P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Analisis Sentimen Ulasan Penumpang Maskapai *Low Cost Carrier* (LCC) Menggunakan Algoritma *XGBoost* Dan *Cosine Similarity*

[1]Noviana Rosianti, [2]Ionia Veritawati
[1,2]Program Studi Teknik Informatika, Universitas Pncasila, Jakarta

E-mail: [1]noviaanarst@gmail.com, [2]ionia.veritawati@univpancasila.ac.id

ABSTRAK

Indonesia menjadi negara dengan kapasitas penerbangan terbesar di ASEAN pada tahun 2025. Hal ini dapat terjadi dikarenakan Indonesia merupakan negara yang mempunyai lebih dari 17.000 pulau sehingga menjadikan transportasi sebagai sarana untuk menghubungkan antar wilayah di Indonesia. Salah satu transportasi udara yang dapat digunakan adalah pesawat. Di Indonesia *market share* pesawat didominasi dengan pesawat jenis *low cost carrier* (LCC). Menurut pengamat penerbangan, dominasi yang terjadi pada *market share* bukanlah sebuah monopoli melainkan karena pasar penerbangan domestik masih memiliki banyak pemain. Sehingga persaingan maskapai saat ini tidak terbatas pada harga melainkan pada faktor lain seperti keberagaman rute, jadwal penerbangan, dan persepsi penumpang terhadap kualitas layanan. Karena hal inilah diperlukan penilaian terhadap pelayanan yang diberikan. Oleh karena itu, penelitian ini dilakukan untuk membantu maskapai melakukan penilaian terhadap pelayanan yang telah diberikan, melalui analisis sentimen dari data *Citilink* dan *Lion Air* menggunakan metode *XGBoost* untuk klasifikasi kelas berdasarkan ulasan dari kedua maskapai tersebut. Penelitian ini memberikan hasil yaitu, analisis sentimen dapat diterapkan dengan menggunakan algoritma *XGBoost* baik untuk data Lion maupun data *Citilink*. Dari hasil tersebut data *Citilink* memperoleh pemodelan dengan hasil terbaik seperti akurasi sebesar 82% presisi 79% *recall* 70% dan *F1 Score* 74%. Sedangkan data Lion mendapatkan pemodelan terbaik dengan akurasi 87% presisi 67% *recall* 59% dan *F1 score* 0,63%. Selain itu, *Cosine Similarity* dapat membantu menemukan anomali di dalam data sehingga hasil dari akurasi model yang digunakan dapat meningkat.

Kata kunci : ***Low cost carrier, Analisis Sentimen, Anomali, Cosine Similarity, XGBoost***

ABSTRACT

Indonesia is expected to become the country with the most extensive flight capacity in ASEAN by 2025. This is possible because Indonesia has more than 17,000 islands, making transportation a crucial means of connecting different regions within the country. One type of air transportation that can be used is airplanes. In Indonesia, the market share of airplanes is dominated by low cost carriers (LCC). *According to aviation observers, the dominance in market share is not a monopoly because the domestic aviation market still has many players. Thus, the competition among airlines is not limited to pricing but also involves other factors such as route diversity, flight schedules, and passenger perceptions of service quality. Because of this, an assessment of the services provided is necessary. Therefore, this study was conducted to help airlines assess their services through sentiment analysis of Citilink and Lion Air data using the XGBoost method for class classification based on reviews from both airlines.This research presents results indicating that sentiment analysis can be effectively applied using the XGBoost algorithm for both Lion and Citilink data. According to the results, the Citilink data achieved the best model, with an accuracy of* 82%, *a precision of* 79%, *a recall of* 70%, *and an F1 Score of* 74%. *Meanwhile, Lion Data achieved the best model with an accuracy of* 87%, *a precision of* 67%, *a recall of* 59%, *and an F1 score of* 0.63%. *Additionally, Cosine Similarity can help identify anomalies in the data, thus improving the accuracy of the applied model.*

Keyword : ***Low cost carrier*, *Sentiment Analysis*, *Anomaly*, *Cosine Similarity*, *XGBoost***

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

1. PENDAHULUAN

Menurut Salim, transportasi adalah proses memindahkan barang dan penumpang dari satu lokasi ke lokasi lainnya (Putu Decy Arwini & Made Juniastra, 2023). Moda transportasi dibedakan berdasarkan permukaan tempat penggunaannya. Contohnya, transportasi darat meliputi mobil, motor, dan kereta api. Transportasi laut mencakup kapal kontainer, kapal tanker, dan kapal feri. Sementara itu, transportasi udara meliputi pesawat sebagai contohnya (Filla, 2022). Sebagai moda transportasi yang dapat mempersingkat waktu perjalanan, transportasi udara memiliki peran penting dalam menghubungkan berbagai wilayah yang ada di Indonesia.

