Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 3, Maret 2025, Hal. 615-623
DOI: https://doi.org/10.52436/1.jpti.796
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Perbandingan Kinerja Metode Binary Relevance, Classifier Chains, dan Label Powerset dalam Klasifikasi Multi-Label Data Pengaduan

Denny Ariyana[*1], Eka Dyar Wahyuni[2], Nambi Sembilu[3]

[1,2,3]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia
Email: [1]21082010040@student.upnjatim.ac.id, [2]ekawahyuni.si@upnjatim.ac.id, [3]nambi.si@upnjatim.ac.id

Abstrak

Di era digital, aplikasi pengaduan masyarakat menjadi sarana penting dalam meningkatkan komunikasi antara warga dan pemerintah. Aplikasi Wargaku memungkinkan masyarakat menyampaikan keluhan terkait layanan publik, yang menghasilkan data pengaduan bersifat multi-label. Oleh karena itu, diperlukan metode klasifikasi yang optimal untuk meningkatkan akurasi dalam pengelolaan pengaduan. Penelitian ini bertujuan untuk membandingkan tiga metode klasifikasi multi-label, yaitu Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP) dengan Random Forest sebagai base classifier. Metode penelitian mengacu pada kerangka CRISP-DM, yang mencakup pemahaman bisnis, pengolahan data, pemodelan, dan evaluasi. Eksperimen dilakukan dengan skenario pembagian data 80:20, 70:30, dan 60:40, serta preprocessing yang mencakup pembersihan teks dan normalisasi. Evaluasi model menggunakan F1 Score untuk mengukur kinerja klasifikasi. Hasil penelitian menunjukkan bahwa Binary Relevance memiliki performa paling stabil di berbagai skenario. Pada skenario 70:30, metode ini mencapai skor F1 tertinggi sebesar 0,76, diikuti oleh Classifier Chains (0,75) dan Label Powerset (0,74). Pada skenario 80:20, Label Powerset unggul dengan skor 0,75, sedangkan Binary Relevance dan Classifier Chains memperoleh 0,75 dan 0,73. Sementara itu, pada skenario 60:40, Binary Relevance kembali unggul dengan skor 0,74, diikuti Label Powerset (0,74) dan Classifier Chains (0,73). Penelitian ini menunjukkan bahwa tidak ada perbedaan signifikan dalam performa metode, namun Binary Relevance dengan Random Forest cenderung lebih stabil di berbagai skenario. Hasil ini dapat digunakan untuk meningkatkan efektivitas klasifikasi pengaduan masyarakat, sehingga mendukung layanan publik yang lebih responsif dan efisien.

Kata kunci: *aplikasi wargaku, binary relevance, classifier chains, F1 score, klasifikasi multi-label, label powerset, pengaduan masyarakat*

Comparison of Evaluation Results of Binary Relevance, Classifier Chains, and Label Powerset Algorithms for Multi-Label Classification of Complaint Data in the Wargaku Application 2023

Abstract

In the digital era, citizen complaint applications serve as a crucial medium for communication between citizens and the government. Wargaku, an application launched in 2023, allows the public to submit complaints regarding public services. The complaint data is multi-label, requiring an optimal classification method. This study compares three multi-label classification approaches: Binary Relevance (BR), Classifier Chains (CC), and Label Powerset (LP), using Random Forest as the base classifier. The research follows the CRISP-DM framework, encompassing business understanding, data processing, modeling, and evaluation. The data is split into 80:20, 70:30, and 60:40 scenarios, with preprocessing including text cleaning and normalization. Model evaluation is conducted using the F1 Score metric. The results show that Binary Relevance performs best in the 70:30 scenario, achieving the highest F1 Score of 0.758, followed by Classifier Chains (0.748) and Label Powerset (0.737). In the 80:20 scenario, Label Powerset outperforms the others with an F1 Score of 0.748, while Binary Relevance and Classifier Chains achieve 0.745 and 0.732, respectively. In the 60:40 scenario, Binary Relevance remains superior with an F1 Score of 0.744, followed by Label Powerset (0.739) and Classifier Chains (0.733). In conclusion, Binary Relevance with Random Forest demonstrates the most stable performance across different scenarios. However, the differences in F1 Scores among the methods are not significant, making the choice of approach dependent on data complexity and system requirements. This study is expected to enhance the effectiveness of complaint classification to support more responsive public services.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Keywords: *aplikasi wargaku, binary relevance, classifier chains, F1 score, klasifikasi multi-label, label powerset, pengaduan masyarakat*

