

OPTIMALISASI MODEL KLASIFIKASI MENGGUNAKAN ALGORITMA NAIVE BAYES DENGAN SMOTE PADA DATA SENTIMEN KOMENTAR KOMPAS TV DI YOUTUBE

Deris Setiadi¹, Martanto², Fatihanursari Dikananda³, Mulyawan⁴.

Program Studi Teknik Informatika¹³
Program Studi Manajemen Informatika²
Program Studi Sistem Informasi⁴

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

(*) Corresponding Author : drissstiadi@gmail.com
Published : 15 Juli 2026

Abstract—The rapid development of digital media particularly YouTube has transformed comment sections into a major source of public opinion on social and political issues. User-generated comments, which are often spontaneous, informal, and unstructured, present significant challenges for sentiment analysis, especially when the dataset exhibits imbalanced class distribution. This study aims to analyze sentiment in user comments on the KOMPAS TV YouTube channel using the Naïve Bayes algorithm and to evaluate the effect of the Synthetic Minority Oversampling Technique (SMOTE) on model performance. Comment data were collected using the youtube-comment-downloader library and subsequently subjected to several preprocessing steps, including text cleaning, normalization, tokenization, stopword removal, and stemming. The comments were automatically labeled using the IndoBERT model to generate three sentiment categories: positive, negative, and neutral. Text representation was conducted using TF-IDF, and sentiment classification was performed using Multinomial and Bernoulli Naïve Bayes. Experiments were executed under two scenarios: without SMOTE and with SMOTE applied to address class imbalance. The findings indicate that the comment dataset is dominated by neutral and negative sentiments, necessitating class-balancing techniques to improve model recognition of minority classes. In the scenario without SMOTE, the Naïve Bayes model produced relatively high accuracy but demonstrated low recall and F1-scores for minority classes. The application of SMOTE significantly improved model performance for positive and neutral sentiments, particularly in terms of recall and F1-score, without substantially compromising overall model stability. These results suggest that SMOTE is effective in enhancing the model's sensitivity to minority classes, although it may introduce noise in high-dimensional feature spaces. Overall, this study provides empirical evidence regarding the implementation of oversampling techniques in Indonesian-language sentiment analysis and contributes a computational framework for understanding public opinion in digital media environments.

Keywords: Sentiment Analysis, Naïve Bayes, SMOTE, YouTube, TF-IDF, IndoBERT.

Abstrak—Perkembangan media digital, khususnya YouTube, telah menjadikan ruang komentar sebagai salah satu sumber utama opini publik terhadap isu sosial dan politik. Komentar yang bersifat spontan, informal, dan cenderung tidak terstruktur memunculkan tantangan dalam proses analisis sentimen, terutama ketika data memiliki distribusi kelas yang tidak seimbang. Penelitian ini bertujuan untuk menganalisis sentimen komentar pada kanal YouTube KOMPAS TV menggunakan algoritma Naïve Bayes serta mengevaluasi pengaruh penerapan Synthetic Minority Oversampling Technique (SMOTE) terhadap performa model. Data komentar dikumpulkan menggunakan library youtube-comment-downloader, kemudian melalui tahap pra-pemrosesan meliputi pembersihan teks, normalisasi, tokenisasi, penghapusan stopword, dan stemming. Selanjutnya, komentar diberi label secara otomatis menggunakan model IndoBERT untuk menghasilkan tiga kategori sentimen: positif, negatif, dan netral. Representasi fitur teks dilakukan menggunakan TF-IDF, kemudian model diklasifikasikan menggunakan Multinomial dan Bernoulli Naïve Bayes. Eksperimen dilakukan pada dua skenario, yaitu tanpa SMOTE dan dengan SMOTE untuk mengatasi ketidakseimbangan kelas. Hasil penelitian menunjukkan bahwa dataset komentar YouTube memiliki dominasi kelas netral dan negatif sehingga diperlukan teknik penyeimbangan agar model mampu mengenali kelas minoritas secara lebih efektif. Pada skenario tanpa SMOTE, model Naïve

Bayes cenderung menghasilkan akurasi tinggi, namun recall dan F1-score kelas minoritas rendah. Penerapan SMOTE terbukti meningkatkan performa pada kelas positif dan netral, terutama dalam metrik recall dan F1-score, tanpa menurunkan stabilitas model secara signifikan. Temuan ini menunjukkan bahwa SMOTE efektif dalam meningkatkan sensitivitas model terhadap kelas minoritas meskipun terdapat risiko peningkatan noise pada fitur berdimensi tinggi. Secara keseluruhan, penelitian ini memberikan kontribusi empiris terkait implementasi metode oversampling pada analisis sentimen berbahasa Indonesia serta menawarkan pendekatan komputasional yang dapat digunakan untuk memetakan opini publik di platform digital.