Hal ini didukung dengan kondisi geografis Indonesia yang mempunyai lebih dari 17.000 pulau. Selain itu, pada tahun 2024 menteri Perhubungan dalam pembukaan rapat koordinasi teknis menyebutkan bahwa layanan transportasi udara mempunyai peran penting dalam menggerakkan perekonomian Indonesia (Biro Komunikasi dan Informasi Publik, 2024). Gambar 1 yang berada dibawah ini merupakan grafik *market share* yang diperoleh semua maskapai pada tahun 2023.

Gambar 1. *Market Share* 2023 (Direktorat Angkutan Udara Direktorat Jenderal Perhubungan Udara, 2023).

Dari gambar 1 dapat diketahui bahwa pasar penumpang pesawat tertinggi didominasi oleh Lion Group lalu diikuti oleh Citilink. Meskipun demikian, beberapa maskapai yang berada di bawah Lion Grup mengalami penurunan pasar penumpang. Seperti maskapai Lion Air, yang memiliki pangsa pasar penumpang tertinggi, mengalami penurunan sebesar 5,5%. Meskipun mengalami kenaikan, Citilink hanya mencatat peningkatan sebesar 0,8% dibandingkan tahun sebelumnya (Direktorat Angkutan Udara Direktorat Jenderal Perhubungan Udara, 2023). Berdasarkan gambar 1 dapat dilihat bahwa pasar tertinggi didominasi oleh maskapai yang memiliki konsep *low cost carrier* (LCC). Dikutip dari Industri Kontan, pengamat penerbangan Alvin Lie menyatakan, persaingan antar maskapai saat ini tidak lagi bertumpu pada harga, hal ini disebabkan oleh tarif yang diberikan masing-masing maskapai relatif serupa. Kondisi ini merupakan dampak dari tidak adanya revisi terhadap Tarif Batas Atas (TBA) dan Tarif Batas Bawah (TBB) dalam lima tahun terakhir. Akibatnya, persaingan antar maskapai lebih berfokus pada faktor lain seperti keberagaman rute, jadwal penerbangan, dan persepsi penumpang terhadap kualitas layanan, khususnya dalam hal ketepatan jadwal, serta kebijakan bagasi (Fitri & Sulistiowati, 2024).

Meskipun demikian, pelayanan serta kenyamanan terus menjadi tantangan bagi maskapai penerbangan, terutama karena adanya keluhan yang diajukan oleh para penumpang. Pelayanan dan kenyamanan ini sangat berdampak pada loyalitas pelanggan, seperti yang disebutkan oleh Kotler P dan Keller KL, loyalitas pelanggan dapat dicapai oleh perusahaan yang mempunyai kemampuan dalam memberikan kepuasan kepada penumpang (Violin et al., 2021). Untuk mencapai tingkat kepuasan penumpang yang tinggi, maskapai perlu melakukan penilaian terhadap kualitas pelayanan yang diberikan. Penilaian ini memerlukan ulasan dari para penumpang, yang saat ini dapat ditemukan melalui berbagai *platform*. Akan tetapi, terdapat tantangan dalam melakukan pengolahan ulasan, hal ini disebabkan oleh banyaknya ulasan

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

yang ada serta banyak terdapat kata ataupun kalimat yang berbahasa asing dan slang. Selain itu, terdapat tantangan mengenai ulasan anomali, yang mana ulasan ini dapat merugikan maskapai.

Dalam transportasi udara analisis sentimen juga telah diterapkan untuk menilai pelayanan dan harga yang ditawarkan, seperti yang dilakukan pada penelitian (Ramadhansyah et al., 2024), (Triyana et al., 2024), dan (Daryanti & Tri Widodo, 2024). Penelitian yang dilakukan oleh (Ramadhansyah et al., 2024) menunjukkan bahwa algoritma *Random Forest* bekerja lebih baik dibandingkan dengan algoritma KNN dalam melakukan analisis sentimen. Sementara itu, algoritma *Random Forest* juga menunjukkan hasil yang baik dibandingkan dengan algoritma *Naïve Bayes* dalam melakukan analisis sentimen terkait harga tiket pesawat domestik (Triyana et al., 2024). Penelitian lain juga membandingkan algoritma seperti *Support Vector Machine*, *Naïve Bayes*, dan *Random Forest* dalam melakukan analisis sentimen. Hasil penelitian ini menunjukkan bahwa algoritma *Support Vector Machine* bekerja lebih baik dibandingkan algoritma lainnya (Daryanti & Tri Widodo, 2024). Penelitian analisis sentimen juga telah berkembang dengan menggunakan metode lain untuk meningkatkan pemahaman data. Salah satu yang pendekatan yang digunakan adalah analisis sentimen berbasis *Lexicon* dan *Cosine Similarity* untuk menemukan anomali di dalam ulasan hotel (Nikolic et al., 2024). Oleh karena itu, penulis mengusulkan untuk melakukan analisis sentimen terhadap dua maskapai yaitu Lion Air dan Citilink serta melakukan pendeteksian anomali menggunakan *cosine similarity*.

