Jurnal Informatika dan Teknologi Interaktif
Volume 2, Nomor 2, Agustus 2025

DOI : 10.63547/jiite.v2i2.92
e-ISSN : 3048-1023

Analisis Perbandingan: SMOTE dan Undersampling pada Klasifikasi Spam Naïve Bayes

Hafizh Dzaky[1], Yusup Ardiyanto[2], Rivaldo Jeffmarvin[3], Rivaldo Jeffmarvin[3], Apriliyanto Dwi Saputra[4], Deri Irawan[5], Jason Bernard Ardianto[6]

[1,2,3,4,5]Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Purwokerto, Indonesia
[6]Jurusan Teknik Informatika, Sekolah Tinggi Ilmu Komputer Yos Sudarso, Purwokerto, Indonesia
surel: [1]hdag304001@gmail.com, [2]yusupardiyanto19@gmail.com, [3]rivaldojeffmarvin0@gmail.com, [4]aprildwi999@gmail.com, [5]ideri9067@gmail.com, [6]jasonadrianto1505@gmail.com

Info Artikel

Sejarah artikel:

Diterima 10-07-2025
Revisi 23-07-2025
Diterima 05-08-2025

Kata kunci:

Naïve Bayes
Spam Email
Bahasa Indonesia
SMOTE
Random Undersampling

ABSTRAK

Spam email merupakan masalah serius dalam komunikasi digital, dan sebagian besar riset mengenai deteksi spam masih terfokus pada dataset berbahasa Inggris, sehingga menciptakan celah penelitian untuk bahasa lain seperti bahasa Indonesia. Penelitian ini bertujuan untuk mengisi celah tersebut dengan mengimplementasikan algoritma *Naïve Bayes* untuk klasifikasi spam pada dataset berbahasa Indonesia. Selain itu, penelitian ini juga membandingkan efektivitas dua teknik penyeimbangan data, yaitu *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Random Undersampling* (RUS), untuk menemukan pendekatan yang paling optimal. Metode penelitian mencakup beberapa tahapan, mulai dari pra-pemrosesan data, ekstraksi fitur menggunakan TF-IDF dan N-gram, hingga pelatihan model *Naïve Bayes*. Hasil evaluasi menunjukkan bahwa kedua model memiliki performa yang sangat baik. Model dengan teknik RUS sedikit lebih unggul dengan akurasi 95,74%, presisi 95,92%, dan F1-score 95,73%, dibandingkan model SMOTE yang mencapai akurasi 95,63%. Kesimpulannya, teknik RUS menunjukkan hasil yang lebih stabil dan efisien untuk dataset ini, membuktikan bahwa Naïve Bayes adalah solusi yang kuat untuk deteksi spam berbahasa Indonesia.

Penulis Korespondensi:

Rivaldo Jeffmarvin
Program Studi Informatika Fakultas Ilmu Komputer Universitas Amikom Purwokerto
Email: rivaldojeffmarvin0@gmail.com

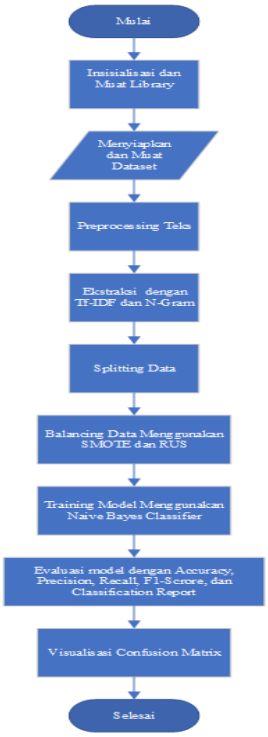
1. PENDAHULUAN

Spam, atau email sampah, merupakan masalah serius dalam komunikasi digital yang menyebabkan berbagai dampak negatif. Dampak tersebut mencakup pemborosan sumber daya teknis seperti *bandwidth* dan ruang penyimpanan, hingga gangguan bagi pengguna melalui pesan yang tidak relevan atau bahkan berbahaya. oleh karena itu, pengembangan sistem penyaringan yang efektif untuk memisahkan email spam dari email yang sah (*ham*) menjadi sangat penting untuk dipelajari [1].

