Optimasi Penanganan Ketidakseimbangan Data pada Klasifikasi Pengaduan Masyarakat Menggunakan Metode *Naïve Bayes*

Fitro Praaidinza Muhammad[1,*], Maimunah[2], Setiya Nugroho[3]

[1, 2, 3]Fakultas Teknik, Teknik Informatika S1, Universitas Muhammadiyah Magelang, Kota Magelang, Indonesia
Email: [1,*]fitropraaidinza776@gmail.com, [2]maimunah@unimma.ac.id, [3]setiya@ummgl.ac.id
[*)] Email Penulis Utama

Abstrak— Pengaduan masyarakat merupakan salah satu bentuk nyata partisipasi publik yang berperan penting dalam proses evaluasi serta peningkatan kualitas pelayanan publik. Data pengaduan yang masuk, apabila diolah dengan baik, dapat menjadi dasar perumusan kebijakan yang lebih responsif terhadap kebutuhan masyarakat. Akan tetapi, salah satu kendala yang dapat muncul dalam pengolahan data pengaduan adalah masalah ketidakseimbangan data. Ketidakseimbangan ini berpotensi menurunkan kinerja algoritma klasifikasi karena model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Penelitian ini dilakukan dengan tujuan untuk mengoptimalkan penanganan ketidakseimbangan data pada klasifikasi pengaduan masyarakat di Dinas Perhubungan Kota Magelang. Algoritma yang digunakan adalah *Naïve Bayes* dengan pembobotan data berbasis TF-IDF. Dataset penelitian diambil dari aplikasi LAPOR! dalam rentang waktu 20 Desember 2020 hingga 10 April 2025 dengan total 350 data awal. Setelah melalui tahap pembersihan, eliminasi duplikasi, serta penghapusan data tidak relevan, tersisa 337 data yang kemudian diberi label manual ke dalam tiga kategori, yaitu MRLL dan PJU, Dalops dan Perparkiran, serta Angkutan dan Terminal. Tahap praproses mencakup penghapusan duplikasi, *case folding*, perbaikan ejaan, penghapusan angka dan tanda baca, *stemming*, serta penghapusan *stopwords*. Selanjutnya, pembobotan TF-IDF hanya diterapkan pada data latih untuk mencegah kebocoran informasi. Data dibagi dengan rasio 80% latih dan 20% uji. Untuk mengatasi ketidakseimbangan kelas, tiga pendekatan digunakan, yakni RUS, SMOTE, dan ADASYN. Semua metode diterapkan hanya pada data latih sebelum proses pelatihan model. Evaluasi dilakukan dengan menggunakan metrik akurasi, presisi, *recall*, dan *f1-score*. Hasil penelitian menunjukkan bahwa *Naïve Bayes* tanpa penanganan ketidakseimbangan hanya mencapai akurasi 86,76% dengan presisi 58%, *recall* 61%, dan *f1-score* 59%. Penerapan SMOTE maupun ADASYN mampu meningkatkan kinerja model pada beberapa metrik, tetapi kombinasi RUS dan *Naïve Bayes* memberikan performa paling optimal, yaitu akurasi 94,12%, presisi 89%, *recall* 96%, dan *f1-score* 92%. Temuan ini membuktikan bahwa strategi undersampling efektif memperbaiki kemampuan model dalam mengenali kelas minoritas.

Kata Kunci: Klasifikasi Pengaduan Masyarakat, Ketidakseimbangan Data, *Naïve Bayes*, RUS, SMOTE, ADASYN

Abstract— Public complaints represent a tangible form of public participation that plays an important role in evaluating and improving the quality of public services. When properly processed, complaint data can serve as the basis for formulating policies that are more responsive to community needs. However, one of the challenges that may arise in processing complaint data is the issue of data imbalance. This imbalance can reduce the performance of classification algorithms, as models tend to be biased toward the majority class while neglecting the minority class. This study aims to optimize the handling of data imbalance in the classification of public complaints at the Transportation Agency of Magelang City. The algorithm employed is Naïve Bayes with TF-IDF-based feature weighting. The dataset was obtained from the LAPOR! application covering the period from December 20, 2020, to April 10, 2025, consisting of 350 initial records. After cleaning, duplicate elimination, and removal of irrelevant entries, 337 records remained, which were manually labelled into three categories: MRLL dan PJU, Dalops dan Perparkiran, and Angkutan dan Terminal. The preprocessing stage included duplicate removal, case folding, spelling correction, removal of numbers and punctuation, stemming, and stopword elimination. TF-IDF weighting was applied only to the training set to prevent information leakage. The data was split into training and testing sets with an 80:20 ratio. To address the imbalance problem, three approaches were applied: RUS, SMOTE, and ADASYN, all conducted solely on the training set before model training. Performance evaluation was carried out using accuracy, precision, recall, and f1-score. The results show that Naïve Bayes without imbalance handling achieved an accuracy of 86.76%, with precision of 58%, recall of 61%, and f1-score of 59%. The application of SMOTE and ADASYN improved some metrics, but the combination of RUS and Naïve Bayes yielded the best performance, with an accuracy of 94.12%, precision of 89%, recall of 96%, and f1-score of 92%. These findings demonstrate that the undersampling strategy is effective in improving the ability of Naïve Bayes to recognize minority classes, despite the potential risk of losing part of the information from majority classes.

Keywords: Public Complaints Classification, Data Imbalance, Naïve Bayes, RUS, SMOTE, ADASYN

1. PENDAHULUAN

Pengaduan masyarakat termasuk dalam salah satu wujud pelayanan publik yang diterapkan sebagai bentuk komitmen dan kontribusi pemerintah kepada masyarakat. Hal ini tertuang dalam Undang-Undang Nomor 25 Tahun 2009 tentang Pelayanan Publik, di mana pelayanan publik diharapkan dapat memiliki akuntabilitas, responsivitas, dan efisiensi [1]. Undang-undang tersebut juga mempercayakan agar penyelenggara pelayanan

publik wajib menyediakan akses kepada masyarakat untuk menyampaikan masukan tentang pelayanan yang telah diberikan [2]. Informasi serta suara dari pengaduan dan masukan tersebut dapat digunakan sebagai bahan rujukan untuk mengetahui permasalahan di masyarakat serta mengevaluasi kebijakan yang telah diterapkan [3]. Saat ini, banyak instansi yang telah menyediakan fasilitas pengaduan masyarakat. Sebagai contoh, pelayanan pengaduan masyarakat telah diterapkan di Satuan Polisi Pamong Praja Kabupaten Kudus, di mana masyarakat dapat melaporkan berbagai hal mengenai ketertiban umum dan ketentraman [4]. Selain itu, pelayanan pengaduan masyarakat juga diterapkan di Dinas Perhubungan Provinsi Kalimantan Barat, di mana masyarakat dapat melaporkan hal-hal seputar lalu lintas, seperti perbaikan lampu lalu lintas, perbaikan jalan, serta pemeliharaan dan perawatan taman kota [5]. Tidak hanya di tingkat provinsi, pelayanan pengaduan masyarakat mengenai lalu lintas juga telah diterapkan di Dinas Perhubungan Kota Magelang.

Pengaduan-pengaduan dari masyarakat Kota Magelang dikirim melalui dua laman web resmi milik pemerintah, yaitu LAPOR! dan Monggo Lapor. Pengaduan-pengaduan tersebut kemudian diteruskan oleh administrator sistem kepada instansi yang berwenang untuk menangani. Pengaduan-pengaduan mengenai perhubungan akan diteruskan ke Dinas Perhubungan Kota Magelang. Setelah itu, setiap pengaduan ditangani oleh salah satu dari keempat seksi yang ada, yaitu Seksi Manajemen Rekayasa Lalu Lintas dan Penerangan Jalan Umum, Seksi Pengendalian Operasional dan Perparkiran, Seksi Pengujian Kendaraan Bermotor, serta Seksi Angkutan dan Terminal. Jumlah pengaduan masyarakat yang diterima oleh Dinas Perhubungan Kota Magelang setiap hari dapat bervariasi, mulai dari satu pengaduan hingga lima puluh pengaduan per hari.

