

ANALISIS SENTIMEN PADA REVIEW BUKU MENGGUNAKAN ALGORITMA NAÏVE BAYES

Dinda Ayu Muthia
Manajemen Informatika
AKADEMI MANAJEMEN INFORMATIKA DANKOMPUTER
AMIK BSI Bekasi
Jl. Cut Mutiah No. 88, Bekasi
dinda.dam@bsi.ac.id

ABSTRACT

Nowadays consumers are increasingly making their opinions and experiences online. Reading those reviews are time-consuming, but, if only few reviews were read, the evaluation would be biased. Sentiment analysis aims to solve this problem by automatically classifying user reviews into positive or negative opinions. Naive Bayes classifier is a popular machine learning technique for text classification, because it is so simple, efficient and it has a great performance in many domains. However, it has a lack that it is highly sensitive to the high number of feature. Therefore, in this research the concatenation of feature selection methods is used, that is Information gain and Genetic algorithm that could increase the accuracy of Naive Bayes classifier. This research turns out text classification in the form of positive or negative from book reviews. The measurement is based on the accuracy of Naive Bayes before and after adding the feature selection method. Evaluation was performed using 10 fold cross validation. Whereas the measurement of accuracy was measured by using confusion matrix and ROC curve. The result of this research is the improvement of accuracy of Naive Bayes from 75.50% to 84.50%.

Keyword: *Sentiment analysis, Review, Book, Naive Bayes, Text Classification.*

1. PENDAHULUAN

1.1. Latar Belakang

Saat ini konsumen yang menulis opini dan pengalaman secara online terus meningkat. Membaca review tersebut secara keseluruhan bisa memakan waktu, namun, jika hanya sedikit review yang dibaca evaluasi akan bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan review pengguna menjadi opini positif atau negatif (Z. Zhang et al., 2011).

Ada beberapa penelitian yang sudah dilakukan dalam melakukan klasifikasi sentimen terhadap review yang tersedia secara onlinediantaranya, analisa sentiment pada komentar review film dan review multi domain lainnya seperti review buku, DVD, barang elektronik, dan lain-lain yang ada di Amazon.com menggunakan pengklasifikasi *Lexicon Labeling, Heuristic Labeling, Self-labeled instance, Self-learned Features, Oracle Labeling, Naive Bayes, Support Vector Machine, dan Maximum Entropy* yang dilakukan oleh He & Zhou pada tahun 2011. Analisa sentimen pada reviewfilm dari IMDB, review produk seperti GPS, buku, dan kamera dari Amazon.com menggunakan

pengklasifikasi *Support Vector Machine* dan *Artificial Neural Network* yang dilakukan oleh Moraes pada tahun 2013. Analisa sentimen pada opini review film menggunakan pengklasifikasi *Support Vector Machine* dan *Particle Swarm Optimization* yang dilakukanoleh Basari pada tahun 2013. Pengklasifikasian sentimen pada review restoran di internet yang ditulis dalam bahasa Canton menggunakan pengklasifikasi *Naive Bayes* dan *Support Vector Machine* yang dilakukan oleh Z. Zhang pada tahun 2011. Analisa sentimen pada *reuters* dan teks bahasa China menggunakan pengklasifikasi *Naive Bayes* dan dua metric pengevaluasi fitur yaitu *Multi-class Odds Ratio (MOR)* dan *Class Discriminating Measure (CDM)* yang dilakukan Chen pada tahun 2009. Klasifikasi sentimen pada review online tempat tujuan perjalanan menggunakan pengklasifikasi *Naive Bayes, Support Vector Machine, dan Character Based N-gram Model* yang dilakukan oleh Ye, Zhang, dan Law pada tahun 2009.

Pengklasifikasi *Naive Bayes* sangat sederhana dan efisien, (Chen et al., 2009). Di samping kesederhanaannya, pengklasifikasi *Naive Bayes* adalah teknik *machine learning* yang populer untuk klasifikasi teks, dan

memiliki performa yang baik pada banyak domain (Ye, Zhang, & Law, 2009). Namun, *Naïve Bayes* memiliki kekurangan yaitu sangat sensitif dalam pemilihan fitur (Chen et al., 2009). Terlalu banyak jumlah fitur, tidak hanya meningkatkan waktu penghitungan tapi juga menurunkan akurasi klasifikasi (Uysal & Gunal, 2012).