2. LANDASAN TEORI

Analisis Sentimen

Analisis sentimen dapat disebut juga sebagai *opinion mining* merupakan teknik yang digunakan untuk mengetahui opini ataupun pandangan seorang penulis ataupun pengguna terhadap topik tertentu bersifat negatif atau positif. Analisis sentimen juga dapat diartikan sebagai tahapan untuk memperoleh informasi dari teks, dengan menggunakan teknik pemrosesan bahasa alami, sehingga sikap penulis dapat diidentifikasi sebagai sentimen positif, negatif atau netral (Nandwani & Verma, 2021).

Maskapai *Low-cost carrier* (LCC)

Maskapai bertarif rendah atau *low cost carrier* adalah maskapai penerbangan yang menerapkan tiket lebih murah dibandingkan dengan maskapai *full service* (Ningrum et al., 2022). Pada maskapai *low cost carrier* (LCC), efisiensi operasional merupakan faktor utama yang menentukan keberhasilan, karena secara langsung mempengaruhi kemampuan mereka dalam mempertahankan pangsa pasar dan meningkatkan loyalitas pelanggan (Finansyah & Gunawan, 2019).

Algoritma *XGBoost*

XGBoost merupakan algoritma yang bekerja dengan cara memperbaiki beberapa pohon keputusan yang saling bergantung satu sama lain (Kurnia et al., 2023). Prinsip utama dari algoritma ini adalah melakukan penyesuaian parameter pembelajaran yang dilakukan secara berulang untuk meminimalkan *loss function*, yang berperan sebagai mekanisme evaluasi model. Prediksi akhir dalam algoritma *XGBoost* diperoleh dengan menjumlahkan hasil prediksi dari setiap pohon regresi yang telah dibangun (Septiana Rizky et al., 2022).

Deteksi Anomali

Anomali merupakan pola yang tidak sesuai dari perilaku yang seharusnya muncul pada data. Data anomali dapat dikenali dengan membandingkan karakteristiknya dengan data yang dianggap sebagai kondisi normal

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

(Zulfikar et al., 2023). Oleh karena itu, deteksi anomali sangat penting dilakukan untuk meningkatkan hasil analisis. Salah satu metode yang telah diterapkan untuk mendeteksi anomali adalah penggunaan *Cosine Similarity*. *Cosine similarity* bekerja dengan menghitung tingkat kemiripan antar teks, menggunakan rumus 1.

$$\textit{Cosine Similarity}\ (A,B) = \frac{A \cdot B}{|A||B|} \qquad (1)$$

Dimana:
A: Representasi vektor teks (dokumen A)
B: Representasi vektor teks (dokumen B)
$(A \cdot B)$: Hasil perkalian *dot product* antara vektor A dan Vektor B
$|A|$: Panjang vektor A
$|B|$: Panjang vektor B

Nilai yang diperoleh kemudian dibandingkan dengan nilai *threshold* untuk menentukan apakah teks tersebut termasuk anomali (Nikolic et al., 2024). Penentuan nilai *threshold* ini dapat dilakukan dengan menggunakan metode *Interquartile Range* (IQR), yang terdapat pada rumus 2 (Novoa-Paradela et al., 2024).

$$\text{IQR} = Q_3 - Q_1 \qquad (2)$$

Dimana:
IQR = mengukur rentang tengah dari suatu data
Q_1= kuartil pertama dari suatu data (0,25 kuantil)
Q_3 = kuartil ketiga dari suatu data (0,75 kuantil)

Hyperparameter

Hyperparameter tuning dilakukan untuk menemukan kombinasi parameter terbaik (Nugraha & Sasongko, 2022). Salah satu yang banyak digunakan adalah *GridSeacrhCV*. *GridSeacrhCV* bekerja dengan melakukan pencarian parameter terbaik dengan cara menguji semua kombinasi parameter yang ada serta melakukan validasi pada semua kombinasi parameter (Nugraha & Sasongko, 2022).

Teknik *Resampling*

Teknik resampling adalah teknik yang dapat digunakan untuk melakukan penanganan pada data *imbalance*. Pada teknik terdapat dua jenis teknik yaitu, *oversampling* dan *undersampling*. Teknik *oversampling* terdiri dari *random oversampling* dan SMOTE. Sedangkan untuk teknik *undesrsampling* biasanya menggunakan *random undersampling* (Hasanah et al., 2024). *Synthetic Minority Oversampling Technique* atau yang biasa disebut dengan SMOTE sebagai salah satu teknik yang dapat digunakan untuk menyeimbangkan kelas bekerja dengan cara membuat data berdasarkan data dari kelas minoritas (Angkoso et al., 2024).

3. METODOLOGI

Gambar 2 merupakan tahapan penelitian yang dilakukan pada penelitian ini.

Gambar 2. Tahapan Penelitian.

Uraian mengenai tahapan penelitian yang tercantum dalam Gambar 2 disajikan pada bagian berikut.