Berdasarkan informasi yang diperoleh dari Kepala Subbagian Umum dan Kepegawaian di Dinas Perhubungan Kota Magelang, sebagian dari proses pengelolaan pengaduan masyarakat di Dinas Perhubungan Kota Magelang masih belum dilaksanakan secara maksimal. Meski Dinas Perhubungan Kota Magelang telah memanfaatkan dua laman web resmi pemerintah dalam mengirim dan menerima pengaduan, klasifikasi dan pengiriman pengaduan masyarakat ke salah satu dari empat seksi yang ada masih dilakukan secara manual melalui aplikasi *WhatsApp*. Hal ini tentunya cukup memakan waktu bagi pihak yang bertanggung jawab dalam mengelola pengaduan di Dinas Perhubungan Kota Magelang, terutama jika terdapat banyak pengaduan masyarakat yang masuk dalam satu waktu. Selain itu, klasifikasi dan pengiriman pengaduan melalui *WhatsApp* ini juga menyebabkan pencarian dan pemetaan ulang pengaduan cukup sulit karena pihak pengelola pengaduan harus memeriksa percakapan *WhatsApp*. Terakhir, terdapat ketidakseimbangan pada data pengaduan yang telah diterima oleh Dinas Perhubungan Kota Magelang, di mana hampir semua data pengaduan diterima oleh Seksi Manajemen Rekayasa Lalu Lintas dan Penerangan Jalan Umum serta Seksi Pengendalian Operasional dan Perparkiran.

Penelitian mengenai pengaduan telah dilakukan oleh Sunarti dan rekan-rekan [6]. Penelitian tersebut menemukan bahwa algoritma *Naïve Bayes* dengan pembobotan data menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) telah berhasil diterapkan dalam klasifikasi pengaduan pelayanan di Fakultas Teknik Universitas Muhammadiyah Makassar. Klasifikasi pengaduan tersebut dilakukan dengan sembilan kelas berbeda dan menghasilkan akurasi sebesar 91%. Akan tetapi, beberapa kelas masih memiliki nilai *recall* dan *f1-score* yang cukup rendah. Hal ini disebabkan oleh ketidakseimbangan jumlah data pada setiap kelas sehingga terdapat kelas dengan representasi data latih yang sangat sedikit dibandingkan kelas lainnya.

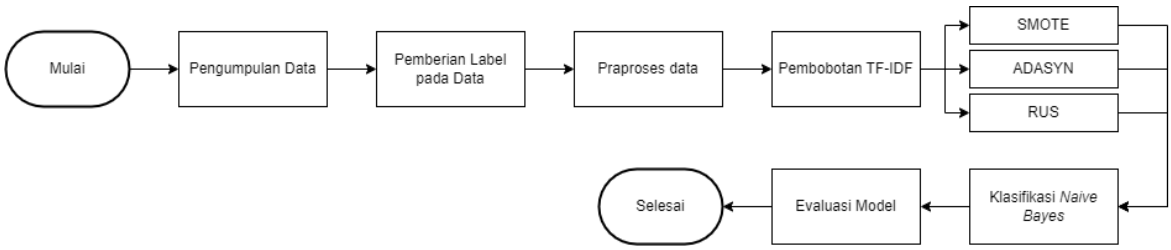
Penelitian lain telah dilakukan oleh Kusuma dan rekan-rekan mengenai penanganan ketidakseimbangan data pada klasifikasi pengaduan masyarakat [7]. Penelitian tersebut berfokus pada analisis mengenai teknik *oversampling* menggunakan kombinasi metode *Adaptive Synthetic Sampling* (ADASYN) dan *Synthetic Minority Over-sampling Technique* (SMOTE). Metode klasifikasi pengaduan masyarakat yang digunakan dalam penelitian tersebut adalah klasifikasi *Support Vector Machine*, *Random Forest*, dan *Naïve Bayes*. Kesimpulan yang didapat dalam penelitian tersebut adalah bahwa kombinasi metode SMOTE dan ADASYN berhasil meningkatkan akurasi, presisi, *recall*, dan *f1-score* pada model dengan algoritma SVM dan *Random Forest*. Akan tetapi, teknik *oversampling* tersebut justru menurunkan akurasi, presisi, *recall*, dan *f1-score* pada model dengan algoritma *Naïve Bayes*. Meski begitu, terdapat peningkatan waktu proses pada model dengan algoritma *Naïve Bayes*.

Penelitian serupa lainnya juga dilakukan oleh Atmaja dan Wahyuni mengenai analisis sentiment berbasis aspek pada sistem layanan pengaduan masyarakat di Kota Surabaya [8]. Penelitian tersebut dilakukan menggunakan metode *Latent Dirichlet Allocation* untuk mengidentifikasi aspek-aspek utama pengaduan dan *Naïve Bayes* untuk melakukan klasifikasi. Penelitian tersebut juga membandingkan tiga puluh skenario berbeda untuk ditemukan model dengan performa yang terbaik. Hasil dari penelitian tersebut menunjukkan bahwa kombinasi skenario terbaik adalah praproses data menggunakan *stopword removal*, perbandingan rasio 90:10 untuk data latih dan data uji, serta *Random Under-Sampling* untuk penanganan ketidakseimbangan data. Kombinasi skenario tersebut menghasilkan akurasi data latih sebesar 84%, akurasi data uji sebesar 80%, serta *presisi*, *recall*, dan *f1-score* sebesar 79%.

Penelitian saat ini bertujuan untuk menangani ketidakseimbangan data dengan metode SMOTE, ADASYN, dan RUS pada klasifikasi pengaduan masyarakat di Dinas Perhubungan Kota Magelang menggunakan metode *Naïve Bayes*. Pembobotan data dalam penelitian ini dilakukan menggunakan metode TF-IDF. Penelitian ini menggunakan 350 data pengaduan yang ditujukan kepada Dinas Perhubungan Kota Magelang sejak tanggal 20 Desember 2020 hingga 10 April 2025 melalui laman web LAPOR!.

2. METODE PENELITIAN

Perencanaan langkah-langkah penelitian harus dilakukan sebelum memulai penelitian untuk membantu penelitian menjadi lebih teratur dan sistematis. Langkah-langkah tersebut kemudian diikuti oleh peneliti dalam melakukan penelitiannya. Langkah-langkah yang diikuti pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart* Alur Penelitian

2.1 Pengumpulan Data

Pengumpulan data dilakukan melalui wawancara dengan Kepala Subbagian Kepegawaian di Dinas Perhubungan Kota Magelang. Pengumpulan data dilakukan dengan mengambil data pengaduan dari aplikasi LAPOR!. Pengambilan data dilakukan menggunakan metode *web-scraping*, yaitu teknik pengambilan data dari halaman web secara otomatis menggunakan skrip pemrograman [9]. Data yang dikumpulkan berupa 350 data pengaduan yang ditangani oleh berbagai seksi, dengan mayoritas data ditangani oleh seksi MRLL dan PJU serta Dalops dan Perparkiran.

2.2 Pemberian Label pada Data

Data yang didapatkan belum diberikan label sesuai dengan kelasnya. Oleh karena itu, perlu dilakukan pemberian label pada data agar data dapat diklasifikasikan. Pemberian label data diberikan berdasarkan informasi mengenai seksi-seksi pada Dinas Perhubungan Kota Magelang serta pengaduan-pengaduan yang ditangani oleh setiap seksi tersebut. Data yang telah diberikan label kemudian akan menjalani praproses data agar dapat dibaca dan diolah oleh komputer.