Berbagai penelitian telah menjawab tantangan ini, seperti studi yang dilakukan oleh Al Amien dkk [2]. Penelitian yang lebih baru bahkan telah mencapai akurasi 98% dengan menggabungkan Naïve Bayes dan teknik *Synthetic Minority Over-sampling Technique* (SMOTE). Namun, riset-riset tersebut memiliki keterbatasan fundamental karena menggunakan dataset berbahasa Inggris, seperti dataset *lingspam* yang digunakan pada penelitian tahun 2022, sehingga membuka celah penelitian untuk bahasa lain. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan menguji penerapan algoritma Naïve Bayes pada dataset email berbahasa Indonesia. Lebih lanjut, penelitian ini akan membandingkan efektivitas dua teknik penyeimbangan data, yaitu *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Random Under Sampling* (RUS), untuk menemukan metode paling optimal dalam konteks linguistik Indonesia. [2]

2. METODE

Metode yang digunakan dalam penelitian ini adalah metode *Naive Bayes Classifier* dengan teknik SMOTE dan RUS dalam mendeteksi pesan email dan pesan sms spam. Kedua metode tersebut akan dibandingkan performanya menggunakan dataset berbahasa indonesia.



Gambar 1. Alur Sistem

2.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah gabungan koleksi pesan SMS dan email dalam format CSV (spam.csv) dengan dua kelas: ham (bukan spam) yang berjumlah 2406 dan spam yang berjumlah 2524 data dengan total 4926 data [3] [4].

Tabel 1. Dataset Pesan Spam dan Ham

Kategori	Pesan
spam	Secara alami tak tertahankan identitas perusahaan Anda sangat sulit untuk ...
spam	Jangan punya uang, dapatkan CD perangkat lunak dari sini! Kompatibilitas ...
spam	Simpan uang Anda beli untuk mendapatkan barang ini di sini Anda belum ...
ham	Kalau mau bikin model/controller mending per apa y?
ham	ya masuk aja, belum ada tugas/quiz kok. cuma perlu ngejar dikit. Materiny ...

2.2. *Preprocessing*

Tahapan *preprocessing data* dalam penelitian ini mencakup pembersihan data (cleaning), mengubah teks menjadi *lowercase* (*case folding*), tokenisasi, penghapusan data tidak relevan (*stopwords removal*), dan stemming menggunakan pustaka sastrawi [5].

2.2.1. *Data Cleaning*

Untuk menjamin integritas dan kualitas himpunan data, langkah esensial yang mengikuti fase pemahaman data (*data understanding*) adalah pembersihan data (*data cleaning*). Tahapan ini secara sistematis mengatasi berbagai permasalahan yang inheren dalam data mentah, meliputi imputasi atau pengisian nilai atribut yang kosong, penyesuaian terhadap nilai ekstrem atau pencilan (*outliers*), dan penyelesaian berbagai bentuk ketidakkonsistenan data yang teridentifikasi. Perlu ditegaskan bahwa kualitas data yang rendah, yang ditandai oleh adanya duplikasi, kesalahan input, maupun data yang tidak lengkap, merupakan isu kritikal. Hal ini karena anomali tersebut dapat mengakibatkan bias signifikan pada hasil analisis dan pada akhirnya memengaruhi akurasi pengambilan keputusan, sehingga penanganannya menjadi tantangan utama dalam pengolahan data [6].

2.2.2. *Case Folding*

Tahapan pra-pemrosesan data diawali dengan normalisasi teks untuk menjamin uniformitas dan mengurangi kompleksitas fitur. Salah satu teknik normalisasi esensial yang diaplikasikan dalam penelitian ini adalah *case folding*. Proses ini secara sistematis mengubah seluruh karakter alfabet pada data mentah ke dalam satu kasus (*case*) yang seragam. Secara operasional, penelitian ini menerapkan konversi penuh ke format huruf kecil (*lowercase*). Langkah ini sangat penting untuk mencegah ambiguitas leksikal, memastikan bahwa token seperti 'Sistem' dan 'sistem' dianggap sebagai satu entitas yang sama, yang pada akhirnya meningkatkan akurasi dan efisiensi analisis [7].