1. Pengumpulan data

Pengumpulan data dilakukan dengan mengambil data dari dua maskapai yaitu Citilink dan Lion Air pada *website* Tripadvisor.com. Pengumpulan ini dilakukan untuk mendapatkan informasi mengenai tanggal ulasan, tanggal perjalanan, dan ulasan penumpang. Pengumpulan ini dilakukan dengan rentang waktu untuk maskapai Lion Air dimulai dari Januari 2016 sampai dengan April 2025. Sedangkan Citilink dimulai dari Agustus 2016 sampai dengan April 2025.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

2. *Preprocessing* data

Pada kedua data dilakukan *preprocessing* secara terpisah, tetapi proses beserta *corpus* yang digunakan sama. Proses ini diawali dengan melakukan penghapusan pada atribut yang memiliki *missing value*, penghapusan data duplikat, penggabungan kolom judul ulasan dengan deskripsi ulasan, melakukan penghapusan terhadap emoji, melakukan *case folding* yaitu mengubah semua huruf pada teks menjadi huruf kecil, melakukan normalisasi pada kata tidak baku, melakukan tokenisasi yaitu pemecahan teks, melakukan *stopword removal* yaitu menghapus kata yang tidak sesuai, dan melakukan *stemming* yaitu mengubah kata ke bentuk dasarnya dengan menghilangkan imbuhan seperti awalan, akhiran, atau kombinasi tertentu.

3. Pelabelan data

Pelabelan data ini dilakukan dengan menggunakan *Inset Lexicon*, *Inset Lexicon* sendiri merupakan kumpulan data yang berisi kata dalam bahasa Indonesia. Terdapat 3609 kata bersentimen positif dan 6609 kata bersentimen negatif, serta bobot kata dimulai -5 sampai +5. Pembobotan kata

Setelah melakukan pelabelan maka data akan melalui tahapan pembobotan kata menggunakan TF-IDF. Pembobotan ini dilakukan dengan mengetahui seberapa sering kata yang akan diberi bobot muncul dan seberapa jarang kata tersebut muncul.

4. Pendeteksian Anomali

Pendeteksian ini akan dilakukan secara terpisah sesuai nama maskapai yang ada pada data. Pendeteksian anomali dilakukan dengan mengetahui nilai *similarity* antar ulasan. Ulasan dapat dikategorikan sebagai anomali jika *nilai similarity* tidak memenuhi *threshold*. Untuk mengetahui nilai *threshold* maka digunakan rumus 2 yaitu metode *Interquartile Range* (IQR). Data yang terdeteksi sebagai anomali akan dihapus.

5. Pemodelan dengan XGBoost

Pemodelan akan dilakukan secara terpisah untuk masing masing maskapai. Akan tetapi sebelum melakukan pemodelan maka data akan dibagi menjadi data *train* dan data *test*. Pada proses pembagian data ini, data nantinya akan dibagi menjadi beberapa proporsi yaitu, 0,1, 0,2, 0,3, 0,4, dan 0,5. Proses ini dilakukan untuk mengetahui proporsi terbaik dalam pembagian data pada saat pemodelan. Penilaian proporsi terbaik nantinya akan menggunakan nilai akurasi yang didapat dari setiap proporsi.

6. Evaluasi

Dikarenakan pemodelan dilakukan secara terpisah, maka evaluasi juga akan dilakukan secara terpisah untuk masing masing data. Evaluasi ini menggunakan nilai akurasi, *precison*, *recall*, dan *F1 Score* yang dihasilkan dari pemodelan yang sudah dilakukan.

7. Menyimpan model ke *pickle*

Model yang di simpan adalah model yang sudah di evaluasi menggunakan nilai akurasi, *precison*, *recall*, dan *F1 Score.* Penyimpanan ini dilakukan agar model dapat digunakan pada *website*, proses penyimpanan ini nantinya akan menggunakan *pickle*

8. Perancangan web

Setelah melakukan pemodelan maka diperlukan perancangan sistem berbasis website menggunakan *Streamlit,* sehingga model dapat ditampilkan pada Website

4. HASIL DAN PEMBAHASAN

Pengumpulan Data

Pengumpulan data penelitian ini diperoleh dari *website* tripadvisor.com. Data tersebut berasal dari ulasan maskapai *Lion Air* dan *Citilink* dalam bahasa Indonesia, dari penumpang maskapai. Atribut data terdiri dari *Review_Title, Review_Description, Review_Date, Rating_Stars, Travel_Date, Seat_Comfort, Customer Service, Cleanliness, Food And Beverage, Legroom, In-Flight Entertainmen, Value For Money* dan *Check-In And Boarding*. Dari data tersebut diperoleh:

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

- dari maskapai *Lion Air* 833 ulasan
- dari maskapai *Citilink* 631 ulasan

Preprocessing Data

Preprocessing dilakukan dengan langkah yang telah diterangkan pada sub bab sebelumnya. Hasil *preprocessing* data Citilink dan Lion Group ditampilkan pada tabel 1 dan tabel 2.

