

ANALISIS SENTIMEN ULASAN PENUMPANG MASKAPAI PENERBANGAN DI INDONESIA DENGAN ALGORITMA RANDOM FOREST DAN KNN

Handoko¹, Donny Ramadhansyah², Ahmad Asrofiq³, Rahmaddeni⁴, Yogi Yunefri⁵

^{1,2,3,4}STMIK Amik Riau

⁵Universitas Lancang Kuning

(Program Studi Teknik Informatika STMIK Amik Riau)

(Jl. Purwodadi Indah KM.10, Sidomulyo Barat, Tampan, Sidomulyo Barat, Kota Pekanbaru, Riau,
telp. (0761) 7047091, 589561)

(Fakultas Ilmu Komputer, Universitas Lancang Kuning)

(Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp. 0811 753 2015)

e-mail: ¹handoko0193@gmail.com, ²donny021201@gmail.com, ³asrofiqtr12@gmail.com,

⁴rahmaddeni@sar.ac.id, ⁵yogiyunefri@unilak.ac.id

Abstrak

Penelitian ini mendalami pada analisis sentimen ulasan pelanggan terhadap maskapai penerbangan di Indonesia melalui NLP dan Machine Learning. Dalam prosesnya, data ulasan melibatkan serangkaian teknik, termasuk cleansing, case folding, tokenization, filtering, dan stemming, sementara sentimen diberikan label menggunakan lexicon affinn. Visualisasi kata-kata dominan dari ulasan diwujudkan melalui wordcloud untuk memberikan gambaran yang kaya dan intuitif. Ekstraksi fitur melibatkan metode TF-IDF, diikuti oleh proses klasifikasi menggunakan algoritma Random Forest dan K-Nearest Neighbor (KNN). Hasil evaluasi model menunjukkan tingkat akurasi yang memuaskan, dengan Random Forest mencapai 83% dan KNN mencapai 82%. Temuan ini memberikan wawasan yang dalam tentang preferensi pelanggan dan potensial masalah dalam pengalaman penerbangan di Indonesia, memberikan kontribusi pada pemahaman yang lebih holistik terhadap dinamika industri penerbangan.

Kata kunci: Analisis Sentimen, TF-IDF, Natural Language Processing, Random Forest, K-Nearest Neighbor (KNN)

Abstract

This research is an in-depth look at sentiment analysis of customer reviews of airlines in Indonesia through NLP and Machine Learning. In the process, the review data involves a series of techniques, including cleansing, case folding, tokenization, filtering, and stemming, while the sentiments are labeled using an affinned lexicon. Visualization of dominant words from reviews is realized through wordcloud to provide a rich and intuitive overview. Feature extraction involves the TF-IDF method, followed by the classification process using Random Forest and K-Nearest Neighbor (KNN) algorithms. Model evaluation results showed satisfactory accuracy rates, with Random Forest reaching 83% and KNN reaching 82%. The findings provide deep insights into customer preferences and potential problems in the flight experience in Indonesia, contributing to a more holistic understanding of the dynamics of the aviation industry.

Keywords: Sentiment Analysis, TFI-IDF, Natural Language Processing, Random Forest, K-Nearest Neighbor (KNN)

1. PENDAHULUAN

Indonesia adalah sebuah negara yang terdiri dari banyak pulau, meliputi wilayah dari Sabang hingga Merauke. Wilayah ini mencakup sejumlah pulau besar dan kecil yang terhubung oleh selat dan laut. Saat ini, terdapat catatan resmi tentang keberadaan sekitar 13.466 pulau yang masing-masing memiliki koordinat geografisnya sendiri. Hal ini menunjukkan kekayaan geografis dan keindahan

alam yang meliputi wilayah yang luas dan bervariasi di seluruh Indonesia[1]. Oleh karena itu Indonesia memiliki keanekaragaman budaya, tradisi dan keindahan alam yang sangat bagus. Keanekaragaman kuliner juga merupakan salah satu faktor ramainya wisatawan datang berkunjung ke Indonesia. Wisatawan luar yang ingin berkunjung ke Indonesia atau wisatawan lokal yang ingin berkunjung ke salah satu daerah di Indonesia tentunya membutuhkan alat transportasi yang efektif terhadap waktu. Se jauh ini alat transportasi yang paling efektif yaitu transportasi udara salah satunya pesawat.

