JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 8 Tahun 2026
ISSN 3025-0919 (media online)
Hal 2007-2011

Optimasi K-NN Menggunakan SMOTE dan *Cosine Similarity* pada Klasifikasi Topik Skripsi

Nurrofiqi Ankisqiantari[1*], Adi Prasetyo[2]

[1,2]Fakultas Bisnis dan Hukum, Program Studi Bisnis Digital, Universitas PGRI Yogyakarta, Indonesia
Email: [1*]isqia@upy.ac.id, [2]adipras@upy.ac.id
(* : coressponding author)

Abstrak−Penentuan dosen pembimbing skripsi yang linear dengan topik penelitian mahasiswa merupakan tahapan krusial dalam penyelesaian tugas akhir. Pemetaan judul ke bidang keahlian dosen secara manual seringkali kurang efisien dan dihadapkan pada tantangan distribusi topik yang tidak merata (*imbalanced data*). Penelitian ini mengusulkan optimasi model klasifikasi berbasis algoritma *K-Nearest Neighbor* (K-NN) menggunakan 1.598 data historis judul skripsi dari Universitas AMIKOM Yogyakarta. Untuk mengatasi ketimpangan jumlah sampel antar bidang dosen, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diimplementasikan pada fase pelatihan data (*data train*). Proses ekstraksi Informasi teks melalui tahapan *preprocessing* komprehensif, meliputi *case folding*, *stopword removal*, dan *stemming*, yang kemudian dibobotkan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF). Pembagian komposisi data latih dan uji ditetapkan sebesar 80:20. Berbeda dengan pendekatan spasial standar, model ini diuji menggunakan metrik jarak *Cosine Similarity* untuk mengukur kedekatan semantik antar dokumen dengan lebih presisi. Hasil eksperimen menunjukkan bahwa klasifikasi K-NN mencapai performa maksimal pada parameter k=21 dengan tingkat akurasi sebesar 85.94%. Integrasi SMOTE dan metrik *Cosine* terbukti secara signifikasn menaikkan sensitivitas model, ditandai dengan nilai *recall* hingga 95% pada pengenalan kategori bidang dosen minoritas yang sebelumnya sulit diklasifikasikan.

Kata Kunci: *Imbalanced Data*, Judul Skripsi, Bidang Dosen, Klasifikasi Teks, K-NN

Abstract−*Determining a thesis supervisor whose expertise aligns with a student's research topic is a crucial stage in completing a final project. Manual mapping of thesis titles to lecturer expertise fields is often inefficient and faces the challenge of uneven topic distribution (imbalanced data). This study proposes the optimization of a classification model based on the K-Nearest Neighbor (K-NN) algorithm using 1.598 historical thesis title records from AMIKOM Yogyakarta University. To overcome the sample size disparity among lecturer fields, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented during the data training phase. The text information extraction process was carried out through comprehensive preprocessing stages, including case folding, stopword removal, and stemming, followed by word weighting using Term Frequency-Inverse Document Frequency (TF-IDF). The dataset was divided into training and testing sets with an 80:20 ratio. Unlike standard spatial approaches, this model was evaluated using the Cosine Similarity distance metric to measure the semantic closeness between documents more precisely. Experimental result indicate that the K-NN classification achieved maximum performance at parameter k=21 with an accuracy rate of 85.94%. The integration of SMOTE and the Cosine metric proved to significantly increase the model's sensitivity, marked by a surge in the recall value up to 95% in recognizing minority lecturer field categories that were previously difficult to classify.*

Keywords: *Imbalanced Data, Thesis Title, Lecturer's Field of Expertise, Text Classification, K-NN*

1. PENDAHULUAN

Penyelesaian tugas akhir atau skripsi merupakan instrument wajib bagi mahasiswa strata satu (S1) sebagai syarat kelulusan. Dalam proses penyusunannya, peran dosen pembimbing sangat krusial untuk mengarahkan dan menjaga kualitas substansi penelitian. Namun, problematika teknis yang sering terjadi di tingkat program studi adalah proses pemetaan (*plotting*) judul skripsi mahasiswa kepada dosen pembimbing yang masih dilakukan secara manual. Mahasiswa kerap mengalami kebingungan dalam menentukan topik yang relevan dengan spesialisasi dosen, yang berujung pada penumpukan bimbingan pada dosen-dosen tertentu dan ketidaksesuaian (*mismatch*) antara bidang keahlian dosen dengan arah riset mahasiswa. Sistem otomatisasi klasifikasi topik skripsi berbasis kecerdasan buatan menjadi solusi mendesak untuk menyelesaikan inefisiensi administratif ini.

