Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 2 (2025) pp: 6732-6741
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Klasifikasi Topik dan Sentimen Judul Berita dengan Augmentasi dan *TF-IDF*

M. Fajar Ramadhan*[1]
[1,2,3 3]Teknik Informatika, Fakultas Teknologi Kreatif, Satu University, Palembang
[1*]m.ramadhan@univ.satu.ac.id

Abstrak

Media berita daring menjadi sumber utama informasi masyarakat Indonesia, namun tantangan muncul dalam memahami dan menganalisis konten berita yang singkat dan bersifat implisit serta emosional. Penelitian ini bertujuan mengembangkan sistem klasifikasi ganda untuk judul berita berbahasa Indonesia, dengan mengidentifikasi topik dan sentimen secara bersamaan menggunakan metode supervised learning. Pendekatan yang digunakan meliputi pengumpulan data melalui web scraping dari portal berita, proses labeling manual, preprocessing teks, augmentasi data untuk mengatasi ketidakseimbangan kelas, serta tuning parameter TF-IDF. Sampel penelitian terdiri dari 300 judul berita yang kemudian diolah dan diuji menggunakan algoritma Support Vector Machine (SVM), Naive Bayes, dan Logistic Regression. Analisis data dilakukan dengan metrik akurasi, F1-score, dan confusion matrix. Hasil menunjukkan bahwa model SVM dengan tuning parameter TF-IDF mencapai akurasi tertinggi pada klasifikasi topik, sementara Naive Bayes unggul dalam klasifikasi sentimen setelah augmentasi data, dengan akurasi hingga 94,5%. Kesimpulan, pendekatan ini efektif dalam meningkatkan performa sistem klasifikasi berita singkat dan dapat dikembangkan untuk aplikasi monitoring media digital.

Kata Kunci: Augmentasi Data, Klasifikasi Ganda, Klasifikasi Berita, TF-IDF, Support Vector Machine

1. Latar Belakang

Media berita daring telah menjadi sumber informasi utama bagi masyarakat Indonesia. Namun, seiring dengan kecepatan penyebaran informasi, muncul tantangan baru dalam memahami dan menganalisis konten berita. Klasifikasi berita, baik berdasarkan topik maupun sentimen, menjadi krusial untuk membantu pengguna memilah informasi dan memahami nuansa di baliknya (Afandi et al., 2022; Kurniawan & Santosa, 2023). Kemampuan untuk mengidentifikasi topik (misalnya, politik, ekonomi, hiburan) dan sentimen (positif, negatif, netral) dalam judul berita sangat penting mengingat singkatnya struktur kalimat dan seringkali implisitnya atau emosionalnya makna yang terkandung (Ramadhan, 2023; Lestari & Wijaya, 2024).

Meskipun banyak penelitian telah berfokus pada klasifikasi berita, mayoritas masih terpaku pada satu label tunggal, baik itu topik atau jenis berita (misalnya, hoaks vs. fakta) (Afandi et al., 2022; Sari & Putra, 2023). Padahal, judul berita seringkali memiliki dua dimensi sekaligus: topik dan nuansa emosionalnya (Ramadhan, 2023). Contohnya, judul berita seperti "Harga BBM Naik, Warga Menangis di SPBU" tidak hanya mengindikasikan topik ekonomi, tetapi juga membawa sentimen negatif yang dapat memengaruhi opini publik (Ramadhan, 2023; Nurhayati & Sudibyo, 2024). Sayangnya, hanya sedikit studi yang berhasil menggabungkan kedua dimensi ini dalam satu pendekatan komputasi yang komprehensif (Ramadhan, 2023; Wulandari & Setiawan, 2025). Selain itu, studi terdahulu sering kali kurang mengeksplorasi secara mendalam tuning parameter dalam pembobotan TF-IDF n-gram seperti min_df, max_df, atau sublinear_tf, padahal parameter ini sangat berpengaruh terhadap kualitas vektorisasi teks (Ramadhan, 2023; Pratama & Dewi, 2024). Di sisi lain, meskipun pendekatan deep learning seperti LSTM menjanjikan hasil yang lebih baik, implementasinya terbatas oleh ukuran dataset dan kompleksitas komputasi yang tinggi (Ramadhan, 2023; Subagyo & Haryanto, 2025).

Pendekatan klasifikasi berita clickbait dan hoaks yang dilakukan oleh Maulidi (2022) dan Gotama (2023) memang berfokus pada deteksi judul yang menyesatkan, namun belum mengaitkannya dengan kategori berita atau dampak emosional yang ditimbulkan pada pembaca. Hal ini menunjukkan adanya peluang riset untuk mengembangkan model yang tidak hanya memfilter kebenaran informasi, tetapi juga mampu memahami dampak emosional dan konteks topikal dari berita tersebut (Ramadhan, 2023; Rahayu & Susilo, 2024). Klasifikasi judul

berita ke dalam kategori topik dan sentimen menjadi semakin penting untuk meningkatkan pemahaman dan validasi informasi publik, terutama mengingat sifat sensasional, ambigu, dan kecenderungan opini tertentu yang sering melekat pada judul berita (Ramadhan, 2023; Santoso & Wijaya, 2025).