1. PENDAHULUAN

Di era digital saat ini, keberadaan aplikasi pengaduan masyarakat menjadi sarana penting dalam memfasilitasi komunikasi antara warga dan pemerintah. Salah satu platform yang mendukung hal ini adalah Aplikasi Wargaku, yang mulai beroperasi pada tahun 2023. Aplikasi ini memungkinkan masyarakat untuk menyampaikan keluhan dan saran terkait layanan publik. Data yang dihasilkan dari laporan pengaduan bersifat multi-label, artinya satu laporan dapat dikategorikan ke dalam beberapa jenis layanan sekaligus. Oleh sebab itu, diperlukan teknik klasifikasi yang optimal untuk mengelompokkan pengaduan berdasarkan kategori yang relevan [1]. Namun, Aplikasi Wargaku masih menghadapi kendala dalam mengklasifikasikan pengaduan masyarakat secara efektif, seperti ketidakmampuan sistem untuk menangani hubungan antar label yang kompleks, yang dapat menyebabkan ketidakakuratan dalam klasifikasi [7]. Misalnya, sebuah laporan pengaduan tentang jalan rusak dan lampu jalan mati harus diklasifikasikan ke dalam dua kategori yang berbeda, namun sistem seringkali hanya mampu mengklasifikasikan ke dalam satu kategori saja [8]. Hal ini menunjukkan urgensi untuk mengembangkan metode klasifikasi yang lebih canggih dan akurat.

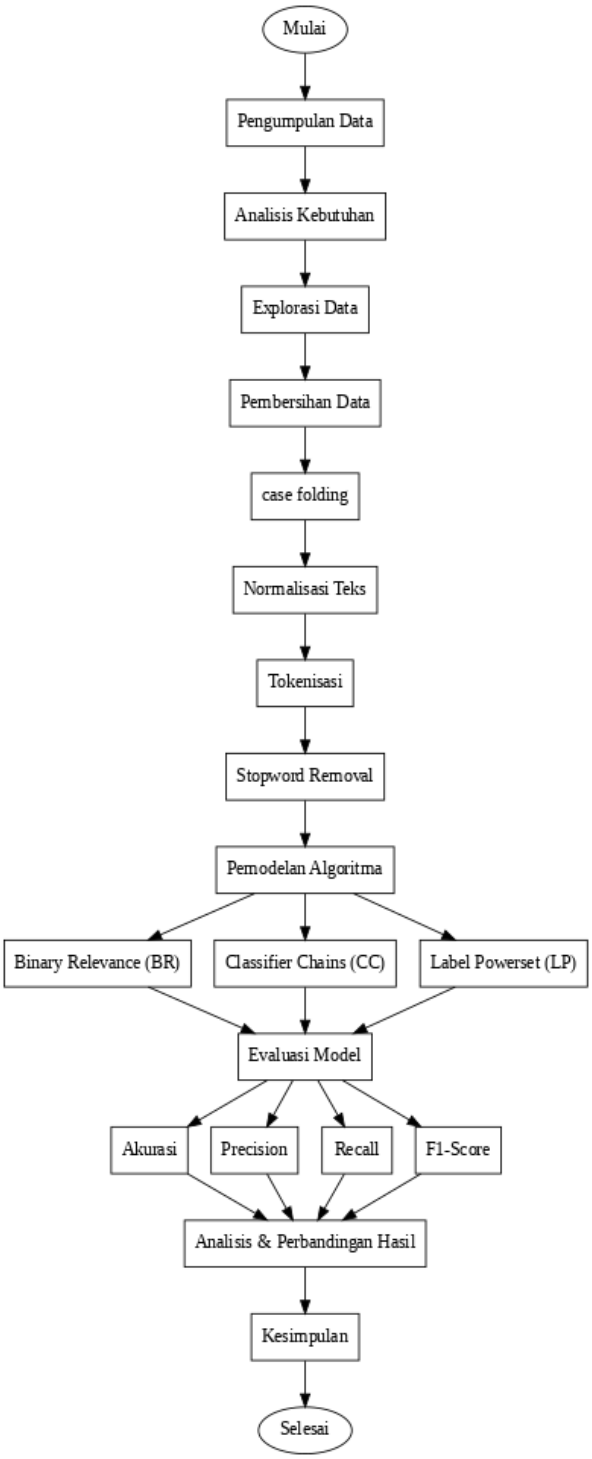
Dalam klasifikasi multi-label, terdapat tiga pendekatan utama, yaitu Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP). Metode Binary Relevance mengubah masalah multi-label menjadi serangkaian klasifikasi biner yang berdiri sendiri tanpa memperhitungkan keterkaitan antar label. Meskipun metode ini cukup sederhana, kelemahannya adalah tidak mempertimbangkan hubungan antar label yang mungkin memengaruhi akurasi prediksi [2]. Sementara itu, Classifier Chains memperbaiki kekurangan tersebut dengan membentuk rantai prediksi, di mana label yang telah diklasifikasikan sebelumnya mempengaruhi prediksi label berikutnya. Teknik ini memperhitungkan ketergantungan antar label sehingga sering kali menunjukkan performa yang lebih baik dalam beberapa penelitian [3]. Adapun metode Label Powerset, pendekatan ini mengubah permasalahan multi-label menjadi multi-kelas dengan menjadikan setiap kombinasi label sebagai satu kelas tersendiri. Namun, pendekatan ini kurang efisien ketika jumlah kombinasi label semakin banyak, karena dapat meningkatkan kompleksitas pemrosesan data secara signifikan [4]. Oleh karena itu, pemilihan metode yang tepat sangat penting untuk memastikan akurasi dan efisiensi dalam klasifikasi data pengaduan masyarakat.

