# KLASIFIKASI SENTIMEN ULASAN BUKU BERBAHASA INGGRIS MENGGUNAKAN *INFORMATION GAIN* DAN *NAÏVE BAYES*

## *SENTIMENT CLASSIFICATION ON ENGLISH BOOK REVIEWS USING INFORMATION GAIN AND NAÏVE BAYES*

**Laila Ratnasari Putri[1], Mohamad Syahrul Mubarok[2], Adiwijaya[3]**

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

**[1]lailaratnasari05@gmail.com, [2]msyahrulmubarok@gmail.com, [3]kang.adiwijaya@gmail.com**

**Abstrak**

**Semakin berkembangnya teknologi informasi, mengakibatkan pertumbuhan data mengenai ulasan buku semakin besar dan pesat. Dengan membaca *review* atau ulasan berdasarkan pengalaman pembaca lain, maka kita akan mengetahui kualitas dari buku tersebut. Begitu banyaknya ulasan akan mempersulit pengguna lain untuk menyimpulkan hasil dari ulasan tersebut mengandung opini positif atau negatif. Oleh karena itu, peneliti memberikan solusi dengan menggunakan klasifikasi sentimen ulasan buku. Metode yang digunakan adalah *Information Gain* dan *Naïve Bayes. Information Gain* digunakan sebagai metode pemilihan fitur yang dapat membuat akurasi penelitian menjadi meningkat dengan mengurangi fitur-fitur yang kurang. *Naïve Bayes* digunakan untuk mengatasi masalah ketidakpastian yang terdapat pada pengklasifikasian teks, dan *Naïve Bayes* mengklasifikasikan ulasan, cenderung beropini positif atau negatif berdasarkan nilai probabilitasnya. Berdasarkan skenario pengujian yang telah dilakukan, performa klasifikasi sentimen pada ulasan buku berbahasa inggris menggunakan *Information Gain* dan *Naive Bayes* dari rata-rata *F1-score* menggunakan *5-fold-cross validation* adalah 88,28%.**

**Kata kunci: ulasan buku, analisis sentimen, *naïve bayes, information gain***

**Abstract**

***The growth of information technology resulting in increased number of data about book review. By reading reviews or reviews based on the experience of other readers, we will know the quality of the book. Large number of reviews will make it hard for other user to conclude the relust of the review whether it is a negative or positive opinion. Therefore, the researcher provides a solution by using the book review sentiment classification. The method used is Information Gain and Naïve Bayes. Information Gain is used as a feature selection method that can make research accuracy improved by reducing features that are less significant in the process of classification. Naïve Bayes is used to overcome the problem of the uncertainty contained in the classification of texts, and Naïve Bayes classifies reviews, tends to be positive or negative based on their probability values. Based on the test scenario that has been done, Information Gain can improve the accuracy of Naïve Bayes for book review cases with an average of F1-score with 5-fold-cross validation that is 88.28%.***

**Keywords:** ***book review, sentiment analysis, naïve bayes, information gain.***

## 1. Pendahuluan

### 1.1 Latar Belakang

Di era globalisasi saat ini, pengguna Internet dapat dengan mudah mencari informasi mengenai suatu hal. Selain itu, pengguna juga dapat menuliskan opini mereka mengenai sebuah produk atau jasa. Salah satu informasi yang terdapat di Internet adalah informasi mengenai buku. Pengguna Internet dapat dengan mudah menuliskan opini atau ulasan mereka mengenai sebuah buku. Dengan membaca *review* atau ulasan berdasarkan pengalaman pembaca lain, maka kita akan mengetahui kualitas dari buku tersebut. Namun, begitu banyaknya opini akan mempersulit pengguna lain untuk memperoleh informasi dari opini tersebut.

Solusi dari permasalahan tersebut adalah dengan menggunakan klasifikasi sentimen. Klasifikasi sentimen merupakan salah satu bagian dari klasifikasi teks, dimana sistem akan melakukan pengklasifikasian sentimen atau opini menjadi sentimen positif atau negatif [1]. Dengan melakukan pengklasifikasian teks pada opini ulasan buku, akan membantu pembaca memperoleh informasi kualitas buku secara cepat.

