Journal of Electrical Engineering and Computer (JEECOM)
Vol. xx, No. xx (20xx), DOI: 10.33650/jeecom.v4i2
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Review Komprehensif Penggunaan Data Imbalance Pada Metode Klasifikasi Dalam Machine Learning

Muammar Reza Pahlawan [1], Arief Setyanto [2], M. Rudyanto Arief [3]
[1, 2] Magister Informatika, Universitas AMIKOM Yogyakarta, Sleman, Indonesia
[3] Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Sleman, Indonesia

Article Info

Article history:

Diterima 16 April 2024
Revisi 18 April 2024
Diterbitkan 21 April 2024

Keywords:

Classification Methods
Super Vector Machine (SVM)
K-Nearest Neighbors (KNN)
Random Forest
Imbalance Data

ABSTRAK

Dengan majunya perkembangan teknologi beberapa tahun terakhir, menghadirkan banyak konten digital. Hal ini juga menghadirkan kesempatan dalam bidang penelitian seperti halnya Machine Learning. Salah satu metode dalam Machine Learning adalah klasifikasi. Klasifikasi bertujuan untuk mengelompokkan data sesuai dengan kelasnya. Akan tetapi faktor seperti data imbalance dapat menyebabkan hasil dari metode ini menjadi kurang sesuai dengan yang diharapkan. Penelitian ini menyajikan tinjauan komprehensif tentang metode klasifikasi dalam pengolahan teks, dengan fokus pada penanganan tantangan yang ditimbulkan oleh data yang tidak seimbang. Dengan pertumbuhan eksponensial konten digital, kebutuhan untuk mengkategorikan dan menganalisis data teks secara efektif telah menjadi semakin kritis. Metode klasifikasi memainkan peran penting dalam upaya ini, memfasilitasi tugas seperti analisis sentimen, klasifikasi dokumen, dan pengambilan informasi. Namun, keberadaan *data imbalance*, ditandai oleh distribusi kelas yang condong, menimbulkan hambatan signifikan terhadap keandalan dan efektivitas model klasifikasi. Dengan penelitian ini diharapkan pembaca, dapat mengetahui metode apa saja yang umumnya digunakan dalam metode klasifikasi. Kemampuan metode klasifikasi tersebut pada umumnya ketika dihadapkan pada kasus tertentu seperti data imbalance. Tinjauan ini menyoroti Support Vector Machine (SVM) sebagai metode klasifikasi paling menonjol sebesar 25%, diikuti oleh K-Nearest Neighbours (KNN) dan Random Forest dengan persentase 19%, Decision Tree, dan Naïve Bayes. Metode alternatif yang disesuaikan dengan tujuan penelitian dan tantangan tertentu juga dieksplorasi. Hasil persentase penggunaan metode tersebut didapat dari kumpulan jurnal yang peneliti kumpulkan dan teliti.

Corresponding Author:
Muhammad Reza Pahlawan,
Universitas AMIKOM Yogyakarta, Jl. Ring Road Utara, Sleman and 55281, Indonesia
Email: rezapahlawan@students.amikom.ac.id

1. PENDAHULUAN

Dalam beberapa tahun terakhir, perkembangan teknologi telah menghasilkan ledakan konten digital yang luar biasa, terutama di ranah teks. Fenomena ini memunculkan kebutuhan yang semakin mendesak untuk menganalisis, mengelompokkan, dan mengekstraksi makna dari data teks yang melimpah tersebut. Berbagai aplikasi, mulai dari analisis sentimen hingga pengambilan informasi, telah menjadi fokus utama pengolahan teks, menempatkan metode klasifikasi sebagai pilar utama dalam mendukung tujuan tersebut.

Namun, di tengah kemajuan dan inovasi dalam pengolahan teks, tantangan yang persisten mengintai: data yang memiliki sebaran tidak seimbang. Skenario *data imbalance*, yang ditandai oleh distribusi kelas yang condong di mana beberapa kelas secara signifikan kurang mewakili dibandingkan dengan yang lain, menimbulkan hambatan yang tangguh terhadap efektivitas dan keandalan model klasifikasi. Dalam konteks pengolahan teks, data yang tidak seimbang dapat menyebabkan kinerja model yang bias, kesalahan klasifikasi pada kelas minoritas, dan keterbatasan hasil.

Data imbalance ini sangat mungkin terjadi, apabila sebaran data lebih besar pada salah satu kelas. Data imbalance ini dapat mempengaruhi hasil akhir dari suatu proses Machine Learning. Baik dengan metode klasifikasi ataupun dengan metode lainnya. Sehingga data yang dihasilkan menjadi tidak seperti yang diharapkan ataupun juga dapat menciptakan data yang memiliki keanehan. Diantara masalah yang dapat terjadi adalah masalah distribusi kelas yang tidak seimbang, yang berarti hasil positif jarang terjadi dibandingkan dengan hasil negatif, dan yang menjadi perhatian kami adalah kelas minoritas dibandingkan kelas mayoritas [1].

Metode klasifikasi merupakan salah satu pendekatan penting dalam bidang pembelajaran mesin yang bertujuan untuk mengidentifikasi kategori atau kelas dari sekumpulan data berdasarkan karakteristik atau fitur yang dimiliki. Metode klasifikasi digunakan untuk memprediksi label kelas dari objek atau fenomena berdasarkan informasi yang terdapat dalam dataset. Klasifikasi juga dapat memilah atau mengklasifikasikan data [2]. Pendekatan ini umumnya melibatkan pembelajaran terawasi dimana model pembelajaran mesin dilatih menggunakan dataset yang telah diberi label.