Tabel 2. Hasil Proses *Case Folding*

Data Pesan Asli	Hasil *Case Folding*
Secara alami tak tertahankan identitas perusahaan Anda sangat sulit untuk mengin …	secara alami tak tertahankan identitas perusahaan anda sangat sulit untuk mengin …
Fanny Gunslinger Perdagangan Saham adalah Merrill tetapi Muzo bukan Colza yang dicap …	fanny gunslinger perdagangan saham adalah merrill tetapi muzo bukan colza yang dicap …

2.2.3. Tokenisasi

Dalam alur pemrosesan bahasa alami (NLP), tokenisasi adalah langkah fundamental yang memecah teks menjadi unit-unit terkecil bernama token, seperti kata atau tanda baca, yang mana proses ini wajib dilakukan sebelum tahap penghapusan *stopwords*. Proses *stopwords removal* baru bisa berjalan efektif setelah teks dipecah menjadi token-token individual, memungkinkan sistem untuk mengidentifikasi dan menyingkirkan kata-kata umum yang minim makna kontekstual (misalnya, "yang", "dan", "adalah") demi meningkatkan efisiensi dan fokus analisis pada kata-kata kunci yang lebih informatif [8].

Tabel 3. Hasil Proses Tokenisasi

Hasil *Case Folding*	Hasil Tokenisasi
secara alami tak tertahankan identitas perusahaan anda sangat sulit untuk …	['secara', 'alami', 'tak', 'tertahankan', 'identitas', 'perusahaan', 'anda', 'sangat', 'sulit', 'untuk', …]
fanny gunslinger perdagangan saham adalah merrill tetapi muzo bukan …	['fanny', 'gunslinger', 'perdagangan', 'saham', 'adalah', 'merrill', 'tetapi', 'muzo', 'bukan', …]

2.2.4. *Stopwords Removal*

Stopwords removal merupakan sebuah metode vital dalam pemrosesan awal teks yang bertujuan untuk mengeliminasi kata-kata berfrekuensi tinggi namun memiliki nilai semantik minimal, seperti konjungsi ("dan", "atau") atau verba kopulatif ("adalah"). Proses ini krusial untuk memfokuskan analisis pada unit-unit linguistik yang lebih informatif dan relevan, terutama dalam konteks analisis sentimen, sebagaimana diuraikan oleh Asaad & Abdulhakim [9] mengenai token sebagai unit terkecil dalam teks.

Tabel 4. Hasil Proses *Stopwords Removal*

Hasil Tokenisasi	Hasil *Stopwords Removal*
['secara', 'alami', 'tak', 'tertahankan', 'identitas', 'perusahaan', 'anda', 'sangat', 'sulit', 'untuk', …]	['alami', 'tertahankan', 'identitas', 'perusahaan', 'sulit', 'perusahaan', 'pasar', 'penuh', …]

['fanny', 'gunslinger', 'perdagangan', 'saham', 'adalah', 'merrill', 'tetapi', 'muzo', 'bukan', …]	['fanny', 'gunslinger', 'perdagangan', 'saham', 'merrill', 'muzo', 'colza', 'dicapai', 'esmark', …]

2.2.5. *Stemming*

Dalam ranah pemrosesan bahasa alami (*Natural Language Processing* - NLP), stemming merupakan sebuah proses krusial yang bertujuan untuk mereduksi variasi morfologis kata menjadi bentuk dasarnya atau "akar" kata. Tahapan ini esensial untuk normalisasi teks, di mana imbuhan—baik itu prefiks (awalan), sufiks (akhiran), maupun infiks (sisipan)—dihilangkan dari kata-kata. Implementasi stemming secara efektif berkontribusi pada penurunan dimensi data tekstual, yang pada gilirannya dapat meningkatkan efisiensi dan akurasi pada berbagai aplikasi NLP seperti pencarian informasi, pengelompokan dokumen, atau klasifikasi teks, dengan menyatukan kata-kata yang berbeda bentuk namun memiliki makna dasar yang sama [10].