Dalam klasifikasi sentimen, terdapat masalah ketidakpastian atau *uncertainty reasoning* pada teks. Sebagai contoh apabila terdapat dua kalimat yang memiliki fitur atau kata yang sama, sistem akan mengalami kesulitan dalam proses pengklasifikasian [2]. Masalah tersebut dapat diselesaikan dengan menggunakan salah satu metode pada pengklasifikasian teks yaitu *Naïve Bayes* [2]. *Naïve Bayes* merupakan metode yang

menerapkan *Bayesian theorema*, dimana *Naïve Bayes* mengasumsikan secara *naive* (sederhana) bahwa setiap fitur bersifat independensi, maksudnya adalah setiap fitur dalam data yang sama tidak saling berkaitan [3]. Selain itu, *Naïve Bayes* merupakan metode yang melihat pendapat atau opini tersebut, cenderung beropini positif atau negatif berdasarkan nilai probabilitasnya [4]. *Naïve Bayes* juga memiliki performa yang tinggi dengan kalkulasi yang sederhana [2]. Berdasarkan hal tersebut, penulis menggunakan *Naïve Bayes* pada penelitian klasifikasi sentimen ulasan buku ini.

Permasalahan lain dalam klasifikasi sentimen adalah jika terdapat terlalu banyak jumlah fitur pada suatu *dataset*, akan menyebabkan turunnya akurasi dari klasifikasi [5]. Oleh karena itu, dibutuhkan *feature selection* yang digunakan untuk memilih fitur yang signifikan, dan membuang atribut yang tidak relevan, sehingga dimensi data akan berkurang dan akurasi akan meningkat [6]. *Information Gain* merupakan salah satu teknik pada *feature selection*, dimana suatu kata diukur dengan menghitung jumlah informasi kata tersebut ada atau tidak pada suatu dokumen [7]. Pada penelitian ini, *Information Gain* digunakan sebagai *feature selection* untuk memilih fitur yang signifikan dan menghilangkan fitur-fitur yang tidak penting sehingga proses klasifikasi sentimen ulasan buku lebih akurat.

### 1.2 Perumusan Masalah

Adapun Masalah yang akan diteliti adalah:

1. Bagaimana implementasi dari kombinasi metode *Information Gain* pada *feature selection* dan *Naïve Bayes* pada klasifikasi sentimen ulasan buku?
2. Berapa performa yang dihasilkan dari penelitian klasifikasi pada ulasan buku?

### 1.3 Tujuan

Tujuan dari penelitian Tugas Akhir ini adalah:

1. Mengimplementasikan kombinasi *Information Gain* untuk *feature selection* dan *Naïve Bayes* untuk klasifikasi sentimen ulasan buku.
2. Mengetahui performa dari penerapan klasifikasi sentimen menggunakan *Information Gain* dan *Naïve Bayes.*

### 1.4 Batasan Masalah

Adapun batasan masalah yang digunakan pada penelitian ini adalah:

1. Data diambil dari *Goodreads* [8]. Data tersebut berisi ulasan atau *review* buku. *Dataset* tersebut telah diberi label berupa 2 *class* yaitu positif dan negatif.
2. Data ulasan buku menggunakan bahasa Inggris.

## 2. Perancangan Sistem

### 2.1 Dataset

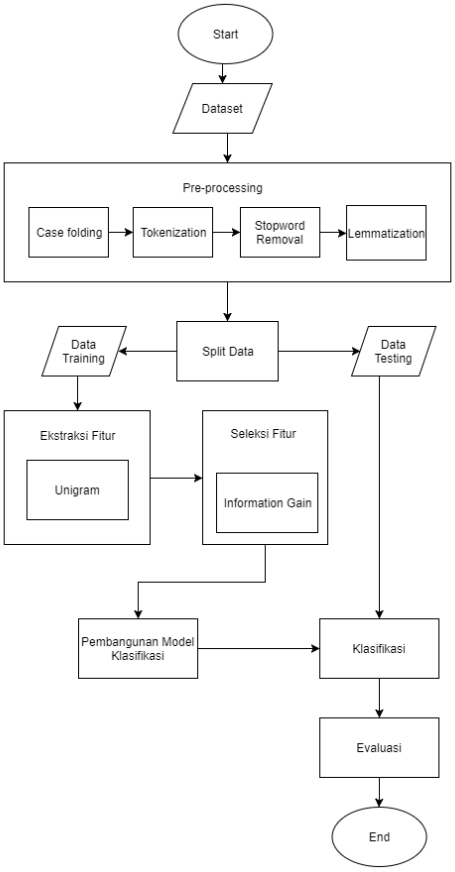
Penelitian ini menggunakan data *review* buku dalam bahasa Inggris yang diambil dari situs *Goodread*s [8]. Situs tersebut merupakan situs jaringan sosial, dimana pengguna dapat menuliskan *review* atau opini mereka mengenai suatu buku bacaan, sehingga dapat memberikan rekomendasi kepada pengguna lain. Penelitian ini menggunakan 1000 *record* data yang terbagi menjadi 500 opini positif dan 500 opini negatif. Contoh dari *dataset review* buku untuk ulasan buku positif dan ulasan buku negatif ditunjukan pada Tabel 1 [8].