Dataset merupakan kumpulan data yang sistematis, yang biasanya diorganisir dalam format tabelar, di mana baris-baris mewakili sampel individu dan kolom mewakili variabel atau atribut dari sampel tersebut. Dalam konteks penelitian ilmiah, dataset digunakan sebagai bahan dasar untuk analisis statistik, pembelajaran mesin, dan berbagai metode pengujian hipotesis lainnya. Pengumpulan, pemrosesan, dan analisis dataset yang efektif merupakan elemen kunci dalam memastikan integritas dan validitas hasil penelitian.

Dataset yang tidak seimbang atau imbalanced dataset memiliki dampak signifikan terhadap performa metode klasifikasi dalam pembelajaran mesin. Dalam dataset semacam ini, jumlah sampel untuk setiap kelas tidak seimbang; satu atau beberapa kelas mungkin jauh lebih dominan daripada kelas lainnya. Hal ini bisa menyebabkan berbagai masalah dalam pelatihan model klasifikasi, terutama terkait dengan bias dan overfitting terhadap kelas mayoritas. Machine learning sering kali tidak dapat diandalkan ketika diterapkan pada data imbalance [3]. Metode kalsidikasi juga dapat bekerja dengan baik pada kumpulan data imbalance, akan tetapi mungkin tidak berfungsi pada kumpulan data imbalance lainnya [4].

Menangani tantangan yang ditimbulkan oleh *data imbalance* adalah hal yang sangat penting untuk memajukan bidang pengolahan teks dan memanfaatkan potensi penuh dari metode klasifikasi. Meskipun banyak penelitian telah dilakukan tentang teknik klasifikasi dalam pengolahan teks, masih ada kebutuhan untuk sebuah tinjauan komprehensif yang secara khusus mengkaji penggunaan metode klasifikasi dalam konteks data yang tidak seimbang.

Sebagai tanggapan terhadap kesenjangan dalam literatur tersebut, makalah ini menyajikan sebuah tinjauan literatur sistematis yang bertujuan untuk memberikan analisis komprehensif tentang metode klasifikasi dalam pengolahan teks, dengan fokus pada penanganan tantangan data yang tidak seimbang. Dengan mensintesis penelitian yang ada dan mengidentifikasi temuan kunci, studi ini bertujuan untuk memberikan pemahaman tentang metodologi yang umum digunakan, tren yang muncul, dan pendekatan inovatif untuk mengurangi efek data yang tidak seimbang dalam tugas klasifikasi teks.

Melalui pemeriksaan yang teliti terhadap artikel ilmiah yang diterbitkan selama beberapa tahun terakhir, makalah ini berusaha untuk memberikan wawasan tentang evolusi tren penelitian, metodologi yang umum digunakan, dan aplikasi teknik klasifikasi dalam konteks data yang tidak seimbang. Dengan menjelaskan keadaan seni terkini dan mengidentifikasi bidang-bidang untuk penelitian masa depan, studi ini bertujuan untuk memberikan kontribusi pada kemajuan metodologi pengolahan teks dan mendorong pengembangan model klasifikasi yang lebih kokoh dan andal untuk aplikasi dunia nyata.

2. METODE

Suatu studi Tinjauan Literatur Sistematis (SLR) bertujuan untuk mengidentifikasi studi yang relevan, mengumpulkan data yang diperlukan, dan selanjutnya menilai dan menggabungkan temuan untuk mendapatkan wawasan yang lebih dalam tentang subjek penelitian [5]. Terlepas dari topik spesifik, fokus disiplin, atau pendekatan teoritis, sebuah SLR mengikuti proses terstruktur yang terdiri dari enam komponen yang berbeda dan penting, diuraikan sebagai berikut.

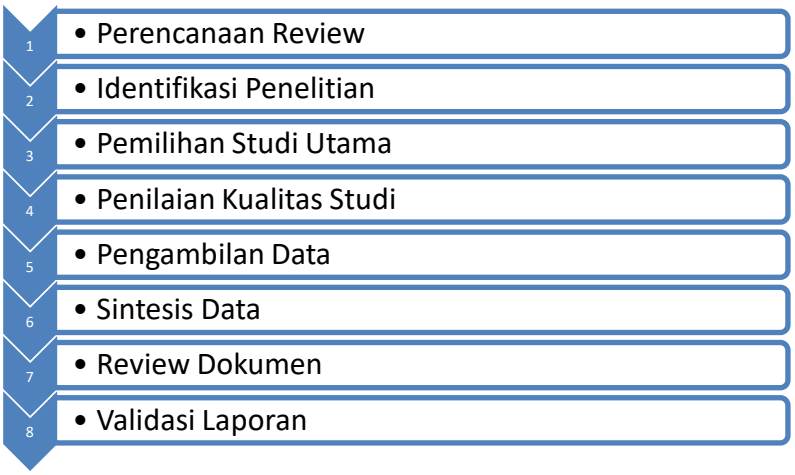


Diagram 1 Skema Metode Penelitian

2.1. Pertanyaan Penelitian

Untuk melakukan penelitian dengan menggunakan teknik tinjauan literatur sistematis (SLR), diperlukan untuk merumuskan serangkaian pertanyaan penelitian (RQs) [6]. Pertanyaan-pertanyaan yang disajikan dalam Tabel 1 memainkan peran penting dalam membentuk kerangka kerja yang jelas, terarah, dan efektif untuk upaya penelitian. Pendekatan yang teliti ini bertujuan untuk meningkatkan dan menyempurnakan proses penelitian, memfasilitasi fokus dan efisiensi yang lebih besar.