2.3 Praproses Data

Praproses data bertujuan untuk membersihkan serta menyiapkan data agar dapat dibaca serta diolah oleh komputer [10]. Langkah-langkah praproses data yang telah dilakukan dalam penelitian ini dimulai dari penghapusan data duplikasi [11] dan *case folding* atau pengubahan semua huruf menjadi huruf kecil [12]. Kemudian, praproses data dilanjutkan dengan tokenisasi, yang terdiri dari penghapusan angka, tanda baca, dan spasi berlebih [13]. Praproses data diakhiri dengan normalisasi atau perbaikan penulisan kata, *stemming* serta tokenisasi untuk mengubah data-data pengaduan menjadi tiap-tiap kata dasar, dan penghapusan *stopwords* [14]. Setelah itu, data-data yang telah dibersihkan dan dipisahkan menjadi tiap-tiap kata dapat dihitung bobotnya.

2.4 Pembobotan TF-IDF

Pembobotan TF-IDF dalam penelitian ini dilakukan untuk menentukan tingkat kepentingan tiap kata yang ada dalam dokumen berdasarkan jumlah kemunculannya. Metode ini dipilih karena memiliki akurasi dan *recall* yang cukup tinggi [15], [16]. Hasil dari pembobotan TF-IDF akan melalui proses *oversampling* dan *undersampling* untuk ditangani ketidakseimbangannya. Rumus dari pembobotan TF-IDF dapat dilihat pada persamaan (1).

$$TF - IDF = TF_{(t,d)} \times IDF_{(t)} \quad (1)$$

Persamaan (1) menunjukkan bahwa pembobotan TF-IDF dilakukan dengan mengalikan hasil *Term Frequency* (TF) dan hasil *Inverse Document Frequency* (IDF). TF adalah frekuensi kemunculan kata dalam sebuah data atau dokumen, di mana kata yang sering muncul akan memiliki nilai *term frequency* yang lebih tinggi. IDF adalah adalah frekuensi kemunculan kata dalam semua data atau dokumen, di mana kata yang lebih jarang muncul akan memiliki nilai *inverse document frequency* yang lebih rendah. Rumus TF dapat dilihat pada persamaan (2), sedangkan rumus IDF dapat dilihat pada persamaan (3).

$$TF_{(t,d)} = \frac{jumlah\ sebuah\ kata\ dalam\ sebuah\ dokuman}{jumlah\ semua\ kata\ dalam\ dokuman\ tersebut} \quad (2)$$

$$IDF_{(t)} = \frac{jumlah\ semua\ dokumen}{dokumen\ dengan\ kata\ t} \quad (3)$$

Pembobotan IDF hanya dilakukan berdasarkan data latih. Hal ini dilakukan untuk mencegah kebocoran data, di mana informasi mengenai data uji tidak sengaja masuk ke data latih. Oleh karena itu, pembobotan IDF

hanya dilakukan berdasarkan data yang ada dalam data latih, sedangkan data uji dianggap tidak ada dalam pembobotan tersebut [17]–[19].

2.5 Penanganan Ketidakseimbangan Data

Ketidakseimbangan data atau ketidakseimbangan kelas adalah kasus di mana jumlah data pada setiap kelas memiliki perbedaan yang signifikan. Jumlah kelas pada suatu data yang tidak seimbang umumnya jauh lebih kecil dibandingkan jumlah kelas pada data lain. Kelas dengan jumlah data yang lebih kecil merupakan kelas minoritas, sedangkan kelas dengan jumlah data yang lebih besar merupakan kelas mayoritas [20]–[22]. Ketidakseimbangan data dapat menyebabkan bias, di mana model klasifikasi cenderung mengklasifikasikan data ke dalam kelas mayoritas dan mengabaikan kelas minoritas [23]. Penanganan ketidakseimbangan data dilakukan dengan metode SMOTE, ADASYN, dan RUS. Hasil dari ketiga metode tersebut akan dimasukkan dalam algoritma klasifikasi *Naïve Bayes*.

SMOTE, atau *Synthetic Minority Over-sampling Technique*, merupakan metode penanganan ketidakseimbangan data yang berfokus pada *oversampling*, yaitu meningkatkan jumlah data dalam kelas minoritas. SMOTE bekerja dengan membuat dan menyisipkan sampel baru dari kelas minoritas untuk menyeimbangkan data berdasarkan tetangga terdekat antara data-data minoritas [24], [25]. Metode ini memiliki beberapa kelebihan, salah satunya adalah tidak adanya kehilangan informasi yang diakibatkan oleh pengurangan data. SMOTE juga dapat meningkatkan akurasi pada kelas minoritas [26].

ADASYN atau *Adaptive Synthetic Sampling* juga merupakan metode penanganan ketidakseimbangan data yang berfokus pada *oversampling*. Metode ini dapat dikatakan sebagai pengembangan dari metode SMOTE [27]. *Oversampling* dalam metode ADASYN dilakukan berdasarkan kesulitan pembelajaran model. Data sintesis yang dihasilkan dalam metode ini berasal dari data minoritas yang lebih sulit dipelajari dibandingkan dengan data minoritas lain [28].

Random Under-Sampling adalah metode penanganan ketidakseimbangan data yang berfokus pada *undersampling*, yaitu pengurangan jumlah data pada kelas mayoritas. Metode ini bekerja dengan cara memilih data-data pada kelas mayoritas secara acak lalu menghapusnya hingga jumlahnya sama dengan data pada kelas minoritas. Penghapusan data tersebut dapat menyebabkan hilangnya informasi penting pada data-data mayoritas dan mempengaruhi akurasi keseluruhan pada model [29]. Meski begitu, metode ini memiliki beberapa kelebihan, seperti mengurangi sampel data mayoritas yang berlebihan dan mengurangi waktu pelatihan model [25].

Sebagai contoh, data latih awal dalam penelitian ini terdiri dari 120 data pada kelas MRLL dan PJU, 140 data pada kelas Dalops dan Perparkiran, serta 9 data pada kelas Angkutan dan Terminal sebagai kelas minoritas. Metode RUS memilih data pada kelas MRLL dan PJU serta Dalops dan Perparkiran secara acak dan mengurangi jumlahnya hingga menjadi sama dengan kelas minoritas, yaitu 9 data pada setiap kelas. Dengan demikian, distribusi data latih menjadi seimbang dengan total 27 data.

Penyeimbangan data Pada metode SMOTE dilakukan dengan menambahkan data sintetis pada kelas minoritas. Oleh karena itu, kelas Angkutan dan Terminal yang awalnya hanya berjumlah 9 data ditingkatkan hingga 140 data agar setara dengan kelas mayoritas. Data sintetis dibangkitkan berdasarkan kemiripan antara data minoritas dengan memanfaatkan pendekatan tetangga terdekat sehingga data baru yang dihasilkan tetap merepresentasikan karakteristik kelas tersebut.

Sementara itu, penambahan data sintetis pada metode ADASYN juga dilakukan pada kelas minoritas, namun dengan pendekatan adaptif. Data minoritas yang berada pada area yang lebih sulit dipelajari oleh model akan mendapatkan jumlah data sintetis yang lebih banyak dibandingkan data yang lebih mudah dipelajari. Setelah penambahan data sintetis dengan metode tersebut dilakukan, jumlah data pada kelas Angkutan dan Terminal meningkat hingga mendekati jumlah kelas mayoritas, yaitu sekitar 140 data.

Proses pembobotan TF-IDF serta penanganan ketidakseimbangan data menggunakan metode RUS, SMOTE, dan ADASYN diterapkan hanya pada data latih. Pendekatan ini bertujuan untuk mencegah terjadinya kebocoran data serta memastikan bahwa model tidak memperoleh informasi dari data uji selama proses pelatihan [30]. Dengan demikian, hasil evaluasi yang diperoleh dapat merepresentasikan performa model secara lebih objektif. Pendekatan ini merupakan bagian dari upaya optimasi dalam pelatihan model, khususnya dalam mengurangi bias akibat ketidakseimbangan data serta mencegah kebocoran data.