Sejumlah penelitian telah dilakukan untuk mengevaluasi efektivitas metode klasifikasi multi-label dalam berbagai konteks. Hasil studi yang dilakukan oleh Read et al. [5] menunjukkan bahwa pendekatan Classifier Chains sering kali lebih unggul dibandingkan Binary Relevance, terutama dalam skenario yang memiliki hubungan kuat antar label. Studi lainnya mengungkapkan bahwa meskipun Label Powerset dapat memberikan hasil yang optimal pada dataset dengan jumlah label yang kecil, performanya cenderung menurun ketika diterapkan pada dataset yang lebih kompleks [6]. Selain itu, penelitian oleh Zhang et al. [9] mengusulkan penggunaan seleksi fitur spesifik label untuk meningkatkan akurasi klasifikasi multi-label, sementara Hanafi et al. [10] mengeksplorasi penggunaan Mutual Information dan k-Nearest Neighbor dalam klasifikasi multi-label pada teks terjemahan. Namun, masih terdapat celah penelitian dalam konteks klasifikasi multi-label untuk data pengaduan masyarakat, terutama dalam hal menangani dataset yang besar dan kompleks dengan hubungan antar label yang tidak terduga.

Penelitian ini berfokus pada analisis dan perbandingan kinerja algoritma Binary Relevance, Classifier Chains, dan Label Powerset dalam mengklasifikasikan data pengaduan masyarakat yang diperoleh dari Aplikasi Wargaku pada tahun 2023. Melalui evaluasi ini, diharapkan dapat diperoleh wawasan yang lebih mendalam mengenai metode yang paling efektif dalam menangani klasifikasi multi-label untuk data pengaduan, sehingga dapat membantu meningkatkan respons dan kualitas layanan publik kepada masyarakat. Penelitian ini juga memberikan kontribusi dengan menguji ketiga metode klasifikasi multi-label pada dataset baru yang diperoleh dari Aplikasi Wargaku, serta mengusulkan pendekatan evaluasi yang berbeda dengan mempertimbangkan kompleksitas dan ketergantungan antar label dalam dataset pengaduan masyarakat. Tujuan penelitian ini adalah untuk menganalisis dan membandingkan kinerja algoritma Binary Relevance, Classifier Chains, dan Label Powerset dalam mengklasifikasikan data pengaduan masyarakat yang bersifat multi-label, sehingga dapat memberikan rekomendasi metode klasifikasi yang paling efektif untuk meningkatkan akurasi dan efisiensi dalam penanganan pengaduan masyarakat.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2. METODE PENELITIAN



Gambar 1. Metodologi Penelitian

Penelitian ini dilakukan secara sistematis dengan beberapa tahapan untuk memastikan keakuratan dalam klasifikasi multi-label terhadap data pengaduan di Dinas Kependudukan Kota Surabaya. Tahapan pertama dimulai dengan pengumpulan data yang diperoleh dari sumber resmi Dinas Kependudukan Kota Surabaya. Dataset yang digunakan terdiri dari 10.000 laporan pengaduan masyarakat yang dikumpulkan selama tahun 2023. Data ini mencakup berbagai kategori layanan publik seperti infrastruktur, kesehatan, pendidikan, dan administrasi kependudukan. Setelah pengumpulan data, tahap analisis kebutuhan dilakukan untuk memahami

p-ISSN: 2775-4227
e-ISSN: 2775-4219

karakteristik dataset, termasuk distribusi label, frekuensi kemunculan label, serta hubungan antar label. Analisis ini membantu dalam menentukan metode klasifikasi multi-label yang paling sesuai untuk diterapkan.

Setelah analisis kebutuhan, data dieksplorasi lebih lanjut untuk memahami pola dan distribusinya. Eksplorasi data meliputi visualisasi distribusi label, analisis frekuensi kata, serta identifikasi outlier atau data yang tidak lengkap. Tahap ini penting untuk memastikan bahwa data siap diproses lebih lanjut. Selanjutnya, data melalui proses pembersihan agar lebih siap untuk digunakan dalam pemodelan. Pembersihan data mencakup beberapa tahap penting, dimulai dari case folding untuk mengubah seluruh teks menjadi huruf kecil guna menjaga konsistensi. Selanjutnya, dilakukan normalisasi teks untuk menghilangkan karakter khusus, angka, serta menyamakan variasi kata dengan makna serupa (misalnya, kata "jalan" dan "jln" disamakan menjadi "jalan"). Setelah itu, proses tokenisasi dilakukan untuk memisahkan teks menjadi kata-kata individu, yang kemudian diikuti dengan stopword removal guna menghapus kata-kata umum yang tidak memiliki makna signifikan dalam klasifikasi, seperti "dan", "atau", serta kata penghubung lainnya.