Tabel 5. Hasil Proses *Stemming* dengan Sastrawi

Hasil *Stopwords Removal*	Hasil *Stemming*
['alami', 'tertahankan', 'identitas', 'perusahaan', 'sulit', 'perusahaan', 'pasar', 'penuh', …]	['alam', 'tertah', 'identitas', 'perusaha', 'sulit', 'perusaha', 'pasar', 'penuh', 'suqestions', …]
['fanny', 'gunslinger', 'perdagangan', 'saham', 'merrill', 'muzo', 'colza', 'dicapai', 'esmark', …]	['fanny', 'gunslinger', 'perdagang', 'saham', 'rrill', 'muzo', 'colza', 'capa', 'esmark', …]

2.3. Ekstraksi Fitur

TF-IDF (Term Frequency–Inverse Document Frequency) adalah metode representasi teks yang memberikan bobot pada setiap kata berdasarkan frekuensinya dalam dokumen dan keseluruhan korpus. Untuk menangkap konteks yang lebih kaya, metode ini dikombinasikan dengan teknik n-gram, yaitu pengelompokan kata berdasarkan urutan kemunculan, seperti unigram (kata tunggal) atau bigram (dua kata berurutan). Penggunaan TF-IDF dengan n-gram memungkinkan model klasifikasi mendeteksi pola frasa yang umum pada spam, sehingga meningkatkan akurasi deteksi [11].

2.4. Pembagian Data

Dalam bidang machine learning, proses splitting data dilakukan untuk memisahkan dataset menjadi dua bagian utama, yaitu data pelatihan (training data) dan data pengujian (testing data). Data pelatihan berfungsi sebagai sumber informasi untuk membangun dan melatih model berdasarkan pola-pola yang telah diketahui sebelumnya. Sementara itu, data pengujian digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data yang belum pernah dilihat sebelumnya secara akurat. Dalam penelitian ini data dibagi menjadi 80% data latih dan 20% data uji [12].

2.5. Penyeimbangan Data

Dalam penelitian ini, digunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) sebagai metode penyeimbangan data pada tahap praproses. SMOTE bekerja dengan mensintesis contoh baru pada kelas minoritas dengan cara menginterpolasi antara sampel yang berdekatan dalam ruang fitur. Pendekatan ini bertujuan untuk mengatasi masalah ketidakseimbangan kelas yang sering menyebabkan bias pada model klasifikasi. Dengan menerapkan SMOTE, model diharapkan dapat belajar secara lebih adil terhadap kedua kelas, sehingga meningkatkan akurasi dan stabilitas dalam proses klasifikasi [13].

Random Undersampling (RUS) merupakan salah satu teknik penyeimbangan data yang digunakan untuk menangani permasalahan ketidakseimbangan kelas dalam dataset. Metode ini bekerja dengan cara mengurangi jumlah sampel dari kelas mayoritas secara acak, sehingga distribusi data antar kelas menjadi lebih seimbang. Dengan menghapus sebagian data dari kelas yang dominan, model pembelajaran mesin dapat terhindar dari kecenderungan untuk mempelajari pola dari kelas mayoritas secara berlebihan, yang sering kali menyebabkan performa yang buruk dalam mengklasifikasikan kelas minoritas. Meskipun sederhana, RUS cukup efektif dan efisien dari segi waktu pemrosesan, namun tetap memiliki risiko kehilangan informasi penting dari kelas mayoritas akibat penghapusan data [14].

2.6. Algoritma Naive Bayes

Naïve Bayes Classifier merupakan salah satu algoritma klasifikasi yang banyak digunakan dalam pembelajaran mesin, terutama untuk data teks seperti deteksi spam. Algoritma ini bekerja berdasarkan prinsip probabilitas, dengan mengasumsikan bahwa setiap fitur dalam data bersifat independen satu sama lain. Meskipun

asumsi ini bersifat sederhana atau "naif", model ini tetap terbukti efektif dalam berbagai studi karena kemampuannya dalam menangani data berskala besar serta efisiensi dalam proses komputasi [15].

2.7. Evaluasi Performa Model

Model dievaluasi menggunakan metriks accuration, precision, recall, dan f1-score. Selain itu confusion matrix digunakan untuk memvisualisasikan hasil kinerja model.

3. HASIL DAN PEMBAHASAN

Hasil evaluasi performa kedua model, yaitu algoritma Naïve Bayes yang dikombinasikan dengan teknik penyeimbangan data SMOTE dan RUS, menunjukkan bahwa keduanya mampu menghasilkan klasifikasi yang cukup akurat dalam mendeteksi pesan spam berbahasa Indonesia. Evaluasi dilakukan dengan menggunakan beberapa metrik pengukuran standar, seperti akurasi, presisi, recall, dan F1-score, untuk menilai efektivitas masing-masing metode. Selain itu, analisis visual melalui confusion matrix digunakan untuk memahami distribusi kesalahan klasifikasi secara lebih mendalam. Pembahasan pada bagian ini difokuskan pada perbandingan kinerja antara model yang menggunakan SMOTE dan model yang menggunakan RUS, guna mengidentifikasi pendekatan mana yang lebih unggul dalam konteks dataset yang digunakan.