Tingkatan lain yang umumnya ditemukan dalam pendekatan klasifikasi sentimen adalah pemilihan fitur. Pemilihan fitur bisa membuat pengklasifikasi baik lebih efisien/efektif dengan mengurangi jumlah data yang dianalisa, maupun mengidentifikasi fitur yang sesuai untuk dipertimbangkan dalam proses pembelajaran. Ada dua jenis utama metode pemilihan fitur dalam *machine learning*: *wrapper* dan *filter*. *Wrapper* menggunakan akurasi klasifikasi dari beberapa algoritma sebagai fungsi evaluasinya. Metode *filter* terdiri dari *Document Frequency*, *mutual information*, *Information Gain*, dan *Chi-Square*. Tidak ada dari keempat metode tersebut yang secara luas diterima sebagai metode penyeleksi fitur terbaik untuk klasifikasi sentimen atau kategorisasi teks, namun, *Information Gain* sering lebih unggul dibandingkan yang lain. *Wrapper* mengevaluasi fitur secara berulang dan menghasilkan akurasi klasifikasi yang tinggi. Salah satu metode *wrapper* yang bisa digunakan dalam pemilihan fitur adalah *Genetic algorithm* (GA).

Umumnya metode pemilihan fitur yang lebih disukai adalah filter dikarenakan waktu pemrosesannya yang relatif rendah. *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 (Uysal & Gunal, 2012). Untuk mengurangi kerumitan perhitungan dilakukan pemilihan fitur dengan menghitung *Information Gain* (Zhang et al., 2011).

Pada penelitian ini pengklasifikasi *Naïve Bayes* dengan *Information Gain* dan *Genetic algorithm* sebagai metode pemilihan fitur akan diterapkan untuk mengklasifikasikan teks pada komentar dari review suatu buku untuk meningkatkan akurasi analisa sentimen.

1.2. Identifikasi Masalah

Pengklasifikasi *Naïve Bayes* sangat sederhana, efisien dan merupakan teknik *machine learning* yang populer untuk klasifikasi teks, serta memiliki performa yang

baik pada banyak domain. Namun, *Naïve Bayes* memiliki kekurangan yaitu sangat sensitif pada fitur yang terlalu banyak, yang mengakibatkan akurasi klasifikasi menjadi rendah.

1.3. Rumusan Masalah

Seberapa besar efek metode pemilihan fitur *Information Gain* dan *Genetic Algorithm* pada akurasi analisa sentimen pada review buku menggunakan pengklasifikasi *Naïve Bayes*?

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengintegrasikan metode pemilihan fitur *Information Gain* dan *Genetic algorithm* dalam menganalisa sentimen pada review buku menggunakan pengklasifikasi *Naïve Bayes*.

1.5. Manfaat Penelitian

1. Membantu para calon pembaca dalam mengambil keputusan saat ingin membaca suatu buku agar bisa mengurangi waktu dalam membaca review dan komentar dari suatu buku.
2. Membantu para calon penonton dalam mengambil keputusan saat ingin menonton film yang dibuat berdasarkan buku-buku *best-seller* agar bisa mengurangi waktu dalam membaca review yang umumnya membandingkan mana yang lebih bagus ceritanya, versi buku atau versi filmnya.
3. Membantu para pengembang sistem yang berkaitan dengan review film, baik dari sumber IMDB, Goodreads, maupun dari sosial media lainnya seperti Twitter, Blog, dan lain-lain.

1.6. Kontribusi Penelitian

Mengklasifikasikan teks analisa sentimen pada review suatu film dengan menggunakan pengklasifikasi *Naïve Bayes* yang menerapkan metode pemilihan fitur *Information gain* dan *Genetic algorithm*.

2. TINJAUAN PUSTAKA

2.1. Review

Menurut Reddy (V, Somayajulu, & Dani, 2010) dalam banyak kasus keputusan yang kita buat dipengaruhi oleh opini dari orang lain. Sebelum kesadaran akan internet menjadi tersebar luas, banyak dari kita yang biasanya menanyakan opini teman atau tetangga mengenai alat elektronik ataupun film sebelum benar-benar membelinya.