Kata Kunci: Analisis Sentimen, Naïve Bayes, SMOTE, YouTube, TF-IDF, IndoBERT.

INTRODUCTION

Perkembangan teknologi informasi dan komunikasi telah mendorong meningkatnya penggunaan media sosial sebagai sarana untuk menyampaikan opini, kritik, dan aspirasi publik. Salah satu platform yang активно digunakan adalah YouTube, yang tidak hanya berfungsi sebagai media berbagi video, tetapi juga sebagai ruang diskusi melalui fitur komentar. Komentar-komentar tersebut mencerminkan berbagai persepsi masyarakat terhadap isu tertentu, termasuk isu pemerintahan.

Analisis sentimen merupakan salah satu teknik dalam *text mining* yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini publik ke dalam kategori tertentu, seperti positif, negatif, atau netral. Teknik ini menjadi penting dalam memahami kecenderungan opini masyarakat secara otomatis dan dalam skala besar [1].

Dalam implementasinya, berbagai algoritma klasifikasi telah digunakan untuk analisis sentimen, salah satunya adalah *Naïve Bayes*. Algoritma ini banyak digunakan karena memiliki keunggulan dalam hal kesederhanaan, efisiensi, serta kemampuan dalam menangani data teks berdimensi tinggi [2]. Meskipun demikian, performa *Naïve Bayes* sangat dipengaruhi oleh kualitas dan distribusi data yang digunakan.

Salah satu permasalahan utama dalam analisis sentimen adalah ketidakseimbangan data (*imbalanced data*), di mana jumlah data pada setiap kelas tidak seimbang. Kondisi ini menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas dan kurang mampu mengenali kelas minoritas, sehingga menurunkan performa model secara keseluruhan [3].

Untuk mengatasi permasalahan tersebut, digunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)*, yang bekerja dengan menghasilkan data sintetis pada kelas minoritas untuk menyeimbangkan distribusi data. Beberapa penelitian menunjukkan bahwa penerapan SMOTE dapat meningkatkan

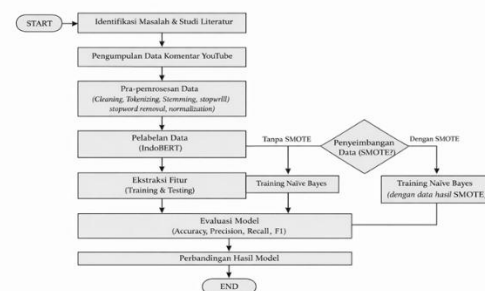
performa model klasifikasi, terutama pada metrik recall dan F1-score [4].

Namun demikian, penerapan SMOTE dalam konteks analisis sentimen berbahasa Indonesia, khususnya pada data komentar YouTube yang bersifat tidak terstruktur dan dinamis, masih relatif terbatas. Selain itu, belum banyak penelitian yang secara khusus melakukan perbandingan performa algoritma *Naïve Bayes* sebelum dan sesudah penerapan SMOTE dalam konteks opini publik terhadap isu pemerintahan.

Berdasarkan uraian tersebut, penelitian ini dilakukan untuk menganalisis pengaruh penerapan SMOTE terhadap performa algoritma *Naïve Bayes* dalam klasifikasi sentimen komentar YouTube, sehingga diharapkan dapat memberikan kontribusi dalam meningkatkan kualitas analisis sentimen pada data media sosial.

MATERIALS AND METHODS

Tahapan penelitian disusun untuk memberikan gambaran menyeluruh mengenai alur penelitian dari awal hingga akhir. Penyajian dalam bentuk diagram bertujuan untuk mempermudah pemahaman terhadap hubungan antar proses penelitian.



