

Analisis Sentimen Terhadap Kurikulum Kampus Merdeka dari Komentar di Platform X Menggunakan Metode Naive Bayes dan Random Forest

Prastya Firmansyah | 221080200076

Program Studi Informatika

Fakultas Sains dan Teknologi UMSIDA 2026

Latar Belakang

- *Perkembangan media sosial mendorong munculnya banyak opini publik terhadap kebijakan pemerintah, salah satunya Kurikulum Kampus Merdeka (MBKM). Berikut alasan penelitian ini dilakukan:*
- Media sosial, khususnya platform X, menjadi ruang ekspresi opini publik secara real-time dan terbuka
- Kebijakan MBKM mendapat beragam respons positif maupun negatif dari masyarakat
- Analisis manual ribuan tweet tidak efisien dan rentan subjektivitas
- Dibutuhkan pendekatan otomatis berbasis NLP untuk klasifikasi opini
- Studi terdahulu melaporkan performa Naive Bayes dan Random Forest yang bervariasi (74%–97%), tergantung karakteristik dataset

Rumusan Masalah & Tujuan

- Berdasarkan latar belakang tersebut, penelitian ini difokuskan pada pertanyaan dan tujuan berikut:
- Bagaimana hasil klasifikasi sentimen masyarakat terhadap Kurikulum MBKM di Platform X?
- Bagaimana akurasi Naive Bayes dan Random Forest dalam analisis sentimen?
- Metode mana yang paling efektif berdasarkan nilai evaluasi (positif, negatif, netral)?
- TUJUAN: Mengklasifikasi sentimen, membandingkan performa NB vs RF, menentukan algoritma terbaik

Metodologi Penelitian

- *Penelitian dilakukan secara sistematis melalui beberapa tahapan, mulai dari pengumpulan data hingga evaluasi model klasifikasi:*
- Pengumpulan Data: Web scraping Platform X (Selenium), bertahap multi-sesi hingga pertengahan 2026 → >2.400 tweet unik
- Random Undersampling: Menyeimbangkan kelas → 1.051 tweet (351 Netral, 350 Positif, 350 Negatif)
- Preprocessing: Case folding, tokenisasi, stopword removal, stemming (Sastrawi)
- Ekstraksi Fitur: TF-IDF
- Pembagian Data: 80% latih (840) | 20% uji (210 tweet)
- Evaluasi: Akurasi, Presisi, Recall, F1-Score

Dataset & Distribusi Sentimen

- *Data dikumpulkan bertahap (multi-sesi) dari Platform X menggunakan lima kata kunci terkait Program MBKM, lalu diseimbangkan melalui random undersampling:*
- **Terkumpul: >2.400 tweet unik | Setelah undersampling: 1.051 tweet seimbang**
- **Kata kunci: MBKM, Kampus Merdeka, Merdeka Belajar, Program MBKM, Merdeka Belajar Kampus Merdeka**
- **Netral: 351 tweet (33,4%) | Positif: 350 tweet (33,3%) | Negatif: 350 tweet (33,3%)**
- **Sebagian label menggunakan pendekatan lexicon, divalidasi via spot-check manual (kesesuaian 70–85%)**

Metode: Naive Bayes & Random Forest

- *Dua algoritma machine learning digunakan dan dibandingkan dalam penelitian ini. Keduanya memiliki pendekatan yang berbeda dalam melakukan klasifikasi teks:*
- NAIVE BAYES: Probabilistik berbasis Teorema Bayes | $P(C|X) = P(X|C) \times P(C) / P(X)$
- Multinomial NB, Alpha = 1.0 (Laplace Smoothing) | Sederhana, cepat, efektif
- RANDOM FOREST: Ensemble learning gabungan banyak Decision Tree
- Majority Voting + Gini Impurity + Bootstrap Sampling (n_estimators = 200)
- Akurasi tinggi, tahan overfitting, mengukur feature importance

Evaluasi Model

- *Performa kedua model diukur menggunakan empat metrik evaluasi standar dalam klasifikasi teks. Referensi penelitian terdahulu turut dijadikan acuan perbandingan:*
- Akurasi = $(TP + TN) / (TP + TN + FP + FN)$
- Presisi = $TP / (TP + FP)$ | Recall = $TP / (TP + FN)$
- F1-Score = Harmonic Mean dari Presisi dan Recall
- Referensi: Zhafira et al. (2021) – NB mencapai akurasi 97% pada topik Kampus Merdeka
- Referensi: Admojo et al. (2024) – Random Forest 88% vs Naive Bayes 74%

Hasil Evaluasi Model

- *Kedua model diuji pada 210 data uji (stratified split 80:20). Random Forest unggul pada seluruh metrik evaluasi:*
- **Naive Bayes:** Accuracy 55,50% | Precision 54,35% | Recall 55,43% | F1-Score 53,93%
- **Random Forest:** Accuracy 63,64% | Precision 65,14% | Recall 63,66% | F1-Score 62,55%
- Random Forest unggul +8,14 poin accuracy, +10,79 poin precision, +8,23 poin recall, dan +8,62 poin F1-score dibanding Naive Bayes
- Diuji pula pada rasio 70:30 dan 60:40 – Random Forest konsisten unggul (62,74% dan 61,48%)

