

ANALISIS SENTIMEN MASKAPAI PENERBANGAN MENGGUNAKAN METODE NAIVE BAYES DAN SELEKSI FITUR INFORMATION GAIN

Arif Bijaksana Putra Negara¹, Hafiz Muhandi², Indira Melinda Putri³

^{1,2,3}Program Studi Informatika Fakultas Teknik, Universitas Tanjungpura, Indonesia
Email: ¹arifbpn@informatika.untan.ac.id, ²hafiz.muhandi@informatika.untan.ac.id, ³indiraputri196@gmail.com,

(Naskah masuk: 22 April 2019, diterima untuk diterbitkan: 27 April 2020)

Abstrak

Zaman sekarang tren masyarakat untuk memesan tiket pesawat sudah melalui situs-situs *booking online*. Pegipegi.com merupakan salah satu *website* yang menyediakan pemesanan tiket dan menyediakan fitur ulasan bagi pengunjung untuk menyampaikan opini. Pengunjung lain yang membaca ulasan-ulasan tersebut dapat memperoleh gambaran secara lebih objektif mengenai maskapai penerbangan. Ulasan pengguna yang terdapat pada website pegipegi.com saat ini sudah sangat banyak sehingga hal ini menyulitkan dan memakan waktu untuk membaca secara keseluruhan. Oleh karena itu dirancang analisis sentimen guna membantu mengklasifikasi ulasan kedalam kategori positif atau negatif sehingga dapat memberikan rekomendasi maskapai penerbangan berdasarkan jumlah kategori ulasan. Metode yang diterapkan untuk klasifikasi sentimen adalah Naïve Bayes dengan seleksi fitur *Information Gain*. Adapun tujuan dari penelitian ini adalah mengetahui pengaruh dari pemilihan fitur *Information Gain* terhadap akurasi klasifikasi dan membuktikan bahwa metode Naïve Bayes dengan *Information Gain* dapat digunakan untuk klasifikasi analisis sentimen. Hasil pengujian yang telah dilakukan menunjukkan bahwa nilai rata-rata akurasi, *precision*, *recall* setelah penambahan *Information Gain* menunjukkan hasil yang lebih baik sebesar 0,865 jika dibandingkan sebelum penambahan *information gain* yakni sebesar 0,81.

Kata kunci: analisis sentimen, Naïve Bayes, *Information Gain*, akurasi

SENTIMENT ANALYSIS ON AIRLINES USING NAÏVE BAYES METHOD AND FEATURE SELECTION INFORMATION GAIN

Abstract

Nowadays people tend to order airplane tickets through online booking sites. Pegipegi.com is a website that provides ticket reservations and a review section for visitors to express their opinions. Other visitors who read the reviews can get a more objective picture of airlines. The user reviews contained on the pegipegi.com website are currently very large so this makes it difficult and time consuming to read in its entirety. Therefore sentiment analysis is designed to help classify reviews into positive or negative categories so that they can provide airline recommendations based on the number of review categories. The method applied for sentiment classification is Naïve Bayes with the *Information Gain* feature selection. The purpose of this study was to determine the effect of selecting the *Information Gain* feature on classification accuracy and prove that the Naïve Bayes method with *Information Gain* can be used for the classification of sentiment analysis. The results of the tests that have been done show that the average value of accuracy, *precision*, *recall* after adding *Information Gain* shows better results of 0.865 compared to the addition of *information gain* which is equal to 0.81.

Keywords: sentiment analysis, Naïve Bayes, *Information Gain*, accuracy

1. PENDAHULUAN

Zaman sekarang tren masyarakat untuk memesan tiket melalui situs-situs booking secara online, baik reservasi hotel, pesawat terbang, kereta api, bus, hiburan dan sebagainya. Layanan situs yang menyediakan berbagai pemesanan tiket secara online diantaranya adalah tiket.com, utiket.com, booking.com, blibli.com, traveloka.

Salah satu layanan pemesanan tiket secara online adalah pegipegi.com. Pegipegi.com merupakan perusahaan yang menyediakan layanan pemesanan tiket berbasis online bagi pengguna sehingga dapat memberikan kemudahan dalam mengatur serta merancang perjalanan mulai dari segi transportasi maupun akomodasi. Situs tersebut juga dilengkapi dengan fitur yang sangat membantu

pengunjung dalam menentukan pesawat terbang mana yang akan dipilih sebagai transportasi menuju ke suatu daerah.

Fitur yang dimaksud adalah adanya ulasan yang memuat berbagai komentar pengunjung yang menggunakan pemesanan tiket pesawat terbang yang ada di pegipegi.com. Pengunjung dapat memperoleh gambaran secara lebih objektif dengan adanya ulasan-ulasan tersebut. Sehingga memudahkan pengguna dalam memilih pesawat terbang sebagai transportasi. Keberadaan ulasan-ulasan yang dituliskan oleh para pengunjung situs tersebut kemudian akan dianalisis sehingga dapat dihasilkan sebuah keluaran yang bisa bermanfaat. Salah satu model analisis yang dapat dilakukan adalah analisis sentimen.