Tabel 1 Contoh ulasan buku positif dan negatif

| No | Kalimat | Sentimen |
|---|---|---|
| 1 | *Read this in one sitting-Max Wolfe is an engaging protagonist, likeable and human while the plot was intriguing and thrilling. Great read.* | Positif |
| 2 | *Great book almost as funny as Peter Kaye in person. Full of fun photos and extra bits to make a great light hearted read.* | Positif |
| 3 | *Junk. Awful book. Far too long. Waste of time. Not worthy of the Clancy series* | Negatif |
| 4 | *WORST BOOK EVER!* | Negatif |
| 5 | *So bad. I hate it.* | Negatif |
| 6 | *A Sensational and enlightening book* | Positif |

### 2.2 Gambaran Umum Sistem

Sistem yang dibangun oleh peneliti adalah sistem yang dapat melakukan proses klasifikasi sentimen ulasan buku berbahasa inggris menggunakan *Information Gain* dan *Naïve Bayes*. Gambaran umum dari sistem yang dibangun dapat dilihat pada Gambar 1.




Gambar 1 Gambaran umum sistem

### 2.2.1 *Pre-processing*

*Pre-processing* merupakan tahap dimana data dari dokumen teks yang tidak terstruktur diolah menjadi data terstruktur. Berikut penjelasan dari tahapan dalam *pre-processing*:

1. *Case Folding*

   Pada tahap *case folding*, seluruh huruf kapital dirubah kedalam huruf kecil (*lower case*) dan karakter *non-alphabet* dihilangkan [9]. Contoh proses *case folding* adalah sebagai berikut:

   Teks: A Sensational and enlightening book.

   Hasil *Case folding*: a sensational and enlightening book

2. *Tokenization*

   *Tokenization* merupakan tahap *pre-processing* dimana kalimat dipecah menjadi beberapa token atau kata [9]. Contoh proses *tokenization* adalah sebagai berikut:

   Teks: a sensational and enlightening book

   Hasil *Tokenization*: ‘a’, ‘sensational’, ‘and’, ‘enlightening’, ‘book’

3. *Stopword Removal*

   *Stopword removal* merupakan tahapan dimana kata yang dianggap tidak penting dihilangkan [10]. Contoh proses *stopword removal* adalah sebagai berikut:

   Teks: ‘a’, ‘sensational’, ‘and’, ‘enlightening’, ‘book’

   Hasil *Stopword Removal*: ‘sensational’, ‘enlightening’, ‘book’

4. *Lemmatization*

   *Lemmatization* merupakan tahapan *pre-processing*, dimana kata diubah menjadi kata dasar [11]. Dalam tahap ini, proses pengubahan kata mengikuti struktur bahasa yang digunakan. Contoh proses *lemmatization* adalah sebagai berikut:

   Teks: ‘sensational’, ‘enlightening’, ‘book’

   Hasil *Lemmatization*: ‘sensational’, ‘enlightening’, ‘book’

### 2.2.2 *Feature Extraction*

Pada tahap ini, sistem akan menghitung jumlah kemunculan *term*, hal ini direpresentasikan ke dalam model *bag-of-words* [12]. *Bag-of-word* merupakan sebuah model dimana sebuah teks direpresentasikan sebagai kumpulan dari kata-kata teks tersebut, tanpa memikirkan urutan kata, namun memperhatikan

keanekaragaman [13]. Pada penelitian ini proses yang diterapkan adalah *unigram.* Dimana, data yang sebelumnya masih berbentuk teks, pada tahap ini data tersebut dirubah menjadi bentuk vektor [14].