Pelabelan Data

Pada penelitian pelabelan dilakukan dengan menggunakan *Inset Lexicon*. Misal pada ulasan yang memiliki nilai lebih dari 0 akan mendapat sentimen positif, sedangkan ulasan yang mendapat nilai kurang dari 0 akan mendapat sentimen negatif. Jika tidak memenuhi kedua kriteria tersebut maka akan diberikan sentimen netral. Hasil pelabelan ini dapat dilihat pada tabel 1 dan tabel 2.

Tabel 1. Hasil *Preprocessing* dan *Labelling* Data Citilink

Teks Asli	Citilink terbaik Penerbangan denganMaskapai Citilink⭐⭐⭐⭐⭐. Pelayanan pramugari sangat ramah dan sigap. Makanan enak, terutama nasi gorengnya. Hiburan juga lengkap. Sangat nyaman! Saya pasti akan terbang dengan Maskapai *Citilink* lagi. ✈️"
Cleaning	Citilink terbaik Penerbangan denganMaskapai Citilink Pelayanan pramugari sangat ramah dan sigap Makanan enak terutama nasi gorengnya Hiburan juga lengkap Sangat nyaman Saya pasti akan terbang dengan Maskapai Citilink lagi
Case folding	citilink terbaik penerbangan denganmaskapai citilink pelayanan pramugari sangat ramah dan sigap makanan enak terutama nasi gorengnya hiburan juga lengkap sangat nyaman saya pasti akan terbang dengan maskapai citilink lagi
Normalisasi	citilink terbaik penerbangan denganmaskapai citilink pelayanan pramugari sangat ramah dan sigap makanan enak terutama nasi gorengnya hiburan juga lengkap sangat nyaman saya pasti akan terbang dengan maskapai citilink lagi
Tokenisasi	['citilink', 'terbaik', 'penerbangan', 'denganmaskapai', 'citilink', 'pelayanan', 'pramugari', 'sangat', 'ramah', 'dan', 'sigap', 'makanan', 'enak', 'terutama', 'nasi', 'gorengnya', 'hiburan', 'juga', 'lengkap', 'sangat', 'nyaman', 'saya', 'pasti', 'akan', 'terbang', 'dengan', 'maskapai', 'citilink', 'lagi']
Stopwords	['citilink', 'terbaik', 'penerbangan', 'denganmaskapai', 'Citilink', 'pelayanan', 'pramugari', 'ramah', 'sigap', 'makanan', 'enak', 'nasi', 'gorengnya', 'hiburan', 'lengkap', 'nyaman', 'terbang', 'maskapai', 'citilink']
Stemming	['citilink', 'baik', 'terbang', 'denganmaskapai', 'citilink', 'layan', 'pramugari', 'ramah', 'sigap', 'makan', 'enak', 'nasi', 'goreng', 'hibur', 'lengkap', 'nyaman', 'terbang', 'maskapai', 'citilink']
Nilai Polaritas	13
Label	Positif

Tabel 2. Hasil *Preprocessing* dan *Labelling* Data Lion

Teks Asli	Sangat buruk Sangat delay!!!!!! Dan sering terlambat Bagasi sedikit dan pelayanan buruk sekali tidak ada snack di pesawat dan kurang Sopan
Cleaning	Sangat buruk Sangat delay Dan sering terlambat Bagasi sedikit dan pelayanan buruk sekali tidak ada snack di pesawat dan kurang Sopan
Case folding	sangat buruk sangat delay dan sering terlambat bagasi sedikit dan pelayanan buruk sekali tidak ada snack di pesawat dan kurang sopan
Normalisasi	sangat buruk sangat delay dan sering terlambat bagasi sedikit dan pelayanan buruk sekali

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

	tidak ada snack di pesawat dan kurang sopan
Tokenisasi	['sangat', 'buruk', 'sangat', 'delay', 'dan', 'sering', 'terlambat', 'bagasi', 'sedikit', 'dan', 'pelayanan', 'buruk', 'sekali', 'tidak', 'ada', 'snack', 'di', 'pesawat', 'dan', 'kurang', 'sopan']
Stopwords	['buruk', 'delay', 'terlambat', 'bagasi', 'pelayanan', 'buruk', 'snack', 'pesawat', 'sopan']
Stemming	['buruk', 'delay', 'lambat', 'bagasi', 'layan', 'buruk', 'snack', 'pesawat', 'sopan']
Nilai Polaritas	-15
Label	Negatif

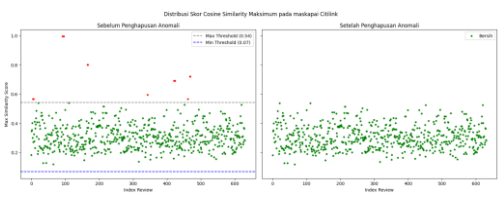
Pembobotan Kata

Setelah pelabelan, data melalui tahapan pembobotan kata menggunakan TF-IDF. Pembobotan ini dilakukan dengan mengetahui seberapa sering kata yang akan diberi bobot muncul dan seberapa jarang kata tersebut muncul.