Tabel 1. Rumusan Masalah (*Research Question*)

TABLE I. RUMUSAN MASALAH

ID	Rumusan Masalah
RQ1	Apa metode yang digunakan peneliti untuk mengumpulkan data terkait klasifikasi?
RQ2	Apa saja metode yang digunakan dalam bidang klasifikasi?
RQ3	Apa temuan yang muncul dari penyelidikan tentang klasifikasi dalam konteks penelitian?

2.2. Strategi Penelitian

Peneliti melakukan pencarian intensif untuk makalah-makalah melalui basis data terkemuka seperti ScienceDirect, IEEE, Springer, Semantic Scholar, Google Scholar, dan Elsevier. Penelitian ini dipandu oleh beberapa kata kunci, yang mencakup terminologi dalam bahasa Indonesia dan bahasa Inggris, untuk memastikan pengambilan materi yang relevan secara komprehensif dan inklusif sebagai berikut

- “Klasifikasi”
- “Machine Learning”
- “Data Imbalance”
- “Decision Tree”
- “KNN” dan “K Nearest Neighbor”
- “SVM” dan “Support Vector Machine”
- “Random Forest”
- “Naïve Bayes”

2.3. Pemilihan Studi

Menentukan kriteria adalah hal yang penting untuk mengevaluasi naskah. Peneliti menggunakan dua jenis kriteria yang berbeda yang relevan untuk pemilihan naskah: kriteria inklusi dan kriteria eksklusi. Berikut adalah kriteria inklusi spesifik yang diimplementasikan dalam kerangka kerja penelitian ini:

- Penelitian yang di terbitkan dari 2019 sampai 2024.
- Penelitian diterbitkan menggunakan Bahasa Indonesia ataupun Inggris.

- Topik utama dalam penelitian adalah klasifikasi.

Sedangkan kriteria eksklusi dalam penelitian ini adalah sebagai berikut:

- Penelitian yang tidak termasuk dalam metode inklusi.
- Penelitian tidak secara jelas menunjukkan metode yang digunakan.
- Penelitian tidak berhasil memenuhi tujuan penelitian.

2.4. Penilaian Kualitas

Untuk memperoleh pemahaman menyeluruh tentang kualitas keseluruhan studi, evaluasi kualitas yang teliti sangat penting. Proses evaluasi ini krusial untuk memastikan relevansi dan kesesuaian data yang diidentifikasi untuk dimasukkan dalam penelitian. Dalam lingkup penelitian ini, data yang terkumpul akan mengalami evaluasi yang ketat yang dipandu oleh serangkaian kriteria yang telah ditentukan sebelumnya yang bertujuan untuk menilai kualitasnya. Penggunaan kriteria evaluasi kualitas ini memastikan tinjauan yang sistematis dan tidak memihak, sehingga meningkatkan kekokohan dan keandalan studi.

Penelitian yang dipilih dan digunakan dalam penelitian ini harus memenuhi kriteria sebagai berikut:

- Apakah penelitian diterbitkan antara tahun 2019 sampai 2024?
- Apakah ditulis menggunakan bahasa Indonesia atau Inggris?
- Apakah klasifikasi adalah topik utama yang diteliti?

2.5. Pengumpulan Data

Pada tahap ini, data yang diekstraksi dari makalah yang diteliti menyoroti aspek-aspek penting, termasuk tahun publikasi penelitian, dataset yang digunakan dalam studi yang sedang ditinjau, metodologi yang digunakan untuk pengumpulan data, pendekatan khusus yang diadopsi untuk klasifikasi dalam penelitian yang diteliti, dan implikasi hasil klasifikasi terhadap studi tersebut. Selanjutnya, semua data yang relevan dimasukkan dengan hati-hati ke dalam dokumen spreadsheet. Dokumentasi yang teliti ini menjadi dasar untuk analisis menyeluruh dari data yang terkumpul. Organisasi informasi yang sederhana ini memfasilitasi studi yang efisien dan sistematis yang sesuai dengan praktik penelitian ilmiah yang telah mapan.

2.6. Penyusunan Data

Pada tahap proses penelitian ini, para peneliti telah mengumpulkan sejumlah 100 penelitian, dengan teliti memeriksa judul dan abstrak untuk relevansi. Proses penyaringan awal ini menyempitkan pilihan kami menjadi 59 penelitian. Namun, evaluasi kami tidak berakhir di sana. Dengan menggunakan kriteria inklusi dan eksklusi yang ketat, kami dengan hati-hati memilih 59 penelitian yang sangat sejalan dengan tujuan penelitian kami. Makalah yang memenuhi kriteria inklusi kami dimasukkan ke dalam tinjauan literatur kami, sementara yang memenuhi kriteria eksklusi dikecualikan. Proses pemilihan yang teliti ini berakhir dengan satu set akhir 44 penelitian, yang kemudian mengalami tinjauan dan analisis yang cermat. Data yang terkumpul dan temuan utama dari makalah-makalah ini menjadi subjek pemeriksaan menyeluruh, dan sintesis mereka disajikan secara sistematis dalam Tabel 2. Pendekatan ini sesuai dengan standar penelitian ilmiah, memastikan eksplorasi literatur yang komprehensif dan terorganisir dengan baik.