### 2.2.3 *Feature Selection*

Metode *feature selection* yang digunakan pada penelitian ini adalah *Information Gain.* Metode ini digunakan untuk memilih fitur-fitur yang dianggap penting pada proses klasifikasi, sekaligus mereduksi dimensi inputan. Fitur yang diterapkan merupakan fitur yang telah diekstraksi yang direpresentasikan dengan model *bag-of-words* menggunakan *unigram.* Fitur pada penelitian ini adalah kata. Selanjutnya kata tersebut diseleksi menggunakan perhitungan *Information Gain* dengan menggunakan persamaan (1) dan (2). Jika nilai *Gain* kata lebih besar dari *threshold*, kata tersebut dipilih dan disimpan ke dalam sistem [15].

### 2.2.4 *Classification*

Proses klasifikasi pada penelitian ini menggunakan *Naïve Bayes. Naïve Bayes* menerapakan teorema Bayesian [16]. *Naïve Bayes* adalah salah satu metode klasifikasi yang digunakan untuk menentukan sebuah kalimat bernilai positif atau negatif berdasakan probabilitasnya [4]. Apabila nilai probabilitas kelas positif untuk kalimat tersebut lebih besar dibandingkan *class* negatif, maka kalimat tersebut merupakan sentimen positif. Namun, apabila probabilitas kelas positif lebih kecil daripada kelas negatif, maka kalimat tersebut merupakan sentimen negatif [4], [17].

### 2.2.5 Evaluasi

Setelah semua proses dari tahap *pre-processing* teks hingga klasifikasi menggunakan *Naïve Bayes* dilakukan, selanjutnya dilakukan evaluasi hasil kinerja klasifikasi. Evaluasi yang dilakukan menggunakan *5-fold cross validation.* Kumpulan dokumen yang ada, urutannya akan diacak sebelum masuk dalam sebuah *fold* [18]. Hal ini dilakukan untuk menghindari pengelompokan dokumen yang berasal dari satu kategori tertentu pada suatu *fold.* Selanjutnya, untuk mengetahui performa sistem yang dibangun pada penelitian ini, menggunakan *Confusion Matrix* [19].

## 2.3 Skenario Pengujian

Skenario pengujian yang akan dilakukan pada penelitian Tugas Akhir ini adalah:

1. Pengujian pengaruh jumlah *dataset* terhadap model klasifikasi

   Pengujian ini dilakukan untuk melihat pengaruh jumlah *dataset* yang berbeda terhadap model klasifikasi. Skenario dari pengujian ini adalah dengan mengubah jumlah *dataset* yang digunakan. Jumlah *dataset* yang digunakan pada pengujian ini terdapat pada Tabel 2.

Tabel 2 Skenario pengujian pengaruh jumlah dataset

| Jumlah *Dataset* | Jumlah Data *Class* Positif | Jumlah Data *Class* Negatif |
|---|---|---|
| 200 | 100 | 100 |
| 400 | 200 | 200 |
| 800 | 400 | 400 |
| 1000 | 500 | 500 |

2. Pengaruh *lemmatization* terhadap proses klasifikasi

   Pengujian pengaruh *lemmatization* terhadap proses klasifikasi dilakukan dengan melakukan pengujian pada tahapan pre-processing. Tujuan dari pengujian ini adalah untuk melihat *lemmatization* dapat menghasilkan klasifikasi yang lebih baik. Adapun skenario pengujian yang dilakukan adalah sebagai berikut:

   a. *Case folding*, *tokenization*, dan *stopword removal*.
   b. *Case folding*, *tokenization*, *stopword removal*, dan *lemmatization*.

3. Pengaruh tanpa *Information Gain* dan *Information Gain* dengan *threshold* terhadap performa *Naïve Bayes.*

   Pengujian ini dilakukan untuk melihat seberapa pengaruh *Information Gain* terhadap performa klasifikasi. Pada penelitian ini, fitur yang dipilih untuk proses klasifikasi adalah fitur yang telah memenuhi batas (*threshold*) yang telah ditentukan [20]. *Threshold* tersebut juga akan dilakukan pengujian pada penelitian ini dengan menggunakan nilai yang berbeda-beda. Pengujian *threshold* dilakukan untuk menemukan parameter yang tepat untuk klasifikasi. Pengujian pengaruh *Information Gain* terhadap performa *Naïve Bayes* ini, memiliki skenario yaitu

   a. Tanpa *Information Gain*

b. Menggunakan *Information Gain* dengan *threshold* = 0,5
c. Menggunakan *Information Gain* dengan *threshold* = 0,8
d. Menggunakan *Information Gain* dengan *threshold* = 0,9
e. Menggunakan *Information Gain* dengan *threshold* = 0,95
f. Menggunakan *Information Gain* dengan *threshold* = 0,96
g. Menggunakan *Information Gain* dengan *threshold* = 0,97
h. Menggunakan *Information Gain* dengan *threshold* = 0,98