Deteksi Anomali

a. Citilink

Setelah data melalui proses *preprcosessing* dan pelabelan, dilakukan deteksi anomali, dengan memberikan bobot untuk setiap kata menggunakan TF-IDF, dilanjutkan dengan menghitung *similarity* pada setiap ulasan. Penentuan *threshold* dilakukan dengan menggunakan rumus 2 yaitu metode *interquartil range* (IQR). Data Citilink mempunyai nilai ambang batas atas 0,54 dan batas bawah 0,07, sehingga ulasan yang tidak memenuhi nilai tersebut akan terdeteksi sebagai anomali. Dibawah ini adalah gambar 3 yang merupakan visualisasi hasil dari deteksi anomali.

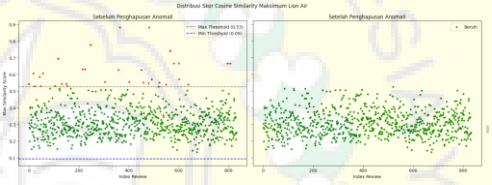


Gambar 3. Hasil Deteksi Anomali pada data Citilink.

Dari hasil deteksi anomali pada gambar 3 diketahui bahwa data Citilink memiliki 12 ulasan yang terdeteksi sebagai anomali. Sebelum proses dilanjutkan, data yang sudah terdeteksi sebagai anomali dihapus.

b. Lion

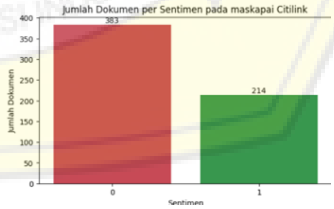
Proses deteksi anomali yang sama dilakukan pada data Lion, dimana dta ini mempunyai nilai ambang batas atas 0,53 dan batas bawah 0,09, sehingga ulasan yang tidak memenuhi nilai tersebut akan terdeksi sebagai anomali. Gambar 4 merupakan hasil proses pendeteksian anomali, dari gambar tersebut diketahui sebanyak 35 ulasan terdeteksi sebagai anomali.



Gambar 4. Hasil Deteksi Anomali pada data Lion.

Pemodelan dengan *XGBoost* dan Evaluasi

Pemodelan ini diawali dengan melakukan pengubahan bentuk pada label yaitu atribut “Indonesia Sentiment”, agar dapat diproses pada saat pemodelan. Proses pemodelan ini diawali dengan melakukan pengecekan proporsi data sebagaimana yang ditampilkan pada gambar 5.



Gambar 5. Proporsi Data Citilink

Pada gambar 5, yaitu proporsi sentimen pada data Citilink, terlihat data Citilink tidak memerlukan teknik *resampling* pada saat pemodelan, sehingga pemodelan dapat langsung dilakukan, dengan menggunakan

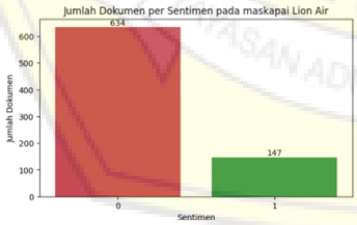
https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

GridSearchCV. Hasil evaluasi dari pemodelan data Citilink dapat dilihat pada tabel 3. Dari tabel 3 diketahui akurasi dari pemodelan yang sudah dilakukan mencapai 0,82.

Tabel 3. Ealuasi Pemodelan Data Citilink.

	Presisi	***Recall***	***F1-Score***
0	0,84	0,90	0,87
1	0,79	0,70	0,74
Akurasi			0,82
Macro avg	0,82	0,80	0,80
Weighted avg	0,82	0,82	0,82

Pada gambar 6, yaitu proporsi data Lion, terlihat kelas minoritas yaitu kelas positif mempunyai proporsi sebesar 18% dari total data. Sementara itu, pada penelitian yang dilakukan oleh Kumar dkk menyebutkan beberapa pembagian kelas untuk data *imbalance*. Pada kelas pertama yaitu kelas *extreme* yaitu untuk data yang kelas minoritas kurang dari 1% dari total keseluruhan data. Kelas *moderate* untuk data yang proporsi kelas minoritasnya berada pada rentang 1-20% dari keseluruhan *dataset*. Sedangkan kelas *mild* untuk data yang proporsi kelas minoritasnya berada direntang 20-40% dari keseluruhan *dataset* (Kumar et al., 2022). Oleh karena itu, data untuk maskapai Lion dimodelkan kembali dengan menggunakan SMOTE, dikarenakan data tersebut termasuk pada kelas *mild*.



Gambar 6. Proporsi Data Lion.

Tabel 4 adalah hasil evaluasi dari pemodelan dengan menggunakan SMOTE. Dari tabel tersebut diketahui akurasi dari pemodelan yang dilakukan sebesar 0,87.