Menurut Liu (2013), analisis sentimen pada review adalah proses menyelidiki review produk di internet untuk menentukan opini atau perasaan terhadap suatu produk secara keseluruhan. Tujuan dari analisis sentimen adalah untuk menentukan perilaku atau opini dari seorang penulis dengan memperhatikan suatu topik tertentu (Basari et al., 2013). Menurut Thelwall analisis sentimen diperlakukan sebagai suatu tugas klasifikasi yang mengklasifikasikan orientasi suatu teks ke opini positif atau negatif. Terdapat beberapa metode klasifikasi yang umumnya digunakan untuk analisis sentimen antara lain *Support Vector Machine*, *Naïve Bayes*, *Character Based N-gram Model*, *Artificial Neural Evolution*. Pengklasifikasian *Naïve Bayes* merupakan algoritma yang populer untuk klasifikasi teks, dan memiliki performa yang baik pada banyak domain (Zhang Ziqiong et al., 2009).

Salah satu masalah pada klasifikasi sentimen adalah banyaknya atribut yang digunakan pada sebuah dataset (Wang et al., 2011). Pada umumnya atribut dari klasifikasi sentimen teks sangat besar, dan jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari *classifier*. Atribut yang banyak membuat nilai akurasi menjadi rendah, maka atribut yang ada harus dipilih dengan algoritma yang tepat.

Penelitian yang sudah dilakukan dalam pengklasifikasian sentimen menggunakan klasifikasi *Naïve Bayes*, diantaranya adalah penelitian oleh Ziqiong Zhang (2011), membandingkan antara *Support Vector Machine* dan *Naïve Bayes* sehingga mendapatkan hasil yang terbaik adalah metode *Naïve Bayes*. Pengklasifikasian sentimen digunakan untuk menentukan ulasan restoran Cantonese online ke dalam kelas positif dan negatif. Sedangkan penelitian oleh Vinita Chandani et al (2018), melakukan komparasi tiga algoritma klasifikasi yaitu *Support Vector Machine*, *Naïve Bayes*, dan *Artificial Neural Network* dan mengkomparasi empat algoritma *feature selection* yaitu *Information Gain*, *Chi Square*, *Forward Selection*, dan *Backward Elimination*. Hasil dari komparasi algoritma klasifikasi adalah *Support Vector Machine*

mendapatkan nilai akurasi terbaik, sedangkan untuk komparasi algoritma *feature selection* didapatkan *information gain* dengan hasil terbaik.

Metode pengklasifikasi *Naïve Bayes* merupakan metode yang sangat sederhana dan efisien. Disamping kesederhanaannya, *Naïve Bayes* masih memiliki beberapa kekurangan yakni sangat sensitif dalam pemilihan fitur. Jumlah fitur yang terlalu banyak tidak hanya

dapat meningkatkan waktu perhitungan tetapi juga dapat menurunkan akurasi klasifikasi. Tingkatan lain yang umumnya ditemukan dalam pendekatan klasifikasi sentimen adalah pemilihan fitur. Pemilihan fitur bisa membuat pengklasifikasi baik lebih efisien atau efektif dengan mengurangi jumlah data yang dianalisis. Beberapa metode pemilihan fitur yang sering digunakan antara lain *Document Frequency*, *Mutual Information*, *Information Gain*, dan *Chi-Square*. Menurut Uysal & Gunal (2012), *Information Gain* sering lebih unggul dibanding yang lainnya. *Information Gain* mengukur berapa banyak informasi kehadiran dan ketidakhadiran dari suatu kata yang berperan untuk membuat keputusan klasifikasi yang benar dalam class apapun. *Information Gain* adalah salah satu pendekatan filter yang sukses dalam pengklasifikasian teks.

Berdasarkan latar belakang diatas ulasan pengguna juga bermanfaat bagi perusahaan pegipegi.com untuk mendapatkan kepercayaan dari konsumen. Namun jumlah ulasan sangat banyak sehingga akan menyulitkan dan memakan waktu untuk membaca secara keseluruhan. Oleh karena itu dirancang sistem yang secara otomatis akan mengelompokkan ulasan pegipegi.com. yaitu Analisis sentimen menggunakan metode klasifikasi *Naïve Bayes* dengan *Information Gain* sebagai metode pemilihan fitur.

2. LITERATURE REVIEW

Berikut adalah beberapa penelitian tentang Analisis Sentimen sebelumnya yang pernah dilakukan oleh beberapa peneliti.