Tabel 2. Jurnal yang di Review (*Reviewd Paper*)

TABLE II. JURNAL YANG DI REVIEW

Penelitian	Tahun	Bahasa	Metode
[4]	2024	Inggris	KNN
[1]	2023	Inggris	SVM
[7]	2023	Inggris	Logistic Regression, Linear Discriminant Analysis, SVM, Random Forest
[8]	2023	Inggris	KNN
[9]	2023	Inggris	Random Forest
[10]	2023	Inggris	SVM
[11]	2023	Indonesia	Decision Tree

TABLE II. JURNAL YANG DI REVIEW

Penelitian	Tahun	Bahasa	Metode
[12]	2023	Inggris	LLM
[3]	2023	Indonesia	KNN, Naïve Bayes
[13]	2022	Inggris	Random Forest
[14]	2022	Inggris	Decision Tree
[15]	2022	Inggris	SVM
[16]	2022	Inggris	CNN
[17]	2022	Inggris	Random Forest
[18]	2022	Inggris	SVM
[19]	2022	Indonesia	Random Forest
[20]	2022	Indonesia	KNN
[21]	2022	Inggris	KNN-SVM
[22]	2022	Inggris	Chi-Square, Random Forest
[23]	2022	Inggris	KNN, SVM
[24]	2022	Inggris	SVM
[25]	2022	Inggris	Decision Tree
[26]	2022	Inggris	Random Forest
[27]	2022	Inggris	KNN
[28]	2022	Inggris	Naïve Bayes
[29]	2022	Indonesia	Decision Tree
[30]	2021	Inggris	Decision Tree
[31]	2021	Inggris	SVM
[32]	2021	Inggris	SVM
[33]	2021	Inggris	SVM
[34]	2021	Inggris	Naïve Bayes
[35]	2021	Inggris	KNN
[36]	2020	Inggris	CNN
[37]	2020	Inggris	NN
[38]	2020	Inggris	Random Forest
[39]	2020	Inggris	Naïve Bayes
[40]	2020	Inggris	C4.5, SVM, Naïve Bayes
[41]	2020	Inggris	Random Forest
[42]	2019	Inggris	KNN
[43]	2019	Inggris	NN
[44]	2019	Inggris	KNN
[2]	2019	Indonesia	KNN
[45]	2019	Inggris	SVM
[46]	2019	Inggris	SVM
[47]	2019	Inggris	Random Forest

3. HASIL DAN PEMBAHASAN

3.1. Tahun Penerbitan

Sebagai komponen dari analisis, makalah-makalah yang terkumpul akan dikategorikan dan diklasifikasikan berdasarkan tahun publikasinya. Metodologi kronologis ini akan memfasilitasi pemeriksaan yang rinci tentang perkembangan tren penelitian dan kemajuan dalam pengolahan teks, terutama dalam konteks

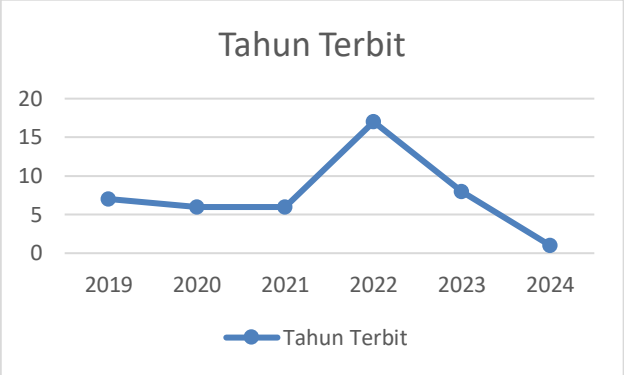
klasifikasi. Tubuh literatur setiap tahun akan menjalani pemeriksaan yang teliti, dengan fokus pada menemukan temuan signifikan dan metodologi yang digunakan.

Trend publikasi selama bertahun-tahun mengungkapkan fluktuasi dalam jumlah makalah yang dipublikasikan setiap tahunnya. Pada tahun 2019, terdapat 7 penelitian yang dipublikasikan, menunjukkan awal yang moderat dalam lanskap penelitian. Tahun berikutnya, 2020, mengalami sedikit penurunan dengan 6 penelitian, menunjukkan kelanjutan yang relatif stabil. Namun, pada tahun 2021, jumlah publikasi tetap konsisten dengan 6 penelitian, menunjukkan datar dalam output penelitian.

Terjadi peningkatan yang signifikan pada tahun 2022, di mana jumlah publikasi melonjak menjadi 17 penelitian, menandai peningkatan aktivitas penelitian yang signifikan. Lonjakan publikasi ini menunjukkan minat dan keterlibatan yang meningkat dalam komunitas penelitian selama periode tersebut. Kemudian, pada tahun 2023, terjadi penurunan ringan dengan 8 penelitian yang dipublikasikan, menandakan penurunan yang ringan dibandingkan tahun sebelumnya.

Pada tahun 2024, terjadi penurunan substansial, dengan hanya 1 penelitian yang dipublikasikan. Penurunan tajam ini mungkin mencerminkan berbagai faktor seperti pergeseran prioritas penelitian, tren yang muncul, atau pengaruh eksternal yang memengaruhi output penelitian.

Secara keseluruhan, data mengilustrasikan tren fluktuatif dalam publikasi makalah selama bertahun-tahun, menunjukkan periode pertumbuhan, stabilitas, dan penurunan dalam lanskap penelitian.



Gambar 1. Tahun Terbit Jurnal

3.2. Metode Klasifikasi

Distribusi metode klasifikasi di antara penelitian-penelitian yang terkumpul memberikan wawasan berharga tentang teknik-teknik yang umum digunakan dalam penelitian pengolahan teks. Di antara metode yang diidentifikasi, Support Vector Machine (SVM) muncul sebagai yang paling sering digunakan, dengan 13 penelitian mengadopsi pendekatan ini. Popularitas SVM menegaskan efektivitasnya dalam menangani berbagai tugas klasifikasi teks, karena kemampuannya untuk menemukan hiperruangan optimal untuk memisahkan titik-titik data ke dalam kelas-kelas yang berbeda.