Penerapan *Machine Learning* dalam klasifikasi teks telah banyak digunakan untuk memetakan dokumen ke dalam kategori spesifik secara otomatis (Kowsari et al., 2019). Algoritma *K-Nearest Neighbor* (K-NN) merupakan salah satu pendekatan yang popular dalam penambangan

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 8 Tahun 2026
ISSN 3025-0919 (media online)
Hal 2007-2011

teks karena arsitekturnya yang tangguh terhadap data yang dinamis. K-NN bekerja dengan mengklasifikasikan teks baru berdasarkan kedekatannya dengan dokumen di dalam himpunan data latih. Kinerja algoritma ini sangat bergantung pada skema representasi vektor dokumen, seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) (Qaiser et al., 2018) serta metrik perhitungan jarak yang digunakan. Penelitian sebelumnya pada dataset judul skripsi yang sama membuktikan bahwa K-NN konvensional (menggunakan metrik spasial *Euclidean*) hanya mampu mencapai akurasi sebesar 85.14% masih tertinggal dari algoritma klasifikasi lainnya (Duhita et al., 2022).

Salah satu faktor utama penyebab rendahnya performa K-NN pada klasifikasi data akademik adalah distribusi kelas yang sangat timpang (*imbalanced data*). Dalam pembelajaran mesin, data yang tidak seimbang seringkali menghasilkan model klasifikasi yang bias, dimana algoritma cenderung memprediksi kelas mayoritas secara presisi namun gagal mengenali kelas minoritas (Thabtah et al., 2020). Analisis pada data historis judul skripsi memperlihatkan bahwa tren topik penelitian cenderung menumpuk pada bidang "Multimedia", sementara bidang "Game" memiliki jumlah sampe yang sangat terbatas. Karena algoritma K-NN beroperasi berdasarkan pemungutan suara mayoritas (*majority vote*), model menjadi rentan terdistorsi oleh kelas dominan sehingga memicu kesalahan klasifikasi pada judul-judul berlabel minoritas.

Untuk mengatasi celah tersebut, penelitian ini mengusulkan optimasi komprehensif pada algoritma K-NN. Pertama, ketimpangan data ditangani menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) pada fase pelatihan. SMOTE bekerja dengan mensitesis sampel data buatan berdasarkan interpolasi fitur dari data minoritas, yang terbukti secara komprehensif mampu meningkatkan batas keputusan (*decision boundary*) pengklasifikasi tanpa menyebabkan *overfitting* (Fern´andez et al., 2018). Kedua, kualitas ekstraksi teks dioptimalkan melalui *stemming* berstandar leksikon bahasa Indonesia untuk menyeragamkan akar kata. Ketiga, metrik jarak *Euclidean* diganti dengan *Cosine Similarity* yang terbukti lebih superior dalam mengukur kedekatan semantic antar dokumen teks berdasarkan sudut vektor, terlepas dari perbedaan panjang teks (Gunawan et al., 2018).

Penelitian ini bertujuan mengevaluasi effektivitas integrasi SMOTE dan *Cosine Similarity* pada algoritma K-NN. Optimalisasi ini diharapkan mampu mendongkrak tingkat akurasi global dan memperbaiki nilai sensitivitas (*recall*) pada kelas minoritas, sehingga sistem rekomendasi dapat memetakan keahlian dosen secara berimbang.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan membandingkan hasil performa algoritma sebelum dan sesudah optimasi. Berdasarkan Gambar 1 Alur Penelitian ini dirancang secara sekuensial mulai dari pengumpulan data mentah hingga evaluasi model akhir.