Zhang et al., (2011), mengklasifikasi ulasan menggunakan metode *Naive Bayes* dan *Support Vector Machine* serta menggunakan enam presentasi fitur n-gram kehadiran atau frekuensi untuk memeriksa efek dari pengklasifikasi dan opsi fitur pada bahasa Cantonese klasifikasi sentimen. Hasil pengujian menunjukkan bahwa Metode *Naive Bayes* mendapatkan nilai akurasi yang lebih baik dengan hanya memperhitungkan *feature selection* (pemilihan fitur) dan tidak memperhitungkan frekuensi fitur. Sedangkan pada *Support Vector Machine* menghasilkan hasil yang lebih baik dengan ada atau tidaknya unigram, bigram frekuensi dan frekuensi trigram. Data latih yang digunakan berjumlah 900 ulasan positif dan 900 ulasan negatif yang diambil secara acak pada situs openrice.com dalam bahasa Cantonese. Pada penelitian ini *feature*

selection digunakan untuk menghitung jumlah bit informasi yang diperoleh kategori untuk prediksi dan mengetahui kehadiran dan ketidakhadiran suatu *term* dalam dokumen. Berdasarkan pengujian yang telah dilakukan menunjukkan bahwa, akurasi terbaik dihasilkan oleh Naive Bayes dari masing-masing representasi fitur lebih baik jika dibandingkan dengan *Support Vector Machine*.

Chandani, V., et al (2015) dengan mengkomparasi metode klasifikasi *Machine Learning* seperti Naive Bayes, *Support Vector Machine*, dan *Artificial Neural Network* dan seleksi fitur seperti *Information Gain*, *Chi Square*, *Forward Selection* dan *Backward Elimination*. Hasil komparasi metode SVM mendapat hasil yang terbaik dengan akurasi 81.10% dan AUC 0.904. Hasil dari komparasi seleksi fitur *Information Gain* mendapatkan hasil yang paling baik dengan rata – rata akurasi 84.57% dan rata – rata AUC 0.899. Hasil integrasi metode klasifikasi terbaik dan metode seleksi fitur terbaik menghasilkan akurasi 81.50% dan AUC 0.929. Hasil ini mengalami kenaikan jika dibandingkan hasil eksperimen yang menggunakan SVM tanpa seleksi fitur. Hasil dari pengujian metode seleksi fitur terbaik adalah *Information Gain* mendapatkan hasil terbaik untuk digunakan pada metode Naive Bayes, *Support Vector Machine*, *Artificial Neural Network* [8].

Mulajati & Fajriya (2017), menjelaskan bahwa tujuan penelitian tersebut adalah melakukan klasifikasi sentimen dengan metode *Text Mining*, sehingga diperoleh suatu informasi yang dianggap penting dan bermanfaat untuk berbagai bidang. Penelitian ini menggunakan metode klasifikasi Naive Bayes dan *Text Mining*. Data latih yang digunakan berupa review bahasa Inggris yang berjumlah masing-masing 915 data latih dan 228 data test. Pengujian menggunakan confusion matrix untuk evaluasi proses. *Cross validation* digunakan untuk menemukan prediksi akurasi terbaik. Pengujian dilakukan dari 100 iterasi dan hasil akurasi tertinggi pada iterasi ke-45 yang berjumlah 84%, namun memiliki nilai precision dan recall yang rendah yaitu 6% dan 29%. Setelah dilakukan observasi, model klasifikasi terbaik adalah iterasi ke-67 dengan akurasi 82%, 21% recall dan 32% precision.

Tien Rahayu Tulili, Muhammad Farman Andrijasa (2018) *Informatics Engineering* di Politeknik Negeri Samarinda dalam jurnal berjudul *Sentiment Classification of Microblogging in Indonesia Airlines Services using Support Vector Macine* menjelaskan bahwa tujuan penelitian untuk mengklasifikasi setiap komentar menjadi opini positif, negatif, atau netral dengan mendukung metode mesin vektor. Metode yang digunakan untuk klasifikasi adalah *Support Vector Macine* (SVM) dengan library LIBSVM dan LIBLINEAR untuk mempermudah penerapannya. Data diperoleh dari

twitter dengan teknik pengumpulan data yaitu *crawler*.

3. METODE PENELITIAN

Beberapa tahapan yang dilakukan pada penelitian, diantaranya dimulai dari pengumpulan data berupa file csv yang digunakan sebagai data latih dan berjumlah 1000 ulasan. Data ulasan tersebut di ambil dari *website* pegipegi.com pada masing-masing maskapai penerbangan. setelah data latih terkumpul, maka selanjutnya adalah pengumpulan data kamus untuk *correcting word*. Data set yang telah terkumpul akan dilakukan analisis sentimen pada masing-masing maskapai. Analisis yang dilakukan terdiri dari beberapa proses yaitu *preprocessing*, TF-IDF, *Information Gain*, dan Naive Bayes. Output dari hasil analisis berupa klasifikasi positif dan negatif. Tahap selanjutnya adalah desain dan perancangan sistem, implentasi dan pengujian sistem, analisis hasil pengujian dan yang terakhir melakukan penarikan kesimpulan.