Maskapai penerbangan merupakan perusahaan yang menawarkan layanan penerbangan untuk penumpang dan barang. Mereka memiliki sejumlah pesawat dan tim yang terampil dalam berbagai bidang terkait penerbangan[2]. Beberapa maskapai yang paling relevan di Indonesia misalnya seperti Garuda Indonesia, Lion Air, Batik Air dan Air Asia merupakan yang paling sering digunakan untuk penerbangan domestik. Setiap armada tersebut pastinya memiliki fasilitas atau pelayanan yang berbeda beda. Oleh karena itu, setiap fasilitas atau pelayanan yang diberikan dari pihak internal armada tersebut akan menghasilkan beberapa ulasan dari setiap penumpang. Dalam dataset yang bisa ditemukan di Kaggle atau sumber data serupa, terdapat informasi yang dapat memberikan gambaran lebih rinci mengenai ulasan-ulasan dari para penumpang terkait pengalaman mereka dengan maskapai penerbangan di Indonesia. Analisis terhadap ulasan-ulasan ini bisa memberikan wawasan yang lebih mendalam mengenai preferensi dan pengalaman yang diinginkan oleh penumpang, serta membantu perusahaan maskapai penerbangan untuk meningkatkan kualitas layanan mereka.

Analisis sentimen berdasarkan ulasan maskapai penerbangan dapat membantu penumpang memilih maskapai yang sesuai dengan kebutuhan mereka. Selain itu, dengan menggunakan ulasan penumpang, analisis sentimen memungkinkan untuk mengumpulkan umpan balik cepat dari penumpang. Hal ini dapat membantu perusahaan maskapai penerbangan dalam meningkatkan kualitas layanan mereka di masa depan. Analisis sentimen sendiri adalah teknik untuk menemukan dan memahami pendapat atau sentimen yang ada dalam teks[3]. Seringkali, teknik ini digunakan dalam pengolahan bahasa alami (NLP) untuk mendapatkan wawasan dari teks.

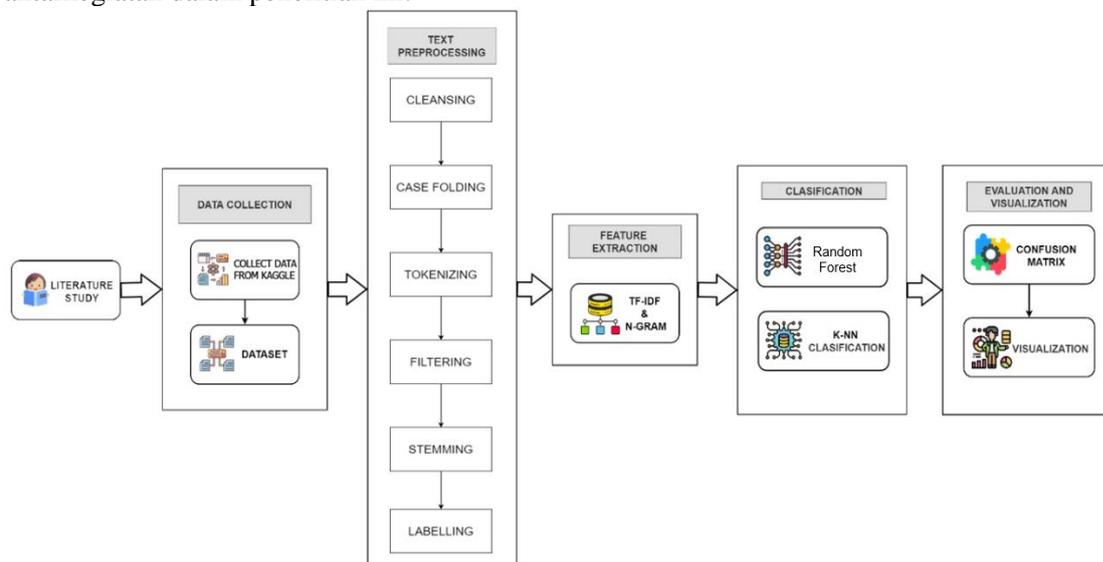
Random Forest adalah metode pembelajaran ensemble yang digunakan untuk tugas klasifikasi dan regresi. Pendekatan ini melibatkan pembangunan beberapa pohon keputusan dan menggabungkannya untuk menghasilkan prediksi yang lebih akurat dan stabil. Nama "Random" dalam Random Forest berasal dari fakta bahwa setiap pohon dalam model ini dibangun menggunakan subset acak dari fitur-fitur dan subset acak dari data pelatihan. Pendekatan ini membantu mengurangi kemungkinan overfitting dan meningkatkan kinerja keseluruhan model[4]. Sementara itu, K-Nearest Neighbors (KNN) adalah Algoritma jiran paling dekat (k-nearest neighbor, disingkat k-NN) merupakan teknik agar objek dapat terklasifikasi berdasarkan data pembelajaran yang terdekat dengannya. Data pembelajaran digambarkan ke dalam ruang berukuran banyak, di mana setiap dimensi menunjukkan setiap atribut atau karakteristik data. Untuk klasifikasi data baru, label k tetangga terdekat dicari; label terbanyak yang ditemukan menjadi label data baru jika $k = 1$. Banyak orang menggunakan jarak Euklides[6].