Confusion Matrix & Pembahasan

- *Analisis confusion matrix model Random Forest (70 data uji per kelas):*
- Kelas Negatif paling akurat dikenali: 54 dari 70 data uji benar
- Kelas Netral juga baik: 50 dari 70 data uji benar
- Kelas Positif paling sulit: hanya 32 dari 70 benar, sering tertukar Netral/Negatif
- Penyebab akurasi moderat: kelas diseimbangkan via undersampling, data latih berkurang (840 tweet), dan noise label dari pendekatan lexicon (kesesuaian 70–85%)

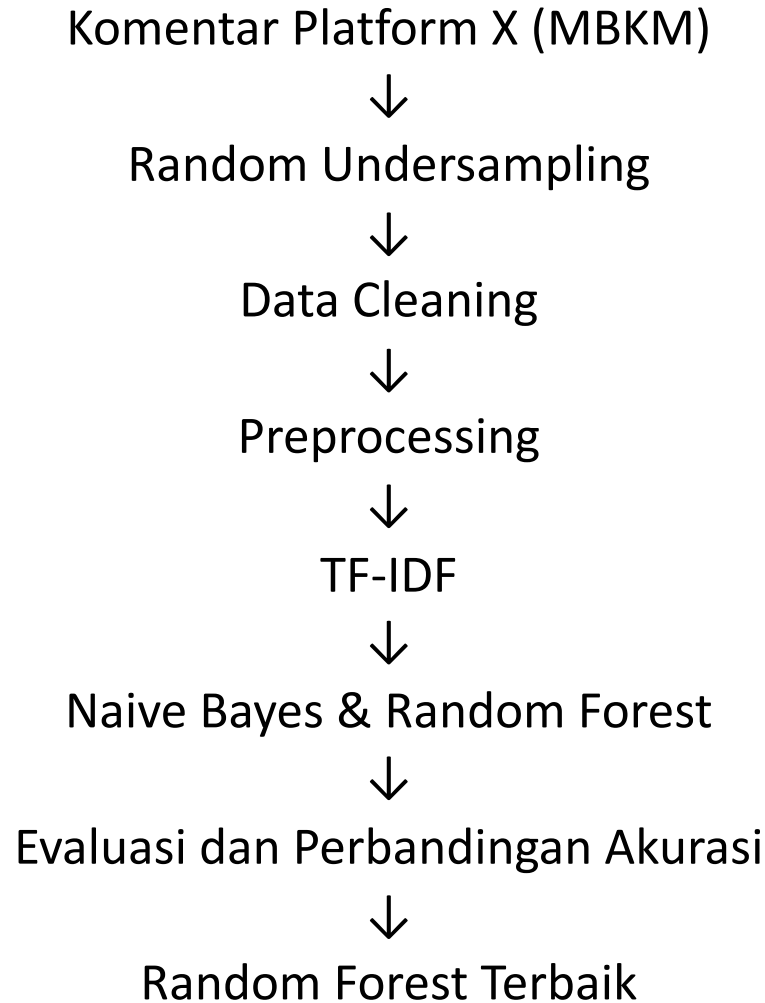
Penelitian Terdahulu

- Zhafira dkk. (2021) – Naive Bayes + TF-IDF, Akurasi 97%
- Admojo dkk. (2024) – Naive Bayes 74% vs Random Forest 88%
- Nugroho dkk. (2025) – Naive Bayes 89% vs Random Forest 93%
- Kesimpulan: Naive Bayes dan Random Forest sama-sama efektif untuk analisis sentimen.

Research Gap

- Belum ada penelitian yang membandingkan Naive Bayes dan Random Forest pada sentimen MBKM di Platform X.
- Sebagian penelitian menggunakan YouTube atau Google Play Store.
- Dataset MBKM di media sosial masih terbatas.
- Penelitian ini menggunakan data Platform X yang dikumpulkan bertahap hingga pertengahan 2026 dengan 3 kelas sentimen (positif, negatif, netral).

Kerangka Berpikir Penelitian



Simpulan & Saran

- *Berdasarkan hasil pengujian terhadap 1.051 tweet berbahasa Indonesia:*
- Random Forest terbukti lebih unggul dari Naive Bayes: Accuracy 63,64% vs 55,50%
- Ensemble learning pada Random Forest lebih mampu menangkap pola hubungan antar kata yang kompleks dibanding asumsi independensi fitur pada Naive Bayes
- Performa moderat dipengaruhi keseimbangan kelas (undersampling), volume data latih terbatas, dan noise label dari pendekatan lexicon
- **Saran:** Perbanyak data berlabel manual & uji algoritma lain (SVM, deep learning) untuk performa lebih optimal

Penutup

- *Penelitian ini diharapkan memberikan kontribusi nyata dalam bidang analisis sentimen kebijakan pendidikan, khususnya Program MBKM yang belum banyak diteliti menggunakan pendekatan ini:*
- Mengisi gap: belum ada studi perbandingan NB vs RF pada sentimen MBKM di Platform X
- Memberikan insight opini publik terhadap kebijakan pendidikan pemerintah
- Kontribusi: Baseline akurasi, rekomendasi Random Forest sebagai metode terbaik, metodologi yang dapat direplikasi
- Implementasi: Python (Scikit-learn, Sastrawi, Pandas, Numpy)
- Pembimbing: Yulian Findawati, ST., M.MT | UMSIDA 2026