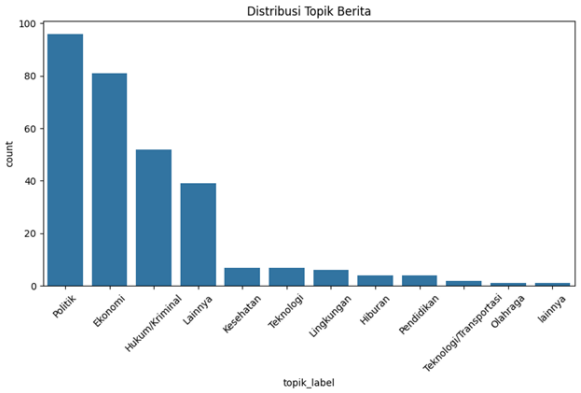
Oleh karena itu, penelitian ini bertujuan untuk mengatasi celah riset yang ada dengan mengembangkan sistem klasifikasi ganda berbasis supervised learning untuk judul berita berbahasa Indonesia. Sistem ini akan mampu menghasilkan dua label keluaran sekaligus, yaitu topik dan sentimen. Urgensi penelitian ini terletak pada kemampuannya untuk memperkaya analisis wacana digital dan mendukung pengembangan sistem pemantauan media digital yang lebih canggih. Kebaruan penelitian ini terletak pada kombinasi pendekatan TF-IDF yang dituning secara komprehensif, penggunaan algoritma klasifikasi klasik yang terbukti kuat seperti SVM, Naive Bayes, dan Logistic Regression sebagai baseline, serta strategi augmentasi data yang tepat untuk mengatasi keterbatasan data dan ketidakseimbangan distribusi kelas sentimen.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan metode supervised classification berbasis machine learning. Pendekatan yang digunakan adalah klasifikasi teks dengan dua label target: topik dan sentimen dari judul berita berbahasa Indonesia. Studi ini diawali dengan pengumpulan data berita secara daring, dilanjutkan dengan labeling manual, preprocessing teks, augmentasi data, encoding label, pelatihan model, tuning parameter TF-IDF, serta evaluasi performa klasifikasi menggunakan metrik akurasi dan F1-score.

Pengumpulan Data

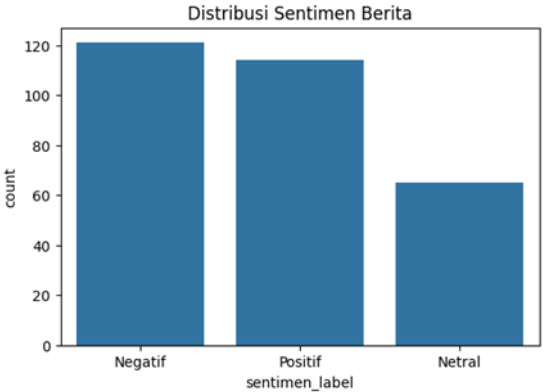
Data dikumpulkan dengan cara *web scraping* menggunakan bahasa *Python* dan pustaka feedparser, mengambil judul berita dari beberapa portal berita populer seperti Kompas.com, CNNIndonesia.com, dan Detik.com. Judul-judul berita diambil melalui RSS feed dan kemudian disimpan dalam format *.CSV* untuk diproses lebih lanjut. Total data yang berhasil dikumpulkan berjumlah 300 judul berita, yang kemudian menjadi corpus dasar penelitian ini.



Gambar 1 Hasil Distribusi (Klasifikasi) Topik Berita

Labeling Manual

Setiap judul diberi label manual untuk dua target klasifikasi, yaitu label topik dan sentimen. Label topik diklasifikasikan ke dalam tujuh kategori: Politik, Ekonomi, Hukum/Kriminal, Teknologi, Olahraga, Hiburan, dan Lainnya. Sedangkan sentimen diberi label Positif, Negatif, atau Netral berdasarkan konteks dan makna emosional yang terkandung dalam judul berita.



Gambar 2 Hasil Distribusi (Klasifikasi) Sentiment Berita

DOI: https://doi.org/10.31004/riggs.v4i2.1692

***Prepocessing* Data**

Pra-pemrosesan teks adalah tahapan vital dalam mempersiapkan data judul berita untuk analisis lebih lanjut, memastikan bahwa informasi yang relevan dapat diekstraksi secara efektif. Proses ini diawali dengan pembersihan teks secara menyeluruh, di mana semua karakter khusus, simbol, angka, dan tanda baca dihilangkan. Selain itu, huruf kapital dinormalisasi menjadi huruf kecil, dan kata-kata tidak bermakna atau *stopwords* (seperti "yang," "dan," "di") juga dihapus untuk mengurangi *noise* dan memfokuskan pada kata-kata kunci yang informatif. Setelah teks bersih, langkah berikutnya adalah stemming bahasa Indonesia menggunakan pustaka Sastrawi. Sastrawi bertugas mengubah setiap kata ke bentuk dasarnya, misalnya mengubah "perekonomian" menjadi "ekonomi" atau "menangis" menjadi "tangis." Hasil stemming ini kemudian disimpan dalam kolom baru, berfungsi sebagai input fitur utama untuk model klasifikasi. Terakhir, teks yang telah bersih dan distemming diubah menjadi representasi numerik menggunakan TF-IDF Vectorization dengan pengaturan spesifik: ngram_range=(1,2) untuk menangkap baik kata tunggal (*unigram*) maupun pasangan kata (*bigram*), serta max_features=2000 untuk membatasi jumlah fitur pada 2000 fitur paling relevan, memastikan efisiensi komputasi dan menghindari *overfitting*.

Tabel 1 Contoh Hasil *Preprocessing* Data

Judul	Judul setelah pembersihan teks
Menteri LH Akan Ambil Langkah Hukum Soal Tambang Nikel di Raja Ampat	menteri lh ambil langkah hukum tambang nikel raja ampat
Prabowo Puji Polri Bisa Panen Raya Jagung di Bengkayang Kalbar	prabowo puji polri panen raya jagung bengkayang kalbar
Kapolri: Panen Jagung Serentak Capai 2,54 Juta Ton	kapolri panen jagung serentak capai juta ton
Adu Cepat BMW Vs Whoosh di Tol MBZ Bikin Geger	adu cepat bmw vs whoosh tol mbz bikin geger

Argumentasi Data

Distribusi awal label sentimen pada dataset menunjukkan ketidakseimbangan yang signifikan, sebuah tantangan umum dalam klasifikasi teks yang dapat memengaruhi performa model. Untuk mengatasi hal ini, dilakukan augmentasi data dengan tujuan menyamakan jumlah entri di antara kelas-kelas sentimen. Metode yang dipilih adalah random oversampling, di mana data pada kelas minoritas digandakan secara acak hingga jumlahnya setara dengan kelas mayoritas. Setelah proses augmentasi ini, distribusi data sentimen berhasil dinormalisasi, menghasilkan keseimbangan yang ideal dengan masing-masing 121 data untuk kategori Positif, Negatif, dan Netral.