Setelah data diproses dan dibersihkan, tahap berikutnya adalah pemodelan algoritma dengan menggunakan tiga pendekatan utama dalam klasifikasi multi-label, yaitu Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP). Pendekatan Binary Relevance mengubah setiap label menjadi masalah klasifikasi biner yang independen, sehingga setiap label diprediksi secara terpisah. Sementara itu, Classifier Chains menggunakan prediksi label sebelumnya sebagai fitur tambahan untuk label berikutnya, sehingga mempertimbangkan ketergantungan antar label. Label Powerset mengubah masalah multi-label menjadi klasifikasi multi-kelas dengan mempertimbangkan kombinasi label yang muncul dalam data. Untuk implementasi, penelitian ini menggunakan Python dengan bantuan library seperti Scikit-learn untuk pemodelan, Pandas untuk manipulasi data, dan NLTK untuk pemrosesan teks. Selain itu, Google Colab digunakan sebagai platform komputasi untuk memastikan konsistensi dan kemudahan dalam menjalankan eksperimen.

Model yang telah dibangun kemudian dievaluasi menggunakan beberapa metrik utama dalam klasifikasi multi-label, yaitu akurasi, precision, recall, dan F1-score. Akurasi digunakan untuk mengukur sejauh mana prediksi model sesuai dengan label yang sebenarnya, sementara precision menunjukkan proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif yang dibuat oleh model. Recall mengukur sejauh mana model mampu menemukan semua label yang benar dalam dataset, dan F1-score digunakan sebagai metrik keseimbangan antara precision dan recall. Selain itu, uji validitas dan reliabilitas data dilakukan untuk memastikan bahwa dataset yang digunakan konsisten dan representatif. Uji validitas dilakukan dengan memeriksa apakah data yang digunakan benar-benar mencerminkan masalah yang ingin dipecahkan, sementara uji reliabilitas dilakukan dengan memastikan bahwa hasil klasifikasi konsisten ketika model dijalankan pada dataset yang sama.

Setelah evaluasi dilakukan, hasil dari berbagai pendekatan algoritma dibandingkan untuk menentukan metode terbaik dalam klasifikasi multi-label pengaduan masyarakat. Tahap terakhir dari penelitian ini adalah menarik kesimpulan berdasarkan hasil analisis dan evaluasi model, serta memberikan rekomendasi untuk penelitian selanjutnya agar akurasi dan efisiensi model dapat ditingkatkan. Dengan demikian, penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem klasifikasi otomatis yang lebih akurat dan efisien dalam pengelolaan pengaduan masyarakat di Dinas Kependudukan Kota Surabaya.

3. HASIL DAN PEMBAHASAN

3.1. Hasil

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini terdiri dari 691 data pengaduan yang diperoleh dari Aplikasi Wargaku tahun 2023. Data ini berasal dari Dinas Kependudukan Kota Surabaya dan mencakup berbagai keluhan masyarakat terkait layanan administrasi kependudukan. Data yang diterima dalam format PDF kemudian dikonversi untuk keperluan dokumentasi dan analisis. Informasi dalam dataset mencerminkan variasi pengaduan yang bersifat multi-label, sehingga diperlukan metode klasifikasi yang optimal untuk mengolah dan mengevaluasi data tersebut sesuai dengan tujuan penelitian.

2. Analisis Kebutuhan

Pada tahap analisis kebutuhan, dilakukan identifikasi terhadap aspek yang diperlukan dalam penelitian, baik dari segi data maupun sistem yang digunakan. Data yang dibutuhkan dalam penelitian ini merupakan data pengaduan dari aplikasi Wargaku Kota Surabaya selama periode Januari hingga Desember 2023. Data tersebut akan diproses dan diklasifikasikan berdasarkan label yang telah ditentukan, sehingga dapat membantu Dinas Kependudukan dalam proses pelabelan untuk keperluan penyusunan laporan bulanan yang disampaikan kepada Kepala Bidang Pengelolaan Informasi Administrasi Kependudukan. Sementara itu, dari sisi kebutuhan sistem, penelitian ini memanfaatkan kombinasi perangkat keras dan perangkat lunak yang mendukung pengembangan

p-ISSN: 2775-4227
e-ISSN: 2775-4219

aplikasi berbasis web. Perangkat lunak yang digunakan meliputi Google Colab untuk pengolahan data, Visual Studio sebagai lingkungan pengembangan dengan bahasa pemrograman HTML dan CSS, serta framework Flask untuk membangun sistem. Adapun perangkat keras yang digunakan adalah laptop Asus Vivobook dengan spesifikasi prosesor AMD Ryzen 5, RAM 8 GB, dan sistem operasi Windows 11. Dengan pemenuhan kebutuhan data dan sistem ini, penelitian dapat berjalan dengan optimal dalam mencapai tujuan klasifikasi data pengaduan.