2.6 Klasifikasi *Naïve Bayes*

Klasifikasi *Naïve Bayes* dilakukan dengan menggunakan data latih yang telah diproses, dengan rasio data latih dan data uji sebesar 80:20. Klasifikasi tersebut dibagi menjadi tiga skenario, di mana masing-masing skenario menggunakan salah satu dari ketiga hasil penanganan ketidakseimbangan data yang telah dijelaskan. Kelas-kelas yang ada dalam klasifikasi ini yaitu MRLL dan PJU, Dalops dan Perparkiran, serta Angkutan dan Terminal. Algoritma ini dipilih karena merupakan algoritma yang cukup sederhana dengan akurasi dan kecepatan yang tinggi [31]. Selain itu, Selain itu, algoritma *Naïve Bayes* juga hanya membutuhkan data pelatihan dalam jumlah kecil [32].

Proses klasifikasi menggunakan *Naïve Bayes* dalam penelitian ini dilakukan secara berurutan. Pertama, data pengaduan yang telah melalui tahap praproses diubah menjadi angka atau vektor menggunakan pembobotan TF-IDF. Kedua, model dilatih menggunakan data latih yang telah diseimbangkan, dengan mempelajari pola kemunculan kata pada masing-masing kelas. Ketiga, pada tahap pengujian, setiap data uji dianalisis dengan melihat kemunculan kata-kata di dalamnya dan mencocokkannya dengan pola yang telah dipelajari berdasarkan data latih. Terakhir, data uji dimasukkan ke dalam kelas yang paling sesuai berdasarkan hasil perhitungan model.

2.7 Evaluasi Model

Evaluasi model dilakukan dengan melihat presisi, akurasi, *recall*, dan *f1-score* dari ketiga skenario yang telah dirancang. Akurasi, presisi, *recall*, dan *f1-score* merupakan metrik evaluasi dalam pembelajaran mesin, khususnya dalam algoritma klasifikasi. Akurasi merupakan persentase prediksi yang benar dari semua hasil prediksi. Presisi adalah persentase prediksi positif yang benar dari semua hasil prediksi positif. *Recall* adalah persentase prediksi yang benar dari satu kelas tertentu. *F1-score* adalah perbandingan antara presisi rata-rata dan *recall*. [33]. Ketiga skenario tersebut akan dibandingkan satu sama lain untuk menemukan skenario dengan performa terbaik. Evaluasi model ini dipilih karena merupakan metode evaluasi standar untuk hasil klasifikasi.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data diunduh pada tanggal 10 April 2025 melalui aplikasi LAPOR!. Pengunduhan dilakukan menggunakan metode *web-scraping*, yaitu teknik pengumpulan data dari internet secara otomatis. Metode tersebut digunakan karena aplikasi LAPOR! tidak menyediakan fitur pengunduhan data pengaduan bagi pengguna dan administrator dari Dinas Perhubungan Kota Magelang. Data pengaduan yang diunduh merupakan data pengaduan dari tanggal 20 Desember 2020 hingga 10 April 2025 dan berjumlah 350 data. Data dari aplikasi tersebut diunduh ke dalam bentuk tabel dengan empat kolom, yaitu *Title*, *Excerpt*, *Category*, dan *Date*. Sampel data yang telah diunduh dapat dilihat pada Tabel 1.

Tabel 1. Sampel Tiga Data dari Dataset Pengaduan

Title	*Excerpt*	*Category*	*Date*
PJU MATI	PJU Jalan Salak 1 RT. 03 RW. 04 kramat selatan , magelang utara , (dpn pos kamling) mati sudah 5 hari... Mohon bantuannya untuk segera di atasi	Penerangan Jalan	Selasa, 2 Mei 2023, 14:47
Macet depan SMP 1 MGL	Banyak mobil ,angkot, yg berhenti di jalan jadi bikin macet parah, tolong di kondisikan agar lebih tertata terutama jam pulang sekolah	Topik Lainnya	Selasa, 2 Mei 2023, 14:44
Kebersihan Terminal Tidar	Mohon dicek kebersihan terminal tidar. Tampak depan terlihat kurang terawat karena banyak rumput liar di sekitar kanstin dan beberapa taman tampak kurang terawat. Mohon bisa menjadi perhatian. Terima kasih.	Lainnya terkait Perhubungan	Rabu, 25 Januari 2023, 14:42

3.2 Pemberian Label pada Data

Pelabelan data dilakukan berdasarkan hasil wawancara dengan Kepala Subbagian Umum dan Kepegawaian di Dinas Perhubungan Kota Magelang. Data-data yang ada dalam dataset tersebut terbagi menjadi tiga kelas, yaitu MRLL dan PJU, Dalops dan Perparkiran, serta Angkutan dan Terminal. Hasil pelabelan data dapat dilihat pada Tabel 2.

Tabel 2. Sampel Tiga Data dari Dataset Pengaduan dengan Label

Title	*Excerpt*	*Category*	*Date*	*Manual_Label*
PJU MATI	PJU Jalan Salak 1 RT. 03 RW. 04 kramat selatan , magelang utara , (dpn pos kamling) mati sudah 5 hari... Mohon bantuannya untuk segera di atasi	Penerangan Jalan	Selasa, 2 Mei 2023, 14:47	MRLL dan PJU
Macet depan SMP 1 MGL	Banyak mobil ,angkot, yg berhenti di jalan jadi bikin macet parah, tolong di kondisikan agar lebih tertata terutama jam pulang sekolah	Topik Lainnya	Selasa, 2 Mei 2023, 14:44	Dalops dan Perparkiran
Kebersihan Terminal Tidar	Mohon dicek kebersihan terminal tidar. Tampak depan terlihat kurang terawat karena banyak rumput liar di sekitar kanstin dan beberapa taman tampak kurang terawat. Mohon bisa menjadi perhatian. Terima kasih.	Lainnya terkait Perhubungan	Rabu, 25 Januari 2023, 14:42	Angkutan dan Terminal

Dataset pengaduan yang berjumlah 350 data juga mengandung data-data duplikat, data pengaduan yang tidak berkaitan dengan Dinas Perhubungan Kota Magelang, dan data yang tidak hanya ditujukan terhadap bidang atau seksi tertentu dengan jumlah 13 data. Oleh karena itu, data yang digunakan dan diberi label berjumlah 337 data. Contoh data tersebut adalah data pengaduan mengenai kualitas udara buruk yang seharusnya menjadi tanggung jawab Dinas Lingkungan Hidup. Data-data tersebut tidak diberi label dan tidak digunakan dalam perancangan model. Contoh-contoh data yang tidak digunakan dalam perancangan model dapat dilihat pada Tabel 3.

Tabel 3. Contoh Data yang Tidak Digunakan

Title	*Excerpt*	*Category*	*Date*
KUALITAS UDARA BURUK, MENJAGA KEBERSIHAN KOTA DARI SISI MANAPUN	Selamat Pagi Bapak/Ibu, semoga Bapak Walikota memperhatikan ini. Mohon bantuannya untuk kepeduliannya terhadap lingkungan sekitar. Mohon untuk terus menggencarkan siaran untuk warga agar bisa menjaga kualitas udara dan ketahanan energi dengan baik…	Topik Lainnya	Kamis, 13 Juli 2023, 2:30
survey pesepeda	Kami pesepeda, para penjaga udara kota ko ga pernah disurvey kebutuhan dalam perencanaan pembangunan?	Lainnya terkait Perhubungan	Rabu, 25 Januari 2023, 14:41
bersepeda bersama anak	Inih keren sekaleee... Andai saja @satlantas_mglta , @Disdik_KotaMGL & @dishubkotamgl bisa kolaborAKSI bangun budaya bersepeda anak anak tentu magelang lebih "menyenangkan".	Topik Lainnya	Selasa, 15 November 2022, 7:06

3.3 Praproses Data

Data yang telah dilabeli menjalani praproses data agar dapat dibaca dan diolah oleh komputer. Praproses data yang dilakukan telah dalam penelitian ini dimulai dari penghapusan data duplikasi, konversi huruf kapital menjadi

huruf kecil, penghapusan angka serta simbol atau tanda baca, penghapusan spasi berlebih, perbaikan ejaan dan penulisan kata, pemisahan teks menjadi tiap-tiap kata dasarnya, serta penghapusan *stopwords*. Hasil dari praproses data dapat dilihat pada Tabel 4.