3.1. Pengumpulan Data

Pengumpulan data dilakukan dengan menggunakan teknik *scraping* berdasarkan url pada *website* pegipegi.com. Data yang dikumpulkan berupa ulasan dari *customer* mengenai pengalaman saat menggunakan salah satu maskapai penerbangan yang dituangkan pada ulasan di web tersebut. Pengumpulan data dibedakan menjadi dua yaitu data latih dan data tes. Data latih yang digunakan berupa kalimat positif dan kalimat negatif. Jumlah kalimat yang digunakan pada penelitian adalah 500 kalimat positif dan 500 kalimat negatif. Data tes dikumpulkan sebanyak 100 ulasan dari masing-masing maskapai yang nantinya akan dianalisis sehingga menghasilkan *output* ulasan positif atau negatif yang diklasifikasikan oleh sistem berdasarkan pembelajaran data latih.

3.2. Preprocessing

Proses *preprocessing* merupakan hal yang penting untuk tahap berikutnya, yaitu mengurangi atribut yang tidak cukup berpengaruh untuk proses klasifikasi. Data yang digunakan pada tahap ini merupakan data yang masih kotor, sehingga dokumen yang dihasilkan dari proses ini diharapkan dapat mempermudah dalam proses klasifikasi. Proses *preprocessing* tersebut dilakukan pada data latih dan data *test*. Data latih yang telah disimpan pada file csv akan di panggil menggunakan fungsi python dan akan dilakukan beberapa proses yaitu *case folding*, *tokenization*, *remove accent mark*, *remove punctuation*, *correcting word*, *stopword*, dan *stemming*. Setelah dilakukan proses tersebut maka file akan disimpan kembali dalam folder *preprocessing* berupa file csv. Data *test* juga dilakukan proses *preprocessing* yang sama seperti data latih, namun tidak disimpan dalam file karena

akan diteruskan ke proses pembobotan tfidf. Berikut adalah tahapan dalam proses ini.

- 1) *Case Folding* adalah mengubah huruf kapital pada semua ulasan yang terdapat pada dokumen data latih dan data *test* diubah menjadi huruf kecil.
- 2) *Tokenization* merupakan pemotongan kata berdasarkan tiap kata yang menyusunnya menjadi potongan tunggal.
- 3) *Remove Accent Mark* menghapus aksan yang ada pada setiap kata. Penghapusan aksan berguna untuk proses stemming apakah kata tersebut perlu dihapus atau tidak.
- 4) *Remove Punctuation* adalah proses membersihkan komponen yang tidak berpengaruh terhadap hasil klasifikasi. Komponen *punctuation* yang tidak diperlukan antara lain titik, titi koma, tanda tanya, tanda seru, kutipan, strip, slash, tanda kurung, dan petik satu.
- 5) *Correcting Word* adalah memperbaiki ejaan yang salah sesuai dengan kamus yang telah disiapkan sebelumnya. Algoritma yang digunakan pada penelitian ini adalah algoritma pengembangan Peter Norvig.
- 6) *Stopword* merupakan proses menghilangkan kata yang tidak sesuai dengan topik dokumen, dikarenakan kata tersebut tidak mempengaruhi akurasi dalam klasifikasi sentimen ulasan.
- 7) *Stemming* merupakan proses untuk mengubah kata-kata yang terdapat dalam suatu ulasan ke dalam kata-kata akarnya dengan menggunakan aturan-aturan tertentu.

3.3. Pembobotan TF-IDF

Pada tahap pembobotan kata, metode *Term Frequency – Inverse Document Frequency* (TF-IDF) digunakan untuk mendapatkan nilai bobot setiap kata pada data yang digunakan. Proses pembobotan kata menggunakan algoritma tfidf. Tfidf menyajikan skor frekuensi kata terutama untuk kata-kata yang menarik, misalnya kata yang sering muncul dalam satu dokumen tetapi tidak untuk semua dokumen. Proses ini dilakukan dengan cara menghitung bobot setiap kata yang ada dalam data *training* dengan menggunakan *sklearn library* dengan dua skema. Skema yang digunakan antara lain *word counts* dengan *CountVectorizer*, *word frequencies* dengan *TfidfVectorizer*.

3.4. Information Gain

Tujuannya adalah mengurangi ukuran dimensi dan meningkatkan akurasi klasifikasi. Proses pemilihan fitur menggunakan metode *Information Gain*. *Information Gain* mengukur berapa banyak informasi kehadiran dan ketidakhadiran dari suatu kata yang berperan untuk membuat keputusan klasifikasi yang benar dalam *class* apapun. Dalam prosesnya *information gain* menjadi penyeleksi fitur yang paling penting agar akurasi dari sistem bisa

lebih baik. Langkah-langkah dalam perhitungan bobot *information gain* sebagai berikut.

- 1) Menghitung nilai *entropy* pada dataset dengan Persamaan 1.

$$Entropy(S) = - \sum_{i=1}^c P(i) \log_2 P(i) \quad (1)$$

Persamaan (1) c merupakan jumlah pada atribut target (jumlah kelas klasifikasi), sedangkan $P(i)$ merupakan proporsi sampel pada kelas i dengan sampel data.