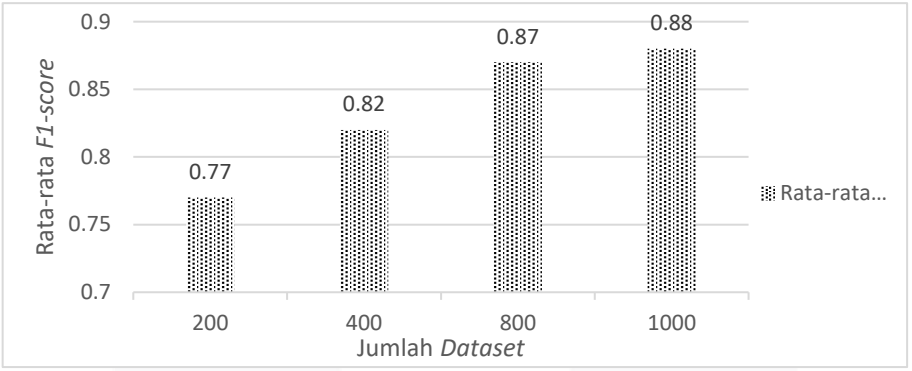
## 3. Analisis Hasil Pengujian

### 3.1 Analisis Pengaruh Jumlah *Dataset* Terhadap Model klasifikasi

Hasil dari pengujian pengaruh jumlah *dataset* terhadap model klasifikasi dibandingkan dan dianalisis menggunakan *confusion matrix*. Analisis terhadap hasil pengujian ini dilakukan dengan membandingkan *F1-score* dari setiap model klasifikasi. Tabel 3 menunjukan hasil evaluasi dari pengujian pengaruh jumlah *dataset* terhadap model klasifikasi untuk setiap *fold.*

Tabel 3 Pengaruh jumlah *dataset* terhadap F1-score

| Jumlah *Dataset* | *Fold* | | | | | Rata-rata F1-score |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| 200 | 0.73 | 0.71 | 0.79 | 0.89 | 0.75 | 0.77 |
| 400 | 0.81 | 0.83 | 0.83 | 0.84 | 0.79 | 0.82 |
| 800 | 0.85 | 0.87 | 0.88 | 0.84 | 0.89 | 0.87 |
| 1000 | 0.85 | 0.88 | 0.90 | 0.88 | 0.9 | **0.88** |