Tabel 4. Sampel Tiga Data Pengaduan Setelah Praproses Data

Title + Excerpt	*Manual_Label*
pju mati pju salak kramat selatan utara pos kamling mati bantu	MRLL dan PJU
macet smp mgl mobil angkot henti bikin macet parah tolong kondisi tata utama jam pulang sekolah	Dalops dan Perparkiran
bersih terminal cek bersih terminal tidar lihat awat rumput liar kanstin taman awat perhati terima kasih	Angkutan dan Terminal

3.4 *Label Encoding* serta Pembagian Data Latih dan Data Uji

Label atau kelas data diubah menjadi bentuk angka terlebih dahulu dengan *label_encoder*. Tahap ini dilakukan agar komputer dapat mengenali label karena komputer tidak dapat mengenali teks kategorikal. Hasil dari tahap ini dapat dilihat pada Tabel 5.

Tabel 5. Hasil *Label_Encoding*

Label Asli	Hasil *Label_Encoding*
Angkutan dan Terminal	0
Dalops dan Perparkiran	1
MRLL dan PJU	2

Data yang telah menjalani proses praproses data dipisah menjadi data latih dan data uji. Perbandingan persentase data latih dan data uji yang digunakan dalam penelitian ini adalah 80:20, dengan data uji sebanyak 0.2 atau 20%. Data latih dimasukkan dalam variabel *X_train*, dengan *y_train* sebagai label atau kelas data tersebut. Data uji dimasukkan dalam variabel *X_train*, dengan *y_test* sebagai kelas sebenarnya dari data uji tersebut. Hasil prediksi algoritma *Naïve Bayes* terhadap data uji lalu dibandingkan dengan *y_test* tersebut untuk dievaluasi. Jumlah data latih dan data uji dapat dilihat pada Tabel 6.

Tabel 6. Jumlah Data Latih dan Data Uji

Kelas	Jumlah Data Latih	Jumlah Data Uji
Angkutan dan Terminal	9	3
Dalops dan Perparkiran	140	32
MRLL dan PJU	120	33

3.5 Pembobotan Kata Menggunakan TF-IDF

Setiap kata atau frasa pada data latih dihitung bobot TF-IDF-nya dengan rumus yang ada pada persamaan (1), (2), dan (3). Hasil dari bobot ini digunakan untuk menentukan prioritas kata atau frasa dalam proses klasifikasi. Kata dengan bobot lebih tinggi akan "diperhatikan" lebih dahulu dibandingkan dengan kata yang berbobot lebih rendah. Sebagai contoh, frasa "parkir lampir" pada Tabel 7 memiliki bobot TF-IDF yang paling tinggi, yaitu sebesar 3,814. Hal itu berarti model klasifikasi lebih memperhatikan frasa tersebut dalam klasifikasi pengaduan karena dianggap lebih penting. Pengaduan yang memiliki frasa tersebut kemungkinan besar diklasifikasikan ke dalam kelas "Dalops dan Perparkiran". Contoh hasil TF-IDF dari data dapat dilihat pada Tabel 7.

Tabel 7. Contoh Hasil TF-IDF

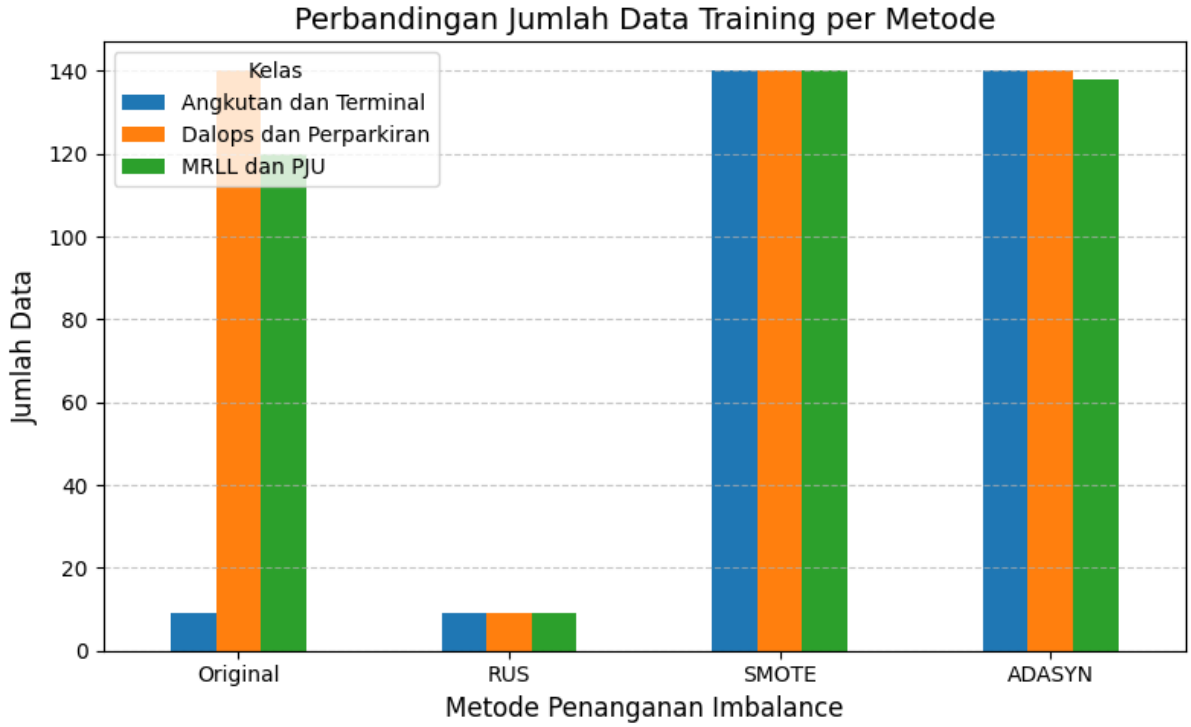
Kata/Frasa	TF	IDF	TF-IDF
parkir lampir	0.6459	5.9053	3.8144
listrik	0.6698	5.4998	3.6838
pju lampu	0.6680	5.2121	3.4815

3.6 Penanganan Ketidakseimbangan Data

Data-data latih yang telah dihitung bobotnya menggunakan TF-IDF kemudian menjalani tiga proses penanganan ketidakseimbangan data, yaitu RUS, SMOTE, dan ADASYN. Hasil proses penanganan ketidakseimbangan data dapat dilihat pada Tabel 8 dan Gambar 2.

Tabel 8. Jumlah Data Latih Sebelum dan Setelah RUS, SMOTE, serta ADASYN

Kelas	Asli	RUS	SMOTE	ADASYN
MRLL dan PJU	120	9	140	138
Dalops dan Perparkiran	140	9	140	140
Angkutan dan Terminal	9	9	140	140



Gambar 2. Grafik Perbandingan Jumlah Data Sebelum dan Setelah RUS, SMOTE, serta ADASYN

Tabel 8 dan Gambar 2 menunjukkan jumlah data latih sebelum dan sesudah dilakukan penanganan ketidakseimbangan data menggunakan metode RUS, SMOTE, dan ADASYN. Jumlah data latih awal adalah 269 data yang terdiri dari 120 data pada kelas MRLL dan PJU, 140 data pada kelas Dalops dan Perparkiran, serta 9 data pada kelas Angkutan dan Terminal. Kondisi ini menunjukkan bahwa kelas Angkutan dan Terminal merupakan kelas minoritas yang hanya mencakup 3,346% dari total data latih. Setelah dilakukan RUS, jumlah data pada setiap kelas menjadi seimbang, yaitu masing-masing 9 data sesuai dengan kelas minoritas. Sementara itu, penerapan SMOTE menghasilkan distribusi data yang juga seimbang, di mana setiap kelas memiliki 140 data latih sesuai dengan kelas mayoritas. Berbeda dengan kedua metode tersebut, ADASYN menghasilkan distribusi data yang tidak sepenuhnya sama antar kelas. Pada metode ini, jumlah data pada kelas MRLL dan PJU menjadi 138, sedangkan kelas Angkutan dan Terminal mencapai 140 data. Perbedaan ini menunjukkan bahwa ADASYN tidak sepenuhnya menyamakan jumlah data antar kelas, melainkan menyesuaikan jumlah data sintetis berdasarkan tingkat kesulitan pembelajaran pada masing-masing kelas. Penelitian yang dilakukan oleh Ahmed dan rekan-rekan juga membuktikan hal tersebut, di mana data latih yang awalnya berjumlah 896, 64, 3200, dan 2240 untuk tiap kelas menjadi berjumlah 3237, 3199, 3200, dan 3152 setelah dilakukan ADASYN [34].

