

## ANALISIS SENTIMEN PENGGUNA *ONLINE TRAVEL AGENT* (OTA) PADA PERUSAHAAN PEGIPEGI.COM MENGGUNAKAN *RANDOM FOREST*

Ayu Rahma Lestari<sup>1\*</sup>, Rukun Santoso<sup>2</sup>, Suparti<sup>3</sup>

<sup>1,2,3</sup>Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro

\*e-mail : [ayurahmalstr@gmail.com](mailto:ayurahmalstr@gmail.com)

DOI: 10.14710/J.GAUSS.12.4.616-624

### Article Info:

Received: 2023-05-08

Accepted: 2024-07-20

Available Online: 2024-07-29

### Keywords:

*Application; Pegipegi; Sentiment Analysis, Random Forest*

**Abstract:** The presence of the internet makes online applications increasingly attractive to the public in supporting their daily activities. Online applications have developed rapidly, including online travel agent (OTA) companies such as Pegipegi. Pegipegi is a platform designed to meet the community's tertiary needs, such as providing accommodations for vacations. Pegipegi has an application that can be downloaded through the Google Playstore. Google Playstore provides a review feature as a medium for communication between application owners and consumers to express opinions that felt when using the application. The reviews submitted can be used as data to carry out sentiment analysis. Data collection was carried out on 11 December 2021 – 11 December 2022. A total of 2926 reviews obtained. Sentiment analysis was able to proceed by a classification method. This research used Random Forest to classify opinions on positive and negative sentiments. Random Forest is a classification model based on the majority vote of all decision trees. Classification using Random Forest produces an accuracy of 92.27% and AUC-ROC of 82.35%. Based on this accuracy and AUC-ROC value, the Random Forest algorithm has a good model performance in classifying the opinions of Pegipegi application users because it has a good accuracy and AUC-ROC value.

## 1. PENDAHULUAN

Perkembangan teknologi di era digital yang terus-menerus berkembang pesat telah membawa berbagai perubahan di masyarakat. Kemajuan teknologi dapat dilihat pada teknologi informasi dan komunikasi seperti Internet. Kehadiran internet menjadikan aplikasi *online* semakin diminati masyarakat dalam menunjang aktivitas sehari-hari. Aplikasi *online* mengalami perkembangan yang cukup pesat termasuk pada perusahaan *online travel agent* (OTA) seperti Pegipegi. Pegipegi merupakan suatu *platform* yang dirancang untuk memenuhi kebutuhan tersier masyarakat yaitu menyediakan akomodasi untuk berlibur. Pegipegi terus melakukan perubahan demi perubahan untuk menjadi OTA terbaik. Hal ini dapat ditemukan melalui aplikasi Pegipegi yang dapat diunduh melalui *Google Playstore*. *Google Playstore* adalah toko *online* besar yang menyediakan jutaan aplikasi, buku, dan permainan untuk Android (Olabenjo, 2016). Fitur *review* yang disediakan di *Google Playstore* memungkinkan pengguna untuk memberikan testimoninya tentang aplikasi yang mereka gunakan, baik positif maupun negatif. *Review* tersebut dapat menjadi masukan untuk pengembang aplikasi agar dapat terus memperbaiki aplikasi mereka. Analisis sentimen diperlukan untuk memproses *review-review* tersebut agar dapat menghasilkan informasi yang berguna bagi pengembang aplikasi. Analisis sentimen dapat dilakukan dengan metode klasifikasi seperti *random forest*. Prinsip kerja dari *random forest* dalam sentimen yaitu menggunakan kombinasi pohon keputusan.

Dalam penelitian ini, analisis sentimen dilakukan pada perusahaan OTA Pegipegi menggunakan algoritma *random forest*. Sistematisa pencarian data dalam penelitian ini

dilaksanakan menggunakan *web scrapping* dari *Google Playstore*. Penelitian ini memiliki tujuan yaitu untuk mengetahui performa kinerja algoritma *random forest* dalam mengklasifikasikan *review* aplikasi Pegipegi berdasarkan nilai akurasi dan nilai AUC-ROC untuk mengetahui pandangan masyarakat terhadap aplikasi Pegipegi.

## 2. TINJAUAN PUSTAKA

Pegipegi merupakan salah satu *platform* yang memberikan layanan pembelian tiket dan penginapan secara *online*. Pegipegi didirikan pada 07 Mei 2012 dan diresmikan langsung oleh Menteri Pariwisata dan Ekonomi Kreatif pada Agustus 2013. Saat ini, Pegipegi telah terkoneksi pada rute penerbangan sebanyak lebih dari 25.000, dan juga pilhan hotel yang lebih dari 25.000, serta untuk rute kereta api sebanyak lebih dari 2.800 yang dapat dibeli langsung lewat *website* ataupun *platform* ponsel (Pegipegi, 2022).