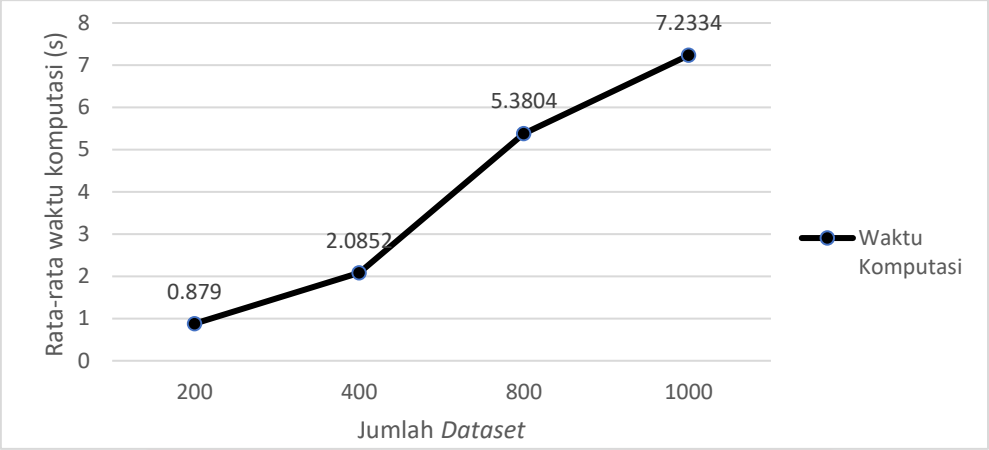



Gambar 2 Pengaruh jumlah dataset terhadap F1-score

Berdasarkan Tabel 3 dan Gambar 2, dapat dilihat bahwa hasil evaluasi dari pengujian ini adalah *dataset* yang berjumlah 1000 data memiliki nilai rata-rata *F1-score* yang tertinggi yaitu sebesar 0,8828. Hasil evaluasi tersebut menunjukan bahwa semakin banyak jumlah *dataset* yang digunakan maka nilai rata-rata *F1-score* yang didapat juga semakin tinggi, ini menyatakan bahwa sistem yang dibangun semakin baik. Hal ini dikarenakan semakin banyak data yang digunakan, maka akan banyak data yang dipelajari oleh sistem, sehingga sistem akan mudah untuk beradaptasi terhadap data yang akan digunakan pada *training* ataupun *testing*. Hal ini juga berdampak pada waktu komputasi sistem. Tabel 4 dan Gambar 3 menunjukan hasil perbandingan waktu komputasi pengaruh jumlah dataset.

Tabel 4 Pengaruh jumlah *dataset* terhadap waktu komputasi

| Jumlah *Dataset* | Waktu Komputasi | | | | | Rata-rata Waktu Komputasi |
|---|---|---|---|---|---|---|
| | Fold ke-1 | Fold ke-2 | Fold ke-3 | Fold ke-4 | Fold ke-5 | |
| 200 | 0.899 | 0.785 | 0.877 | 0.893 | 0.941 | 0.879 |
| 400 | 2.039 | 2.139 | 2.129 | 2.134 | 1.985 | 2.0852 |
| 800 | 5.503 | 5.256 | 5.516 | 5.363 | 5.264 | 5.3804 |
| 1000 | 9.407 | 6.531 | 6.753 | 6.747 | 6.729 | **7.2334** |




Gambar 3 Pengaruh jumlah *dataset* terhadap waktu komputasi

Berdasarkan Tabel 4 dan Gambar 3, waktu komputasi yang diperlukan sistem dengan jumlah *dataset* 200 dan 400 data memiliki selisih waktu sebesar 2 detik. Untuk jumlah *dataset* 400 dan 800, selisih waktu komputasi sistem adalah 3 detik. Sedangkan untuk selisih waktu komputasi dengan jumlah *dataset* 800 dan 1000 adalah 3 detik. Hal ini menunjukan bahwa semakin banyak jumlah data yang digunakan maka sistem akan membutuhkan waktu yang lebih lama untuk proses komputasi, dibandingkan dengan jumlah data yang sedikit.

### 3.2 Analisis Pengaruh *Lemmatization* Terhadap Proses Klasifikasi

Hasil dari pengujian pengaruh lemmatization dievaluasi dengan menggunakan F1-score. Selanjutnya, hasil tersebut dianalisis, semakin tinggi nilai rata-rata F1-score menyatakan bahwa sistem yang dibangun semakin baik. Pada pengujian pengaruh lemmatization data yang digunakan adalah 1000 data. Pengujian ke-1 merupakan pengujian tanpa lemmatization, sedangkan pengujian ke-2 adalah pengujian dengan lemmatization. Hasil dari pengujian ini terdapat pada Tabel 5.

Tabel 5 Hasil evaluasi pengaruh lemmatization

| *Fold* | *F1-score* | |
|---|---|---|
| | Skema Pengujian 1 | Skema Pengujian 2 |
| 1 | 0.83 | 0.85 |
| 2 | 0.85 | 0.87 |
| 3 | 0.88 | 0.90 |
| 4 | 0.85 | 0.88 |
| 5 | 0.92 | 0.90 |
| Rata-rata *F1-score* | 0.8677 | **0.88** |

Berdasarkan Tabel 5 pengujian ke-2 memiliki nilai rata-rata *F1-score* yang tinggi yaitu sebesar 0,8828. Pengujian kedua merupakan pengujian tahapan *pre-processing* yang terdiri dari *case folding, tokenization, stopword removal*, dan *lemmatization.* Dari hasil pengujian dapat diketahui bahwa skema pengujian yang menggunakan *lemmatization* dapat mengkhasilkan klasifikasi yang lebih baik. Hal ini dikarenakan *lemmatization* mengubah setiap kata menjadi kata dasar. Dengan menggunakan *lemmatization*, sistem akan dapat mengenali fitur yang sama, sehingga tidak ada kata yang sebenarnya memiliki bentuk dasar yang sama namun dianggap berbeda.