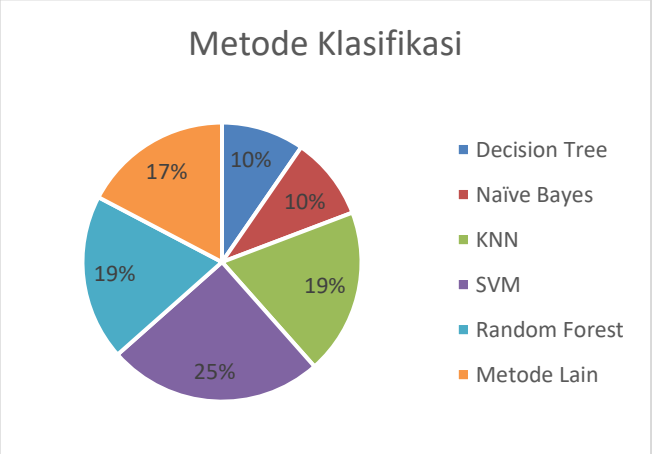
Mengikuti secara dekat di belakang SVM adalah algoritma K-Nearest Neighbors (KNN), yang digunakan dalam 10 penelitian. Kesederhanaan dan sifat intuitif KNN membuatnya menjadi pilihan yang populer, terutama untuk tugas klasifikasi teks di mana titik-titik data menunjukkan hubungan yang kompleks dan non-linear.

Random Forest, sebuah teknik pembelajaran gabungan lainnya, juga menonjol, dengan 10 makalah menggunakan metode ini. Kemampuan Random Forest untuk menangani data berdimensi tinggi dan mengurangi overfitting membuatnya cocok untuk tugas klasifikasi teks dengan beragam set fitur.

Metode Pohon Keputusan dan Naïve Bayes memiliki representasi yang sama, masing-masing muncul dalam 5 penelitian. Pohon keputusan menawarkan interpretabilitas dan kemudahan visualisasi, membuatnya cocok untuk memahami pentingnya fitur dan proses pengambilan keputusan. Naïve Bayes, di sisi lain, dihargai karena kesederhanaan dan efisiensinya, terutama dalam menangani dataset besar dengan ruang fitur berdimensi tinggi.

Selain itu, terdapat 9 penelitian yang mengeksplorasi metode klasifikasi alternatif di luar teknik-teknik tersebut. Ini bisa mencakup pendekatan inovatif yang disesuaikan dengan tantangan pengolahan teks tertentu atau adaptasi dari metode yang sudah ada untuk cocok dengan konteks penelitian yang unik.

Secara keseluruhan, distribusi metode klasifikasi menegaskan keragaman dan adaptabilitas pendekatan yang digunakan dalam penelitian pengolahan teks, mencerminkan pemahaman yang halus tentang kompleksitas yang melekat dalam analisis bahasa dan tugas klasifikasi.



Gambar 2. Metode Klasifikasi

3.3. Kegunaan Klasifikasi

Analisis metode klasifikasi dan penggunaannya di seluruh makalah yang terkumpul mengungkapkan beragam aplikasi dan tujuan dalam ranah pengolahan teks. Teknik klasifikasi berfungsi sebagai alat yang tak tergantikan untuk berbagai tugas dan tujuan, setiap metode disesuaikan untuk mengatasi tantangan dan persyaratan spesifik yang melekat dalam analisis bahasa dan tugas klasifikasi.

Support Vector Machine (SVM), metode klasifikasi yang paling umum diidentifikasi dalam analisis, menemukan aplikasi yang luas di berbagai konteks penelitian. Ketangguhan dan fleksibilitas SVM membuatnya cocok untuk tugas-tugas seperti analisis sentimen, kategorisasi topik, dan klasifikasi dokumen. Kemampuannya untuk membedakan pola yang rumit dan hubungan non-linear dalam data teks memungkinkan peneliti untuk mengekstrak wawasan yang bermakna dan membuat keputusan yang terinformasi.

Algoritma K-Nearest Neighbors (KNN), sebuah metode yang juga banyak digunakan, sangat dihargai karena kesederhanaannya dan efektivitasnya dalam menangani tugas klasifikasi teks. Para peneliti memanfaatkan KNN untuk tugas-tugas seperti pengelompokan dokumen, kategorisasi teks, dan pengambilan informasi, di mana analisis berbasis kedekatan dan pengukuran kesamaan memainkan peran penting dalam mengidentifikasi pola dan kelompok dalam data teks.

Metode Random Forest, Decision Tree, dan Naïve Bayes juga menonjol dalam lanskap klasifikasi, masing-masing melayani tujuan dan aplikasi yang berbeda. Pohon keputusan sangat baik dalam menyediakan model yang dapat diinterpretasi dan mengidentifikasi kepentingan fitur, menjadikannya berharga untuk tugas-tugas seperti pemilihan fitur dan klasifikasi teks dalam domain di mana interpretabilitas sangat penting. Naïve Bayes, yang dikenal karena kesederhanaan dan efisiensinya dalam komputasi, menemukan aplikasi dalam berbagai tugas klasifikasi teks, termasuk penyaringan spam, analisis sentimen, dan kategorisasi dokumen.

Selain itu, eksplorasi metode klasifikasi alternatif menyoroti pendekatan inovatif dan adaptasi yang disesuaikan dengan tujuan dan tantangan penelitian tertentu. Metode-metode ini sering kali menangani kebutuhan atau kendala yang unik dalam domain khusus, seperti pemrosesan bahasa regional, identifikasi dialek, atau analisis teks berbasis domain.