*Google Play* adalah toko *online* yang berisi aplikasi, *game*, film, acara TV, buku, dan konten lainnya. Pada 22 Oktober 2008 toko ini pertama kali dilaunching bermula dinamai *Android Market*. *Google Play* menyediakan fitur tambahan pada setiap aplikasi di dalamnya seperti fitur penilaian, ukuran file, dan fitur komentar yang dapat digunakan oleh para penggunanya.

Analisis sentimen merupakan sebuah teknologi pengolahan data yang digunakan untuk mengekstrak, menganalisis, dan memproses teks dalam suatu entitas seperti produk, layanan, fenomena, atau topik tertentu untuk mendeteksi opini, pandangan, perasaan, subjektivitas, dan aspek lainnya yang terkait dengan entitas tersebut (Liu, 2015).

*Text mining* ialah proses dari kegiatan menganalisa teks yang dilaksanakan untuk mendapatkan informasi dari sebuah susunan kata yang diresume kedalam suatu arsip. *Text mining* tidak hanya dilakukan untuk sekadar mendapatkan informasi saja, tetapi sering juga digunakan untuk menyelesaikan permasalahan seperti klasifikasi dan *clustering*. *Text mining* akan memproses data teks berubah menjadi data angka yang ditujukan untuk dilaksanakannya pemrosesan lanjutan yaitu *preprocessing* data (Feldman dan Sanger, 2007).

*Pre-processing* data adalah proses untuk mempersiapkan data yang masih mentah dan tidak berstruktur menjadi lebih terstruktur. *Pre-processing* data bertujuan untuk merubah data ke dalam suatu format agar lebih mudah dipahami. *Pre-processing* memuat beberapa tahapan lanjutan, yaitu:

1. *Case Folding*  
*Case folding* adalah langka pengubahan semua huruf kapital jadi huruf kecil dengan maksud agar seluruh kata pada dokumen teks seragam.
2. *Remove Punctuation*  
*Remove punctuation* merupakan tahapan penghapusan seluruh karakter yang tidak termasuk *alphabet* seperti tanda baca atau simbol. Tanda baca perlu dihapus karena tidak diperlukan dalam proses analisis sentimen.
3. *Remove Number*  
*Remove number* merupakan tahapan penghapusan angka yang terdapat pada dokumen. Penggunaan angka tidak dapat memberikan penilaian atau menunjukkan suatu perasaan sehingga perlu dihapus.
4. *Remove Emoticon*  
*Remove emoticon* merupakan tahapan penghapusan simbol/emoji yang terdapat pada ulasan. Simbol atau emoji tidak dapat memberikan penilaian sehingga perlu dihapus dalam proses analisis sentimen.
5. Normalisasi Kata  
Normalisasi kata merupakan tahapan pengubahan kata yang tidak baku menjadi kata baku sehingga dapat mempermudah pemahaman terhadap teks. Kata tidak baku yang

dimaksud yaitu kata yang mengandung bahasa daerah maupun singkatan. Perubahan kata tidak baku menjadi baku dilakukan dengan cara mencocokkan dengan kamus KBBI.

Proses pelabelan kelas sentimen menggunakan kamus leksikon untuk mengelompokkan data teks menjadi positif atau negatif. Nilai sentimen dihitung dengan mengurangi jumlah nilai teks positif dengan jumlah nilai teks negatif dalam seluruh kata ulasan. Kata dengan nilai  $\geq 0$  dikategorikan sebagai positif, sedangkan nilai  $< 0$  dikategorikan sebagai negatif. (Santoso dan Nugroho, 2019).

*Feature Selection* ialah metode *pre-processing* yang berguna dalam mengecilkan sejumlah layanan data dan memilih variabel terbaik yang relevan dengan kelas target. Hal ini bertujuan untuk memperoleh nilai kelas target yang akurat dengan memilih variabel yang relevan dan menghindari data yang menyebabkan kesalahan dalam klasifikasi (Ningtyas, 2016). *Feature Selection* memuat beberapa tahapan, yaitu:

1. *Tokenizing*

*Tokenizing* merupakan tahapan pemangkasan *string* inputan didasarkan teks-teks yang membentuknya atau disebut juga pemisahan susunan teks menjadi kata.

2. *Stopwords Removal*

*Stopwords removal* adalah proses untuk menghilangkan konjungsi, preposisi dan kata ganti.