3. Eksplorasi Data

Subbagian ini membahas analisis eksplorasi data, termasuk visualisasi word cloud untuk label. Dengan pendekatan ini, kita dapat memahami pola dan tren yang muncul dalam data keluhan, yang akan membantu dalam proses pelabelan dan analisis lebih lanjut.

Gambar 2. Word Cloud Labels

Proses word cloud labels adalah mempermudah identifikasi kata-kata atau topik yang paling sering muncul dalam suatu kumpulan data, sehingga dapat memberikan gambaran visual yang cepat dan intuitif. Dengan word cloud, analisis terhadap data menjadi lebih efisien karena kata-kata yang paling relevan atau sering dilaporkan akan langsung terlihat dari ukuran dan posisi teksnya. Hal ini membantu dalam pengambilan keputusan yang lebih tepat, mengarahkan fokus pada isu-isu utama, dan mempercepat proses analisis data yang kompleks.

4. Cleaning

Pada tahap ini, elemen-elemen yang tidak diperlukan, seperti karakter khusus, tanda baca yang tidak relevan, dan simbol yang tidak memiliki makna dalam analisis teks, dihilangkan guna memastikan kelancaran proses pengolahan data berikutnya.

Tabel 1. Proses *Cleaning*

Keluhan Data Asli	**Keluhan Setelah *Cleaning***
Selamat Pagi, \n\nsaya ingin bertanya bagaimana prosesnya untuk pembuatan akta perkawinan? apakah bisa melalui website atau harus datang ke kantor kecamatan/keluarahan? mohon bantuan dan bimbingannya secara terperinci langkahlangkahnya, karena saya tidak paham. Terima kasih.	Selamat Pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham Terima kasih

5. Case Folding

Case folding dilakukan untuk menyamakan format huruf dalam teks dengan mengonversi seluruh karakter menjadi huruf kecil. Tahap ini bertujuan untuk menghindari perbedaan yang tidak signifikan antara kata-kata yang memiliki makna serupa tetapi ditulis dengan variasi huruf kapital yang berbeda, seperti "KTP" dan "ktp".

Tabel 2. Proses *Case Folding*

Keluhan Setelah *Cleaning*	**Keluhan Setelah *Case Folding***
Selamat Pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham Terima kasih	selamat pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham terima kasih

6. Normalisasi

Normalisasi bertujuan untuk menyelaraskan berbagai variasi kata yang memiliki makna serupa, seperti mengonversi "ktp-el" menjadi "ktp elektronik" atau membenarkan kesalahan ejaan dalam teks agar lebih konsisten dan mudah diproses.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 3. Proses Normalisasi

Keluhan Setelah Case Folding	Keluhan Setelah Normalized
selamat pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham terima kasih	selamat pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham terima kasih

7. Tokenezing

Proses tokenisasi dilakukan untuk membagi teks menjadi unit-unit kecil berupa kata atau token. Dalam analisis teks, tokenisasi berperan dalam mengidentifikasi setiap kata secara terpisah, sehingga dapat digunakan sebagai dasar untuk tahap pemrosesan selanjutnya.

Tabel 4. Proses Tokenezing

Keluhan Setelah Normalized	Keluhan Setelah Tokenizing
selamat pagi saya ingin bertanya bagaimana prosesnya untuk pembuatan akta apakah bisa melalui website atau harus datang ke kantor mohon bantuan dan secara langkah dua karena saya tidak paham terima kasih	['selamat', 'pagi', 'saya', 'ingin', 'bertanya', 'bagaimana', 'prosesnya', 'untuk', 'pembuatan', 'akta', 'apakah', 'bisa', 'melalui', 'website', 'atau', 'harus', 'datang', 'ke', 'kantor', 'mohon', 'bantuan', 'dan', 'secara', 'langkah', 'dua', 'karena', 'saya', 'tidak', 'paham', 'terima', 'kasih']

8. Stopword Removal

Stopword Removal dilakukan untuk menyaring kata-kata yang dianggap kurang relevan dalam analisis teks.

Tabel 5. Proses Filtering

Keluhan Setelah Tokenizing	Keluhan Setelah Filltering
['selamat', 'pagi', 'saya', 'ingin', 'bertanya', 'bagaimana', 'prosesnya', 'untuk', 'pembuatan', 'akta', 'apakah', 'bisa', 'melalui', 'website', 'atau', 'harus', 'datang', 'ke', 'kantor', 'mohon', 'bantuan', 'dan', 'secara', 'langkah', 'dua', 'karena', 'saya', 'tidak', 'paham', 'terima', 'kasih']	['selamat', 'pagi', 'saya', 'ingin', 'bertanya', 'bagaimana', 'prosesnya', 'untuk', 'pembuatan', 'akta', 'apakah', 'bisa', 'melalui', 'website', 'atau', 'harus', 'datang', 'ke', 'kantor', 'mohon', 'bantuan', 'dan', 'secara', 'langkah', 'dua', 'karena', 'saya', 'tidak', 'paham', 'terima', 'kasih']