Dari penelitian ini penulis menemukan beberapa referensi dari penelitian sebelumnya yang telah menggunakan algoritma Random Forest dan KNN pada kasus analisis sentimen. Dalam penelitian Sentiment Analysis of Social Media Network Using Random Forest Algorithm yang dilakukan oleh Karthika, Murugeswari dan Manoranjithem, Random Forest memberikan akurasi terbaik di antara algoritma lainnya dengan 97% dan Support Vector Machine (SVM) memberikan akurasi 92%. Pada algoritma Random Forest, pemilihan fitur implisit dilakukan selama analisis dataset[4]. Dalam perbandingan akurasi yang dilakukan oleh Anita Wulan Sari, Teguh Iman Hermanto dan Meriska Defriani pada Sentiment Analysis Of Tourist Reviews dengan Algoritma K-Nearest Neighbors dan Support Vector Machine. Vector Machine menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan metode klasifikasi K-Nearest Neighbors. Pada pengujian menggunakan metode Support Vector Machine, nilai akurasi yang didapatkan pada Waterbom Bali sebesar 88%, Mandala Suci Wenara Wana sebesar 83%, Terasering Tegalalang sebesar 89%, Pura Tanah Lot sebesar 88%, dan Pura Uluwatu sebesar 95%. Sedangkan pada metode K-Nearest Neighbors diperoleh nilai akurasi pada Waterbom Bali sebesar 67%, Mandala Suci Wenara Wana sebesar 60%, Terasering Tegalalang sebesar 69%, Pura Tanah Lot sebesar 71%, dan Pura Uluwatu sebesar 81% [7]. Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter yang diteliti oleh Styawati, Nirwana Hendrastut, Auliya Rahman Isnain dan Ari Yanti Rahmadhani menggunakan Support Vector

Machine, Evaluasi hasil menunjukkan bahwa nilai akurasi pada model dengan kernel linear mencapai 98.67%, dengan nilai precision sebesar 98%, recall sebesar 99%, dan F1-Score sebesar 98%. Sementara itu, pada model dengan kernel RBF, nilai akurasi mencapai 98.34%, dengan precision sebesar 97%, recall sebesar 98%, dan F1-Score sebesar 98% [8]. Penelitian yang dilakukan oleh Rasenda, Hendarman Lubis dan Ridwan dalam usaha mengklasifikasikan sentimen analisis dari tweet di Twitter tentang berbagai pendapat mengenai riba pada bunga bank di Indonesia, digunakan teknik penambangan teks. Proses penambangan data dilakukan dengan menerapkan algoritma klasifikasi K-NN untuk memprediksi label dalam dataset. Hasil evaluasi menunjukkan akurasi K-NN sebesar 70,59%, dengan nilai presisi sebesar 69,87%. Selain itu, hasil untuk recall K-NN mencapai 62,32%. Hasil tersebut memberikan gambaran tentang sejauh mana kemampuan algoritma dalam memprediksi dan mengklasifikasikan sentimen masyarakat terkait dengan riba pada bunga bank di Indonesia. [9]. Kemudian komparasi 3 metode yang dilakukan oleh Ratna Andini Husen, Rizki Astuti, Lili Marlia, Rahmaddeni, dan Lusiana Efrizoni menyimpulkan bahwa dalam penelitian mereka, algoritma SVM, Naïve Bayes, dan Logistic Regression telah menghasilkan akurasi tertinggi. Khususnya, penggunaan algoritma SVM pada pembagian data 90:10 menghasilkan nilai akurasi sebesar 88%. Hasil ini menunjukkan bahwa SVM memberikan kinerja yang baik dalam konteks dataset dan pemisahan data tersebut, memperkuat keandalan algoritma tersebut dalam melakukan klasifikasi pada penelitian yang dilakukan. [10].

2. METODE PENELITIAN

Penelitian ini memerlukan suatu metodologi penelitian yang terstruktur, yang mencakup berbagai kegiatan dan proses penelitian secara sistematis. Tahapan penelitian menjelaskan secara rinci bagaimana penelitian dilaksanakan, dari awal hingga akhir. Proses ini dikendalikan oleh suatu kerangka kerja yang diilustrasikan dalam Gambar 1. Dengan memanfaatkan kerangka kerja tersebut, penelitian dapat dilaksanakan secara lebih terstruktur dan sesuai dengan rencana yang telah ditetapkan sebelumnya. Gambar 1 berfungsi sebagai panduan untuk memahami urutan dan hubungan antarkegiatan dalam penelitian ini.



Gambar 1. Tahap Metode Penelitian

2.1. Studi Literatur (*Literature Study*)

Tahap penelitian literatur merupakan langkah untuk mengumpulkan informasi yang akan menjadi referensi dalam mendukung penelitian. Data literatur ini diperoleh dari jurnal dan artikel yang terkait dengan analisis sentimen, klasifikasi, maskapai penerbangan, K-nearest neighbor, dan Random Forest.