3. *Stemming*

*Stemming* yakni sistematika perubahan bentuk sebuah teks jadi teks intinya. tiap teks berimbuhan (awalan dan akhiran) akan dibuang dan dibentuk teks inti.

*Term Weighting* atau pemberian skor teks ialah proses pengubahan data kualitatif jadi data kuantitatif dengan menetapkan skor atau nilai pada setiap kata dengan tujuan untuk mempermudah proses klasifikasi (Nurjanah dkk., 2017). Pemberian skor yang umum dipakai ialah teknik *Term Frequency-Inverse Document Frequency* (TF-IDF) sebagaimana dinyatakan pada Persamaan (1):

$$W_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log_e \frac{N}{df(i)} + 1 \quad (1)$$

dengan  $W_{i,j}$  pemberian skor TF-IDF pada term ke- $i$  dan arsip ke- $j$ ,  $n_{i,j}$  besarnya angka munculnya term ke- $i$  pada arsip ke- $j$ ,  $\sum_k n_{k,j}$  banyaknya muncul sejumlah term ke- $i$  pada arsip ke- $j$ ,  $N$  jumlah arsip keseluruhan dan  $df(i)$  banyaknya dokumen yang mengandung term  $i$ .

*Random forest* merupakan bagian dari pengembangan metode *Classification and Regression Tree* (CART) (Breiman, 2001). *Classification and Regression Tree* (CART) merupakan metode klasifikasi data dengan melakukan eksplorasi terhadap data yang didasarkan pada teknik pohon keputusan. Algoritma CART menggunakan *gini index* sebagai kriteria dalam pemilahan simpul untuk membentuk pohon keputusan (Sa'adah dkk., 2021).

*Random Forest* adalah metode klasifikasi yang menggabungkan banyak pohon klasifikasi untuk meningkatkan akurasi. Pohon-pohon ini dibangun dengan memilih variabel yang dipakai guna *split* dengan *random*, serta setiap pohon akan menghasilkan output yang berbeda-beda. Proses klasifikasi dilakukan setelah semua pohon terbentuk. Metode ini menggunakan proses *majority vote* untuk mendapatkan prediksi pada klasifikasi akhir.

Prosedur untuk membangun *random forest* disebut *bootstrap*, yang terdiri dari mengambil sampel acak dengan pengembalian, membangun pohon tanpa pemangkasan dengan *random feature selection*, dan mengulangi proses ini untuk membentuk *forest* yang terdiri atas banyak pohon (Breiman dan Cutler, 2003).

Nilai *mtry* dapat mempengaruhi kebaikan model yang dihasilkan dan dapat diubah untuk mengatur banyaknya variabel independen yang dipakai menjadi calon pembagi ketika

membentuk pohon. Pemakaian *mtry* yang tepat kemudian memperoleh *random forest* sebagai hubungan antara pohon yang terbilang minim akan tetapi kapabilitas tiap pohon dinilai kuat.

Penentuan proses akhir klasifikasi dengan metode *random forest* dapat menggunakan nilai *Gini index* untuk menentukan variabel *split* yang akan dijadikan *root* (akar) atau *node* (pemilah). Perhitungan *gini index* dan *gini split* terhadap klasifikasi dilakukan menggunakan Persamaan (2) dan (3):

$$\text{Gini Index (Si)} = 1 - \sum_{i=1}^c \pi_i^2 \quad (2)$$

Dengan  $\pi_i$  adalah probabilitas dari *gini index* milik bagian *i* serta *c* sejumlah skor fasilitas yang terhimpun ke pada sebuah bagian didasarkan pada fasilitas data.

$$\text{Gini Split} = \sum_{i=1}^c \left(\frac{n_i}{n}\right) \times \text{Gini Index (Si)} \quad (3)$$

Dengan  $n_i$  jumlah sampel setelah di *split*, *n* jumlah sampel pada *node* yang diberikan, *c* sejumlah skor fasilitas yang tergolong ke pada sebuah bagian didasarkan pada fasilitas data.

*Random Forest* mengindikasikan respon sebuah pengamatan dengan mengkolaborasikan output perkiraan *n tree* pohon dan memilih kategori bagian dinilai lebih dominan terdeteksi selaku output perkiraan.

Sistem pengelompokan dapat mengalami kesalahan dalam sistematika pengelompokan hingga dibutuhkan dilaksanakannya perhitungan kapabilitas dengan menggunakan data uji. Salah satu alat yang penting dalam menganalisis berhasil atau tidaknya suatu *classifier* adalah *confusion matrix*.