9. Modeling

Pada tahap modeling, dilakukan implementasi dan evaluasi tiga pendekatan utama dalam klasifikasi multi-label, yaitu Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP) dengan algoritma Random Forest sebagai base classifier. Setiap metode memiliki karakteristik yang berbeda dalam menangani hubungan antar label dalam data pengaduan aplikasi Wargaku. Binary Relevance bekerja dengan membagi tugas klasifikasi menjadi beberapa model single-label, sehingga lebih fleksibel tetapi tidak mempertimbangkan korelasi antar label. Classifier Chains menggabungkan hasil prediksi sebelumnya sebagai fitur tambahan, yang memungkinkan pemodelan hubungan antar label secara berurutan. Sementara itu, Label Powerset memperlakukan kombinasi label yang ada sebagai satu kelas baru, sehingga mampu menangkap keterkaitan antar label tetapi rentan terhadap data yang jarang muncul (imbalanced class).

Proses pelatihan model dilakukan dengan menggunakan berbagai skenario pembagian data, yaitu 80:20, 70:30, dan 60:40 untuk melihat bagaimana perubahan proporsi data latih dan data uji mempengaruhi performa model. Setiap model dievaluasi menggunakan metrik F1 Score, yang merupakan ukuran keseimbangan antara precision dan recall dalam menangani klasifikasi multi-label. Hasil evaluasi menunjukkan bahwa model Binary Relevance Random Forest memiliki kinerja paling stabil dan unggul pada pembagian data 70:30 dengan F1 Score tertinggi sebesar 0.758. Model Label Powerset memiliki performa yang kompetitif, terutama pada pembagian data 80:20, sementara Classifier Chains cenderung memiliki F1 Score lebih rendah dibandingkan dua metode lainnya di semua skenario.

10. Evaluation

Evaluasi dilakukan untuk mengukur kinerja model dalam skenario klasifikasi multi-label dengan membandingkan tiga pendekatan utama, yaitu Binary Relevance, Classifier Chains, dan Label Powerset. Pada tahap ini, analisis dilakukan dengan membandingkan hasil evaluasi dari masing-masing algoritma, serta menampilkan visualisasi hasil klasifikasi. Evaluasi ini bertujuan untuk menilai efektivitas dan akurasi model dalam mengklasifikasikan data pengaduan dari aplikasi Wargaku tahun 2023. Berikut adalah hasil evaluasi yang diperoleh:

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 6. Tabel Confusion Matriks

Pecah Data	*Model*	*Accuracy*	*Precision*	Recall	F1 *Score*
80:20	BR Random Forest	0.485	0.830	0.676	0.745
	CC Random Forest	0.475	0.814	0.665	0.732
	LP Random Forest	0.53	0.815	0.692	0.748
70:30	BR Random Forest	0.48	0.845	0.687	0.758
	CC Random Forest	0.5	0.839	0.675	0.748
	LP Random Forest	0.52	0.806	0.678	0.737
60:40	BR Random Forest	0.477	0.843	0.665	0.744
	CC Random Forest	0.492	0.840	0.651	0.733
	LP Random Forest	0.535	0.818	0.675	0.739

Tabel di atas menyajikan hasil evaluasi performa model klasifikasi multi-label menggunakan algoritma Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP) dengan metode Random Forest pada berbagai skenario pembagian data. Pembagian data dilakukan dengan rasio 80:20, 70:30, dan 60:40, di mana setiap model dievaluasi berdasarkan metrik akurasi, presisi, recall, dan F1-score.

Pada rasio 80:20, model LP Random Forest menunjukkan performa terbaik dengan akurasi 0.53 dan F1-score 0.748, diikuti oleh BR Random Forest dengan akurasi 0.485 dan F1-score 0.745, serta CC Random Forest dengan akurasi 0.475 dan F1-score 0.732.

Untuk rasio 70:30, model CC Random Forest memperoleh akurasi tertinggi sebesar 0.5, sedangkan LP Random Forest memiliki akurasi 0.52 dan BR Random Forest mencapai 0.48. Dari segi F1-score, BR Random Forest menunjukkan nilai tertinggi sebesar 0.758, yang mengindikasikan keseimbangan antara precision dan recall yang lebih baik dibandingkan model lainnya.

Pada rasio 60:40, LP Random Forest kembali mencatatkan akurasi tertinggi sebesar 0.535, diikuti oleh CC Random Forest dengan 0.492 dan BR Random Forest dengan 0.477. Namun, dalam metrik F1-score, BR Random Forest memperoleh skor tertinggi sebesar 0.744, sedikit lebih unggul dibandingkan LP Random Forest yang mendapatkan 0.739.