2.2. Sub Pengambilan data (*Data Collection*)

Penulis menggunakan Kaggle untuk mendapatkan data ulasan penumpang pesawat berdasarkan

maskapai penerbangan, dan mereka juga menggunakan data dari www.airlinequality.com. Dataset ini berisi berbagai ulasan penumpang tentang pengalaman mereka setelah menggunakan layanan penerbangan maskapai tersebut. Data tentang maskapai memiliki review yang paling banyak dan sebanding dengan data tentang maskapai lainnya. Dengan 4 maskapai penerbangan: Garuda Indonesia, AriAsia, Lion Air, dan Batik Air, setiap maskapai memiliki 100 review.

2.3. Pemrosesan Awal Teks (*Text Preprocessing*)

Text preprocessing pada sentimen analisis melibatkan serangkaian langkah-langkah untuk membersihkan dan mempersiapkan data teks sehingga dapat proses dengan baik oleh algoritma analisis sentimen[11]. Proses pembersihan data (*cleansing*) adalah tahap pertama, Proses ini menghilangkan karakter tidak diinginkan seperti, simbol, atau tulisan miring dari teks serta tanda baca, URL, atau karakter khusus lainnya yang tidak relevan, Tujuan dari pembersihan data adalah untuk meningkatkan kualitas data dan membuat analisis sentimen lebih mudah. Langkah kedua yaitu *Case Folding*, Ketika melakukan preprocessing teks, tahapan ini hampir selalu disertakan, Ketika kita berurusan dengan data yang tidak selalu terstruktur dan konsisten, tantangan dalam analisis dan pengelolaan data menjadi lebih kompleks, Karena itu, *case folding* berfungsi untuk membuat penggunaan huruf kapital sama. Langkah selanjutnya adalah tokenisasi (*tokenization*), yang dilakukan untuk membagi teks menjadi kata-kata atau token. Proses ini memecah kalimat menjadi bagian yang lebih kecil, yang dimaksudkan untuk mempermudah pengolahan teks dengan mengidentifikasi unit penting (token). Setelah tokenisasi dilakukan selanjut *Filtering* untuk menghilangkan kata-kata pengisi (*stop words*) dan kata-kata lain yang tidak relevan, menyaring kata-kata umum yang tidak berkontribusi pada analisis sentimen dan memusatkan perhatian pada kata-kata yang lebih signifikan adalah strategi untuk memperoleh hasil sentimen yang lebih akurat. Dan yang terakhir adalah melakukan pelabelan (*Labelling*) pada ulasan penumpang maskapai penerbangan yang akan menandai atau memberikan label pada teks berdasarkan sentimen (positif dan negatif). Kategorisasi teks ke dalam kelas sentimen yang sesuai, Penting untuk melatih model dalam tugas analisis sentimen dengan menggunakan data berlabel[11].

2.4. Fitur Ekstraksi (*Feature Extraction*)

Kumpulan data yang tidak terstruktur biasanya terdiri dari teks dan dokumen. Namun, ketika pengklasifikasi menggunakan pemodelan matematika, urutan teks yang tidak terstruktur ini harus diubah menjadi ruang fitur terstruktur. Pertama, karakter dan kata yang tidak perlu harus dibuang dari data. Setelah data dibersihkan, teknik ekstraksi fitur formal dapat digunakan[[12]. Teknik pembobotan yang umum digunakan dalam pencarian informasi dan data mining adalah TF-IDF (*Term Frequency Inverse Document Frequency*), di mana TF mengacu pada frekuensi kata dalam sebuah teks, dan IDF mengacu pada indeks frekuensi teks terbalik. Salah satu prinsip utama TF-IDF adalah bahwa kata/dokumen yang muncul lebih sering dalam satu dokumen daripada yang muncul lebih sedikit di dokumen lain harus diberi prioritas yang lebih tinggi karena mereka dapat diklasifikasikan dengan lebih baik[13]. N-gram merujuk kepada serangkaian n kata yang muncul dalam susunan tertentu dalam sebuah kalimat atau sejumlah teks. Meskipun tidak sepenuhnya mencerminkan representasi utuh dari suatu teks, N-gram dapat dijadikan fitur yang memberikan gambaran tentang struktur dan konteks teks tersebut[12].

2.5. Klasifikasi (*Classification*)

Selanjutnya, data yang telah melalui *Text PreProcessing* dan pembobotan menggunakan TF-IDF sebelumnya akan dibagi menjadi data pelatihan dan data pengujian. Setelah model data instruksi diperoleh, metode Random Forest dan K-Nearest Neighbor akan digunakan untuk mengklasifikasikan data. Setelah itu, model tersebut akan diuji menggunakan data uji untuk mendapatkan hasil klasifikasi. Jupyter Notebook dan bahasa pemrograman Python digunakan untuk melakukan klasifikasi.