**Tabel 1. Confusion Matrix**

		Kelas Prediksi	
		Positif	Negatif
Kelas Aktual	Positif	(TP)	(FN)
	Negatif	(FP)	(TN)

Salah satu metrik yang umum dipakai guna mengukur tingkat akurasi dalam proses klasifikasi adalah nilai akurasi keseluruhan. Akurasi merupakan tingkat ketepatan antara nilai prediksi dengan nilai sebenarnya. Makin tinggi angka kepastiannya, dinilai kinerjanya terkelompok makin baik. Angka kepastiannya bisa diperhitungkan dengan Persamaan (4):

$$\text{Accuracy} = \frac{\text{predeteksi data benar}}{\text{total data}} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (4)$$

Selain itu, untuk memperkuat evaluasi kinerja algoritma, dilakukan pengujian menggunakan Area Under the Curve (AUC). AUC adalah ukuran yang menggambarkan luas area di bawah kurva Receiver Operating Characteristic (ROC). Nilai AUC berkisar antara 0 hingga 1 dan memberikan gambaran tentang tingkat kecocokan model dengan data yang digunakan. Semakin besar nilai AUC, semakin baik kemampuan variabel yang diuji dalam memprediksi kejadian tertentu (Maskoen dan Purnama, 2018). Nilai AUC-ROC sering digunakan sebagai metrik evaluasi pada kelas data yang tidak seimbang dan cenderung memberikan hasil yang konsisten (Wardhani dkk., 2019). Dalam konteks klasifikasi data mining, nilai AUC dapat dibagi ke dalam beberapa kategori, seperti klasifikasi sangat baik dengan rentang 0,90 hingga 1, klasifikasi baik dengan rentang 0,80 hingga 0,90, klasifikasi cukup dengan rentang 0,70 hingga 0,80, klasifikasi buruk dengan rentang 0,60 hingga 0,70, dan klasifikasi yang salah dengan rentang 0,50 hingga 0,60 (Sari dkk., 2020). Nilai AUC-ROC dapat dilihat pada Tabel 2.

**Tabel 2.** Nilai AUC-ROC

Ukuran	Rumus
TPR (True Positive Rate)	$TPR = \frac{TP}{(TP + FN)}$
FPR (False Positive Rate)	$FPR = \frac{FP}{(FP + TN)}$

*Word cloud* adalah metode visualisasi data teks yang menampilkan kumpulan kata dalam *Term Document Matrix* dengan ukuran kata yang semakin besar menunjukkan frekuensi yang semakin tinggi (Pradana, 2020).

### 3. METODE PENELITIAN

Data yang dipakai dalam penelitian ini ialah data kualitatif, yakni data hasil *web scrapping* dari *google playstore*. Data yang digunakan sebanyak 2926 *review* aplikasi Pegipegi dengan rentang waktu pengambilan data dari 11 Desember 2021 – 11 Desember 2022.

Langkah-langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut:

1. *Web Scrapping* data
2. *Pre-processing* data
3. Pelabelan Kelas Sentimen
4. *Feature Selection*
5. Pembobotan TF-IDF
6. Tahap Klasifikasi
  - a. Mengklasifikasikan data jadi data *training* dan data *testing* 70%:30%
  - b. Menentukan parameter yang digunakan pada klasifikasi *random forest* yaitu nilai *mtry* dan nilai *nree*.
  - c. Membangun model klasifikasi *random forest* dengan menggunakan perhitungan *gini Index* dan *gini split*.
7. Evaluasi model klasifikasi dengan *confusion matrix*
8. *Wordcloud*

### 4. HASIL DAN PEMBAHASAN

Data yang dipakai ialah *review* pengguna aplikasi Pegipegi pada kolom komentar di *Google Playstore* dengan cara *web scrapping* menggunakan bahasa pemrograman *Python* dengan *Google Collaboratory*. Jumlah data yang diperoleh dari proses *web scrapping* adalah sebanyak 2926 dengan pengambilan dibatasi dari rentang waktu 11 Desember 2021 – 11 Desember 2022. Setelah dilakukan *pre-processing*, data yang digunakan sebanyak 2747 *review*.

Data *review* aplikasi pegipegi yang terdapat pada *Google Playstore* merupakan data teks yang tidak terstruktur dan memiliki berbagai macam tipe. Data tersebut terdiri atas simbol, angka, singkatan, dan emoji yang beragam pada setiap kalimatnya sehingga perlu diolah menggunakan *text mining* dengan bahasa pemrograman *Python* pada *Google Collaboratory*. Proses ini merupakan tahap transformasi data tidak tersusun jadi informasi yang tersusun sehingga memudahkan untuk melakukan langkah penelitian selanjutnya.