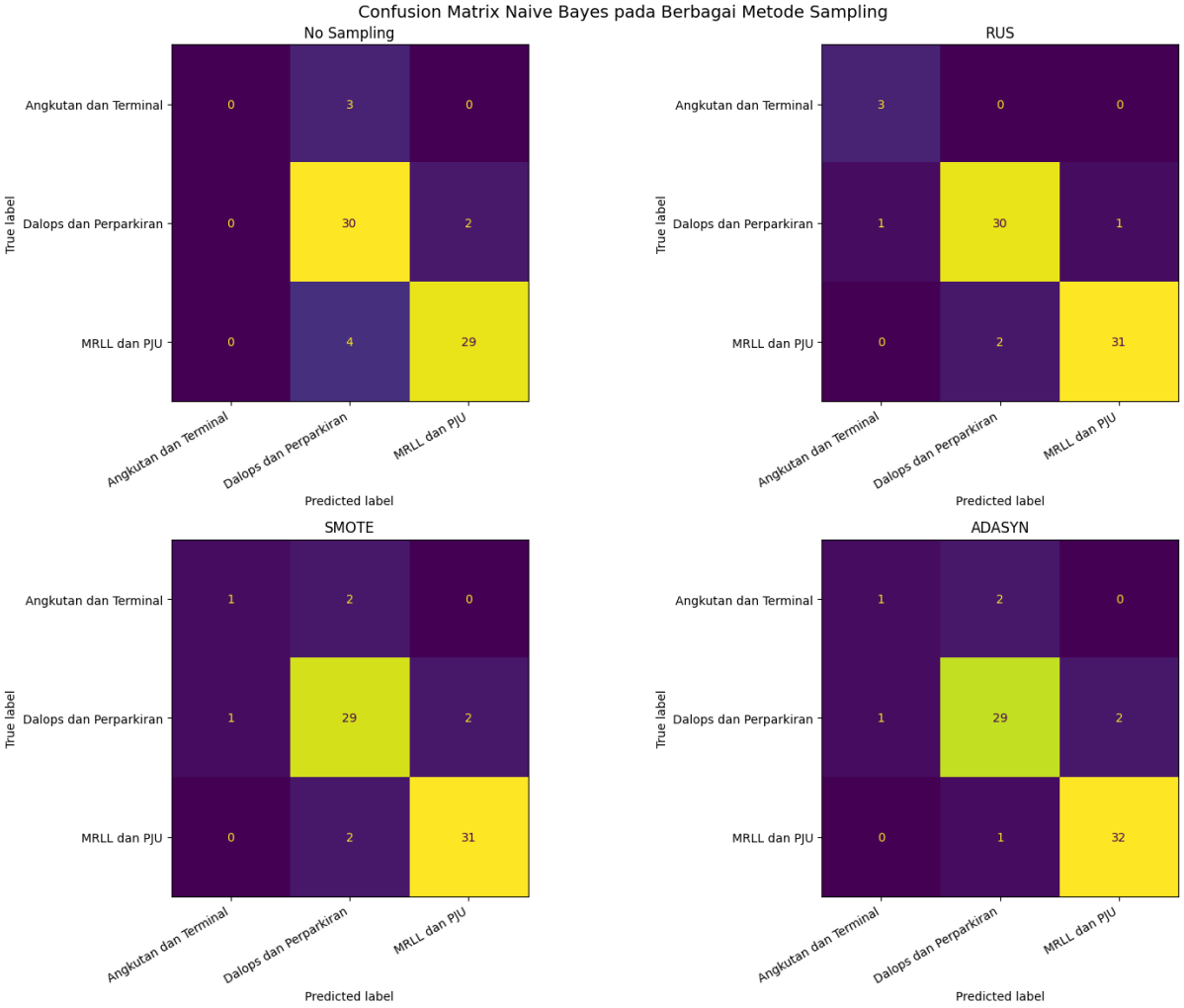
3.7 Klasifikasi *Naïve Bayes* dan Evaluasi Model

Data-data latih yang telah menjalani proses RUS, SMOTE, dan ADASYN dimasukkan ke dalam model algoritma *Naïve Bayes*. Model yang telah dilatih menggunakan data latih lalu diuji menggunakan data uji untuk dievaluasi performanya. Evaluasi dilakukan dengan membandingkan akurasi, presisi, *recall*, dan *f1-score* pada model klasifikasi untuk ditentukan model yang terbaik. Performa model klasifikasi dapat dilihat pada Tabel 9.

Tabel 9. Perbandingan Evaluasi Model Klasifikasi *Naïve Bayes*

	Akurasi	Presisi	*Recall*	*F1-Score*
Naïve Bayes	86,76%	58%	61%	59%
RUS + *Naïve Bayes*	94,12%%	89%	96%	92%
SMOTE + *Naïve Bayes*	89,71%	77%	73%	74%

Performa model klasifikasi pengaduan menggunakan *Naïve Bayes* pada tabel 9 menunjukkan bahwa model klasifikasi *Naïve Bayes* tanpa penanganan ketidakseimbangan data menghasilkan akurasi yang cukup tinggi, yaitu 86,76%. Akan tetapi, presisi, *recall*, dan *f1-score* dari model klasifikasi tersebut cukup rendah, yaitu 58%, 61%, dan 59%. Hal ini disebabkan kurangnya contoh atau representasi data pada kelas Angkutan dan Terminal, yang merupakan kelas minoritas dan hanya mencakup 3,346% dari jumlah total data latih. Hal tersebut menyebabkan model klasifikasi tidak dapat mempelajari data pada kelas tersebut dalam tahap pelatihan data sehingga model salah dalam mengklasifikasikan data uji pada kelas tersebut. Hal ini merupakan contoh bias. Hal ini dapat dilihat dalam matriks konfusi pada Gambar 3, yang menunjukkan jumlah data uji yang diprediksi sesuai kelas aslinya dan yang diprediksi sebagai kelas lain. Dalam matriks konfusi model klasifikasi tanpa penanganan ketidakseimbangan data, model tidak mampu melakukan klasifikasi secara tepat terhadap data uji dalam kelas Angkutan dan Terminal sama sekali.



Gambar 3. Matriks Konfusi Evaluasi Model Klasifikasi

Tabel 9 juga menunjukkan bahwa penggunaan SMOTE, ADASYN, dan RUS untuk menangani ketidakseimbangan data meningkatkan performa model secara signifikan, terutama dalam presisi, *recall*, dan *f1-score*. Tabel tersebut juga menunjukkan bahwa penggunaan metode RUS menghasilkan performa paling tinggi dalam klasifikasi pengaduan, dengan nilai akurasi sebesar 94,12%, presisi sebesar 89%, *recall* sebesar 96%, dan *f1-score* sebesar 92%. Hal ini berarti model klasifikasi dapat mengenali mayoritas data pada tiap kelas sesuai dengan kelas data yang sebenarnya. Hal ini juga dibuktikan pada Gambar 3, di mana model klasifikasi dengan metode penanganan ketidakseimbangan data SMOTE dan ADASYN dapat melakukan klasifikasi secara tepat terhadap mayoritas data uji dalam kelas Dalops dan Perparkiran serta MRLL dan PJU. Akan tetapi, kedua model tersebut hanya mampu melakukan klasifikasi secara tepat terhadap satu dari tiga data uji pada kelas Angkutan dan Terminal. Hal ini berbeda dengan model klasifikasi menggunakan metode penanganan ketidakseimbangan data RUS, di mana model tidak hanya dapat melakukan klasifikasi secara tepat terhadap mayoritas data uji dari kelas Dalops dan Perparkiran serta MRLL dan PJU, tetapi juga dapat melakukan klasifikasi secara tepat terhadap semua data uji dalam kelas Angkutan dan Terminal.