Dengan berkembangnya ketersediaan dan popularitas akan sumber yang kaya opini seperti *website* review online dan blog pribadi, kesempatan baru dan tantangan muncul semenjak orang-orang sekarang bisa dan menggunakan secara aktif informasi dan teknologi untuk mencari dan memahami opini orang lain.

Menurut Yessenov (Yessenov, 2009) ada beberapa contoh *website* yang bisa mereview produk, seperti Amazon, atau situs review film seperti Rotten Tomatoes yang memungkinkan untuk memberikan rating pada produk, biasanya beberapa skala ditentukan sama dengan review personal yang dibuat.

2.2. Sentiment Analysis (Analisa Sentimen)

Menurut Feldman (Feldman, 2013) analisa sentimen (*Opinion mining*) didefinisikan sebagai tugas menemukan opini dari penulis tentang entitas tertentu. Menurut Tang dalam Haddi (Haddi, Liu, & Shi, 2013), analisa sentimen pada review adalah proses menyelidiki review produk di internet untuk menentukan opini atau perasaan terhadap suatu produk secara keseluruhan. Menurut Thelwall dalam Haddi, analisa sentimen diperlakukan sebagai suatu tugas klasifikasi yang mengklasifikasikan orientasi suatu teks ke dalam positif atau negatif. Menurut Mejova dalam Basari (Basari et al., 2013), tujuan dari analisa sentimen adalah untuk menentukan perilaku atau opini dari seorang penulis dengan memperhatikan suatu topik tertentu. Perilaku bisa mengindikasikan alasan, opini atau penilaian, kondisi kecenderungan (bagaimana si penulis ingin mempengaruhi pembaca).

2.3. Pemilihan Fitur (Feature Selection)

Menurut Gorunescu (Gorunescu, 2011) pemilihan fitur digunakan untuk menghilangkan fitur yang tidak relevan dan berulang, yang mungkin menyebabkan kekacauan, dengan menggunakan metode tertentu. Menurut John, Kohavi, dan Pflieger dalam Chen (Chen et al., 2009) ada dua jenis metode pemilihan fitur dalam *machine learning*, yaitu *wrapper* dan *filter*.

Menurut Chen (Chen et al., 2009) *wrapper* menggunakan akurasi klasifikasi dari beberapa algoritma sebagai fungsi evaluasinya. Menurut Gunal (Gunal, 2012) salah satu metode *wrapper* yang bisa digunakan dalam pemilihan fitur adalah *Genetic algorithm* (GA).

1. Genetic Algorithm

Menurut Han (Han & Kamber, 2007) *Genetic algorithm* berusaha untuk menggabungkan ide evolusi alam. Secara umum, pembelajaran genetika dimulainya sebagai berikut:

- a. Sebuah populasi awal dibuat terdiri dari acak. Setiap aturan bis diwakili oleh string bit. Sebagai contoh sederhana, misalkan bahwa sampel dalam satu set pelatihan yang diberikan dijelaskan oleh dua atribut Boolean, A_1 dan A_2 , dan bahwa ada dua kelas, C_1 dan C_2 . Aturan "*If A_1 And Not A_2 Then C_2* " dapat dikodekan sebagai string bit "100," di mana dua bit paling kiri mewakili atribut A_1 dan A_2 , masing-masing, dan bit paling kanan mewakili kelas. Demikian pula, aturan "*If Not A_1 And Not A_2 Then C_1* " dapat dikodekan sebagai "001." Jika atribut memiliki nilai-nilai, di mana >2 , maka bit dapat digunakan untuk mengkodekan nilai-nilai atribut itu. Kelas dapat dikodekan dengan cara yang sama.
- b. Berdasarkan gagasan ketahanan dari yang paling sesuai, populasi baru terbentuk terdiri dari aturan yang paling sesuai dalam populasi saat ini, serta keturunan aturan ini. Biasanya, fitness aturan dinilai dengan akurasi klasifikasi pada satu set sampel pelatihan.
- c. Keturunan diciptakan dengan menerapkan operator genetika seperti *crossover* dan mutasi. Dalam *crossover*, substring dari sepasang aturan ditukarkan untuk membentuk pasangan aturan baru. Dalam mutasi, bit yang dipilih secara acak dalam aturan string dibalik.
- d. Proses menghasilkan populasi baru berdasarkan aturan populasi sebelumnya berlanjut sampai populasi, P , berkembang di mana setiap aturan dalam P memenuhi ambang batas *fitness* yang sudah ditentukan.