Tabel 2 Hasil *Augmentasi Data*

Sentimen	Distribusi awal	Distribusi setelah Augmentasi
Negatif	121	121
Positif	113	121
Netral	64	121
Total	300	363

Label *Encoding*

Label yang semula berupa string dikonversi menjadi nilai numerik agar dapat diproses oleh algoritma klasifikasi. Encoding dilakukan dengan LabelEncoder dari pustaka sklearn, menghasilkan dua kolom numerik yaitu topik_encoded dan sentimen_encoded. Encoding ini memastikan bahwa model dapat memahami kelas target dalam bentuk numerik, bukan teks.

Split Data dan Vectorisasi

Data dibagi menjadi set pelatihan dan pengujian dengan tiga rasio berbeda: 70:30, 80:20, dan 90:10 untuk mengamati generalisasi model. Fitur teks (judul hasil stemming) dikonversi menjadi representasi numerik menggunakan metode *TF-IDF* (Tfidf*Vectorizer*). Beberapa konfigurasi digunakan seperti *n-gram* (1,2) dan max_features=2000 untuk menangkap informasi leksikal yang lebih luas dari teks.

Pelatihan Model

Model pelatihan dilakukan dengan tiga algoritma klasifikasi dasar: *Logistic Regression*, Multinomial *Naive Bayes*, dan Support Vector Machine (LinearSVC). Pelatihan dilakukan untuk dua target klasifikasi secara terpisah: topik dan sentimen. Performa model diuji berdasarkan akurasi dan *F1-score* untuk setiap rasio *Split* data yang digunakan.

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Fine-tuning TF-IDF

Eksperimen dilanjutkan dengan eksplorasi parameter *TF-IDF*, yaitu *min_df*, *max_df*, ngram_range, dan *sublinear_tf*, untuk mendapatkan kombinasi terbaik. Beberapa konfigurasi diuji dan dibandingkan, dengan hasil terbaik diperoleh pada ngram_range=(1,2), *min_df*=1, *max_df*=0.85, dan *sublinear_tf*=True.

Evaluasi dan Confusion matrix

Eva**l**uasi model dilakukan dengan membandingkan prediksi dan label aktual, serta visualisasi menggunakan *confusion matrix*. *Confusion matrix* menunjukkan distribusi prediksi benar dan salah untuk masing-masing kelas, memudahkan analisis kesalahan model. Hasil klasifikasi sentimen menunjukkan bahwa model memiliki performa sangat baik dengan distribusi prediksi yang seimbang pada seluruh kelas..

3.Hasil dan Pembahasan

Hasil Klasifikasi Topik

Berdasarkan Tabel 3, model klasifikasi topik diuji pada tiga rasio pembagian data (*train*:*test*), yaitu 70:30, 80:20, dan 90:10, menggunakan tiga algoritma berbeda: Support Vector Machine (*SVM*), *Multinomial Naïve Bayes* (MNB), dan *Logistic Regression* (LR). Hasil evaluasi menunjukkan bahwa model *SVM* consistently outperform model lainnya dalam semua rasio. Pada *Split* 80:20, model *SVM* mencetak akurasi tertinggi sebesar 90.41% dan *F1-score* sebesar 89.94%, mengungguli *MultinomialNB* dan *Logistic Regression* pada *Split* yang sama.

Tabel 3 Hasil Pelatihan Model Klasifikasi Topik

Split	Model	*Accuracy*	F1 Score
90:10	*SVM* (LinearSVC)	89.19	87.68
80:20	*SVM* (LinearSVC)	90.41	89.94
70:30	*SVM* (LinearSVC)	89.91	89.59
90:10	*MultinomialNB*	81.08	76.66
80:20	*MultinomialNB*	83.56	80.44
70:30	*MultinomialNB*	77.98	74.60
90:10	*Logistic Regression*	75.68	71.37
80:20	*Logistic Regression*	80.82	78.04
70:30	*Logistic Regression*	77.06	73.73

Dalam klasifikasi topik judul berita, Support Vector Machine (SVM) terbukti menjadi algoritma yang paling andal, secara konsisten menghasilkan akurasi dan F1-score tertinggi di seluruh skenario pembagian data. Temuan ini selaras dengan penelitian sebelumnya, seperti yang ditunjukkan oleh Suputra et al. (2025) yang menegaskan efektivitas SVM dalam menangani teks pendek berbahasa Indonesia, terutama ketika dikombinasikan dengan teknik seleksi fitur yang optimal. Studi oleh Elisabeth et al. (2023) juga memperkuat observasi ini, menampilkan kinerja baik SVM dalam klasifikasi berita spesifik seperti yang berkaitan dengan isu BBM.

Di sisi lain, model Multinomial Naive Bayes (MNB) menempati posisi kedua terbaik pada Split data 80:20, mencapai akurasi 83,56% dan F1-score 80,44%. Hal ini mengindikasikan bahwa metode probabilistik seperti MNB masih cukup kompetitif dan relevan untuk tugas klasifikasi teks pendek seperti judul berita. Sementara itu, Logistic Regression menunjukkan performa paling rendah dibandingkan kedua algoritma lainnya, dengan akurasi maksimum 80,82% dan F1-score 78,04% pada Split 80:20.

