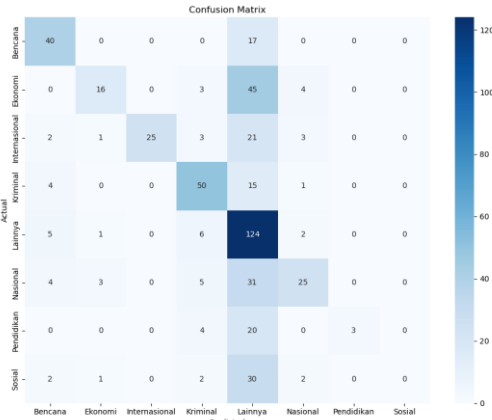
Hasil penelitian menunjukkan bahwa sistem klasifikasi judul berita online berbasis web berjalan sesuai rancangan. Proses scraping dan integrasi data berhasil menghasilkan dataset

ditunjukkan pada gambar 7, data ini sudah terstruktur dan siap digunakan untuk pelatihan model.



Gambar 8 Perbandingan Akurasi Model

Implementasi preprocessing dan ekstraksi fitur menghasilkan representasi TF-IDF dengan lebih dari 33 ribu kata unik. Pada gambar 8 hasil Evaluasi performa model menunjukkan akurasi tinggi pada data training, yaitu 93–98% untuk klasifikasi kategori maupun sentimen, baik pada dataset asli maupun seimbang. Sebaliknya, akurasi pada data testing berada di kisaran 54–57%, mengindikasikan adanya overfitting.



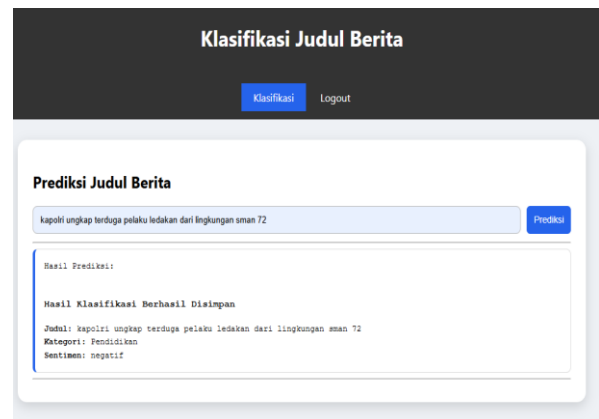
Gambar 9 Confusion Matrix

Confusion matrix menunjukkan bahwa sebagian besar prediksi benar terkonsentrasi pada kategori mayoritas, sedangkan kesalahan klasifikasi lebih banyak terjadi pada kategori minoritas. Hal ini diperkuat oleh nilai precision, recall, dan F1-score yang lebih rendah pada kategori Pendidikan dan Sosial, sehingga kategori tersebut masih sulit diprediksi secara akurat.



Gambar 10 WordCloud Kategori Bencana

Visualisasi WordCloud pada kategori bencana menunjukkan kata-kata dominan yang terkait dengan bencana, seperti banjir, jalan, warga, korban, dan akibat. Kata-kata ini mencerminkan peristiwa bencana, lokasi terdampak, serta dampak yang dirasakan masyarakat, sehingga tetap selaras dengan konteks berita.



Gambar 11 Antarmuka Web Sistem Klasifikasi Berita

Implementasi antarmuka web yang ditampilkan pada gambar memungkinkan pengguna untuk memasukkan judul berita dan secara instan memperoleh hasil prediksi terkait kategori serta sentimen dari judul berita tersebut. Berdasarkan contoh pada gambar, saat pengguna memasukkan judul berita "kapolri ungkap terduga pelaku ledakan dari lingkungan sman 72", sistem menghasilkan output di mana judul berita tersebut diklasifikasikan ke dalam Kategori Pendidikan dengan Sentimen negatif.

Hasil penelitian menunjukkan bahwa pendekatan TF-IDF dan Multinomial Naïve Bayes efektif sebagai baseline classifier untuk judul berita berbahasa Indonesia. Tingginya akurasi pada data training menandakan model mampu mempelajari pola kata dengan baik, namun perbedaan signifikan dengan akurasi data testing menegaskan adanya overfitting. Penerapan analisis sentimen sebagai pendekatan balancing membantu meningkatkan stabilitas prediksi pada kategori minoritas meskipun peningkatannya tidak signifikan. Judul berita terbukti cukup informatif untuk klasifikasi

otomatis, tetapi sifatnya yang pendek dan ambigu membatasi performa model. Keterbatasan penelitian meliputi ukuran dataset yang relatif kecil, penggunaan algoritma baseline, serta perbedaan preprocessing antara tahap pelatihan dan prediksi web. Meskipun demikian, sistem yang dikembangkan layak digunakan sebagai prototipe awal dan fondasi untuk pengembangan lanjutan.

CONCLUSION

Penelitian ini berhasil merancang dan membangun sistem klasifikasi judul berita online berbahasa Indonesia berbasis web menggunakan metode TF-IDF dan algoritma Naïve Bayes. Hasil penelitian menunjukkan bahwa kombinasi preprocessing teks, TF-IDF, dan Naïve Bayes mampu melakukan klasifikasi kategori dan sentimen judul berita secara otomatis.

Meskipun akurasi pada data testing masih terbatas akibat dataset yang tidak seimbang dan karakteristik judul berita yang singkat, penerapan analisis sentimen sebagai pendekatan balancing terbukti membantu meningkatkan kestabilan model. Sistem ini layak digunakan sebagai sistem awal dan dapat dikembangkan lebih lanjut dengan dataset yang lebih besar, metode preprocessing yang lebih kaya, serta algoritma klasifikasi yang lebih kompleks pada penelitian selanjutnya.

REFERENCE

- Fauzi, M. A. (2017). Indonesian news classification using naïve Bayes and two-phase feature selection. *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*. <https://ijeecs.iaescore.com/index.php/IJECS/article/download/9456/7915>
- Ionendri, N. A., Candra, F., & Rizal, A. (2025). News classification using natural language processing with TF-IDF and multinomial naïve Bayes. *J. Appl. Comput. Sci. Technol.*, 6(1). <https://doi.org/10.20961/itsmart.v6i1.11310>
- Nisa, N. B., Nanda, R. P., Kudadiri, Z. H., Alfahri, B. A., & Furqan, M. (2025). Klasifikasi berita detik.com terkait teknologi informasi menggunakan TF-IDF dan naïve Bayes. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, 8(3). <https://ojs.serambimekkah.ac.id/jnkti/article/download/9171/pdf>
- Rahman, A., Wiranto, W., & Doewes, A. (2017). Online news classification using multinomial

naïve Bayes. *ITSMART: Jurnal Ilmiah Teknologi Dan Informasi*, 6(1). <https://doi.org/10.20961/itsmart.v6i1.11310>

Santoso, H. A. (2020). Hoax classification and sentiment analysis of Indonesian news. *TELKOMNIKA (J. Telekom. & Elektron.)*, 18(5). <https://telkomnika.uad.ac.id/index.php/TELKOMNIKA/article/viewFile/14744/7887>

Sudrajat, A., Wulandari, N., & Syafwan, E. (2022). Indonesian language hoax news classification based on naïve Bayes. *Journal of Applied Intelligent System*, 7(1). <https://doi.org/10.33633/jais.v7i1.5985>