Genetic algorithm mudah disejajarkan dan telah digunakan untuk klasifikasi seperti masalah optimasi lainnya. Dalam data mining, algoritma genetika dapat digunakan untuk mengevaluasi fitness algoritma lainnya.

Menurut Chen (Chen et al., 2009) metode filter terdiri dari *Information gain*, *Term frequency*, *Chi-square*, *Expected cross*

entropy, Odds ratio, The weight of evidence of text, Mutual information, dan Gini index. Menurut Santosa (Santosa, 2007), ada pula metode Gain ratio. Menurut Moraes (Moraes, Valiati, & Gavião Neto, 2013), ada pula metode Document Frequency. Tidak ada dari keempat metode tersebut yang secara luas diterima sebagai metode penyeleksi fitur terbaik untuk klasifikasi sentimen atau kategorisasi teks, namun, information gain sering lebih unggul dibandingkan yang lain.

1. Information Gain

Tahapan dalam proses perhitungan Information gain sebagai berikut:

- a. Cari nilai entropi sebelum pemisahan dengan rumus berikut:

$$Entropi(y) = -\sum P_i \log_2 P_i$$

P_i : proporsi data y dengan kelas i

- b. Cari nilai entropi setelah pemisahan berdasarkan atribut A dengan rumus berikut:

$$Total\ entropi = \sum_{c \in nilai(A)} \frac{y_c}{y} entropi\ y_c$$

- c. Cari nilai information gain dengan rumus berikut:

$$gain(y, A) = entropi(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entropi(y_c)$$

2.4. Algoritma Naïve Bayes

Menurut Markov (Markov & Daniel, 2007) tahapan dalam algoritma Naïve Bayes:

- 1. Hitung probabilitas bersyarat/likelihod:

$$P(x / C) = P(x_1, x_2, \dots, x_n / C)$$

C = class

x = vektor dari nilai atribut n

$P(x_i/C)$ = proporsi dokumen dari class C yang mengandung nilai atribut x_i

- 2. Hitung probabilitas prior untuk tiap class:

$$P(C) = \frac{N_j}{N}$$

N_j = jumlah dokumen pada suatu class

N = jumlah total dokumen

- 3. Hitung probabilitas posterior dengan rumus:

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)}$$

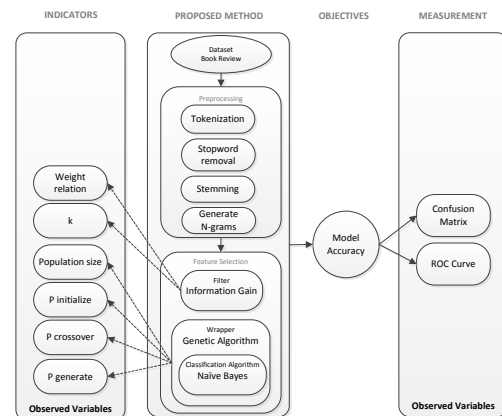
Menurut Santoso (Santoso, 2007) dengan kata-kata yang lebih umum, rumus Bayes bisa diberikan sebagai berikut:

$$Posterior = \frac{likelihood \times prior}{evidence}$$

2.5. Kerangka Pemikiran

Penelitian ini dimulai dari adanya masalah dalam klasifikasi teks pada review film menggunakan Naïve Bayes, di mana pengklasifikasi tersebut memiliki kekurangan

yaitu sangat sensitif pada fitur yang terlalu banyak, yang mengakibatkan akurasi klasifikasi menjadi rendah. Dataset yang digunakan dalam penelitian ini berupa review buku-buku best-seller yang di tahun 2014 diangkat menjadi sebuah film, diantaranya buku Divergent, Horns, Hunger Games: Mockingjay, dan Vampire Academy yang didapat dari situs khusus review buku yaitu www.goodreads.com yang terdiri dari 100 review positif dan 100 review negatif. Preprocessing yang dilakukan dengan tokenization, stopwords removal, stemming, dan generate N-grams. Metode pemilihan fitur yang digunakan adalah information gain dan genetic algorithm, sedangkan pengklasifikasi yang digunakan adalah Naïve Bayes. Pengujian 10 foldcross validation akan dilakukan, akurasi algoritma akan diukur menggunakan confusion matrix dan hasil olahan data dalam bentuk kurva ROC. RapidMiner versi 5.3 digunakan sebagai alat bantu dalam mengukur akurasi data eksperimen. Gambar 1. menggambarkan kerangka pemikiran yang penulis usulkan dalam penelitian ini.