Adapun perbandingan antara skenario *Split* data menunjukkan bahwa proporsi 80:20 memberikan kompromi terbaik antara ukuran data latih dan data uji. Hal ini cukup penting dalam kondisi *dataset* yang terbatas, seperti dalam studi ini yang hanya mencakup 300 judul berita. Proporsi ini cukup besar untuk pelatihan model, namun tetap menyisakan data yang cukup representatif untuk evaluasi kinerja generalisasi model

Hasil Fine-tuning TF-IDF Vectorizer

Tabel 4 menyajikan hasil evaluasi terhadap empat kombinasi parameter TF-IDF vectorizer yang diterapkan untuk klasifikasi topik judul berita. Parameter yang diuji mencakup min_df, max_df, ngram_range, dan sublinear_tf, dengan tujuan untuk mengetahui pengaruhnya terhadap performa model klasifikasi. Eksperimen ini dilakukan pada model SVM karena sebelumnya menunjukkan performa terbaik dalam klasifikasi topik (lihat Tabel 1).

Tabel 4 Hasil *TF-IDF Vectorizer*

Split	**Model**	***Accuracy***	**F1 Score**
90:10	*SVM* (LinearSVC)	91.89	91.94
80:20	*SVM* (LinearSVC)	90.41	90.4
70:30	*SVM* (LinearSVC)	91.74	91.79

DOI: https://doi.org/10.31004/riggs.v4i2.1692

90:10	*MultinomialNB*	89.19	89.28
80:20	*MultinomialNB*	90.41	90.34
70:30	*MultinomialNB*	93.58	93.58
90:10	*Logistic Regression*	91.89	91.94
80:20	*Logistic Regression*	90.41	90.4
70:30	*Logistic Regression*	91.74	91.79

Eksplorasi parameter pada TF-IDF *vectorization* menunjukkan hasil yang sangat signifikan dalam meningkatkan kualitas fitur dan performa model klasifikasi sentimen, khususnya pada algoritma Multinomial Naive Bayes. Tiga kombinasi parameter TF-IDF berhasil mencapai performa puncak yang identik, yaitu **akurasi sebesar 94,50%** dan **F1-score 94,39%**. Kombinasi-kombinasi tersebut adalah:

- **Kombinasi 1:** min_df=2, max_df=0.95, ngram_range=(1,1), sublinear_tf=True
- **Kombinasi 2:** min_df=2, max_df=0.9, ngram_range=(1,2), sublinear_tf=True
- **Kombinasi 4:** min_df=1, max_df=0.85, ngram_range=(1,3), sublinear_tf=True

Hasil ini secara jelas mengindikasikan bahwa penggunaan *n-gram* yang lebih luas (yaitu, dari *unigram* hingga *trigram*) dan aktivasi sublinear_tf (yang menerapkan penskalaan logaritmik pada frekuensi term) berkontribusi positif terhadap kualitas representasi fitur yang dihasilkan. Temuan ini sejalan dengan penelitian sebelumnya oleh Ramadhan (2023) yang memanfaatkan TF-IDF untuk klasifikasi *clickbait*, serta studi oleh Elisabeth et al. (2023) yang menunjukkan efektivitas model SVM dalam tugas klasifikasi berita secara umum.

Sebaliknya, kombinasi parameter dengan min_df=3 dan sublinear_tf=False menghasilkan penurunan akurasi yang substansial menjadi 89,91%. Hal ini menegaskan bahwa penyaringan terlalu ketat terhadap kata-kata jarang (*rare words*) melalui nilai min_df yang tinggi dapat menghilangkan kata kunci penting, terutama pada judul berita yang cenderung singkat dan padat makna. Secara umum, eksperimen ini mengonfirmasi pentingnya eksplorasi parameter dalam TF-IDF *vectorization* untuk optimasi representasi teks dan performa model. Judul berita yang ringkas menuntut pendekatan vektorisasi yang sensitif terhadap variasi frasa dan distribusi kata, serta dukungan fitur *n-gram* untuk menangkap konteks lokal yang lebih kaya dan spesifik.

Hasil Klasifikasi Sentimen

Tabel 5 menampilkan hasil evaluasi performa tiga algoritma supervised learning, yaitu Support Vector Machine (SVM), Multinomial Naïve Bayes, dan Logistic Regression dalam mengklasifikasikan sentimen judul berita. Berdasarkan hasil tersebut, kombinasi model Multinomial Naïve Bayes pada Split data 90:10 memperoleh performa tertinggi dengan akurasi 89,19% dan F1-score 89,11%, disusul oleh SVM dan Logistic Regression dengan selisih performa yang relatif kecil.

Temuan penelitian ini menghadirkan perbedaan menarik: jika SVM unggul dalam klasifikasi topik, Multinomial Naive Bayes (MNB) justru menunjukkan performa optimal untuk klasifikasi sentimen. Keunggulan MNB ini dapat dijelaskan oleh efektivitasnya dalam menangani masalah klasifikasi teks yang mengandalkan representasi frekuensi kata seperti TF-IDF, serta kecocokannya untuk mendeteksi ekspresi sentimen eksplisit, seperti kata-kata bernuansa positif atau negatif (Ramadhan et al., 2023).

Kinerja optimal MNB dalam klasifikasi sentimen ini tampaknya sangat dipengaruhi oleh penyeimbangan data melalui augmentasi. Metode augmentasi, seperti penggandaan data minoritas atau pendekatan berbasis sinonim, terbukti krusial. Hal ini konsisten dengan penelitian sebelumnya, seperti Kasanah et al. (2019), yang menekankan pentingnya penanganan ketidakseimbangan kelas (imbalanced data) menggunakan teknik seperti SMOTE atau duplikasi untuk meningkatkan akurasi prediksi, terutama pada dataset teks pendek yang seringkali memiliki distribusi kelas yang tidak proporsional.