Secara keseluruhan, teknik klasifikasi adalah alat yang tak tergantikan dalam penelitian pengolahan teks, memfasilitasi berbagai tugas mulai dari analisis sentimen dan kategorisasi dokumen hingga pengambilan informasi dan pemodelan topik. Beragamnya metode yang tersedia menekankan sifat dinamis dari analisis bahasa dan klasifikasi, dengan para peneliti terus-menerus menjelajahi pendekatan dan teknik baru untuk mengekstrak wawasan yang bermakna dari data teks.

4. KESIMPULAN

Tinjauan literatur sistematis yang disajikan dalam makalah ini memberikan analisis komprehensif tentang metode klasifikasi yang digunakan dalam penelitian pengolahan teks, terutama dalam konteks analisis bahasa dan tugas klasifikasi. Melalui pemeriksaan teliti terhadap 100 penelitian yang diterbitkan selama beberapa tahun, studi ini memberikan wawasan berharga tentang evolusi tren penelitian, metodologi yang umum digunakan, dan aplikasi teknik klasifikasi.

Analisis mengungkapkan beragamnya metode klasifikasi yang digunakan di berbagai konteks penelitian, dengan Support Vector Machine (SVM) muncul sebagai pilihan paling menonjol di antara para peneliti dengan persentase sebesar 25%. Ketangguhan dan fleksibilitas SVM membuatnya cocok untuk berbagai tugas klasifikasi teks, termasuk analisis sentimen, kategorisasi topik, dan klasifikasi dokumen. Selain itu, studi ini menyoroti penggunaan metode klasifikasi lain seperti K-Nearest Neighbors (KNN), Random Forest, Decision Tree, dan Naïve Bayes, masing-masing metode tersebut juga menjadi metode yang cukup sering digunakan dan peneliti temui dalam proses tinjauan ini. Random forest dan K-Nearest Neighbors memiliki persentase penggunaan yang sama dengan angka penggunaan sebesar 19%. Mulai dari pengelompokan dokumen hingga pemilihan fitur dan dari penyaringan spam hingga analisis sentimen, metode-metode ini memainkan peran penting dalam mengekstrak wawasan yang bermakna dari data teks. Selain itu, pemeriksaan terhadap metode klasifikasi alternatif menegaskan pendekatan inovatif dan adaptasi yang disesuaikan dengan tujuan dan tantangan penelitian yang spesifik.

Demikian dengan penelitian ini pembaca dapat mengetahui umum nya metode klasifikasi apa yang digunakan dalam penelitian beserta dengan persentase nya dalam penelitian ini. Sehingga dapat dijadikan sebagai acuan untuk metode yang dapat digunakan dlama penelitian berikut nya khususnya penelitian yang menggunakan data imbalance dalam datasetnya.

UCAPAN TERIMA KASIH

Atas nama para penulis, kami ingin mengucapkan terima kasih kepada pembimbing kami, Dr. Arief Setyanto, S.Si., M.T., dan M. Rudyanto Arief, S.Kom, M.T., atas bimbingan mereka selama penelitian ini, memberikan saran, pendapat, dan memotivasi kami selama proses studi ini.

REFERENSI

[1] C.-A. Tsai and Y.-J. Chang, "Efficient Selection of Gaussian Kernel SVM Parameters for Imbalanced Data," *Genes*, vol. 14, no. 3, p. 583, Feb. 2023, doi: 10.3390/genes14030583.

[2] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *RESTI*, vol. 3, no. 2, pp. 196–201, Aug. 2019, doi: 10.29207/resti.v3i2.945.

[3] I. Binanto, N. F. Sianipar, F. Dea, M. N. Primadani, and T. W. Kartikasari, "KLASIFIKASI SENYAWA KELADI TIKUS MENGGUNAKAN ALGORITMA KNN, GAUSSIAN NAÏVE BAYES DENGAN MENERAPKAN IMBALANCE DATA BORDERLINE SMOTE," *SNST*, vol. 13, no. 1, p. 377, Nov. 2023, doi: 10.36499/psnst.v13i1.9005.

[4] A. Fazli and J. Poshtan, "Wind turbine fault detection and isolation robust against data imbalance using KNN," *Energy Science & Engineering*, vol. 12, no. 3, pp. 1174–1186, Mar. 2024, doi: 10.1002/ese3.1706.

[5] M. Wahyu Ade Saputra, E. Utami, and A. Yaqin, "Unlocking Insights: A Literature Review on Enhanced Confix Stripping and Nazief & Adriani Algorithm Modifications for Makassar Language Text Stemming," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 603–610, Mar. 2024, doi: 10.38124/ijisrt/IJISRT24MAR437.

[6] Prema Adhitya Dharma Kusumah, Kusrini Kusrini, and Kusnawi Kusnawi, "Optimizing Data Security: A Literature Review on the Implementation of Beaufort Cipher for Vigenère Affine Cipher," Feb. 2024, doi: 10.5281/ZENODO.10685974.

[7] P. Thölke *et al.*, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *NeuroImage*, vol. 277, p. 120253, Aug. 2023, doi: 10.1016/j.neuroimage.2023.120253.

[8] Sukamto, Hadiyanto, and Kurnianingsih, "KNN Optimization Using Grid Search Algorithm for Preeclampsia Imbalance Class," *E3S Web Conf.*, vol. 448, p. 02057, 2023, doi: 10.1051/e3sconf/202344802057.