Metode penanganan ketidakseimbangan data RUS justru dapat menghilangkan banyak informasi pada kelas mayoritas sehingga mempengaruhi akurasi klasifikasi. Namun, hal ini belum berarti metode tersebut akan berdampak buruk pada performa model klasifikasi. Hasil dari penelitian yang dilakukan memberikan bukti bahwa penggunaan metode RUS dalam klasifikasi pengaduan masyarakat dengan algoritma *Naïve Bayes* justru dapat meningkatkan performa model klasifikasi, terutama dalam presisi dan *recall*. Selain itu, seperti yang telah dijelaskan dalam paragraf sebelumnya, model klasifikasi *Naïve Bayes* dengan metode RUS dapat melakukan klasifikasi secara tepat terhadap semua data uji dalam kelas minoritas. Bukti lain dapat dilihat pada penelitian yang dilakukan oleh Primadya dan rekan-rekan serta penelitian oleh Anargya dan rekan-rekan. Penelitian pertama berfokus pada klasifikasi data serangan pada serangan *Internet of Things* atau IoT, sedangkan penelitian kedua berfokus pada klasifikasi data serangan pada serangan *Internet of Vehicles* atau IoV. Kedua penelitian tersebut menghasilkan kesimpulan bahwa metode RUS dapat meningkatkan performa model klasifikasi. Selain itu, penelitian yang dilakukan Primadya dan rekan-rekan tersebut menegaskan bahwa penggunaan metode RUS dapat menurunkan akurasi, tapi juga dapat meningkatkan nilai presisi dan *recall* [35], [36].

4. KESIMPULAN

Penanganan ketidakseimbangan data pada klasifikasi pengaduan masyarakat di Dinas Perhubungan Kota Magelang menggunakan metode *Naïve Bayes* dilakukan menggunakan metode SMOTE, ADASYN, dan RUS. Kemudian, performa model klasifikasi *Naïve Bayes* dengan menggunakan ketiga metode tersebut dievaluasi dengan dibandingkan satu sama lain untuk ditemukan metode yang paling baik. Metode yang terbaik ditentukan dengan melihat nilai akurasi, presisi, *recall*, dan *f1-score* tertinggi dari ketiga model klasifikasi tersebut. Hasil penelitian yang telah dilakukan menunjukkan bahwa model klasifikasi yang menggunakan metode RUS memiliki performa terbaik, dengan nilai akurasi sebesar 94,12%, presisi sebesar 89%, *recall* sebesar 96%, dan *f1-score* sebesar 92%. Oleh karena itu, metode ini dapat disimpulkan sebagai metode yang efektif untuk menangani ketidakseimbangan data dan meningkatkan performa model klasifikasi *Naïve Bayes* dalam klasifikasi pengaduan masyarakat di Dinas Perhubungan Kota Magelang. Selain itu, hasil dari penelitian ini juga menunjukkan bahwa metode *Naïve Bayes* memiliki kelemahan dalam melakukan klasifikasi terhadap dataset yang tidak seimbang, khususnya terhadap data dalam kelas Angkutan dan Terminal, yang merupakan kelas minoritas.

Penelitian berikutnya dapat membahas hal-hal yang belum diteliti dalam penelitian kali ini. Contohnya, penelitian berikutnya dapat dilakukan menggunakan metode penanganan ketidakseimbangan data yang lain dalam klasifikasi pengaduan, seperti metode *Random Over-Sampling*, *Edited Nearest Neighbors*, atau metode-metode lainnya. Selain itu, penelitian berikutnya juga dapat dilakukan menggunakan metode klasifikasi yang lain untuk mendapatkan performa model yang lebih baik. Terakhir, penelitian berikutnya juga dapat berfokus pada mengembangkan algoritma perbaikan ejaan kata sehingga perbaikan ejaan kata dapat menjadi lebih akurat.

Penelitian yang dilakukan kali ini memiliki beberapa keterbatasan. Penelitian ini hanya berfokus pada penanganan ketidakseimbangan data pada klasifikasi pengaduan masyarakat di Dinas Perhubungan Kota Magelang menggunakan algoritma *Naïve Bayes*. Data yang digunakan dalam penelitian hanya mencakup 350 data pengaduan dari tanggal 20 Desember 2020 hingga 10 April 2025 yang diunduh dari aplikasi LAPOR!. Selain itu, metode penanganan ketidakseimbangan data yang digunakan hanya metode RUS, SMOTE, dan ADASYN. Terakhir, penelitian yang dilakukan merupakan penelitian eksperimental dan tidak mencakup perancangan aplikasi.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Dinas Perhubungan Kota Magelang atas izin yang diberikan dalam pelaksanaan penelitian ini serta data dan informasi yang telah diberikan untuk mendukung penelitian ini. Penulis juga mengucapkan terima kasih kepada pihak-pihak lain yang telah membantu penelitian serta penulisan artikel ilmiah ini.

DAFTAR PUSTAKA

[1] W. A. Setyarini, "Survei Kepuasan Masyarakat terhadap Pelayanan Pengaduan Masyarakat Lapor Hendi Tahun 2021," *J. Riptek*, vol. 16, no. 2, pp. 90–96, 2022, doi: 10.35475/riptek.v16i2.157.

[2] S. E. R. Putri Gunawan and D. Hertati, "Inovasi Pelayanan Pengaduan Masyarakat Melalui Aplikasi Wargaku Berbasis Android di Dinas Komunikasi dan Informatika Kota Surabaya," *J. Ilm. Univ. Batanghari Jambi*, vol. 22, no. 3, p. 1360, 2022, doi: 10.33087/jiubj.v22i3.2462.

[3] Y. Sansena, "Implementasi Sistem Layanan Pengaduan Masyarakat Kecamatan Medan Amplas Berbasis Website," *J. Ilm. Teknol. Inf. Asia*, vol. 15, no. 2, p. 91, 2021, doi: 10.32815/jitika.v15i2.611.

[4] T. Wijayanti, F. Nugraha, and A. P. Utomo, "Rancang Bangun Sistem Manajemen Pengelolaan Pengaduan Masyarakat Di Kabupaten Kudus," *J. Comput. Inf. Syst. Ampera*, vol. 3, no. 1, pp. 56–65, 2022, doi: 10.51519/journalcisa.v3i1.141.

[5] E. Meilinda, R. Sabaruddin, and D. Fitriani, "Model Prototype Sebagai Metode Pengembangan Perangkat Lunak Pada Sistem Informasi Pengaduan Umum," *J. Khatulistiwa Inform.*, vol. 9, no. 2, pp. 86–91, 2021.

[6] Sunarti, Ridwang, and M. A. M. Hayat, "Klasifikasi Pengaduan Pelayanan Fakultas Teknik Universitas Muhammadiyah Makassar menggunakan Natural Language Processing," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, pp. 572–579, 2024.

[7] I. G. N. A. Kusuma, I. M. Pradipta, I. M. A. Santosa, and I. M. A. Dharmendra, "Penanganan Ketidakseimbangan Data Pada Klasifikasi Pengaduan Masyarakat," *J. Teknol. Inf. dan Komput.*, vol. 9, no. 5, pp. 489–496, 2023, doi: 10.36002/jutik.v9i5.2643.

[8] F. Atmaja and E. D. Wahyuni, "Analisis Sentimen Berbasis Aspek pada Sistem Layanan Pengaduan Masyarakat di Kota Surabaya Menggunakan Metode Latent Dirichlet Allocation dan Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 527–534, 2025, doi: https://doi.org/10.36040/jati.v9i1.12438.