3.1. *Classification Report*

3.1.1. Model dengan SMOTE

```
=== Evaluasi Model SMOTE ===
Akurasi  : 0.9563
Presisi  : 0.9581
Recall   : 0.9563
F1 Score : 0.9563

=== Classification Report (Per Kelas) ===
              precision    recall  f1-score   support

         ham       0.99      0.92      0.95       481
        spam       0.93      0.99      0.96       504

    accuracy                           0.96       985
   macro avg       0.96      0.96      0.96       985
weighted avg       0.96      0.96      0.96       985
```

Gambar 2. *Classification Report Model SMOTE*

Model Naïve Bayes yang diterapkan pada data yang telah diseimbangkan menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE) menunjukkan performa klasifikasi yang baik. Berdasarkan hasil evaluasi, model ini memperoleh nilai akurasi sebesar 95,63%, dengan presisi 95,81%, recall 95,63%, dan F1-score 95,63%. Pada level kelas, nilai presisi untuk kelas spam tercatat sebesar 0,93 dan recall 0,99, menandakan bahwa sebagian besar pesan spam berhasil dikenali dengan benar, meskipun masih terdapat beberapa pesan non-spam yang salah diklasifikasikan sebagai spam. Untuk kelas ham, presisi mencapai 0,99, sedangkan recall berada pada angka 0,92, mengindikasikan masih terdapat sebagian kecil pesan non-spam yang diklasifikasikan sebagai spam. Hasil ini menunjukkan bahwa model dengan SMOTE cenderung lebih sensitif terhadap deteksi spam, tetapi sedikit mengorbankan presisi pada kelas non-spam.

3.1.2. Model dengan RUS

```
=== Evaluasi Model RUS ===
Akurasi  : 0.9574
Presisi  : 0.9592
Recall   : 0.9574
F1 Score : 0.9573

=== Classification Report (Per Kelas) ===
              precision    recall  f1-score   support

         ham       0.99      0.92      0.95       481
        spam       0.93      0.99      0.96       504

    accuracy                           0.96       985
   macro avg       0.96      0.96      0.96       985
weighted avg       0.96      0.96      0.96       985
```

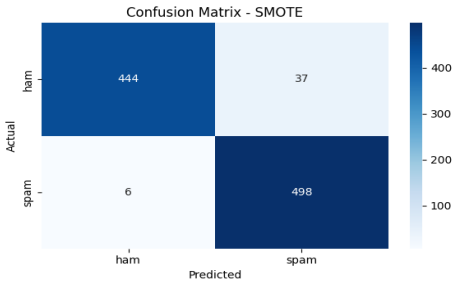
Gambar 3. *Classification Report Model RUS*

Model Naive Bayes yang dikombinasikan dengan teknik Random Under Sampling (RUS) menunjukkan performa yang sangat kompetitif dan sedikit lebih unggul dibandingkan SMOTE. Hasil evaluasi menunjukkan akurasi sebesar 95,74%, presisi 95,92%, recall 95,74%, dan F1-score 95,73%. Nilai presisi dan recall pada kelas spam masing-

masing adalah 0,93 dan 0,99, menunjukkan tingkat keberhasilan yang hampir identik dengan model SMOTE dalam mendeteksi spam. Untuk kelas ham, nilai presisi tetap tinggi di angka 0,99, dan recall sebesar 0,92. Performa model dengan RUS dapat dikatakan lebih stabil, dengan sedikit penurunan false negative dibandingkan SMOTE. Ini menunjukkan bahwa teknik undersampling seperti RUS dapat menjadi alternatif yang efisien dan efektif dalam menangani ketidakseimbangan data, khususnya untuk kasus klasifikasi teks berbahasa Indonesia.