Hasil perbandingan label yang di hasilkan oleh model yang paling bagus akan di sajikan dalam tabel di bawah ini:

Tabel 7. Hasil Perbandingan Label Actual dan Prediksi BR Random Forest

Keluhan	Label Actual	Label Prediksi
Processed_Text_522	0, 0, 1, 0, 0	0, 0, 1, 0, 0
Processed_Text_738	0, 1, 0, 0, 1	0, 1, 1, 0, 1
Processed_Text_741	1, 1, 1, 0, 0	0, 1, 1, 0, 0

Tabel 7 menunjukkan hasil perbandingan antara label aktual dan label prediksi menggunakan metode Binary Relevance dengan algoritma Random Forest pada data keluhan yang telah diproses (Processed_Text). Data keluhan ini merupakan hasil dari teks asli yang telah melalui tahapan preprocessing, seperti penghapusan karakter khusus, stopword removal, stemming, dan tokenisasi, sehingga teks menjadi lebih bersih dan siap digunakan dalam model klasifikasi.

Sebagai contoh, Processed_Text_522 berasal dari keluhan yang setelah diproses tetap mempertahankan informasi penting, menghasilkan label aktual (0, 0, 1, 0, 0). Model berhasil memprediksi label ini dengan sangat baik, menunjukkan kemampuannya dalam mengenali pola yang sesuai dengan data yang telah dilatih.

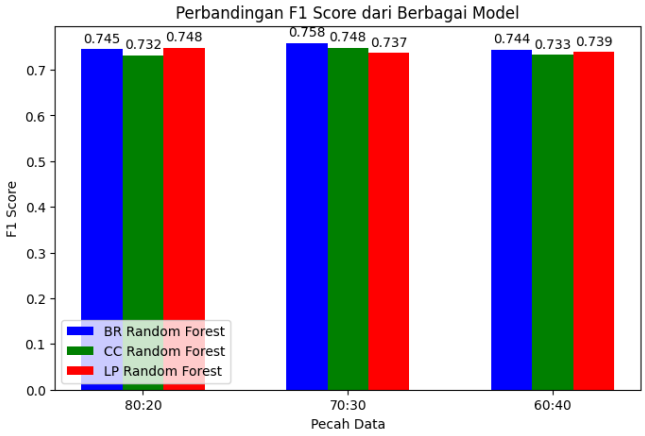
Pada Processed_Text_738, model memprediksi label (0, 1, 1, 0, 1), sementara label aktualnya adalah (0, 1, 0, 0, 1). Meskipun terdapat perbedaan pada label ketiga, model tetap mampu menangkap sebagian besar pola yang ada. Hal ini menunjukkan bahwa model memiliki sensitivitas yang baik terhadap fitur yang relevan, meskipun dalam beberapa kasus, variasi dalam data membuat prediksi sedikit berbeda.

Sementara itu, Processed_Text_741 memiliki label aktual (1, 1, 1, 0, 0), tetapi model memprediksi (0, 1, 1, 0, 0). Perbedaan ini menunjukkan bahwa dalam beberapa kasus, model mungkin lebih konservatif dalam memberikan label positif, yang bisa jadi merupakan strategi untuk menghindari overfitting. Namun, prediksi pada label lainnya tetap konsisten dengan label aktual, yang menunjukkan bahwa model mampu memahami struktur data dengan baik.

Secara keseluruhan, hasil ini mencerminkan bahwa model Random Forest dengan pendekatan Binary Relevance memiliki performa yang cukup baik dalam mengenali pola label pada data pengaduan. Meskipun terdapat beberapa perbedaan dalam prediksi, hal ini bisa disebabkan oleh kompleksitas data dan variasi fitur dalam teks pengaduan. Dengan optimasi lebih lanjut, seperti penyesuaian parameter atau peningkatan fitur, model ini berpotensi memberikan akurasi yang lebih tinggi dalam klasifikasi multi-label.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

3.2. Analisis Dan Perbandingan



Gambar 3. Grafik Perbandingan F1 Score

F1 Score dipilih sebagai metrik evaluasi utama dalam penelitian ini karena mampu memberikan keseimbangan antara Precision dan Recall. Dalam konteks klasifikasi multi-label pada data pengaduan aplikasi Wargaku, Precision menunjukkan seberapa akurat model dalam mengidentifikasi label yang benar, sedangkan Recall mengukur sejauh mana model dapat mengenali semua label yang seharusnya diklasifikasikan. Mengingat bahwa data pengaduan memiliki berbagai label yang harus diklasifikasikan secara akurat dan lengkap, F1 Score menjadi metrik yang lebih representatif dibandingkan Accuracy yang hanya mengukur proporsi prediksi yang benar tanpa mempertimbangkan ketidakseimbangan antara Precision dan Recall.