- 2) Menghitung nilai *entropy* dengan A merupakan atribut, dan v adalah suatu nilai yang mungkin untuk atribut A , serta $Entropy(S_v)$ yang merupakan *entropy* untuk setiap partisi j dengan Persamaan 2.

$$Entropy(S, A) = - \sum_{j=1}^v \frac{|D_j|}{D} Entropy(S_v) \quad (2)$$

Persamaan (2), $|D|$ adalah jumlah sampel data dan $|D_j|$ adalah jumlah sampel data untuk nilai partisi j . v adalah suatu nilai yang mungkin untuk atribut A .

- 3) Langkah terakhir untuk mendapatkan nilai bobot *information gain* adalah menggunakan Persamaan 3.

$$Entropy(S, A) = -Entropy(D) - |Entropy(S, A)| \quad (3)$$

Persamaan (3) nilai *information gain* diperoleh dari pengurangan nilai Persamaan 1 dengan nilai Persamaan 1.

Selanjutnya dengan menggunakan bobot *information gain* tersebut setiap kata diranking sehingga menghasilkan fitur yang terbaik. Bobot tersebut digunakan untuk klasifikasi data uji yang nantinya diambil fitur dengan nilai tertinggi.

3.5. Klasifikasi Naïve Bayes

Metode Naive Bayes merupakan metode yang digunakan untuk pengklasifikasian. Metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes yaitu dengan memprediksi probabilitas dimasa depan berdasarkan pengalaman dimasa sebelumnya. Konsep dasar teori *bayes* yaitu peluang bersyarat $P(X|C)$, dimana X adalah *posterior* dan C adalah *prior*. *Prior* adalah pengetahuan kita tentang karakteristik suatu parameter atau pengalaman dimasa lalu, sedangkan *posterior* adalah karakteristik yang diduga pada kejadian yang datang. Perhitungan naïve bayes menggunakan Persamaan 4.

$$p(C_j|d_i) = \frac{p(C_j|d_i)p(C_j)}{p(d_i)} \quad (4)$$

Persamaan 4 diilustrasikan data training yang dikategorikan menjadi beberapa kategori $C_j = \{C_1, C_2, C_a, \dots, C_k\}$, dan *prior probability* untuk setiap kategori adalah $p(C_j)$, dimana $j = 1, 2, 3, \dots, k$. Sedangkan untuk simbol $d_i = (W_1, \dots, W_j, \dots, W_m)$, dan kata yang ada dalam dokumen berupa W_j , dimana $j = 1, 2, 3, \dots, m$, masuk ke dalam kategori C_j .

Untuk mengklasifikasi dokumen d_i , dilakukan dengan menghitung nilai probabilitas dari semua dokumen (*posterior probability*). Sehingga *posterior probability* suatu dokumen pada suatu kategori dapat dihitung dengan Persamaan 4.

Metode Naïve Bayes dalam proses pengklasifikasian memiliki dua tahapan yaitu tahap pelatihan dan tahap klasifikasi. Proses pelatihan dilakukan pertama untuk mendapatkan nilai *prior*. Dalam penelitian ini proses pelatihan yang digunakan adalah ulasan yang berjumlah 1000 kalimat dari beberapa maskapai. Setelah dilakukan perhitungan nilai probabilitas, maka selanjutnya adalah perhitungan untuk data *testing* yang belum diketahui nilai probabilitasnya. Nilai probabilitas dari data *testing* dikategorikan ke dalam dua kelas yaitu positif dan negatif.

Langkah-langkah perhitungan naïve bayes pada penelitian ini adalah sebagai berikut.

- 1) Pertama, menghitung nilai prior pada masing-masing kelas, yaitu dengan cara menghitung label kelas yang terdapat pada data latih dan dibagi dengan total data latih.
- 2) Kedua, menghitung likelihood, yakni probabilitas pada masing-masing atribut.
- 3) Ketiga, menghitung nilai posterior.
- 4) Terakhir, menentukan label kelas dengan melakukan perbandingan antar nilai posterior. Label kelas dengan nilai posterior tertinggi akan menjadi label kelas data yang diuji.

3.6. Validasi dan Evaluasi Sistem

Pengujian yang dilakukan terdiri dari pengujian pengaruh fitur *information gain*, pengujian pengaruh jumlah data latih terhadap model klasifikasi, dan validasi menggunakan *5-fold cross validation*. Hasil pengujian pengaruh fitur *information gain* akan dibahas melalui *confusion matrix* untuk menunjukkan seberapa baik model yang terbentuk. Pengukuran dengan *confusion matrix* di sini akan menampilkan perbandingan dari hasil akurasi model naïve bayes sebelum ditambahkan metode pemilihan fitur dan setelah ditambahkan metode pemilihan fitur. Pengujian pengaruh data latih terhadap model klasifikasi menghasilkan nilai *f1-score* dengan perbandingan sebelum dan setelah menggunakan *information gain*. Nilai akurasi akan ditampilkan untuk menjadi perbandingan data latih dengan jumlah tertentu yang digambarkan dalam bentuk grafik. Validasi menggunakan *5 fold cross validation*. Pada saat melakukan validasi, urutan dari kumpulan dokumen yang ada akan diacak. Hal ini bertujuan untuk menghindari adanya pengelompokan dokumen yang berasal dari kategori tertentu.