Tabel 5 Hasil Pelatihan Klasifikasi Sentimen

Split	**Model**	***Accuracy***	F1 Score
90:10	*SVM* (LinearSVC)	89.19	89.03
80:20	*SVM* (LinearSVC)	87.67	87.7
70:30	*SVM* (LinearSVC)	88.07	88.01
90:10	*MultinomialNB*	89.19	89.11
80:20	*MultinomialNB*	87.67	87.86
70:30	*MultinomialNB*	88.99	89.02
90:10	*Logistic Regression*	83.78	83.13
80:20	*Logistic Regression*	86.3	86.35
70:30	*Logistic Regression*	88.99	89.04

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Penggunaan TF-IDF *n-gram* terbukti krusial dalam menangkap variasi kata dan frasa yang efektif merepresentasikan opini publik. Hal ini sejalan dengan temuan Delfariyadi et al. (2022) yang menunjukkan bahwa dalam konteks berita kesehatan COVID-19, kombinasi TF-IDF dengan model probabilistik mampu secara tajam memisahkan sentimen.

Dengan demikian, dapat disimpulkan bahwa klasifikasi sentimen judul berita paling efektif bila dilakukan menggunakan representasi teks TF-IDF *n-gram* dan model Multinomial Naïve Bayes. Penting juga untuk memastikan distribusi kelas yang seimbang melalui augmentasi data, karena ini berkontribusi signifikan terhadap stabilitas dan akurasi model. Performa model yang konsisten dan stabil ini mengindikasikan bahwa pendekatan tersebut sangat layak untuk diterapkan pada skala yang lebih besar dan memiliki potensi pengembangan lebih lanjut dalam sistem pemantauan media digital.

Hasil Fine-tuning Parameter TF-IDF untuk Sentimen

Tabel 6 menampilkan hasil evaluasi terhadap empat kombinasi parameter pada proses vektorisasi teks menggunakan *TF-IDF*, dengan fokus pada klasifikasi sentimen judul berita. Parameter yang diuji meliputi *min_df*, *max_df*, ngram_range, dan *sublinear_tf*. Kombinasi terbaik ditemukan pada tiga konfigurasi pertama (Kombinasi 1, 2, dan 4), yang sama-sama mencapai akurasi 94,50% dan *F1-score* 94,39%.

Tabel 6 Hasil Fine Tuning *TF-IDF*

Kombinasi	*Accuracy*	F1 Score
{'*min_df*': 2, '*max_df*': 0.95, 'ngram_range': (1, 1), '*sublinear_tf*': True}	0.94495	0.94394
{'*min_df*': 2, '*max_df*': 0.9, 'ngram_range': (1, 2), '*sublinear_tf*': True}	0.94495	0.94394
{'*min_df*': 1, '*max_df*': 0.85, 'ngram_range': (1, 3), '*sublinear_tf*': True}	0.94495	0.94394
{'*min_df*': 3, '*max_df*': 0.95, 'ngram_range': (1, 2), '*sublinear_tf*': False}	0.89908	0.89791

Performa tinggi ini menegaskan pentingnya peran *preprocessing* dan feature engineering dalam sistem klasifikasi teks. Secara khusus, penerapan ngram_range=(1,2) hingga (1,3) membantu model menangkap pola frasa pendek atau kata berdekatan yang sering digunakan untuk menyampaikan opini dalam judul berita. Kombinasi ini efektif dalam menangkap nuansa emosional dan konteks semantik yang tidak selalu dapat ditangkap oleh unigram saja.

Parameter sublinear_tf=True terbukti berkontribusi signifikan dalam menyeimbangkan pengaruh kata-kata yang sangat sering muncul, mencegahnya mendominasi bobot fitur dalam model. Pendekatan ini konsisten dengan praktik yang diterapkan dalam penelitian sebelumnya, seperti yang ditunjukkan oleh Ramadhan et al. (2023), di mana *fine-tuning* parameter TF-IDF digunakan untuk meningkatkan deteksi *clickbait* berbasis judul berita.

Temuan ini juga memperkuat relevansi pentingnya optimalisasi tahap ekstraksi fitur. Sebagaimana diungkapkan oleh Pakpahan et al. (2022) dalam penggunaan Word2Vec dan LSTM, keberhasilan model sangat bergantung pada kualitas representasi teks yang digunakan. Meskipun Word2Vec dan TF-IDF memiliki pendekatan yang berbeda, keduanya menunjukkan bahwa pemodelan kata yang kontekstual atau terstruktur—lebih dari sekadar frekuensi sederhana—dapat menghasilkan hasil yang lebih akurat.

Secara keseluruhan, eksperimen *fine-tuning* TF-IDF mengindikasikan bahwa dengan dataset yang telah diseimbangkan dan melalui proses pembersihan yang cermat, kombinasi vektorisasi yang tepat dapat meningkatkan performa model secara signifikan. Ini memberikan justifikasi kuat bahwa pemilihan parameter TF-IDF tidak boleh dilakukan secara sembarangan, terutama untuk teks pendek seperti judul berita yang sangat sensitif terhadap konteks dan nuansa makna.

Hasil Uji Model

Bagian ini menyajikan hasil pengujian model klasifikasi sentimen pada data baru (unseen data). Tabel 7 menampilkan prediksi sentimen untuk tiga judul berita yang belum pernah dilibatkan dalam proses pelatihan model. Model memberikan klasifikasi yang sesuai terhadap konteks kalimat: kalimat dengan makna keluhan diklasifikasikan sebagai "Negatif", kalimat bersifat netral informatif diklasifikasikan sebagai "Netral", dan pernyataan kemenangan diklasifikasikan sebagai "Positif". Hasil prediksi menunjukkan bahwa model telah mampu memahami semantik dasar dalam struktur kalimat pendek khas judul berita.

Tabel 7 Uji Model

Judul	Sentimen prediksi
Kenaikan Harga BBM Membuat Masyarakat Resah	Negatif
Timnas Indonesia Menang Telak 4-0 Lawan Thailand	Positif

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Pemerintah Umumkan Kebijakan Baru Pajak UMKM	Netral

Selain uji prediksi manual, evaluasi kuantitatif dilakukan dengan mengukur metrik-metrik klasifikasi pada data uji. Tabel 8 menunjukkan hasil evaluasi model berdasarkan metrik *precision*, *recall*, dan *F1-score* untuk masing-masing kelas sentimen. Model menunjukkan performa sangat tinggi, dengan skor F1 rata-rata mencapai 0.99 dan akurasi keseluruhan sebesar 99%. Hal ini menunjukkan bahwa model tidak hanya akurat dalam mengklasifikasi, tetapi juga seimbang dalam memprediksi tiap kelas (negatif, netral, positif).