Berdasarkan hasil evaluasi yang disajikan dalam tabel, model dengan pendekatan Binary Relevance, Classifier Chains, dan Label Powerset dibandingkan menggunakan tiga skenario pembagian data yang berbeda (80:20, 70:30, dan 60:40). Dari hasil yang diperoleh, terlihat bahwa perbedaan nilai F1 Score antara model relatif kecil, namun Binary Relevance memiliki nilai yang cenderung lebih stabil di berbagai skenario pembagian data. Hal ini menunjukkan bahwa metode Binary Relevance cukup efektif dalam menangani klasifikasi multi-label pada data pengaduan aplikasi Wargaku.

4. KESIMPULAN

Penelitian ini membandingkan Binary Relevance (BR), Classifier Chains (CC), dan Label Powerset (LP) menggunakan Random Forest untuk klasifikasi multi-label data pengaduan pada Aplikasi Wargaku. Hasil evaluasi menunjukkan bahwa Binary Relevance memiliki performa paling stabil di berbagai skenario, dengan F1 Score tertinggi 0,76 pada pembagian data 70:30. Label Powerset unggul pada skenario 80:20 (F1 Score 0,75), sedangkan Classifier Chains cenderung memiliki performa lebih rendah dibandingkan dua metode lainnya.

Untuk implementasi sistem klasifikasi di dunia nyata, Binary Relevance dengan Random Forest direkomendasikan karena kestabilannya dalam menangani berbagai skenario data. Label Powerset bisa menjadi alternatif jika hubungan antar label dalam pengaduan perlu diperhitungkan. Sebaliknya, Classifier Chains kurang direkomendasikan karena performanya yang lebih rendah.

Penelitian selanjutnya dapat mengeksplorasi metode deep learning, ensemble learning, atau optimasi hyperparameter untuk meningkatkan akurasi klasifikasi. Selain itu, pengujian pada dataset yang lebih besar dan lebih beragam akan membantu memastikan generalisasi model dalam sistem pengaduan berbasis multi-label.

DAFTAR PUSTAKA

[1] J. Read, B. Pfahringer, G. Holmes, dan E. Frank, “Classifier Chains: A Review and Perspectives,” arXiv preprint arXiv:1912.13405, 2019.

[2] M. Arslan dan C. Cruz, “Imbalanced Multi-label Classification for Business-related Text with Moderately Large Label Spaces,” arXiv preprint arXiv:2306.07046, 2023.

[3] F J. Wainer, “Comparison of 14 different families of classification algorithms on 115 binary datasets,” arXiv preprint arXiv:1606.00930, 2016.

[4] N. B. Putri dan A. W. Wijayanto, “Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing,” Komputika: Jurnal Sistem Komputer, vol. 11, no. 1, pp. 59–66, 2022.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[5] J. Read, B. Pfahringer, G. Holmes, dan E. Frank, "Classifier Chains for Multi-label Classification," Machine Learning, vol. 85, no. 3, pp. 333–359, 2011.

[6] R. Alifarahman, "Klasifikasi Multi-label Dokumen Pertanyaan Medis dengan Pendekatan Berbagai Macam Teknik Machine Learning," Medium, 2020. [Online]. Available: https://riswandaali.medium.com/klasifikasi-multi-label-dengan-pendekatan-berbagai-macam-teknik-machine-learning-55d3bf8dee60.

[7] I. Akbar, M. Faisal, and T. Chamidy, "Kinetik: Game Technology, Information System," Computer Network, Computing, Electronics, and Control Journal, vol. 4, no. 3, pp. 119–128, 2019, [Online]. Available: https://kinetik.umm.ac.id/index.php/kinetik/article/view/1901

[8] J. Zhang, K. Liu, X. Yang, H. Ju, and S. Xu, "Multi-label learning with Relief-based label-specific feature selection," Applied Intelligence, vol. 53, no. 15, pp. 18517–18530, 2023, doi: 10.1007/s10489-022-04350-1.

[9] A. Hanafi, A. Adiwijaya, and W. Astuti, "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 9, no. 3, pp. 357–364, Sep. 2020, doi: 10.32736/sisfokom.v9i3.980.

[10] Manueke, "Implementation of Multi-Label Classification to Determine Scientific Articles Keyword in Journals," 2022. [Online]. Available: https://ejournal.unsrat.ac.id/index.php/informatika

[11] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 5, pp. 1774–1785, 2018.

[12] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1–13, 2007.

[13] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 8, pp. 1819–1837, 2014.

[14] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," Machine Learning, vol. 73, no. 2, pp. 133–153, 2008.

[15] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," Machine Learning, vol. 104, no. 1, pp. 55–98, 2016.

[16] Y. Zhang and J. G. Schneider, "Multi-label output codes using canonical correlation analysis," in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011, pp. 873–881.

[17] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in European Conference on Principles of Data Mining and Knowledge Discovery, 2001, pp. 42–53.

[18] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," Machine Learning, vol. 39, no. 2-3, pp. 135–168, 2000.

[19] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 995–1000.

[20] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," Pattern Recognition, vol. 37, no. 9, pp. 1757–1771, 2004.