2.6. Evaluasi dan Visualisasi

Pada tahap evaluasi ini, confusion matrix digunakan untuk melihat hasil perbandingan klasifikasi data ulasan antara algoritma Random Forest (RF) dan K-Nearest Neighbors (KNN). Confusion matrix membantu dalam membandingkan hasil klasifikasi data antara kedua algoritma

tersebut. Kedua algoritma tersebut dapat mengukur akurasi, presisi, recall, dan F1-Score selama proses evaluasi.

3. HASIL DAN PEMBAHASAN

3.1. Pengambilan Data (Data Collection)

Data ulasan maskapai dikumpulkan menggunakan web scrapping dari situs web <https://www.airlinequality.com/> menggunakan pustaka Python BeautifulSoup (Scarping Web. Setiap ulasan mencakup informasi seperti nama maskapai, peringkat, judul ulasan, tanggal, dan atribut lainnya. Data ulasan yang didapat dari tahun 2016 sampai dengan tahun 2023.

Tabel 1. Hasil Data Maskapai Yang ada di Indonesia

Airline Name	Jumlah
AirAsia	100
Batik Air	100
Citilink	60
Garuda Indonesia	100
Lion Air	100
Sriwijaya Air	55
Wings Air	21

Dari data yang didapat, jumlah dari ulasan tidak sama, jadi yang diambil adalah jumlah ulasan 100 dan dibawah 100 dihapus.

Tabel 2. Data Ulasan Maskapai 100 Ulasan

Airline Name	Jumlah
AirAsia	100
Batik Air	100
Garuda Indonesia	100
Lion Air	100

3.2. Pemrosesan Text (Text PreProcessing)

Beberapa library yang dibutuhkan untuk membersihkan data adalah NLTK (Natural Language Toolkit): NLTK adalah salah satu librarby pemrosesan bahasa alami yang paling populer. Berbagai fungsi yang ditawarkan oleh NLTK termasuk *Cleansing*, *Case Folding*, *Tokenization*, *Filtering* dan *Stemming*[14]. Hasil preprocessing teks untuk dataset ditunjukkan di bawah ini.

Out[18]:

	Airline Name	Review
0	AirAsia	I recently flew from Cebu to gold coast and I ...
1	AirAsia	We had visa problems with Vietnam entry visa T...
2	AirAsia	I purchased a Premium Economy seat and paid fo...
3	AirAsia	Low cost airline so expectations should be se...
4	AirAsia	I tried to book directly with AirAsia and ther...
...
455	Lion Air	Singapore to Jakarta Obviously this years old...
456	Lion Air	Lion Air is the worst airline in Indonesia I'v...
457	Lion Air	Jakarta to Denpasar Due to take off pm Delay ...
458	Lion Air	Singapore to Jakarta I was supposed to board o...
459	Lion Air	Jambi to Jakarta Lion Air is infamous for dela...

400 rows × 2 columns

Gambar 2. Cleansing

Out[20]:

	Airline Name	Review
0	AirAsia	i recently flew from cebu to gold coast and i ...
1	AirAsia	we had visa problems with vietnam entry visa t...
2	AirAsia	i purchased a premium economy seat and paid fo...
3	AirAsia	low cost airline so expectations should be se...
4	AirAsia	i tried to book directly with airasia and ther...
...
455	Lion Air	singapore to jakarta obviously this years old...
456	Lion Air	lion air is the worst airline in indonesia i'v...
457	Lion Air	jakarta to denpasar due to take off pm delay ...
458	Lion Air	singapore to jakarta i was supposed to board o...
459	Lion Air	jambi to jakarta lion air is infamous for dela...

400 rows × 2 columns

Gambar 3. Case Folding

Out[21]:

	Airline Name	Review
0	AirAsia	[i, recently, flew, from, cebu, to, gold, coas...
1	AirAsia	[we, had, visa, problems, with, vietnam, entry...
2	AirAsia	[i, purchased, a, premium, economy, seat, and...
3	AirAsia	[low, cost, airline, so, expectations, should, ...
4	AirAsia	[i, tried, to, book, directly, with, airasia, ...
...
455	Lion Air	[singapore, to, jakarta, obviously, this, year...
456	Lion Air	[lion, air, is, the, worst, airline, in, indon...
457	Lion Air	[jakarta, to, Denpasar, due, to, take, off, pm...
458	Lion Air	[singapore, to, jakarta, i, was, supposed, to, ...
459	Lion Air	[jambi, to, jakarta, lion, air, is, infamous, ...

400 rows × 2 columns

Gambar 4. Tokenization

Out[22]:

	Airline Name	Review
0	AirAsia	recently flew cebu gold coast tow away large a...
1	AirAsia	visa problems vietnam entry visa air asia staf...
2	AirAsia	purchased premium economy seat paid lie flat b...
3	AirAsia	low cost airline expectations set accordingly ...
4	AirAsia	tried book directly airasia way add hold bagga...
...
455	Lion Air	singapore jakarta obviously years old aircraft...
456	Lion Air	lion air worst airline indonesia 've flight ja...
457	Lion Air	jakarta Denpasar due take pm delay information...
458	Lion Air	singapore jakarta supposed board pm got delaye...
459	Lion Air	jambi jakarta lion air infamous delays took fl...