Gambar 1. Alur Penelitian

Gambar tersebut menunjukkan alur penelitian analisis sentimen berbasis data komentar YouTube yang dilakukan secara sistematis dari tahap awal hingga evaluasi model. Proses dimulai dari tahap identifikasi masalah dan studi literatur, yang bertujuan untuk memahami konteks penelitian serta menemukan celah penelitian yang relevan.

Selanjutnya dilakukan pengumpulan data komentar YouTube sebagai sumber utama data penelitian.

Tahap berikutnya adalah pra-pemrosesan data, yang meliputi proses cleaning (pembersihan data), tokenizing (pemecahan teks menjadi kata), stemming (mengubah kata ke bentuk dasar), stopword removal (menghapus kata umum yang tidak bermakna), dan normalisasi. Setelah data bersih, dilakukan pelabelan data menggunakan IndoBERT, yaitu model berbasis deep learning untuk menentukan sentimen (misalnya positif, negatif, atau netral).

Kemudian masuk ke tahap ekstraksi fitur, di mana data dibagi menjadi data latih (training) dan data uji (testing) untuk keperluan pemodelan. Pada tahap ini, terdapat percabangan terkait penyeimbangan data menggunakan SMOTE (Synthetic Minority Over-sampling Technique). Jika tanpa SMOTE, data langsung digunakan untuk training model Naïve Bayes. Sementara itu, jika menggunakan SMOTE, data terlebih dahulu diseimbangkan, kemudian digunakan untuk melatih model Naïve Bayes.

Setelah model dilatih, dilakukan evaluasi model menggunakan metrik performa seperti accuracy, precision, recall, dan F1-score. Hasil dari kedua pendekatan (dengan dan tanpa SMOTE) kemudian dibandingkan pada tahap perbandingan hasil model untuk menentukan metode terbaik. Proses penelitian diakhiri pada tahap end, yang menandakan selesainya seluruh rangkaian analisis.

RESULTS AND DISCUSSION

Evaluasi performa model dilakukan dengan mengukur nilai Precision, Recall, F1-Score, dan Support terhadap data uji yang telah diberi label otomatis menggunakan model IndoBERT. Selain itu, digunakan confusion matrix untuk menggambarkan distribusi hasil prediksi antar kelas sehingga dapat memberikan gambaran yang lebih komprehensif terhadap kinerja model.

Evaluasi Model Naïve Bayes Tanpa SMOTE Hasil evaluasi model Naïve Bayes tanpa penerapan teknik Synthetic Minority Over-sampling Technique (SMOTE) menunjukkan bahwa model memiliki performa yang kurang seimbang antar kelas. Secara keseluruhan, model menghasilkan nilai akurasi sebesar 0,84.

Dilihat dari metrik per kelas, model menunjukkan performa sangat baik pada kelas negatif, dengan nilai precision sebesar 0,84, recall 1,00, dan f1-score 0,91, yang mengindikasikan bahwa hampir seluruh data

negatif berhasil terdeteksi dengan baik. Namun, pada kelas netral, performa model menurun drastis dengan recall hanya 0,05 dan f1-score 0,10, meskipun precision mencapai 1,00.

Sementara itu, pada kelas positif, model gagal melakukan klasifikasi dengan baik, ditunjukkan oleh nilai precision, recall, dan f1-score yang semuanya 0,00. Hal ini mengindikasikan bahwa model tidak mampu mengenali data pada kelas minoritas.

Nilai macro average f1-score yang hanya sebesar 0,34 semakin menegaskan adanya ketidakseimbangan performa model. Kondisi ini disebabkan oleh distribusi data yang tidak seimbang (imbalanced dataset), dimana kelas negatif mendominasi jumlah data dibandingkan kelas lainnya.

Evaluasi model Naive Bayes tanpa SMOTE:

	precision	recall	f1-score	support
negatif	0.66	0.98	0.79	338
netral	0.36	0.04	0.07	136
positif	0.00	0.00	0.00	42
accuracy			0.65	516
macro avg	0.34	0.34	0.28	516
weighted avg	0.52	0.65	0.53	516

Akurasi (tanpa SMOTE): 0.6492248062015504

Gambar 2. Naïve Bayes tanpa Smote

Evaluasi Model Naïve Bayes dengan SMOTE Setelah dilakukan penyeimbangan data menggunakan metode SMOTE, performa model mengalami peningkatan yang signifikan dengan akurasi mencapai 0,95. Model mampu mengklasifikasikan seluruh kelas dengan sangat baik dan seimbang.