Tabel 4. Evaluasi Pemodelan Data Citilink

	Presisi	***recall***	***F1-Score***
0	0,91	0,93	0,92
1	0,67	0,59	0,63
Akurasi			0,87
Macro avg	0,79	0,76	0,77
Weighted avg	0,86	0,87	0,86

Hasil Pemodelan dan Perancangan *Website*

Model disimpan dengan pickle agar dapat digunakan di *website*, pembuatan *website* ini dilakukan dengan menggunakan *Streamlit*. Gambar 7 merupakan halaman awal penggunaan *website* yang dimulai dengan memasukkan data teks yang akan diprediksi. Kemudian pada gambar 8 dapat dilihat hasil dari proses, yang berisi data anomali yang dibuang. Gambar 9 merupakan hasil akhir menunjukkan visualisasi dari perkembangan sentimen. Dari gambar 9 dapat diketahui pada data terbaru tahun 2025 maskapai Citilink mendapatkan 8 ulasan bersentimen negatif, sementara itu belum terdapat ulasan positif pada tahun yang sama. Sedangkan untuk maskapai Lion mendapatkan 7 ulasan negatif pada tahun 2025 dan belum mendapatkan ulasan positif pada tahun yang sama.

Gambar 9 juga menampilkan visualisasi rasio dengan hasil, rasio ulasan positif tertinggi untuk maskapai Citilink terjadi padah tahun 2016 -2017, pencapaian ini menunjukkan pelayanan yang baik. Sementara pada tahun 2021-2022 tercatat rasio negatif tertinggi akibat pembatasan COVID-19. Setelahnya, rasio negatif menurun meski jumlah ulasan berkurang. Untuk Lion, rasio ulasan positif 2015–2025 tidak stabil, dengan puncak pada 2018, namun penurunan jumlah ulasan membuat capaian tersebut belum mencerminkan peningkatan layanan.

Gambar 7. Tampilan Halaman *Home*.

Gambar 8. Tampilan Halaman Anomali.

Gambar 9. Halaman Visualisasi.

5. KESIMPULAN

Kesimpulan pada penelitian ini adalah Algoritma *XGBoost* dapat diterapkan untuk melakukan analisis sentimen. Pada data *Citilink* akurasi mencapai 0,82. Sedangkan untuk *Lion Air* akurasi pada pemodelan dengan melakukan penanganan data *imbalance* mencapai akurasi 0,87. Selain itu, Penerapan *Cosine Similarity* berhasil dilakukan, dari data *Citilink* terdeteksi 12 data sebagai anomali sedangkan data *Lion Air* terdeteksi 35 data sebagai anomali.

6. UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada seluruh pihak yang telah membantu pada saat penyusunan penelitian ini.

DAFTAR PUSTAKA

Angkoso, C. V., Thrisna, M. A. N., Satoto, B. D., & Kusumaningsih, A. (2024). Optimasi Klasifikasi Sentimen Menggunakan Random Forest dengan Preprocessing K-Means Clustering dan SMOTE. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)* , *10*(3). https://doi.org/https://doi.org/10.26418/jp.v10i3

Biro Komunikasi dan Informasi Publik. (2024, September 26). *Transportasi Udara Punya Peran Kunci Gerakkan Ekonomi Bangsa*. Kementerian Perhubungan Republik Indonesia. https://www.dephub.go.id/post/read/transportasi-udara-punya-peran-kunci-gerakkan-ekonomi-bangsa

Daryanti, & Tri Widodo. (2024). Analisis Sentimen Ulasan Penumpang Maskapai Penerbangan Indonesia Menggunakan Support Vector Machine , Naive Bayes, dan Random Forest. *The Indonesian Journal of Computer Science*, *13*(5), 8437–8445. https://doi.org/10.33022/ijcs.v13i5.4452

Direktorat Angkutan Udara Direktorat Jenderal Perhubungan Udara. (2023). *Statistik Angkutan Udara tahun 2023*. https://hubud.kemenhub.go.id/hubud/website/assets/file/bukustatistik/9a4eaeb68ab0e5bb3be1a8b5202ccb90.pdf

Filla, F. (2022). PEMILIHAN MODA TRANSPORTASI DARAT DALAM MENDUKUNG EFEKTIVITAS PENGIRIMAN BARANG. *Jurnal Bisnis, Logistik Dan Supply Chain (BLOGCHAIN)*, *2*(2), 52–57. https://doi.org/10.55122/blogchain.v2i2.526

Finansyah, F. F., & Gunawan, J. (2019). Analisis Perbedaan Perilaku Pelanggan dan Pengguna Media Sosial Maskapai Penerbangan Low-Cost Indonesia. *JURNAL SAINS DAN SENI ITS* , *8*(2), 290–295. https://doi.org/10.12962/j23373520.v8i2.48506

Fitri, A. N., & Sulistiowati, T. (2024, November 19). Pasar Maskapai Penerbangan Ketat, Begini Pandangan Pengamat Penerbangan. *Kontan*. https://industri.kontan.co.id/news/pasar-maskapai-penerbangan-ketat-begini-pandangan-pengamat-penerbangan