400 rows × 2 columns

Gambar 5. Filtering

Out[23]:

	Airline Name	Review
0	AirAsia	recently flew cebu gold coast tow away large a...
1	AirAsia	visa problems vietnam entry visa air asia staf...
2	AirAsia	purchased premium economy seat paid lie flat b...
3	AirAsia	low cost airline expectations set accordingly ...
4	AirAsia	tried book directly airasia way add hold bagga...
...
455	Lion Air	singapore jakarta obviously years old aircraft...
456	Lion Air	lion air worst airline indonesia ve flight jam...
457	Lion Air	jakarta Denpasar due take pm delay information...
458	Lion Air	singapore jakarta supposed board pm got delaye...
459	Lion Air	jambi jakarta lion air infamous delays took fl...

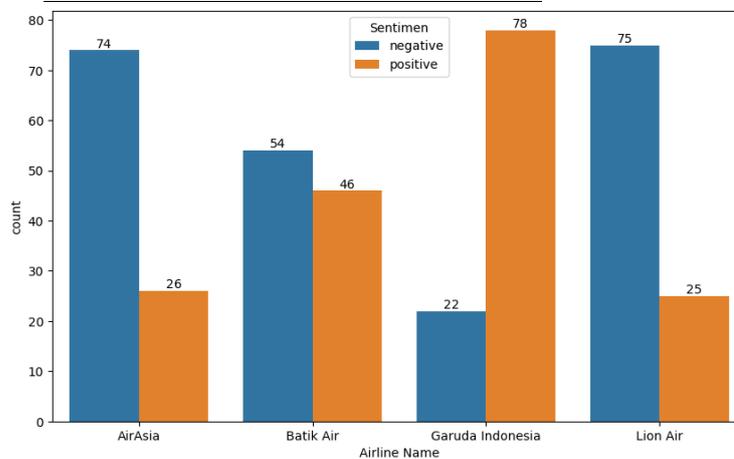
400 rows × 2 columns

Gambar 6. Stemming

Setelah melakukan stemming kemudian pelabelan dilakukan menggunakan lexicon base afinn[15][16], dimana setiap kata diberikan nilai apakah positif dan negative, dari hasil tersebut, dilakukan pelabelan apakah kalimat tersebut termasuk kalimat positif atau negative. Berikut adalah hasil dari pelabelan yang sudah dilakukan.

Tabel 3. Hasil dari Pelabelan yang telah dilakukan

Airline Name	Positif	Negatif
AirAsia	26	74
Batik Air	46	54
Garuda Indonesia	78	22
Lion Air	25	75



Gambar 7. Diagram Bar Hasil dari Pelabelan


```

from sklearn.ensemble import RandomForestClassifier

random_forest = RandomForestClassifier(n_estimators=100, random_state=42)
random_forest.fit(X_train, y_train)
rf_pred = random_forest.predict(X_test)
print("\nRandom Forest Accuracy:", accuracy_score(y_test, rf_pred))
print(classification_report(y_test, rf_pred, zero_division=0))
    
```

Gambar 12. Source Code Random Forest

```

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
knn_pred = knn.predict(X_test)
accuracy = accuracy_score(y_test, knn_pred)
print("\nKNN Accuracy: ", accuracy)
print("\nClassification Report:\n", classification_report(y_test, knn_pred, zero_division=0))
    
```

Gambar 13. Source Code KNN

Confusion matrix digunakan untuk mengevaluasi akurasi algoritma Random Forest (RF) dan K-Nearest Neighbor (K-NN) pada dataset. Ini mencakup True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dari sini, metrik seperti akurasi, presisi, recall, dan F1-Score dapat dihitung untuk memberikan gambaran kinerja model secara lebih rinci. Confusion matrix membantu mengidentifikasi sejauh mana model dapat memprediksi dengan benar dan mengklasifikasikan data.

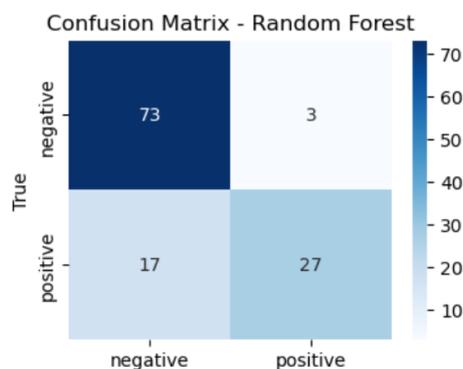
3. 5. Evaluasi (Evaluation)

Pengujian dilakukan pada bagian evaluasi ini dengan menggunakan metode confusion matrix. Selama proses evaluasi, algoritma Random Forest (RF) dan K-Nearest Neighbor (K-NN) mengukur Accuracy, precicion, recall, dan F1-score.