### 3.3 Analisis Pengaruh Fitur Information Gain

*Information Gain* akan memilih fitur-fitur yang akan digunakan oleh algoritma klasifikasi. Fitur yang dipilih adalah fitur yang memenuhi *threshold.* Parameter *threshold* yang digunakan akan menentukan kinerja dari *Naïve Bayes. Threshold* yang tepat, akan meningkatkan performa *Naïve Bayes*. Apabila memilih parameter *threshold* yang terlalu kecil atau terlalu besar maka akan berpengaruh terhadap proses klasifikasi. *Range threshold* yang digunakan pada pengujian *threshold* dipilih berdasarkan rata-rata nilai hasil perhitungan *gain* yang dilakukan. Hasil evaluasi dari pengujian parameter *threshold* terdapat pada Tabel 6.

Tabel 6 Pengaruh *threshold Information Gain* terhadap *F1-score*

| *Threshold* | Rata-Rata *F1-score* |
|---|---|
| 0,5 | 0.8825 |
| 0,8 | 0.8825 |
| 0,9 | 0.8825 |
| 0,95 | 0.8819 |
| **0,96** | **0.8828** |
| 0,97 | 0.8795 |
| 0,98 | 0.8802 |

Berdasarkan Tabel 6, hasil terbaik dari pengujian *threshold* adalah 0,96 dengan nilai sebesar 0,8828. *Threshold* 0,96 merupakan *threshold* yang tepat untuk penelitian ini. Hal ini dikarenakan, pada *threshold* 0,96 fitur yang dipilih oleh sistem merupakan fitur yang memiliki nilai *gain* lebih dari *threshold*, fitur-fitur tersebut adalah fitur yang signifikan terhadap proses klasifikasi. Selain itu, fitur yang kurang dari *threshold* akan dihilangkan, sehingga tidak akan terdapat data-data *noise*.

Selanjutnya, *threshold* tersebut digunakan untuk pengujian pengaruh *Information Gain* terhadap performa *Naïve Bayes*. Analisis dilakukan dengan membandingkan hasil rata-rata *F1-score* menggunakan *Information Gain* dan tanpa *Information Gain*. Hasil evaluasi dari pengujian ini terdapat pada Tabel 7.

Tabel 7 Hasil evaluasi pengaruh *Information Gain* terhadap *F1-score*

| *Fold* | *F1-score* | |
|---|---|---|
| | Tanpa *Information Gain* | Dengan *Information Gain* |
| 1 | 0.848214 | 0.850679 |
| 2 | 0.877358 | 0.877358 |
| 3 | 0.898551 | 0.902913 |
| 4 | 0.881188 | 0.881188 |
| 5 | 0.907317 | 0.901961 |
| Rata-rata *F1-score* | 0.8825 | **0.8828** |

Berdasarkan hasil evaluasi pengujian pengaruh *Information Gain* pada Tabel 7, performa *Naïve Bayes* sebelum menggunakan *Information Gain* adalah 88,25%. Sedangkan performa setelah menggunakan *Information Gain* dengan *threshold* sebesar 0,96 adalah 88,28%. Performa *Naïve Bayes* mengalami peningkatan dengan menggunakan *Information Gain* sebagai *feature selection*. Hal ini dikarenakan *Information Gain* melakukan proses seleksi pada fitur-fitur yang dianggap penting terhadap proses klasifikasi. Proses seleksi pada *Information Gain*, dipilih berdasarkan hasil *Gain* dari perhitungan yang dilakukan, dimana fitur yang memiliki nilai *Gain* lebih dari *threshold* yang dipilih. Semakin tinggi nilai *Gain* yang dimiliki suatu kata, maka semakin signifikan kata tersebut dalam membantu proses klasifikasi. Kata-kata yang tidak terpilih selanjutnya akan dihilangkan oleh *Information Gain*, sehingga menyebabkan dimensi data akan berkurang. Oleh karena itu, dengan berkurangnya dimensi data, performa yang dihasilkan oleh klasifikasi *Naïve Bayes* akan meningkat.