Pelabelan data dilaksanakan secara otomatis dengan menggunakan kamus *lexicon*. Proses pelabelan kelas sentimen menghasilkan sebanyak 2462 dokumen dengan label sentimen positif dan 285 dokumen dengan label negatif.

Pembobotan kata TF-IDF digunakan untuk mengubah kata-kata menjadi angka yang bisa diperhitungkan serta dikelola. Penetapan skor teks TF-IDF bertujuan untuk analisis sentimen pada data jenis kualitatif. Perolehan dari penetapan skor kata bisa diamati di Tabel 3.

**Tabel 3.** Hasil Pembobotan TF-IDF

Sampel Ke-	admin	bantu	kasih	pergi	...	yogyakarta
1	0	0	0	1,9673	...	0
2	0	1,7466	0	0	...	0
3	0	1,1644	1,3327	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2747	0	0	0	0	...	0

Proses klasifikasi di penelitian ini memakai data *training* dan data *testing* dalam pembuatan model *machine learning*. Proporsi terbaik pembagian data *training* 70% dan data *testing* 30%.

Metode *random forest* menggunakan perhitungan gini untuk menetapkan pemisahan *node*. Algoritma ini terus melaksanakan *splitting* hingga penghitungan *gini* mencapai angka = 0. Pembuatan tiap pohon dilakukan secara acak dan tidak dapat diketahui baris data yang digunakan. Contoh dataset sederhana bisa diamati di Tabel 4.

Tabel 4. *Dataset* yang disederhanakan

Review ke-	Bantu	Label Sentimen
1	1,1644	Positif
2	1,7466	Positif
3	0	Negatif
4	0	Positif
5	0	Negatif

Tabel 4 menunjukkan hasil pembobotan TF-IDF *term* “bantu” pada 5 *review* dan label sentimen yang akan dilakukan perhitungan *gini* dari masing-masing *review*. Perhitungan *gini index* dan *gini split term* “bantu” pada *review* pertama adalah sebagai berikut:

$$Gini\ index\ (bantu \leq 1,1644) = 1 - \left[ \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right] = 1 - 0,50 = 0,50$$

$$Gini\ index\ (bantu > 1,1644) = 1 - \left[ \left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2 \right] = 1 - 1 = 0$$

$$Gini\ Split\ (bantu) = \left(\frac{4}{5}\right) (0,50) + \left(\frac{1}{5}\right) (0) = 0,40$$

Perhitungan dari *gini index* dan *gini split term* “bantu” pada seluruh *review* dihitung dengan cara yang sama dan bisa diamati di Tabel 5.

Tabel 5. Perhitungan *Gini Index* dan *Gini Split* pada *term* “bantu” pada seluruh *review*

Bobot TF-IDF Term Bantu	Label		Total	Gini Index	Gini Split
	Neg	Pos			
≤ 0	2	1	3	0,4444444444	0,2666666667
> 0	0	2	2	0	
≤ 1,1644	2	2	4	0,5	0,4
> 1,1644	0	1	1	0	
≤ 1,7466	2	3	5	0,48	0,48





- Pengembangan Teknologi Informasi dan Ilmi Komputer, Vol.1, No.12, Hal: 1750-1757.
- Olabenjo, B. 2016. Applying Naive Bayes Classification to Google Play Apps categorization. Jurnal: *University of Saskatchewan*
- Pegipegi. 2022. About Pegipegi. Tersedia: <https://www.pegipegi.com/team/> (diakses pada tanggal 26 November 2022).
- Pradana, M. G. 2020. Penggunaan Fitur *Wordcloud* dan *Document Term Matrix* dalam *Text Mining*. Jurnal Ilmiah Informatika, Vol. 8, No.1, Hal: 38-43.
- Sa'adah U., Rochayani, M. Y., Lestari, D. W., dan Lusya, D. A. 2021. Kupas Tuntas Algoritma *Data Mining* dan Implementasinya Menggunakan R. Malang: UBPress.
- Santoso, E. B., dan Nugroho, A. 2019. Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook. Jurnal Eksplora Informatika, Vol. 9, No.1, Hal: 60-69.
- Sari, V. R., Firdausi, F., dan Azhar, Y. 2020. Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. EDUMATIC: Jurnal Pendidikan Informatika, Vol. 4 No.2, Hal: 1-9.
- Wardhani, N. W. S., Rochayani, M. Y, Iriany, A., Sulistyono, A. D., dan Lestantyo, P. 2019. *Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data*. Hal: 14-18.