[9] L. Ren, A. S. Seklouli, H. Zhang, T. Wang, and A. Bouras, "An adaptive Laplacian weight random forest imputation for imbalance and mixed-type data," *Information Systems*, vol. 111, p. 102122, Jan. 2023, doi: 10.1016/j.is.2022.102122.

[10] Mohd. Mustaqeem and T. Siddiqui, "A Hybrid Software Defects Prediction Model for Imbalance Datasets Using Machine Learning Techniques: (S-SVM Model)," *J Autonom Intell*, vol. 6, no. 1, p. 559, Jun. 2023, doi: 10.32629/jai.v6i1.559.

[11] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and E. Fitriani, "Penerapan Teknik Random Undersampling untuk Mengatasi Imbalance Class dalam Prediksi Kebakaran Hutan Menggunakan Algoritma Decision Tree," *AJCSR*, vol. 5, no. 1, p. 1, Jan. 2023, doi: 10.38101/ajcsr.v5i1.617.

[12] X. Cai, M. Xiao, Z. Ning, and Y. Zhou, "Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-based Data Augmentation," in *2023 IEEE International Conference on Data Mining (ICDM)*, Shanghai, China: IEEE, Dec. 2023, pp. 956–961. doi: 10.1109/ICDM58522.2023.00107.

[13] A. Z. Zakaria, A. Selamat, L. K. Cheng, and O. Krejcar, "Improving Class Imbalance Detection And Classification Performance: A New Potential of Combination Resample and Random Forest," in *2022 IEEE International Conference on Computing (ICOCO)*, Kota Kinabalu, Malaysia: IEEE, Nov. 2022, pp. 316–323. doi: 10.1109/ICOCO56118.2022.10031922.

[14] I. A. E. Zaeni, W. Primadi, M. K. Osman, D. R. Anzani, D. Lestari, and A. N. Handayani, "Detection of the Imbalance Step Length using the Decision Tree," in *2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia: IEEE, Sep. 2022, pp. 157–162. doi: 10.1109/ICVEE57061.2022.9930456.

[15] M. Yan, J. Wang, D. Li, and J. Meng, "An Improved Imbalanced Data Classification Algorithm Based on SVM," in *2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, Nanjing, China: IEEE, Nov. 2022, pp. 454–459. doi: 10.1109/ICCSI55536.2022.9970637.

[16] Z. Xing, R. Zhao, Y. Wu, and T. He, "Intelligent fault diagnosis of rolling bearing based on novel CNN model considering data imbalance," *Appl Intell*, vol. 52, no. 14, pp. 16281–16293, Nov. 2022, doi: 10.1007/s10489-022-03196-x.

[17] H. Suryono, H. Kuswanto, and N. Iriawan, "Rice phenology classification based on random forest algorithm for data imbalance using Google Earth engine," *Procedia Computer Science*, vol. 197, pp. 668–676, 2022, doi: 10.1016/j.procs.2021.12.201.

[18] Md. A. Sahid, M. Hasan, N. Akter, and Md. M. R. Tareq, "Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction using Machine Learning and Deep Learning," in *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India: IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/TENSYMP54529.2022.9864473.

[19] D. Mualfah, W. Fadila, and R. Firdaus, "Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest," *CoSciTech*, vol. 3, no. 2, pp. 107–113, Aug. 2022, doi: 10.37859/coscitech.v3i2.3912.

[20] S. Maula Chamzah, M. Lestandy, N. Kasan, and A. Nugraha, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) untuk Imbalance Class pada Data Text Menggunakan kNN," *Syntax J. Inf.*, vol. 11, no. 02, pp. 56–67, Nov. 2022, doi: 10.35706/syji.v11i02.6940.

[21] B. Ma, "The Impact of Environmental Pollution on Residents' Income Caused by the Imbalance of Regional Economic Development Based on Artificial Intelligence," *Sustainability*, vol. 15, no. 1, p. 637, Dec. 2022, doi: 10.3390/su15010637.

[22] K. Kurniabudi, A. Harris, V. Veronica, and E. Yanti, "Optimizing Attack Detection for High Dimensionality and Imbalanced Data with SMOTE, Chi-Square and Random Forest Classifier," *ijics*, vol. 6, no. 1, p. 1, Mar. 2022, doi: 10.30865/ijics.v6i1.3890.

[23] M. A. Ganaie and M. Tanveer, "KNN weighted reduced universum twin SVM for class imbalance learning," *Knowledge-Based Systems*, vol. 245, p. 108578, Jun. 2022, doi: 10.1016/j.knosys.2022.108578.

[24] C. Fu, S. Zhou, D. Zhang, and L. Chen, "Relative Density-Based Intuitionistic Fuzzy SVM for Class Imbalance Learning," *Entropy*, vol. 25, no. 1, p. 34, Dec. 2022, doi: 10.3390/e25010034.

[25] F. Budiman, I. A. Saputro, P. Purwanto, and P. N. Andono, "Optimization Of Classification Results By Minimizing Class Imbalance On Decision Tree Algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jakarta, Indonesia: IEEE, Jan. 2022, pp. 6–11. doi: 10.1109/ISMODE53584.2022.9743062.

[26] U. Bradter, J. D. Altringham, W. E. Kunin, T. J. Thom, J. O'Connell, and T. G. Benton, "Variable ranking and selection with random forest for unbalanced data," *Environ. Data Science*, vol. 1, p. e30, 2022, doi: 10.1017/eds.2022.34.

[27] S. A. Bahanshal, R. S. Baraka, B. Kim, and V. Verdhan, "An Optimized Hybrid Fuzzy Weighted k-Nearest Neighbor with the Presence of Data Imbalance," *IJACSA*, vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130476.