Gambar 1 Alur Penelitian

Penelitian ini menggunakan dataset sekunder berupa 1.598 data historis judul skripsi dari Universitas AMIKOM Yogyakarta yang terbagi menjadi lima kelas bidang dosen, yaitu *Multimedia, Networking, Software Engineering, Data Science,* dan *Game*. Data tersebut dibersihkan dari *noise* untuk menghasilkan token kata yang berkualitas. Kemudian data teks dikonversi menjadi representasi vektor numerik. Data dibagi menjadi himpunan data latih (*training data*) sebesar 80% dan data uji (*testing data*) sebesar 20%. Kemudian dilakukan *oversampling* secara eksklusif hanya pada himpunan data latih untuk menyeimbangkan kelas minoritas (Game dan Data Science) dengan kelas mayoritas (Multimedia). Setelah itu dilatih menggunakan model K-NN menggunakan *Cosine Similarity* dengan melakukan iterasi nilai k (1 hingga 29) untuk mencari titik akurasi tertinggi.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 8 Tahun 2026
ISSN 3025-0919 (media online)
Hal 2007-2011

Evaluasi untuk mengukur performa model menggunakan *Confusion Matrix* yang menghasilkan nilai *Accuracy, Precision, Recall,* dan *F1-Score.*

2.1 Pra-pemrosesan Teks dan Ekstraksi Fitur

Data judul skripsi dibersihkan melalui empat tahapan utama. Pertama, *case folding* untuk mengubah seluruh karakter menjadi huruf kecil. Kedua, *punctuation removal* untuk menghilangkan angka dan tanda baca. Ketiga. *stopword removal* menggunakan *library* Sastrawi untuk membuang kata hubung atau kata tidak bermakna (seperti "dan", "yang", "di"). Keempat, *stemming* berstandar Sastrawi diterapkan untuk mengembalikan kata berimbuhan menjadi kata dasar (Safar et al., 2024). Teks bersih tersebut kemudian dibobotkan menggunakan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk menilai tingkat kepentingan suatu kata terhadap keseluruhan dokumen (Al Tawil et al., 2024; Ankisqiantari, 2025; Yusupov et al., 2024).

2.2 Penerapan SMOTE dan *K-Nearest Neighbor*

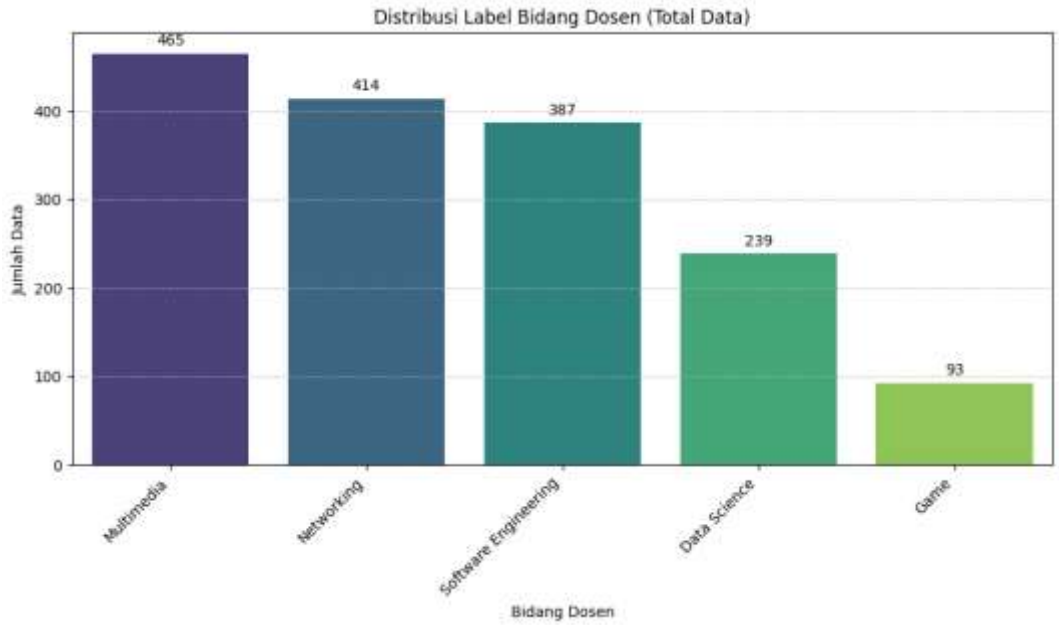
Ketidakseimbangan pada dataset diatasi menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Berbeda dengan duplikasi acak, SMOTE menciptakan data sintesis baru baru di sepanjang garis yang menghubungkan sampel minoritas dengan tetengga terdekatnya (Idwan et al., 2025). Proses ini secara ketat hanya diaplikasikan pada data latih setelah proses *splitting* 80:20 dilakukan, guna mencegah kebocoran data (*data leakage*) ke dalam fase pengujian.