3.2. Distribusi Kelas

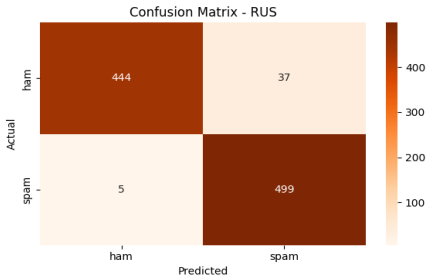
3.2.1. Model dengan SMOTE



Gambar 4. *Confusion Matrix Method SMOTE*

Berdasarkan confusion matrix untuk model dengan teknik SMOTE, diketahui bahwa dari total data uji, terdapat 444 data kelas ham yang berhasil diklasifikasikan dengan benar (true negative) dan 498 data kelas spam yang juga terklasifikasi dengan benar (true positive). Namun, terdapat 37 kasus false positive, yaitu data ham yang salah diklasifikasikan sebagai spam, dan 6 kasus false negative, yaitu data spam yang salah diklasifikasikan sebagai ham. Hal ini mengindikasikan bahwa meskipun model ini sangat sensitif terhadap spam (recall tinggi), namun masih terdapat beberapa pesan non-spam yang tertangkap sebagai spam, yang dapat menyebabkan ketidaknyamanan pengguna.

3.2.2. Model dengan RUS



Gambar 5. *Confusion Matrix Method RUS*

Confusion matrix untuk model dengan teknik RUS menunjukkan distribusi klasifikasi yang hampir serupa dengan model SMOTE. Terdapat 444 data ham yang diklasifikasikan dengan benar dan 499 data spam yang juga diklasifikasikan secara tepat. Adapun jumlah kesalahan klasifikasi terdiri dari 37 false positive dan hanya 5 false negative. Dengan satu kesalahan klasifikasi spam yang lebih sedikit dibandingkan SMOTE, model ini memperlihatkan ketepatan yang lebih tinggi dalam mengenali pesan spam tanpa mengorbankan presisi pada kelas ham. Hal ini mendukung temuan pada evaluasi metrik sebelumnya bahwa model dengan RUS sedikit lebih unggul dalam hal konsistensi dan efisiensi.

4. KESIMPULAN

Penelitian ini dilakukan untuk mengatasi kurangnya studi deteksi spam yang berfokus pada dataset berbahasa Indonesia dengan menguji efektivitas algoritma Naïve Bayes. Secara khusus, penelitian ini membandingkan kinerja model yang menggunakan dua teknik penyeimbangan data yang berbeda: *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Random Under Sampling* (RUS).

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa algoritma Naïve Bayes terbukti sangat efektif untuk melakukan klasifikasi email spam berbahasa Indonesia. Kedua pendekatan penyeimbangan data berhasil menghasilkan model dengan akurasi tinggi. Model yang menggunakan SMOTE mencapai akurasi 95,63%, sementara model dengan RUS menunjukkan performa yang sedikit lebih unggul dengan akurasi 95,74%. Keunggulan model RUS juga terlihat dari jumlah *false negative* yang lebih rendah, yaitu hanya 5 kasus dibandingkan 6 kasus pada model SMOTE, yang mengindikasikan keandalannya yang sedikit lebih baik dalam mengenali spam. Temuan ini mengonfirmasi bahwa meskipun kedua teknik efektif, RUS menawarkan alternatif yang lebih efisien dan stabil untuk menangani ketidakseimbangan kelas pada dataset teks berbahasa Indonesia.

Hasil dari penelitian ini membuka beberapa peluang untuk pengembangan di masa mendatang. Dengan Eksplorasi Algoritma Lain membandingkan kinerja Naïve Bayes dengan algoritma klasifikasi lain seperti Support Vector Machine (SVM), Random Forest, atau bahkan model *deep learning* seperti Long Short-Term Memory (LSTM) untuk melihat apakah ada metode yang dapat memberikan akurasi yang lebih tinggi. Pengembangan Teknik Ekstraksi Fitur: Selain TF-IDF dan N-gram, dapat diuji pula teknik ekstraksi fitur yang lebih canggih seperti *word embedding* (misalnya, Word2Vec atau GloVe) yang mampu menangkap makna semantik dan kontekstual dari kata-kata, yang berpotensi meningkatkan kinerja model. Perluasan Dataset: Menggunakan dataset yang lebih besar dan lebih beragam, yang secara eksklusif terdiri dari email (bukan gabungan dengan SMS), dapat memberikan hasil yang lebih general dan valid untuk kasus penggunaan spesifik pada filter email. Analisis Fitur Spesifik Bahasa Indonesia: Melakukan analisis lebih mendalam terhadap fitur-fitur linguistik yang khas pada spam berbahasa Indonesia, seperti penggunaan singkatan, bahasa gaul, atau pola kalimat tertentu, dapat membantu dalam rekayasa fitur (*feature engineering*) yang lebih baik.