Hasanah, U., Soleh, A. M., & Sadik, K. (2024). Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models. *Jurnal Matematika, Statistika Dan Komputasi*, *21*(1), 88–102. https://doi.org/10.20956/j.v21i1.35552

Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., Shorfuzzaman, M., Alsufyani, A., & Uddin, M. (2022). Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare (Switzerland)*, *10*(7), 1–28. https://doi.org/10.3390/healthcare10071293

Kurnia, D., Mazdadi, M. I., Kartini, D., Nugroho, R. A., Abadi, F., & Korespondensi, P. (2023). SELEKSI FITUR DENGAN PARTICLE SWARM OPTIMIZATION PADA

KLASIFIKASI PENYAKIT PARKINSON MENGGUNAKAN XGBOOST. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *10*(5), 1083–1094. https://doi.org/10.25126/jtiik.2023107252

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 1–19. https://doi.org/10.1007/s13278-021-00776-6

Nikolic, M., Stojanovic, M., & Marjanovic, M. (2024). Anomaly Detection in Hotel Reviews: Applying Data Science for Enhanced Review Integrity. *2024 32nd Telecommunications Forum, TELFOR 2024 - Proceedings of Papers*, 1–4. https://doi.org/10.1109/TELFOR63250.2024.10819036

Ningrum, U., Riananditasari, N., Tinggi, S., Kedirgantaraan, T., & Abstrak, Y. (2022). ANALISIS STRATEGI PEMASRAN MASKAPAI PENERBANGAN BERTARIF RENDAH (LION AIR) TERHADAP DAYA SAING PERUSAHAAN. *Jurnal Flight Attendant Kedirgantaraan*, *4*(2), 139–146. https://doi.org/https://doi.org/10.56521/attendant-dirgantara.v4i1.418

Novoa-Paradela, D., Fontenla-Romero, O., & Guijarro-Berdiñas, B. (2024). Explained anomaly detection in text reviews: Can subjective scenarios be correctly evaluated? *Engineering Applications of Artificial Intelligence*, *133*, 1–14. https://doi.org/https://doi.org/10.1016/j.engappai.2024.108065

Nugraha, W., & Sasongko, A. (2022). Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search Hyperparameter Tuning on Classification Algorithm with Grid Search. *SISTEMASI: Jurnal Sistem Informasi*, *11*(2), 2540–9719. https://doi.org/https://doi.org/10.32520/stmsi.v11i2.1750

Putu Decy Arwini, N., & Made Juniastra, I. (2023). PERAN TRANSPORTASI DALAM DUNIA INDUSTRI. *Jurnal Ilmiah Vastuwidya*, *6*(1), 70–77. https://doi.org/https://doi.org/10.47532/jiv.v6i1.794

Ramadhansyah, D., Asrofiq, A., & Yunefri, Y. (2024). ANALISIS SENTIMEN ULASAN PENUMPANG MASKAPAI PENERBANGAN DI INDONESIA DENGAN ALGORITMA RANDOM FOREST DAN KNN. *ZONAsi: Jurnal Sistem Informasi*, *6*(2), 287–297. https://doi.org/ttps://doi.org/10.31849/zn.v6i2.19177

Septiana Rizky, P., Haiban Hirzi, R., Hidayaturrohman, U., Hamzanwadi Selong Jl TGKH Muhammad Zainuddin Abdul Madjid Pancor, U., & Timur, L. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. *J Statistika*, *15*(2), 228–236. www.unipasby.ac.id

Triyana, D., Muharrom, M., Haromainy, A., & Maulana, H. (2024). IMPLEMENTASI METODE ENSEMBLE MAJORITY VOTE PADA ALGORITMA NAIVE BAYES DAN RANDOM FOREST UNTUK ANALISIS SENTIMEN TWITTER HARGA TIKET PESAWAT DOMESTIK. *Jurnal Mahasiswa Teknik Informatika*, *8*(4), 7885–7894. https://doi.org/https://doi.org/10.36040/jati.v8i4.10475

Violin, V., Hasan, S., & Sufri, M. (2021). Pengaruh Konsep Low-Cost Carrier dan Kualitas Layanan terhadap Kepuasan dan Loyalitas Pelanggan pada Maskapai Lion Airlines di Indonesia. *Journal of Management Science (JMS)*, *2*(1), 150–160. https://doi.org/https://doi.org/10.52103/jms.v3i2.957

Zulfikar, A., Rahmani, F. A., & Azizah, N. (2023). DETEKSI ANOMALI MENGGUNAKAN ISOLATION FOREST BELANJA BARANG PERSEDIAAN KONSUMSI PADA SATUAN KERJA KEPOLISIAN REPUBLIK INDONESIA. *Jurnal Manajemen Perbendaharaan*, *4*(1), 1–15. https://doi.org/10.33105/jmp.v4i1.435