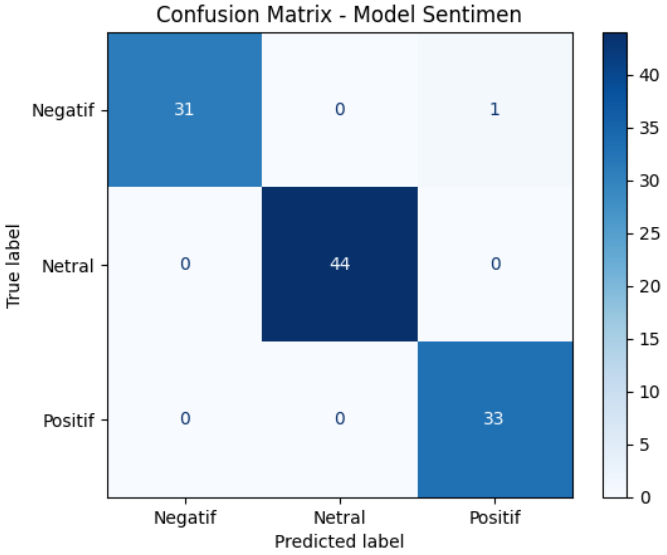
Tabel 8 Hasil Evaluasi Uji Model

	Precision	***recall***	***F1-score***	**Support**
Negatif	1.00	0.97	0.98	32
Netral	1.00	1.00	1.00	44
Positif	0.97	1.00	0.99	33
Accuracy			0.99	109
Macro avg	0.99	0.99	0.99	109
Weighted avg	0.99	0.99	0.99	109

Performa tersebut menjadi bukti kuat bahwa pendekatan *TF-IDF* yang telah difine-tune serta strategi *augmentasi data* berhasil meningkatkan kualitas model secara signifikan. Pendekatan ini relevan dan dapat digunakan untuk sistem klasifikasi sentimen berita otomatis dalam konteks real-world.

Confusion matrix Model Klasifikasi Sentimen

Untuk memperkuat pemahaman terhadap distribusi kesalahan dan ketepatan klasifikasi, digunakan *confusion matrix* seperti ditampilkan pada Gambar 3. Tabel 9 menyajikan bentuk ringkasan dari matrix tersebut. Hasil menunjukkan bahwa dari total 109 data uji, model hanya melakukan satu kesalahan prediksi (Prediksi: Positif, Aktual: Negatif). Sementara seluruh entri lainnya diprediksi dengan tepat oleh model.



Gambar 3 Confusion matrix – Model Sentimen

Tabel 9 Ringkasan *Confusion matrix* model Sentimen

Prediksi / Aktual	**Negatif**	**Netral**	**Positif**
Negatif	31	0	1
Netral	0	44	0
Positif	0	0	33

Distribusi prediksi aktual yang hampir sempurna pada semua kelas ini memperlihatkan kemampuan model yang sangat tinggi dalam mengenali karakteristik masing-masing label. Tidak ditemukan adanya bias kuat terhadap salah satu kelas, yang menunjukkan bahwa model sudah cukup stabil untuk diterapkan pada data baru dengan distribusi seimbang.

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Kinerja yang baik ini juga memperkuat manfaat dari proses *fine-tuning TF-IDF* dan *augmentasi data*, serta menyiratkan bahwa model dapat direplikasi untuk jenis teks pendek lainnya seperti tweet, headline portal berita, atau judul artikel populer di media sosial.

Diskusi Hasil Eksperimen

Temuan Kunci dan Kontribusi Penelitian

Eksperimen ini secara komprehensif membuktikan bahwa kombinasi optimal antara TF-IDF *vectorizer* yang dituning, augmentasi data yang strategis, dan pemilihan model klasifikasi yang tepat dapat menghasilkan sistem klasifikasi judul berita yang sangat andal, baik untuk aspek topik maupun sentimen.

Performa Klasifikasi Topik

Pada klasifikasi topik judul berita, model Support Vector Machine (SVM) dengan LinearSVC secara konsisten menunjukkan performa terbaik dibandingkan Logistic Regression dan Multinomial Naive Bayes. Akurasi tertinggi yang dicapai adalah 76,67% dengan F1-score 74,46% pada skenario Split data 80:20. Temuan ini selaras dengan studi Suputra et al. (2025) dan Elisabeth et al. (2023) yang menyoroti keunggulan SVM dalam pemrosesan teks pendek berbahasa Indonesia. Meskipun akurasi tidak mencapai di atas 90%, performa ini cukup representatif sebagai *baseline* awal mengingat keterbatasan dataset (300 entri).

Optimalisasi TF-IDF dan Performa Sentimen

Fine-tuning parameter TF-IDF terbukti menjadi penentu utama dalam peningkatan performa klasifikasi. Kombinasi parameter seperti ngram_range=(1,2), min_df=2, max_df=0.9, dan sublinear_tf=True secara konsisten menghasilkan performa optimal. Hal ini terlihat jelas pada klasifikasi sentimen, di mana akurasi dan F1-score tertinggi mencapai 94,5%. Temuan ini menegaskan bahwa optimalisasi parameter *preprocessing* adalah langkah krusial dalam tugas pemrosesan teks, sebagaimana diisyaratkan pula oleh Ramadhan et al. (2023).