4. HASIL DAN PEMBAHASAN

Analisis sentimen maskapai penerbangan menggunakan metode Naive Bayes dan seleksi fitur

Information Gain berbasis web yang bertujuan untuk menganalisis *review* maskapai penerbangan dan memberikan rekomendasi dari yang terbaik. Setelah dilakukan analisis pada masing-masing maskapai yang menampilkan jumlah positif dan negatif, selanjutnya memberikan rekomendasi maskapai yang terbaik. Maskapai yang terbaik adalah maskapai penerbangan yang memiliki jumlah kalimat positif yang paling banyak. Berdasarkan hasil analisis urutan dari maskapai yang memiliki jumlah kalimat positif terbanyak tersebut dapat dilihat pada Tabel 1.

Tabel 1. Tabel Hasil Analisis

Nama Maskapai	Hasil Analisis	
	Kalimat Positif	Kalimat Negatif
Wings	93	7
Lion Air	83	17
Garuda Indonesia	81	9
Batik Air	80	20
Sriwijaya	80	20
Nam Air	77	23
Citilink	74	26
Air Asia	70	30
Scoot	3	1

Presentase data tes yang digunakan pada masing-masing maskapai sebanyak 100 kalimat ulasan terkecuali untuk jumlah data tes pada maskapai scoot hanya berjumlah 4 ulasan, dikarenakan ulasan maskapai scoot yang terdapat pada *website* pegipegi.com hanya berjumlah 4 ulasan saja. Hasil uji coba pengujian dipaparkan pada sub bab berikut.

4.1. Pengujian Pengaruh *Information Gain*

Pengujian ini akan dibahas melalui *confusion matrix* untuk menampilkan perbandingan dari hasil akurasi model Naïve Bayes sebelum dan setelah ditambahkan *information gain*. Hasil pengujian pengaruh *information gain* dapat dilihat pada Tabel 2 dan Tabel 3.

Tabel 2. Confusion Matrix Model Naïve Bayes Sebelum Penambahan Seleksi Fitur *Information Gain*

Akurasi Naïve Bayes : 0,81					
	Pred. Negatif	Pred. Positif	Precision	Recall	F1-Score
Actual Negatif	81	19	0,81	0,81	0,81
Actual Positif	19	81	0,81	0,81	0,81

Tabel 2 pengujian sebelum penambahan *information gain* menghasilkan nilai *precision*, *recall*, dan *f1-score* yang sama untuk negatif dan positif yakni 0,81. Akurasi yang dihasilkan sebesar 0,81.

Tabel 3. Confusion Matrix Model Naïve Bayes Setelah Penambahan Seleksi Fitur *Information Gain*

	Akurasi Naïve Bayes : 0,865				F1-Score
	Pred. Negatif	Pred. Positif	Precision	Recall	
Actual Negatif	82	18	0,90	0,82	0,86
Actual Positif	9	91	0,83	0,91	0,87

Tabel 3 menunjukkan pengujian dengan penambahan *information gain* menghasilkan nilai *precision* yakni 0,90 negatif, 0,83 positif, sedangkan *recall* negatif 0,82, positif 0,91. F1-score menghasilkan nilai positif 0,87 dan 0,86 nilai negatif. Akurasi yang dihasilkan sebesar 0,865.

4.2. Pengujian Pengaruh Jumlah Data Latih terhadap Model Klasifikasi

Pengujian ini bertujuan untuk menguji tingkat akurasi klasifikasi jika menggunakan data latih dengan jumlah tertentu dan waktu yang diperlukan untuk komputasi. Hasil pengujian jumlah data latih dapat dilihat pada Tabel 4. Nilai akurasi Naïve Bayes juga ditampilkan ke dalam grafik berdasarkan jumlah data latih yang dapat dilihat pada Gambar 1.

Tabel 4. Hasil Pengujian Jumlah Data Latih

Jumlah data latih	F1-Score		Akurasi
	Dengan <i>information gain</i>	Tanpa <i>information gain</i>	
200	0,814	0,519	0,815
400	0,829	0,674	0,83
600	0,839	0,769	0,84
800	0,834	0,769	0,835
1000	0,864	0,804	0,865