[9] D. Chrisinta and J. E. Simarmata, "Eksplorasi Teknik Web Scraping pada Data Mining: Pendekatan Pencarian Data Berbasis Python," *Fakt. Exacta*, vol. 17, no. 1, pp. 58–68, 2024, doi: 10.30998/faktorexacta.v17i1.22393.

[10] N. Nyoman Eny Perimawati, R. Rudolf Huizen, D. Pramana Hostiadi, and M. Sistem Informasi, "Analisa Pengaruh Pre-Procesing Data Untuk Model Deteksi Akun Palsu Pada Media Sosial," *Pros. Semin. Has. Penelit. Inform. dan Komput. Ed. Maret 2025*, vol. 2, no. 1, p. 2025, 2025.

[11] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023.

[12] Y. T. Handika, S. Defit, and G. W. Nurcahyo, "Text Mining dalam Membandingkan Metode Naïve Bayes dengan C.45 dalam Mengidentifikasi Berita Hoax pada Media Sosial," *Rang Tek. J.*, vol. 5, no. 1, pp. 116–123, 2022.

[13] L. Hermawati, V. Berland, A. Rahmadiah, E. Hutabarat, and D. D. Saputra, "Komparasi Metode Text Mining Terhadap Masalah Pengklasifikasian Narasi Informative & Non Informative Pada twitter @ PLN _ 123," *J. Sistim Inf. dan Teknol.*, vol. 5, no. 1, pp. 109–120, 2023, doi: 10.37034/jsisfotek.v4i2.191.

[14] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023, doi: 10.51454/decode.v3i2.131.

[15] M. H. Mahendra, D. T. Murdiansyah, and K. M. Lhaksmana, "Analisis Sentimen Tweet COVID-19 menggunakan K-Nearest Neighbors dengan TF-IDF dan Ekstraksi Fitur CountVectorizer," *DIKE J. Ilmu Multidisiplin*, vol. 1, no. 2, pp. 37–43, 2023, doi: 10.69688/dike.v1i2.35.

[16] K. Tri Putra, M. Amin Hariyadi, and C. Crysdian, "Perbandingan Feature Extraction TF-IDF Dan BOW Untuk Analisis Sentimen Berbasis SVM," *J. Cahaya Mandalika*, vol. 3, no. 2, p. 1449, 2023.

[17] K. Ishak, "Understanding Data Leakage in Machine Learning: A Focus on TF-IDF," *Summer of Reproducibility 24, UC Santa Cruz OSPO*, 2024. https://ucsc-ospo.github.io/report/osre24/nyu/data-leakage/20240905-kyrillosishak/#:~:text=How Data Leakage Occurs with,IDF (accessed May 21, 2025).

[18] S. Wehnert, V. Sudhi, S. Dureja, L. Kutty, S. Shahania, and E. W. De Luca, "Legal Norm Retrieval with Variations of the BERT Model Combined with TF-IDF Vectorization," *Proc. 18th Int. Conf. Artif. Intell. Law, ICAIL 2021*, pp. 285–294, 2021, doi: 10.1145/3462757.3466104.

[19] C. Yang, R. A. Brower-Sinning, G. Lewis, and C. Kästner, "Data Leakage in Notebooks: Static Detection and Better Processes," *ACM Int. Conf. Proceeding Ser.*, 2022, doi: 10.1145/3551349.3556918.

[20] M. Anjas Aprihartha, D. Zulhan, A. F. Nurfaizal, and T. Nur Alam, "Penyelesaian Masalah Ketidakseimbangan Data Melalui Teknik Oversampling dan Undersampling pada Klasifikasi Siswa Tidak Naik Kelas," *J. Tek. Ibnu Sina*, vol. 9, no. 01, pp. 43–52, 2024.

[21] Y. A. Sir and A. H. H. Soepranoto, "Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas," *J. Komput. dan Inform.*, vol. 10, no. 1, pp. 31–38, 2022, doi: 10.35508/jicon.v10i1.6554.

[22] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *Sistemasi*, vol. 10, no. 2, p. 445, 2021, doi: 10.32520/stmsi.v10i2.1303.

[23] M. 'Ainur Rasyida and M. F. Rizal, "Optimalisasi Klasifikasi Disposisi Pengaduan Masyarakat melalui Kanal

LAPOR Menggunakan Algoritma Naïve Bayes dan Integrasi Ekstensi Chrome," *Integr. Perspect. Soc. Sci. J.*, vol. 2, no. 3, pp. 4115–4121, 2025.

[24] A. Fatkhudin, F. A. Artanto, N. A. Safli, and D. Wibowo, "Decision Tree Berbasis SMOTE Dalam Analisis Sentimen Penggunaan Artificial Intelligence Untuk Skripsi," *REMIK Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 8, no. April, pp. 494–505, 2024, [Online]. Available: https://www.jurnal.polgan.ac.id/index.php/remik/article/view/13531%0Ahttps://www.jurnal.polgan.ac.id/index.php/remik/article/download/13531/2453

[25] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[26] K. Pramayasa, I. M. D. Maysanjaya, and I. G. A. A. D. Indradewi, "Analisis Sentimen Program MBKM Pada Media Sosial Twitter Menggunakan KNN Dan SMOTE," *SINTECH (Science Inf. Technol. J.*, vol. 6, no. 2, pp. 89–98, 2023, doi: 10.31598/sintechjournal.v6i2.1372.

[27] I. Pratama, A. Y. Chandra, and P. T. Presetyaningrum, "Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN," *J. Eksplora Inform.*, vol. 11, no. 1, pp. 38–49, 2022, doi: 10.30864/eksplora.v11i1.578.

[28] D. V. Ramadhanti, R. Santoso, and T. Widiharih, "Perbandingan Smote Dan Adasyn Pada Data Imbalance Untuk Klasifikasi Rumah Tangga Miskin Di Kabupaten Temanggung Dengan Algoritma K-Nearest Neighbor," *J. Gaussian*, vol. 11, no. 4, pp. 499–505, 2023, doi: 10.14710/j.gauss.11.4.499-505.

[29] I. K. Dharmendra, I. M. Agus, W. Putra, and Y. P. Atmojo, "Evaluasi Efektivitas SMOTE dan Random Under Sampling pada Klasifikasi Emosi Tweet," *Informatics Educ. Prof. J. Informatics*, vol. 9, no. 2, pp. 192–193, 2024, doi: https://doi.org/10.51211/itbi.v9i2.3183.

[30] S. Kabane, "Impact of Sampling Techniques and Data Leakage on XGBoost Performance in Credit Card Fraud Detection," *Mach. Learn.*, pp. 1–19, 2024, [Online]. Available: http://arxiv.org/abs/2412.07437

[31] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 68–75, 2022, doi: 10.57152/malcom.v2i2.432.

[32] F. Septianingrum and A. S. Y. Irawan, "Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 799, 2021, doi: 10.30865/mib.v5i3.2983.

[33] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *J. Inform.*, vol. 9, no. 1, pp. 13–17, 2022, doi: 10.31294/inf.v9i1.11931.

[34] G. Ahmed *et al.*, "DAD-Net: Classification of Alzheimer's Disease Using ADASYN Oversampling Technique and Optimized Neural Network," *Molecules*, vol. 27, no. 20, pp. 1–21, 2022, doi: 10.3390/molecules27207085.

[35] N. D. Primadya, A. Nugraha, S. Y. Fahrezi, and A. Luthfiarta, "Optimizing Imbalanced Data Classification: Under Sampling Algorithm Strategy with Classification Combination," *Techné J. Ilm. Elektrotek.*, vol. 23, no. 2, pp. 277–288, 2024, doi: 10.31358/techne.v23i2.435.

[36] M. A. N. Anargya, W. Ghozi, and F. A. Rafrastara, "Random Under Sampling for Performance Improvement in Attack Detection on Internet of Vehicles Using Machine Learning," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 11–19, 2025, doi: 10.30591/jpit.v10i1.8034.