3.5.1. Random Forest

Dalam hal pengujian dengan Machine Learning telah dilakukan akurasi metode Random Forest dengan akurasi 83 %, gambar dibawah menunjukkan matrix confusion serta nilai Accuracy, precicion, recall, dan F1-score yang terkait dengan tes dengan dataset.

Random Forest Accuracy: 0.8333333333333334				
	precision	recall	f1-score	support
negative	0.81	0.96	0.88	76
positive	0.90	0.61	0.73	44
accuracy			0.83	120
macro avg	0.86	0.79	0.80	120
weighted avg	0.84	0.83	0.82	120



Gambar 14. Akurasi Menggunakan Random Forest dan Evaluasi Confusion Matrix

3.5.2. KNN

Dalam hal pengujian dengan Machine Learning telah dilakukan akurasi metode KNN dengan akurasi 82 %, gambar dibawah menunjukkan matrix confusion serta nilai Accuracy, precicion, recall, dan F1-score yang terkait dengan tes dengan dataset.

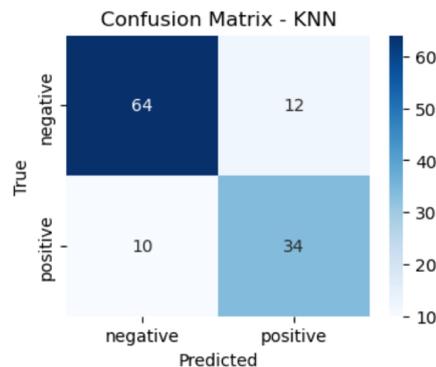
```

KNN Accuracy: 0.8166666666666667

Classification Report:
              precision    recall  f1-score   support

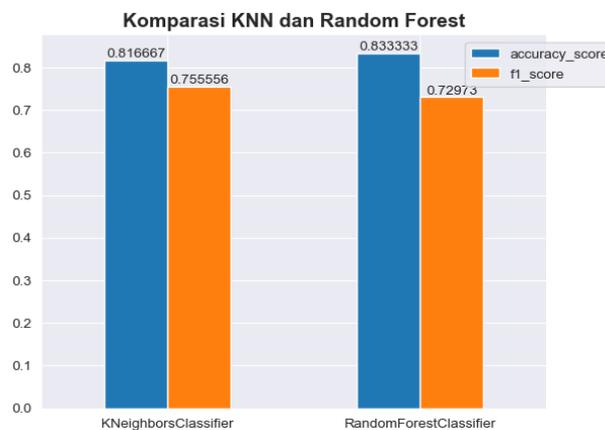
   negative    0.86    0.84    0.85        76
   positive    0.74    0.77    0.76        44

   accuracy    0.82
  macro avg    0.80    0.81    0.80    120
 weighted avg    0.82    0.82    0.82    120
    
```



Gambar 15. Akurasi Menggunakan KNN dan Evaluasi Confusion Matrix

Gambar 16 di bawah ini menunjukkan perbandingan hasil akurasi antara metode K-Nearest Neighbor dan Random Forest.



Gambar 16. Komparasi KNN dengan Random Forest

Tabel 4. Hasil Perbandingan Algoritman Random Forest dan KNN

Algoritma	Positif	Negatif
Random Fores	83%	73%
KNN	82%	75%

3. 6. Pembahasan

Dalam penelitian ini, dua metode yaitu Random Forest dan K-Nearest Neighbors (KNN) digunakan untuk menganalisis sentimen data dari Kaggle yang diambil dari www.airlinequality.com. Hasil menunjukkan bahwa Random Forest sedikit lebih akurat daripada KNN, karena KNN mencari nilai K data terdekat dalam ruang fitur untuk menentukan label data[6]. Kelebihan KNN adalah mudah dipahami dan mudah digunakan. Namun, dia sensitif terhadap skala data dan lambat saat

bekerja pada dataset besar. Namun, Random Forest adalah teknik kelompok yang membangun pohon keputusan selama pelatihan.[4]. Keuntungan Random Forest termasuk kemampuan untuk mengatasi overfitting dan kinerja yang baik pada berbagai jenis dataset. Dalam konteks analisis sentimen, Random Forest memberikan akurasi yang lebih tinggi, yang menunjukkan bahwa model Random Forest lebih cocok untuk dataset dan kondisi analisis yang diberikan. Namun demikian, sifat dataset, tujuan analisis, dan hasil eksperimen yang teliti harus menentukan pemilihan antara KNN dan Random Forest.