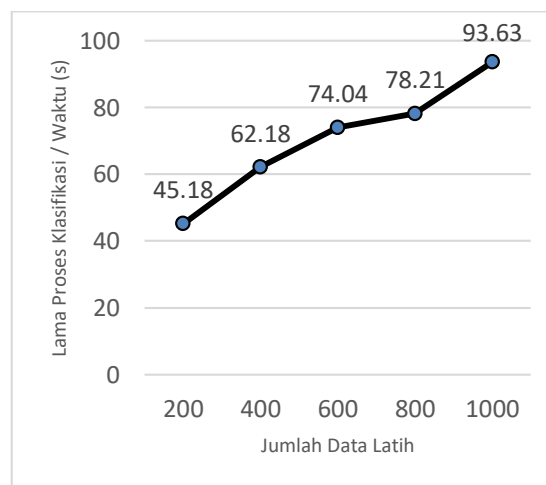
Hasil pengaruh jumlah data latih pada model klasifikasi adalah 0,815 (data latih 200), 0,83 (data latih 400), 0,84 (data latih 600), 0,835 (data latih 800), dan 0,865 (data latih 1000). Hasil pengaruh jumlah data latih dapat dilihat pada tabel 3. Hasil akurasi tertinggi adalah dengan jumlah korpus latih 1000 kalimat dengan nilai F1-score 0,865. Akurasi terendah adalah jumlah korpus latih 200 kalimat dengan nilai F1-score 0,815.

Berdasarkan hasil uji yang dilakukan, dapat dilihat pada Gambar 1 bahwa dengan data latih 200 memerlukan waktu komputasi 45.18 sekon, 400 data latih memerlukan waktu komputasi 62.18 sekon, 600 data latih memerlukan waktu komputasi selama 74.04 sekon, 800 data latih memerlukan waktu komputasi selama 78.21 sekon dan 1000 data latih memerlukan waktu komputasi selama 93.63 sekon.

4.3. Validasi 5-Fold Cross Validation

Pengujian model dengan menggunakan teknik *5-fold cross validation*, dimana proses ini membagi data secara acak ke dalam 5 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Hasil validasi menggunakan *5-fold cross validation* dapat dilihat

pada Tabel 5. Hasil validasi menampilkan nilai akurasi Naïve Bayes dengan *Information Gain* dan tanpa *Information Gain*. Selanjutnya hasil dihitung nilai rata-rata pada masing-masing nilai akurasi.



Gambar 1. Analisis waktu terhadap jumlah data latih

Tabel 5. Hasil Pengujian 5-fold cross validation

Fold	Akurasi Naïve Bayes	
	Dengan <i>information gain</i>	Tanpa <i>information gain</i>
1	0,815	0,765
2	0,855	0,78
3	0,85	0,75
4	0,845	0,79
5	0,82	0,775
Rata-rata	0,838	0,772

Nilai akurasi yang dihasilkan pada setiap fold hasilnya berbeda-beda. Tabel 5 menunjukkan bahwa akurasi tertinggi pada fold-2 dengan *information gain*, sedangkan tanpa *information gain* pada fold-4. Akurasi terendah pada fold-1. Perbedaan akurasi pada setiap fold disebabkan karena karakteristik data uji dan data latih pada setiap fold berbeda-beda. Pada fold-2 menghasilkan akurasi tinggi karena pada data uji dan data latih memiliki kedekatan karakteristik, sedangkan pada fold-1 antara data latih dan data uji memiliki karakteristik yang cukup jauh. Rata-rata akurasi yang diperoleh dengan *information gain* lebih tinggi yakni 0.838 jika dibandingkan tanpa *information gain* yang hanya 0.772. Hal tersebut bisa disimpulkan bahwa dengan penambahan *information gain* nilai akurasi menggunakan cross validation sangat baik jika dibandingkan tanpa *information gain*.

4.4. Analisis Hasil Pengujian Keseluruhan

Hasil klasifikasi keseluruhan yang diperoleh sebelumnya, mendapatkan beberapa kesimpulan yakni nilai akurasi untuk perbandingan setelah dan sebelum penambahan *information gain*. Berdasarkan Tabel 2 dan 3 nilai akurasi terbaik yang dihasilkan adalah setelah penambahan *information gain* yakni sebesar 0.865. Tabel 4 menampilkan nilai f1-score dengan *information gain* dan tanpa *information gain*.

Nilai tersebut juga dibandingkan dengan jumlah data latih tertentu. Kombinasi ini untuk melihat bagaimana nilai akurasi dan *f1-score* jika menggunakan jumlah data latih tertentu. Nilai akurasi terbaik yang diperoleh adalah dengan jumlah data latih 1000 yakni 0.865 untuk nilai akurasi dan 0.864 untuk nilai *f1-score* setelah penambahan *information gain*. Gambar 1 melihat grafik waktu yang dibutuhkan untuk mengklasifikasi dengan jumlah data tertentu. Berdasarkan grafik tersebut diperoleh bahwa jumlah data latih 1000 ulasan membutuhkan waktu yang cukup lama. Hal ini disebabkan proses *processing* yang cukup panjang dan pengambilan fitur untuk *information gain* juga akan cukup banyak, sehingga klasifikasi untuk 1000 ulasan akan memakan waktu sekitar 93.63 sekon. *Cross fold validation* dilakukan untuk melakukan validasi dan menghasilkan nilai akurasi. Pengujian menggunakan 5-fold yang berarti data terbagi menjadi 5 subset. Tabel 5 menampilkan nilai akurasi pada masing-masing fold dan mendapatkan nilai rata-rata tertinggi yakni 0.838. Setelah melakukan tiga pengujian tersebut maka dapat disimpulkan bahwa nilai akurasi dan *f1-score* yang diperoleh dengan hasil terbaik adalah setelah penambahan *information gain*.