REFERENSI

[1] M. Anita, B. Susanto, and L. Larwuy, "Perbandingan Metode Random Forest dan Naïve Bayes dalam Email Spam Filtering," *KUBIK J. Publ. Ilm. Mat.*, vol. 7, no. 2, pp. 88–96, 2023, doi: 10.15575/kubik.v7i2.18933.

[2] H. Mukhtar, J. Al Amien, and M. A. Rucyat, "Filtering Spam Email menggunakan Algoritma Naïve Bayes," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 1, pp. 9–19, 2022, doi: 10.37859/coscitech.v3i1.3652.

[3] https://www.kaggle.com/datasets/gevabriel/indonesian-email-spam/code

[4] https://www.kaggle.com/datasets/bobsteward/dataset-sms-spam-indonesia

[5] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[6] A. T. Yulianto and Andi Riansyah, "Exploratory Data Analysis Berbasis Excel Dalam Analisis Data Untuk Meningkatkan Penjualan Produk Pada Vending Machine," *J. Comput. Sci. Inf. Technol.*, vol. 2, no. 2, pp. 218–226, 2025, doi: 10.70248/jcsit.v2i2.2044.

[7] I. Amelia, A. Mutiara, and I. Santoso, "Analisis Sentimen Opini Publik Terhadap Pengambil Alihan Tmii Oleh Pemerintah Dengan Algoritma Naïve Bayes," *J. IKRAITH-INFORMATIKA*, vol. 7, no. 2, pp. 142–148, 2023, [Online]. Available: https://journals.upi-yai.ac.id/index.php/ikraith-informatika/issue/archive

[8] R. P. Setiawan, B. Irawan, and W. P. Prihartono, "Analisis Sentimen Ulasan Growtopia Di Google Play Store Menggunakan Naïve Bayes Classifier Untuk Identifikasi Kebutuhan Pengguna," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 2, 2025, doi: 10.23960/jitet.v13i2.6415.

[9] U. Bina and S. Informatika, "Analisis Ulasan Konsumen sebagai Data Non-Keuangan dalam Sistem Informasi Akuntansi," vol. 5, no. 1, pp. 64–74, 2025.

[10] I. M. Juniandika and I. B. M. Mahendra, "Analisis Sentimen Aplikasi Zenius Menggunakan Metode Logistic Regression," *JNATIA J. Nas. Teknol. Inf. dan Apl.*, vol. 1, no. 4, pp. 2986–3929, 2023.

[11] Y. Hapsari, S. Mujahidin, and N. Fadhliana, "Analisis Sentimen Isu Vaksinasi Covid-19 pada Twitter dengan Metode Naive Bayes dan Pembobotan TF-IDF Tokenisasi 1-2 Gram," *SPECTA J. Technol.*, vol. 7, no. 2, pp. 573–583, 2023, doi: 10.35718/specta.v7i2.812.

[12] R. Azan and A. Maslan, "Analisis Klasifikasi Email Spam Menggunakan Algoritma Naïve Bayes," *Comasie*, vol. 05, no. 03, pp. 97–106, 2020.

[13] K. Karfindo, R. Turaina, and R. Saputra, "Optimalisasi Klasifikasi Umpan Balik Mahasiswa Terhadap Layanan Kampus dengan Sinergi Random Forest dan Smote," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 6, no. 6, pp. 820–827, 2024, doi: 10.32672/jnkti.v6i6.7269.

[14] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.

[15] N. B. Classifier, "NAIVE BAYES CLASSIFIER UNTUK DETEKSI EMAIL SPAM - Google Scholar," vol. 15, no. 4, pp. 675–680, 2024, [Online]. Available: https://scholar.google.com/scholar?hl=id&as_sdt=0%2C5&q=NAIVE+BAYES+CLASSIFIER+UNTUK+DETEKSI+EMAIL+SPAM&btnG=