## 4. Kesimpulan

1. Semakin banyak jumlah *dataset* yang digunakan pada klasifikasi sentimen, semakin tinggi performa yang dihasilkan oleh sistem. Dalam penelitian ini, performa paling baik adalah dengan menggunakan 1000 data yaitu 88,28%. Hal ini dikarenakan semakin banyak jumlah data yang digunakan, maka akan banyak data yang dipelajari oleh sistem, sehingga sistem akan mudah untuk beradaptasi terhadap data yang akan digunakan pada *training* ataupun *testing*.
2. Dengan menggunakan *lemmatization* dapat meningkatkan performa sistem yang dibangun. Pada penelitian ini, performa dengan menggunakan *lemmatization* lebih tinggi dibandingkan tanpa *lemmatization*.
3. Pemilihan *threshold* yang tepat dapat meningkatkan performa *Naïve Bayes.* pada penelitian ini, *threshold* yang tepat adalah 0.96. Untuk itu, performa sistem yang dihasilkan dengan menggunakan *Information Gain* dengan *threshold* 0.96 lebih baik dibandingkan tanpa *Information Gain.*
4. Klasifikasi sentimen pada ulasan buku berbahasa inggris menggunakan *Information Gain* dan *Naïve Baye*s menghasilkan performa sebesar 88,28%.

**Daftar Pustaka**


[1] E. Indrayuni, "Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *J. Evolusi*, vol. 4, 2016.

[2] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, "Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes," *Am. Inst. Phys. Conf. Proc. 1867*, 2017.

[3] F. Arfiana, "Klasifikasi Kendaraan Roda Empat Menggunakan Metode Naïve Bayes," Universitas Widyatama, 2014.

[4] C. C. Aggarwal, *Data Classification: Algorithms and Application*. CRC Press.

[5] A. Jai, "Analisis Sentimen Review Hotel Dengan Algoritma Support Vector Machines dan Naive Bayes," Universitas Dian Nuswantoro, Semarang, 2016.

[6] I. Sofiana, I. Atastina, and A. Ardiyanti, "Analisis Pengaruh Feature Selection Menggunakan Information Gain dan Chi-Square Untuk Kategorisasi Teks Berbahasa Indonesia," *Telkom Univ.*, 2012.

[7] T. Setiyori, "Penerapan Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal pada Taksonomi Bloom," *J. Sist. Inf.*, vol. 6, Feb. 2017.

[8] O. Chandler and E. Khuri, "Reviews Book," *Goodreads*, 2006. [Online]. Available: https://www.goodreads.com/review/recent_reviews.

[9] I. Monica, S. Mubarok, and Adiwijaya, "Analisis Sentimen level Aspek pada Ulasan Produk menggunakan Multinomial Naïve Bayes," *Univ. Telkom*, 2017.

[10] A. A. H. K. Amin, M. S. Mubarok, and W. Maharani, "Sentiment Analysis Online Product Reviews Menggunakan Naive Bayes Classifier dan Algoritma Apriori," *Univ. Telkom*, 2016.

[11] A. K. Ingason, S. Helgadóttir, H. Loftsson, and E. Rögnvaldsson, "A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI)," 2008.

[12] R. A. Aziz, M. S. Mubarok, and Adiwijaya, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naïve Bayes," *Indones. Symp. Comput. IndoSC*, 2016.

[13] M. H. Syahnur, M. A. Bijaksana, and M. S. Mubarok, "Kategorisasi Topik Tweet di Kota Jakarta, Bandung, dan Makassar dengan Metode Multinomial Naïve Bayes Classifier," *E-Proceeding Eng.*, vol. 3, Aug. 2016.

[14] Adiwijaya, *Aplikasi Matriks dan Ruang Vektor*. Yogyakarta: Graha Ilmu, 2014.

[15] D. A. Bimantoro and S. 'Uyun, "Pengaruh Penggunaan Information Gain Untuk Seleksi Fitur Citra Tanah Dalam Rangka Menilai Kesesuaian Lahan Pada Tanaman Cengkeh," *J. Inform. Sunan Kalijaga*, vol. 2, May 2017.

[16] A. H. R. Z. Arifin, M. S. Mubarok, and Adiwijaya, "Learning Struktur Bayesian Networks Menggunakan Novel Modified Binary Differential Evolution Pada Klasifikasi Data," *Indones. Symp. Comput. IndoSC*, 2016.

[17] Adiwijaya, *Matematika Diskrit dan Aplikasinya*. Bandung: Alfabeta, 2016.

[18] Dyta Anggraeni, "Klasifikasi Topik Menggunakan Metode Naive Bayes dan Maximum Entropy Pada Artikel Media Massa dan Abstrak Tulisan," *Univ. Indones.*, Jan. 2008.

[19] A. Indriani, "Klasifikasi Data Forum Dengan Menggunakan Metode Naïve Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf. SNATI*, Jun. 2014.

[20] A. R. Naufal, R. S. Wahono, and A. Syukur, "Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan," *J. Intell. Syst.*, vol. 1, Dec. 2015.