5. KESIMPULAN

Berdasarkan hasil analisis dan pengujian terhadap sistem analisis sentimen review maskapai penerbangan menggunakan metode Naive Bayes dan seleksi fitur *Information Gain*, maka dapat ditarik kesimpulan sebagai berikut :

1. Penelitian ini membuktikan bahwa metode klasifikasi Naive Bayes dengan seleksi fitur *Information Gain* memiliki pengaruh yang cukup signifikan terhadap nilai akurasi klasifikasi. Rata-rata nilai akurasi yang didapatkan sebesar 80% sebagai nilai yang paling optimal. Tingkat akurasi tersebut didapatkan dengan jumlah data latih 1000 kalimat dan 5 kali *fold validation* dengan diimplementasikannya fitur *Information Gain*.
2. Penelitian ini membuktikan bahwa metode klasifikasi Naive Bayes dengan pemilihan fitur *Information Gain* dapat digunakan dalam menganalisis sentimen ulasan maskapai penerbangan yang ada pada *website* pegipegi.com.
3. Penelitian ini mengembangkan aplikasi analisis sentimen maskapai penerbangan berbasis web yang dapat menampilkan hasil ulasan dalam bentuk positif dan negatif, sehingga dapat memberikan rekomendasi kepada konsumen dalam memilih maskapai penerbangan yang terbaik.
4. Penelitian ini membuktikan bahwa jumlah data latih dalam sistem analisis sentimen memiliki pengaruh terhadap prediksi sistem. Selain jumlah, kualitas data latih juga berperan karena

semakin tinggi kualitas data maka sistem akan mendapatkan *vocabulary* yang semakin besar sehingga akan lebih tepat dalam memprediksi kelas sentimen.

DAFTAR PUSTAKA

- AGUSTA, L. (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. *Proceeding Konferensi Nasional Sistem dan Informatika*. Yogyakarta, 196-201
- BASARI, A., HUSSIN, B., & ANANTA, I. (2013). Opinion Mining of Movie Review Using Hybrid Method of Support Vector Machine and Particle Swarm Optimization.
- CHANDANI, V. W. (2015). Komparasi Algoritma Klasifikasi Machine Learning dan Feature. *Journal of Intelligent System*.
- CHEN, J., HUANG, H., TIAN, S., & QU, Y. (2009). Feature selection for text classification with Naive Bayes. *Expert System with Applications*, 5432-5435.
- FELDMAN, R. (2013). Techiques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82.
- HADDI, E., LIU, X., & SHI, Y. (2013). The Role of Text Pre-processing in sentiment Analysis. *Procedia Computer Science*, 17, 26-32.
- KAMILAH, A. N. (2017). Analisa Sentimen Pelanggan Tokopedia Menggunakan Algoritma Naive Bayes Berdasarkan Review Pelanggan. *Sistem Informasi*. Kediri: Universitas Nusantara PGRI Kediri.
- MULAJATI, M., & HAKIM, R. F. (2017). Sentiment Analysis On Online Review Using Naive Bayes Classifier Method and Text Association (Case Study : Garuda Indonesia Airlines Passenger Reviews On Tripadvisor Site).
- PRATAMA, N. D., SARI, Y., & ADIKARA, P. (2018). Analisa Sentimen Pada Review Konsumen Menggunakan Metode Naive Bayes dengan Seleksi Fitur Chi-Square Untuk Rekomendasi Lokasi Makanan Tradisional. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.
- RADITYATAMA, A. K. (2014). Pengaruh Atribut Merek Daring Pada Citra Merek Perusahaan Daring. *Jurnal Ekonomika dan Bisnis*. Yogyakarta: Universitas Gadjah Mada.
- TULILI, T. R., & ANDRIJASA, M. F. (2018). Sentiment Classification of Microblogging in Indonesia Airlines Services using Support Vector Machine.
- UYSAL, A., & GUNAL, S. (2012). A Novel probabilistic feature selection method for

text clasification . *Knowledge-Based System*, 36, 226-235.

- WANG, S. L. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification, *Expert Systems with Applications*, 38 (7), 8696-8702.
- ZAFIKRI, A. (2010). Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi. *Skripsi. Universitas Sumatera Utara*.
- ZHANG ZIQIONG, YE, Q., & LAW, R. (2009). Expert System with Sentiment clasification ofonline reviews to travel destinations by supervised machine learning approaches. *Expert Systems With Applications*, 36(3), 6527-6535.
- ZHANG, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet Restaurant Reviews Written in Cantonese. *Expert with Application*. 38(6), 7674–7682.