[28] M. Badar, M. Fisichella, V. Iosifidis, and W. Nejdl, "Discrimination and Class Imbalance Aware Online Naive Bayes," 2022, doi: 10.48550/ARXIV.2211.04812.

[29] N. Yudistira, A. F. Putra, Ahmad Saifuddin, and Noverio Athariq Syafaz, "Algoritma Decision Tree Dan Smote Untuk Klasifikasi Serangan Jantung Miokarditis Yang Imbalance," *JLE*, vol. 2, no. 2, pp. 112–122, Dec. 2021, doi: 10.51402/jle.v2i2.48.

[30] J. Yan, Z. Zhang, and H. Dong, "AdaDT: An adaptive decision tree for addressing local class imbalance based on multiple split criteria," *Appl Intell*, vol. 51, no. 7, pp. 4744–4761, Jul. 2021, doi: 10.1007/s10489-020-02061-z.

[31] J.-B. Wang, C.-A. Zou, and G.-H. Fu, "AWSMOTE: An SVM-Based Adaptive Weighted SMOTE for Class-Imbalance Learning," *Scientific Programming*, vol. 2021, pp. 1–18, May 2021, doi: 10.1155/2021/9947621.

[32] S. Park and H. Park, "Performance Comparison of Multi-class SVM with Oversampling Methods for Imbalanced Data Classification," in *Advances on Broad-Band Wireless Computing, Communication and Applications*, vol. 159, L. Barolli, M. Takizawa, T. Enokido, H.-C. Chen, and K. Matsuo, Eds., in Lecture Notes in Networks and Systems, vol. 159. , Cham: Springer International Publishing, 2021, pp. 108–119. doi: 10.1007/978-3-030-61108-8_11.

[33] B. Huang, Y. Zhu, Z. Wang, and Z. Fang, "Imbalanced Data Classification Algorithm Based on Clustering and SVM," *J CIRCUIT SYST COMP*, vol. 30, no. 02, p. 2150036, Feb. 2021, doi: 10.1142/S0218126621500365.

[34] Maradi, A. Y. (2020). Pemanfaatan android untuk sistem kendali robot penembak dengan mikrokontroler. CYCLOTRON, 3(1).

[35] K. Ahlawat, A. Chug, and A. P. Singh, "A Novel Hybrid Sampling Algorithm for Solving Class Imbalance Problem in Big Data," *Adv. Data Sci. Adapt. Data Anal.*, vol. 13, no. 02, p. 2150005, Apr. 2021, doi: 10.1142/S2424922X21500054.

[36] B. Zhao, X. Zhang, H. Li, and Z. Yang, "Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions," *Knowledge-Based Systems*, vol. 199, p. 105971, Jul. 2020, doi: 10.1016/j.knosys.2020.105971.

[37] Herlina, A., Syahbana, M. I., Gunawan, M. A., & Rizqi, M. M. (2022). Sistem Kendali Lampu Berbasis Iot Menggunakan Aplikasi Blynk 2.0 Dengan Modul Nodemcu Esp8266. INSANtek, 3(2), 61-66..

[38] Q. Shu, T. Hu, and S. Liu, "Random Forest Algorithm Based on GAN for Imbalanced Data Classification," *J. Phys.: Conf. Ser.*, vol. 1544, no. 1, p. 012014, May 2020, doi: 10.1088/1742-6596/1544/1/012014.

[39] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 31, no. 9, pp. 3525–3539, Sep. 2020, doi: 10.1109/TNNLS.2019.2944962.

[40] Setyobudi, R. (2023). Utilization of tds sensors for water quality monitoring and water filtering of carp pools using IoT. EUREKA: Physics and Engineering, (6), 69-77.

[41] S. Abdullah and G. Prasetyo, "EASY ENSEMMBLE WITH RANDOM FOREST TO HANDLE IMBALANCED DATA IN CLASSIFICATION," *JFMA*, vol. 3, no. 1, pp. 39–46, Jun. 2020, doi: 10.14710/jfma.v3i1.7415.

[42] G. Zheng, C. A. Wu, and H. Guo, "KNN-based ensemble selection for imbalance learning," *IJCSYSE*, vol. 5, no. 2, p. 82, 2019, doi: 10.1504/IJCSYSE.2019.100025.

[43] Setyobudi, R. (2023). Utilization of tds sensors for water quality monitoring and water filtering of carp pools using IoT. EUREKA: Physics and Engineering, (6), 69-77.

[44] Md. Mahin, Md. J. Islam, B. C. Debnath, and A. Khatun, "Tuning Distance Metrics and K to Find Sub-categories of Minority Class from Imbalance Data Using K Nearest Neighbours," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh: IEEE, Feb. 2019, pp. 1–6. doi: 10.1109/ECACE.2019.8679380.

[45] Prabowo, Y. A., Imaduddin, R. I., Pambudi, W. S., Firmansyah, R. A., & Fahruzi, A. (2021). Identification of automatic guided vehicle (agv) based on magnetic guided sensor for industrial material transfer. In IOP Conference Series: Materials Science and Engineering (Vol. 1010, No. 1, p. 012028). IOP Publishing..

[46] D. Devi, S. K. Biswas, and B. Purkayastha, "Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique," *Connection Science*, vol. 31, no. 2, pp. 105–142, Apr. 2019, doi: 10.1080/09540091.2018.1560394.

[47] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased Random Forest For Dealing With the Class Imbalance Problem," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 30, no. 7, pp. 2163–2172, Jul. 2019, doi: 10.1109/TNNLS.2018.2878400.