Gambar 1. Kerangka Pemikiran

3. METODE PENELITIAN

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:

- a. Pengumpulan Data

Penulis menggunakan data review buku-buku best-seller yang di tahun mendatang diangkat menjadi sebuah film, diantaranya buku Divergent, Horns, Hunger Games: Mockingjay, dan Vampire Academy yang didapat dari situs khusus review buku yaitu www.goodreads.com yang terdiri dari 100 review positif dan 100 review negatif.

- b. Pengolahan Awal Data

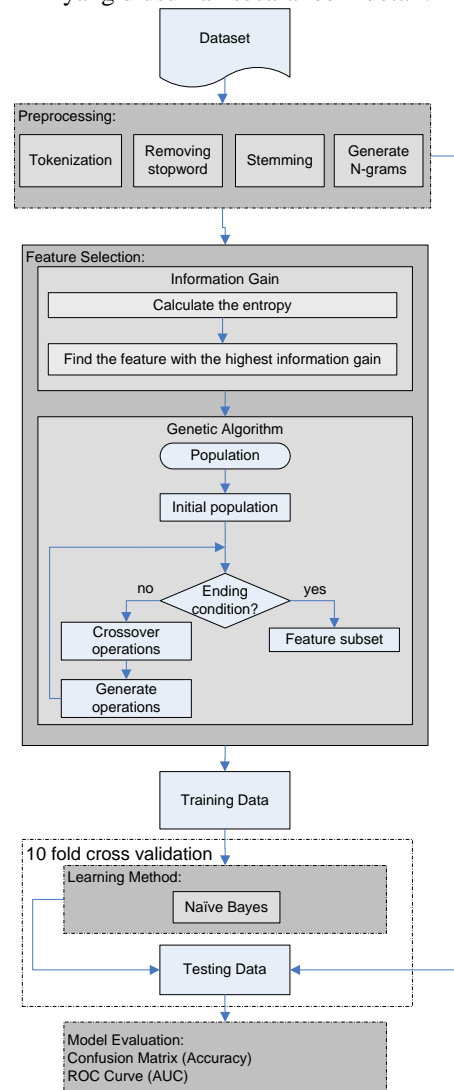
Dataset ini dalam tahap *preprocessing* harus melalui 4 proses, yaitu:

- 1) *Tokenization*
Yaitu mengumpulkan semua kata yang muncul dan menghilangkan tanda baca maupun simbol apapun yang bukan huruf.
- 2) *Stopwords Removal*
Yaitu penghapusan kata-kata yang tidak relevan, seperti *the, of, for, with,* dan sebagainya.
- 3) *Stemming*
Yaitu mengelompokkan kata ke dalam beberapa kelompok yang memiliki kata dasar yang sama, seperti *drug, drugged,* dan *drugs* di mana kata dasar dari semuanya adalah kata *drug*.
- 4) *Generate N-grams*
Yaitu menggabungkan kata sifat yang seringkali muncul untuk menunjukkan sentimen, seperti kata *definitely* dan kata *recommended*. Kata *recommended* memang sudah menunjukkan sentimen bentuk opini positif. Kata *definitely* tidak akan berarti jika berdiri sendiri. Namun jika dua kata tersebut digabung menjadi "*definitely recommended*", maka akan sangat menguatkan opini positif tersebut. Penulis hanya menggunakan penggabungan dua kata, yang disebut 2-grams (bigrams).

Sedangkan untuk tahap *transformation* dengan melakukan pembobotan TF-IDF pada masing-masing kata. Di mana prosesnya menghitung kehadiran atau ketidakhadiran sebuah kata di dalam dokumen. Berapa kali sebuah kata muncul di dalam suatu dokumen juga digunakan sebagai skema pembobotan dari data tekstual.