Model klasifikasi dibangun menggunakan K-NN. Algoritma ini mencari sejumlah k dokumen terdekat dari data uji di dalam ruang vektor (Isnain et al., 2021; Saputra et al., 2024). Untuk mengatasi kelemahan jarak *Euclidean* pada dimensi teks yang besar, pengukuran kedekatan diganti menggunakan *Cosine Similarity* yang menghitung sudut kosinus antar vektor dokumen (Gunawan et al., 2018). Eksperimen pencarian (*hyperparameter tuning*) dilakukan pada nilai k ganjil untuk menghindari hasil seri (*tie*) dalam pemungutan sura kelas.

3. ANALISA DAN PEMBAHASAN

Penelitian ini mengevaluasi efektivitas penggunaan algoritma *K-Nearest Neighbor* (K-NN) yang dioptimasi menggunakan metrik *Cosine Similarity* dan teknik penyeimbangan data SMOTE. Evaluasi difokuskan pada peningkatan akurasi global dan kepekaan model dalam mengklasifikasikan kelas minoritas pada dataset judul skripsi.

3.1 Hasil Penyeimbangan Data dengan SMOTE



JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 8 Tahun 2026
ISSN 3025-0919 (media online)
Hal 2007-2011

Gambar 2 Distribusi Label BIdang Dosen

Eksplorasi awal pada dataset menunjukkan adanya ketidakseimbangan kelas (*imbalanced data*) yang ekstrem. Berdasarkan Gambar 2 kelas "Multimedia" mendominasi populasi dengan 465 sampel, sedangkan kelas "Game" bertindak sebagai minoritas mutlak dengan hanya 93 sampel. Ketimpangan ini berisiko membuat algoritma K-NN menjadi bias karena mengandalkan probabilitas jarak terdekat (*majority vote*). Penerapan SMOTE yang dieksekusi secara ekslusif pada himpunan data latih (setelah rasio *split* 80:20) berhasil menyintesis data buatan untuk kelas-kelas minoritas. Hasilnya, seluruh kelas pada fase pelatihan memiliki jumlah distribusi sampel latih yang seimbang, sehingga model K-NN dapat mengenali pola fitur kata tanpa terdistorsi oleh dominasi kelas tertentu.

3.2 Pengujian Akurasi dan Pencarian Nilai K

Pengujian (*hyperparameter tuning*) dilakukan dengan membandingkan nilai K ganjil dengan rentang 1 hingga 29 menggunakan perhitungan metrik *Cosine Similarity*. Penggunaan metrik ini bertujuan untuk mengukur sudut vektor TF-IDF antar dokumen, yang secara komputasi lebih relevan untuk analisis teks dibandingkan metrik *Euclidean.*

Hasil eksperimen menunjukkan bahwa akurasi algoritma K-NN bersifat pada nilai K yang kecil, namun mulai stabil pada K>15. Performa puncak atau akurasi tertinggi dicapai pada nilai K=21 dengan tingkat akurasi sebesar 85,94%. Hasil ini membuktikan adanya perbaikan performa dibandingkan penelitian sebelumnya (tanpa SMOTE dan *Cosine Similarity*) yang hanya mencapai akurasi maksimal 85,14% pada K=18.

3.3 Analisis Kinerja Klasifikasi menggunakan *Confusion Matrix*

Akurasi global seringkali menyembunyikan kelemahan model pada pengenalan kelas minoritas. Oleh karena itu, performa akhir pada nilai K=21 dibedah menggunakan *Classification Report* pada himpunan data uji (*data testing* yang tidak tercemar oleh sintesis SMOTE).

Tabel 1. Evaluasi Model

Bidang Dosen	Precision	Recall	F1-Score
Data Science	75%	87%	81%
Game	69%	95%	80%
Multimedia	96%	92%	94%
Networking	92%	93%	92%
Software Engineering	82%	68%	74%
Rata-Rata Akurasi		**86%**	

Berdasarkan Tabel 1 evaluasi diatas, integrasi SMOTE memberikan dampak yang sangat massif terhadap sensitivitas model. Kelas "Game", yang pada distribusi aslinya sangat rentan disalahpahami oleh model konvensional, berhasil mencatatkan nilai *Recall* sebesar 95%. Ini mengindikasikan bahwa sistem kini sangat sensitive dalam mendeteksi dan memetakan fitur *keyword* spesifik terkait pengembangan *game*, alih-alih mengklasifikasikannya secara acak ke kelas mayoritas.