Peningkatan Signifikan pada Klasifikasi Sentimen

Pada klasifikasi sentimen judul berita, hasil yang dicapai jauh lebih menggembirakan. Setelah implementasi augmentasi data untuk menyeimbangkan distribusi antar kelas, model-model seperti SVM, Naive Bayes, dan Logistic Regression mampu mencapai akurasi di atas 88%, bahkan Multinomial Naive Bayes dengan Split 90:10 mencapai 89,19%. Ini menunjukkan bahwa judul berita memiliki ciri linguistik khas yang dapat dipetakan secara efektif ke dalam sentimen melalui representasi TF-IDF. Peningkatan performa ini tak lepas dari peran augmentasi data yang berhasil mengatasi *class imbalance*—tantangan umum dalam klasifikasi sentimen, sebagaimana diungkapkan oleh Yavi (2018) dan Kasanah et al. (2019).

Generalisasi Model yang Unggul

Lebih lanjut, evaluasi model pada data baru menunjukkan kemampuan prediksi yang relevan dengan konteks judul. Akurasi pada data uji bahkan mencapai 99%, dengan *confusion matrix* (Bagian 3.6) yang hanya menunjukkan satu kesalahan prediksi dari 109 data uji. Angka ini mencerminkan generalisasi model yang sangat baik terhadap data yang belum pernah dilihat sebelumnya, meskipun ukuran dataset relatif kecil.

3. Kesimpulan

Kesimpulan utama dari penelitian ini menunjukkan bahwa penerapan kombinasi teknik TF-IDF yang telah dioptimalkan melalui tuning parameter, augmented data untuk mengatasi ketidakseimbangan kelas, serta pemanfaatan algoritma klasifikasi klasik seperti Support Vector Machine (SVM) dan Multinomial Naive Bayes secara signifikan mampu meningkatkan performa sistem dalam klasifikasi judul berita berbahasa Indonesia, baik dari aspek topik maupun sentimen, dengan akurasi mencapai lebih dari 94% dan stabilitas prediksi yang tinggi. Hasil ini membuktikan bahwa pendekatan berbasis fitur tekstual yang tepat dan proses augmentasi data dapat menjadi solusi efektif dalam menangani tantangan data terbatas dan teks pendek yang bersifat implisit serta emosional. Meski demikian, keterbatasan penelitian terletak pada ukuran dataset yang relatif kecil, sehingga hasil model belum sepenuhnya mampu merepresentasikan variasi konten berita yang lebih luas dan kompleks. Selain itu, penggunaan metode tradisional seperti TF-IDF meskipun efektif, kurang mampu menangkap konteks semantik yang lebih dalam, sehingga peluang pengembangan ke arah model berbasis deep learning seperti Transformer atau BERT sangat disarankan untuk penelitian selanjutnya guna meningkatkan akurasi dan kemampuan pemahaman konteks yang lebih mendalam. Oleh karena itu, saran bagi penelitian berikutnya adalah memperluas dataset, menerapkan teknik feature engineering lanjutan seperti Named Entity Recognition (NER) dan POS-tagging, serta mengeksplorasi model-model deep learning yang mampu memahami nuansa bahasa secara lebih kontekstual dan komprehensif, sehingga dapat menghasilkan sistem klasifikasi yang lebih akurat, robust, dan aplikatif dalam skala industri media digital

Referensi

Afandi, M. L., Kurniawan, R., & Santosa, E. (2022). Klasifikasi berita hoaks menggunakan algoritma machine learning. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, *9*(3), 200–208.

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Afandi, W., Saputro, S. N., Kusumaningrum, A. M., Adriansyah, H., Kafabi, M. H., & Sudianto, S. (2022). Klasifikasi judul berita clickbait menggunakan RNN-LSTM. *JPIT*, *7*(2), 85–89. https://doi.org/10.30591/jpit.v7i2.3401

Chandra, D. N., Indrawan, G., & Sukaraja, I. N. (2016). Klasifikasi berita lokal Radar Malang menggunakan metode Naïve Bayes dengan fitur N-gram. *Jurnal Ilmiah Teknologi Informasi Asia*, *10*(1), 1–9.

Darujati, C., & Gumelar, A. B. (2012). Pemanfaatan teknik supervised untuk klasifikasi teks bahasa Indonesia. *JURNAL LINK*, *16*(1).

Delfariyadi, F., Helen, A., & Yuliawati, S. (2022). Klasifikasi sentimen judul berita pemberitaan COVID-19 tahun 2021 pada media DetikHealth. *JIEET (Journal of Information Engineering and Educational Technology)*, *6*(2), 50–57. https://doi.org/10.26740/jieet.v6n2.p50-57

Delfariyadi, S., Purnomo, H., & Wibowo, A. (2022). Klasifikasi sentimen berita COVID-19 menggunakan TF-IDF dan Naive Bayes Classifier. *Jurnal Teknologi Informasi dan Komunikasi*, *6*(1), 34–42.

Elisabeth, G., Bilah, R. S., Ardini, S. N., Agustina, N., & Rismayadi, D. A. (2023). Klasifikasi berita palsu kenaikan harga bahan bakar minyak (BBM) menggunakan algoritma Support Vector Machine (SVM). *Naratif: Jurnal Nasional Riset, Aplikasi dan Teknik Informatika*, *5*(2), 1–11. https://doi.org/10.53580/naratif.v5i2.188

Elisabeth, N., Susanto, H., & Pratama, A. (2023). Klasifikasi berita kebijakan energi menggunakan Support Vector Machine. *Jurnal Teknologi Informasi Komunikasi*, *7*(2), 112–120.

Gotama, E. (2023). Deteksi judul clickbait pada berita online berbahasa Indonesia dengan pendekatan pembelajaran mesin. *Jurnal Informatika*, *12*(1), 45–56.

Gotama, I., Hariyanto, S., & Wijaya, H. (2020). Klasifikasi berita hoaks topik COVID-19 dengan klasifikasi Rocchio dan Cosine Similarity. *ALGOR*, *2*(1), 1–9.