4. Kesimpulan

Pada penelitian ini, data ulasan maskapai penerbangan Indonesia dikumpulkan dari berbagai sumber dengan fokus pada ulasan dengan jumlah 400 atau lebih. Proses pemrosesan teks melibatkan langkah-langkah seperti *cleansing*, *case folding*, *tokenization*, *filtering*, dan *stemming* menggunakan library NLTK. Setelah itu, dilakukan pelabelan sentimen menggunakan lexicon base afinn, dan kata-kata dominan dari ulasan divisualisasikan melalui wordcloud untuk masing-masing maskapai. Ekstraksi fitur dilakukan dengan metode TF-IDF menggunakan SciKit-Learn. Model klasifikasi Random Forest dan K-Nearest Neighbor (KNN) digunakan untuk mengklasifikasikan sentimen ulasan. Hasil evaluasi menunjukkan bahwa baik Random Forest (akurasi 83%) maupun KNN (akurasi 82%) memberikan performa yang baik dengan nilai presisi, recall, dan F1-Score yang memuaskan. Komparasi antara keduanya menunjukkan hasil yang kompetitif. Secara keseluruhan, model ini efektif dalam mengklasifikasikan sentimen ulasan maskapai penerbangan, dengan catatan penting untuk memperhatikan kualitas data dan kemungkinan fine-tuning pada model.

Daftar Pustaka

- [1] "Indonesia Memiliki 13.466 Pulau yang Terdaftar dan Berkoordinat." Accessed: Dec. 19, 2023. [Online]. Available: <https://big.go.id/news/2014/05/12/indonesia-memiliki-13-466-pulau-yang-terdaftar-dan-berkoordinat>
- [2] "Apa Itu Maskapai Penerbangan dan Contohnya di Indonesia." Accessed: Dec. 19, 2023. [Online]. Available: <https://sttkd.ac.id/berita/apa-itu-maskapai-penerbangan-dan-contohnya-di-indonesia/>
- [3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.
- [4] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," *IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2019*, pp. 1–5, 2019, doi: 10.1109/INCOS45849.2019.8951367.
- [5] "Algoritme k tetangga terdekat - Wikipedia bahasa Indonesia, ensiklopedia bebas." Accessed: Dec. 20, 2023. [Online]. Available: https://id.wikipedia.org/wiki/Algoritme_k_tetangga_terdekat
- [6] F. R. Irawan, A. Jazuli, and T. Khotimah, "Analisis Sentimen Terhadap Pengguna Gojek Menggunakan Metode K-Nearest Neighbors Sentiment Analysis of Gojek Users Using K-Nearest Neighbor," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 1, pp. 62–68, 2022, doi: 10.33387/jiko.
- [7] A. W. Sari, T. I. Hermanto, and M. Defriani, "Sentiment Analysis Of Tourist Reviews Using K-Nearest Neighbors Algorithm And Support Vector Machine," *Sinkron*, vol. 8, no. 3, pp. 1366–1378, 2023, doi: 10.33395/sinkron.v8i3.12447.
- [8] S. Styawati, N. Hendrastuty, and A. R. Isnain, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, pp. 150–155, 2021, doi: 10.30591/jpit.v6i3.2870.
- [9] R. Rasenda, H. Lubis, and R. Ridwan, "Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 369, 2020, doi: 10.30865/mib.v4i2.2051.
- [10] R. A. Husen, R. Astuti, L. Marlia, R. Rahmaddeni, and L. Efrizoni, "Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma Machine Learning," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 211–218, 2023, doi:

- 10.57152/malcom.v3i2.901.
- [11] C. S. Pavan Kumar and L. D. Dhinesh Babu, *Novel text preprocessing framework for sentiment analysis*, vol. 105. Springer Singapore, 2019. doi: 10.1007/978-981-13-1927-3_33.
- [12] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.
- [13] M. Chiny, M. Chihab, Y. Chihab, and O. Bencharef, "LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 265–275, 2021, doi: 10.14569/IJACSA.2021.0120730.
- [14] M. Wang and F. Hu, "The Application of NLTK Library for Python Natural Language Processing in Corpus Research," *Theory Pract. Lang. Stud.*, vol. 11, no. 9, pp. 1041–1049, Sep. 2021, doi: 10.17507/TPLS.1109.09.
- [15] W. Ahmed, N. A. Semary, K. Amin, and M. Adel Hammad, "Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study," *IJCI. Int. J. Comput. Inf.*, vol. 10, no. 3, pp. 52–57, Nov. 2023, doi: 10.21608/IJCI.2023.236052.1128.
- [16] V. Sheth, U. Tripathi, and A. Sharma, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," *Procedia Comput. Sci.*, vol. 215, pp. 422–431, Jan. 2022, doi: 10.1016/J.PROCS.2022.12.044.
- [17] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Comput. Sci.*, vol. 152, pp. 341–348, Jan. 2019, doi: 10.1016/J.PROCS.2019.05.008.



ZONAsi: Jurnal Sistem Informasi

is licensed under a [Creative Commons Attribution International \(CC BY-SA 4.0\)](https://creativecommons.org/licenses/by-sa/4.0/)