Kelas *negatif* memperoleh F1-score sebesar 0,92, sementara kelas *netral* dan *positif* masing-masing mencapai 0,96 dan 0,97. Nilai recall pada kelas netral dan positif bahkan mencapai 1,00, yang menunjukkan bahwa seluruh data pada kedua kelas tersebut berhasil terdeteksi dengan sempurna.

Selain itu, nilai macro average dan weighted average yang sama-sama mencapai 0,95 menunjukkan bahwa model memiliki performa yang konsisten dan tidak bias terhadap kelas tertentu. Hal ini membuktikan bahwa penerapan SMOTE efektif dalam mengatasi permasalahan ketidakseimbangan data serta meningkatkan kemampuan generalisasi model Naïve Bayes dalam klasifikasi sentimen.

Evaluasi model Naive Bayes dengan SMOTE:

	precision	recall	f1-score	support
negatif	0.88	0.64	0.71	338
netral	0.77	0.95	0.81	338
positif	0.84	0.92	0.88	339
accuracy			0.88	1015
macro avg	0.88	0.88	0.88	1015
weighted avg	0.88	0.88	0.88	1015

Akurasi: 0.8809852216748768

Gambar 3. Naïve Bayes dengan SMOTE

Confusion Matrix Terlihat bahwa model mampu mengklasifikasikan data secara lebih seimbang pada setiap kelas. Jumlah prediksi benar meningkat secara signifikan, khususnya pada kelas netral dan positif.

Penerapan SMOTE berhasil menghasilkan distribusi data yang lebih proporsional, sehingga model dapat mempelajari karakteristik masing-masing kelas dengan lebih optimal. Teknik ini terbukti efektif dalam meningkatkan kemampuan model dalam mengenali kelas minoritas (Chawla et al., 2022).

CONCLUSION

Kesimpulan pada penelitian ini disusun berdasarkan seluruh rangkaian proses yang telah dilakukan, mulai dari pengumpulan data komentar YouTube, tahapan pra-pemrosesan teks, pelabelan otomatis, hingga pelatihan dan evaluasi model Naïve Bayes dengan dan tanpa penerapan SMOTE. Melalui analisis yang sistematis, penelitian ini memberikan gambaran menyeluruh mengenai pengaruh ketidakseimbangan data terhadap performa model klasifikasi sentimen, serta efektivitas teknik oversampling dalam meningkatkan hasil prediksi. Oleh karena itu, bagian ini merangkum temuan utama yang diperoleh sepanjang penelitian dan menjadi dasar bagi rekomendasi pengembangan penelitian selanjutnya.

Ketidakseimbangan kelas memiliki pengaruh besar terhadap performa model Naïve Bayes. Tanpa SMOTE, model cenderung lebih akurat pada kelas mayoritas, sementara kelas minoritas menunjukkan nilai recall dan F1-score yang rendah.

Penerapan SMOTE berhasil menyeimbangkan distribusi kelas sehingga model dapat mempelajari karakteristik setiap kelas secara lebih optimal. Perbaikan paling terlihat pada kelas positif dan netral.

SMOTE meningkatkan kemampuan model dalam mengenali variasi sentimen, meskipun tetap perlu kehati-hatian karena oversampling berpotensi menambah noise pada data berdimensi tinggi.

Kombinasi pra-pemrosesan yang tepat, penggunaan TF-IDF sebagai representasi fitur, serta penerapan SMOTE menghasilkan performa klasifikasi yang lebih stabil dan seimbang dalam analisis sentimen komentar YouTube berbahasa Indonesia.

REFERENCE

- [1] A. G. Ganie, "Presence of informal language, such as emoticons, hashtags

and slang, impact the performance of sentiment-analysis models on social-media text?," *arXiv preprint*, 2023.

- [2] Y. Mao, "Sentiment Analysis Methods, Applications, and Challenges," *Expert Systems with Applications: Artificial Intelligence*, vol. 1, p. 100024, 2024, doi: 10.1016/j.eswaai.2024.100024.

- [3] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *Journal of Big Data*, vol. 11, p. 87, 2024, doi: 10.1186/s40537-024-00943-4.

- [4] J. Pardede and D. P. Pamungkas, "The impact of balanced data techniques on classification model performance," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 401–412, 2024, doi: 10.15294/sji.v11i2.3649.