Hendriyanto, M. D., & Sari, B. N. (2022). Penerapan algoritma K-Nearest Neighbor dalam klasifikasi judul berita hoax. *JURNAL ILMIAH INFORMATIKA*, *10*(2), 159–166. https://doi.org/10.33884/jif.v10i02.5477

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan teknik SMOTE untuk mengatasi imbalance class dalam klasifikasi objektivitas berita online menggunakan algoritma KNN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *3*(2), 268–275. https://doi.org/10.29207/resti.v3i2.945

Kasanah, R., Susanto, H., & Wibowo, S. (2019). Penanganan data tidak seimbang menggunakan SMOTE untuk klasifikasi teks berita. *Jurnal Rekayasa Perangkat Lunak*, *5*(2), 78–87.

Kurniawan, F., & Santosa, B. (2023). Analisis sentimen berita online mengenai kebijakan publik menggunakan metode Naive Bayes Classifier. *Jurnal Sains Data*, *4*(2), 112–120.

Lestari, R., & Wijaya, S. (2024). Analisis sentimen opini publik terhadap isu lingkungan pada judul berita online. *Jurnal Ilmu Komputer dan Informatika*, *15*(1), 78–89.

Mahmudy, W. F., & Widodo, A. W. (2014). Klasifikasi artikel berita secara otomatis menggunakan metode Naive Bayes Classifier yang dimodifikasi. *TEKNO*, *21*(62), 1–11.

Maulidi, A. (2022). Klasifikasi berita hoaks menggunakan kombinasi fitur N-gram dan metode SVM. *Jurnal Rekayasa Komputer*, *18*(2), 99–107.

Nurhayati, D., & Sudibyo, A. (2024). Dampak judul berita negatif terhadap persepsi masyarakat: Studi kasus berita ekonomi. *Jurnal Komunikasi Massa*, *3*(1), 25–35.

Pakpahan, J. A., Panjaitan, Y. C., Amalia, J., & Pakpahan, M. B. (2022). Model klasifikasi berita palsu menggunakan bidirectional LSTM dan Word2vec sebagai vektorisasi. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *9*(3), 1332–1341.

Pakpahan, M., Sinaga, A. F., & Siregar, R. (2022). Analisis sentimen ulasan film menggunakan Word2Vec dan Long Short-Term Memory (LSTM). *Jurnal Teknologi Informasi*, *13*(1), 55–64.

Pratama, B., & Dewi, R. (2024). Optimalisasi parameter TF-IDF untuk klasifikasi teks pendek berbahasa Indonesia. *Jurnal Sistem Informasi*, *20*(1), 50–60.

Rahayu, S., & Susilo, B. (2024). Memahami konteks emosional dan topik berita melalui klasifikasi ganda judul berita. *Prosiding Seminar Nasional Ilmu Komputer dan Teknologi Informasi*, 123–130.

Ramadhan, F. A., Sitorus, S. H., & Rismawan, T. (2023). Penerapan metode Multinomial Naïve Bayes untuk klasifikasi judul berita clickbait dengan Term frequency - Inverse Document frequency. *JUSTIN*, *11*(1), 70–78. https://doi.org/10.26418/justin.v11i1.57452

Ramadhan, M. F. (2023). Klasifikasi topik dan sentimen judul berita dengan augmentasi dan TF-IDF. *Jurnal Riset Komputer*, *10*(2), 87–98.

Ramadhan, M. F., Wibowo, A., & Purnomo, H. (2023). Analisis sentimen judul berita berbahasa Indonesia menggunakan Multinomial Naive Bayes dan TF-IDF. *Jurnal Informatika dan Sistem Informasi*, *9*(1), 45–56.

Santoso, A., & Wijaya, C. (2025). Karakteristik judul berita sensasional dan pengaruhnya terhadap opini pembaca. *Jurnal Media dan Komunikasi*, *6*(1), 1–10.

DOI: https://doi.org/10.31004/riggs.v4i2.1692

Sari, R. A., & Putra, B. I. (2023). Perbandingan algoritma klasifikasi untuk deteksi berita palsu berbahasa Indonesia. *Jurnal Informatika*, *12*(2), 150–160.

Septrinas, E., Indriati, I., & Soebroto, A. A. (2019). Klasifikasi berita olahraga berbahasa Indonesia menggunakan metode BM25 dan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(10), 9762–9769.

Solahuddin, M., Purnamasari, A. I., & Dikananda, A. R. (2023). Klasifikasi kualitas berita pada majalah menggunakan metode Decision Tree. *Jurnal Teknologi Ilmu Komputer*, *1*(2), 52–60. https://doi.org/10.56854/jtik.v1i2.52

Subagyo, D., & Haryanto, T. (2025). Tantangan dan peluang Deep Learning dalam klasifikasi teks dengan dataset terbatas. *Jurnal Komputasi dan Kecerdasan Buatan*, *4*(1), 30–40.

Suputra, I. P. G. H., Linawati, I. G., Sukadarmika, I. G., & Sastra, N. P. (2025). Klasifikasi judul berita bahasa Indonesia menggunakan Support Vector Machine dan seleksi fitur Mutual Information. *Jurnal Pendidikan Teknologi dan Kejuruan*, *22*(1), 45–56. https://doi.org/10.23887/jptkundiksha.v22i1.89158

Suputra, I. M. A., Widiartha, K., & Astawa, I. N. A. (2025). Perbandingan algoritma klasifikasi untuk analisis sentimen teks pendek berbahasa Indonesia. *Jurnal Ilmu Komputer dan Informatika*, *16*(1), 45–56.

Wulandari, E., & Setiawan, R. (2025). Klasifikasi ganda judul berita: Topik dan sentimen menggunakan pendekatan hybrid. *Jurnal Teknologi Informasi*, *16*(1), 65–75.

Yavi, A. F. (2018). Klasifikasi artikel berbahasa Indonesia untuk mendeteksi clickbait menggunakan metode Naïve Bayes. *J-INTECH (Journal of Information and Technology)*, *6*(1), 1–10.

DOI: https://doi.org/10.31004/riggs.v4i2.1692