Disisi lain, penerapan *oversampling* terbukti tidak merusak pola pengenalan pada kelas mayoritas. Kelas "Multimedia" tetap mempertahankan performa prediktif yang sangat superior dengan *precision* mencapai 96% dan *F1-Score* 94%. Secara keseluruhan, keseimbangan performa antarkelas berhasil dicapai, dibuktikan dengan nilai rerata harmonis (*macro average* dan *weighted average*) dari *F1-Score* yang stabil di angka 84% hingga 86%.

4. KESIMPULAN

Penelitian ini membuktikan bahwa optimasi algoritma *K-Nearest Neighbor* (K-NN) menggunakan metrik *Cosine Similarity* dan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) berhasil meningkatkan kualitas klasifikasi teks judul skripsi. Kinerja maksimal model tercapai pada parameter K=21 dengan tingkat akurasi global sebesar 85,94%. Penggantian metrik spasial standar menjadi *Cosine Similarity* terbukti jauh lebih presisi dalam mengukur tingkat kedekatan semantic antarvektor teks hasil pembobotan TF-IDF.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 8 Tahun 2026
ISSN 3025-0919 (media online)
Hal 2007-2011

Keberhasilan paling krusial dari pemodelan ini terletak pada penanganan masalah ketidakseimbangan kelas (*imbalanced data*). Implementasi SMOTE pada fase pelatihan berhasil mendongkrak sensitivitas algoritma terhadap kelas minoritas secara dramatis tanpa merusak akurasi kelas dominan. Hal ini tervalidasi dari kenaikan nilai *Recall* pada kelas "Game" yang mampu menyentuh angka 95%. Angka tersebut memastikan bahwa sistem K-NN yang dibangun tidak lagi melakukan pemungutan suara yang bias dan buta ke arah kelas mayoritas.

REFERENCES

Al Tawil, A., Almazaydeh, L., Qawasmeh, D. S., Qawasmeh, B., Alshinwan, M., & Elleithy, K. (2024). Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT. *Comput. Mater. Continua*, *81*(2), 3395–3412.

Ankisqiantari, N. (2025). *Penanganan Imbalance Data Klasifikasi Teks Lowongan Pekerjaan : Komparasi Algorithmic vs Data Level*. *14*(2), 73–84.

Duhita, W. M. P., Darmawan, M. F. K., Triana, L., & Ankisqiantari, N. (2022). *Perbandingan Algoritma Supervised Learning untuk Klasifikasi Judul Skripsi Berdasarkan Bidang Dosen Comparison of Supervised Learning Algorithm for Classification of Thesis Titles Based on Lecturer*. *8*, 427–437.

Fern´andez, A., Garc´ıa, S., Herrera, F., & Chawla, N. V. (2018). *SMOTE for Learning from Imbalanced Data_ Progress and Challenges, Marking the 15-year Anniversary.pdf* (pp. 863--905).

Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*, *978*, 012120.

Idwan, S., Etaiwi, W., Rafayia, H., & Matar, I. (2025). A comprehensive review of statistical variants and enhancements of SMOTE oversampling method. *International Journal of Data Science and Analytics*, *20*(8), 6887–6904.

Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) algorithm for public sentiment analysis of online learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(2), 121–130.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). *Text Classification Algorithms : A Survey*. 1–68. https://doi.org/10.3390/info10040150

Qaiser, Shahzad, & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*, 25–29. https://doi.org/10.5120/ijca2018917395

Safar, N. Z. M., Mujilahwati, S., Halif, K. M. N. K., & Ghazalli, N. (2024). Optimizing sentiment analysis of Indonesian texts: Enhancing deep learning models with genetic algorithm-based feature selection. *Journal of Soft Computing and Data Mining*, *5*(2), 208–222.

Saputra, N. A., Aeni, K., & Saraswati, N. M. (2024). Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF-ICS$ρ$F. *Vol*, *11*, 21–30.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). *Data imbalance in classification : Experimental evaluation*. *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Yusupov, K., Islam, M. R., Muminov, I., Sahlabadi, M., & Yim, K. (2024). Comparative analysis of machine learning and deep learning models for email spam classification using TF-IDF and word embedding techniques. *International Conference on Broadband and Wireless Computing, Communication and Applications*, 114–122.