- c. Metode Yang Diusulkan
Metode yang penulis usulkan adalah menggabungkan 2 jenis metode pemilihan fitur, yaitu *filter* dan *wrapper*. Dari jenis *filter*, digunakan *information gain* dan dari jenis *wrapper*, digunakan *genetic algorithm* sebagai metode pemilihan fitur agar akurasi pengklasifikasi *Naïve Bayes* bisa meningkat. Penulis menggunakan pengklasifikasi *Naïve Bayes* karena sangat sederhana, efisien dan merupakan teknik *machine learning* yang populer untuk klasifikasi teks, serta memiliki performa yang baik pada banyak domain. *Genetic algorithm* yang penulis terapkan adalah menggunakan *Naïve Bayes* yang diuji di dalam tahap

wrapper. Lihat gambar 2. untuk model yang diusulkan secara lebih detail.



Gambar 2. Model yang diusulkan

Hasil yang dibandingkan adalah akurasi *Naïve Bayes* sebelum menggunakan metode pemilihan fitur dengan akurasi *Naïve Bayes* setelah menggunakan metode pemilihan fitur gabungan, yaitu *Information gain* dan *Genetic algorithm*. Di mana pada *Genetic algorithm*, *Naïve Bayes* diuji di dalam tahap *wrapper*.

- d. Eksperimen dan Pengujian Metode
Untuk eksperimen data penelitian, penulis menggunakan RapidMiner 5 untuk mengolah data.
- e. Evaluasi dan Validasi Hasil
Validasi dilakukan menggunakan *10 fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC untuk mengukur nilai AUC. Dengan *confusion matrix*, akurasi *Naïve Bayes* sebelum menggunakan metode pemilihan fitur dan setelah

menggunakan metode pemilihan fitur. Tabel 1 berikut adalah tampilan *confusion matrix* dan rumus perhitungannya menurut Gorunescu (Gorunescu, 2011):

Tabel 1. *Confusion Matrix*

Classification	Predicted Class		
	Class = Yes	Class = No	
Observed Class	Class = Yes	a (True positive - TP)	b (False Negative - FN)
	Class = No	c (False positive - FP)	d (True negative - TN)

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN}$$

4. HASIL DAN PEMBAHASAN

4.1. Hasil

4.1.1. Klasifikasi Teks Menggunakan Algoritma *Naïve Bayes*

Data training yang digunakan dalam pengklasifikasian teks ini terdiri dari 100 review buku positif dan 100 review buku negatif. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Sebelum diklasifikasikan, data tersebut harus melalui beberapa tahapan proses agar bisa diklasifikasikan dalam proses selanjutnya, berikut adalah tahapan prosesnya:

1. Pengumpulan Data

Data review positif disatukan dalam folder dengan nama pos. Sedangkan data review negatif disatukan penyimpanannya dalam folder dengan nama neg. Tiap dokumen berekstensi .txt yang dapat dibuka menggunakan aplikasi Notepad.

2. Pengolahan Awal Data

Proses yang dilalui terdiri dari *tokenization*, *stopwords removal*, *stemming*, dan *generate N-grams*. Hasil pengolahan data awal dapat dilihat pada tabel 2.

Tabel 2. Pengolahan awal data

Review	Tokenization	Stop words removal	Stemming	Generate N-grams
The sad thing is I could	The sad thing is I could	sad thing I go paragraphs	sad thing i go para graph	sad sad_thing thing thing_i i_go go

d	proba	listin	h list	go_para
prob	bly go	g	exam	graph
ably	on for	exam	pl	paragra
go	severa	ples	unre	ph
on	l more	unrea	alist	paragra
for	paragr	listic	illog	ph_list
sever	aphs	illogi	plot	list
al	listing	cal	point	list_exa
more	all the	plot	nons	mpl
para	examp	points	ens	exampl
grap	les of	nonse	idea	exampl_
hs	unreal	nsical	reaso	unrealist
listin	istic	ideas	n whi	unrealist
g all	illogic	reaso	i t	unrealist
the	al plot	ns	book	_illog
exam	points	why I		illog
ples	nonse	t		illog_plo
of	nsical	book		t plot
unre	ideas			plot_poi
alistic,	as			nt point
illogi	well			point_no
cal	as			nsens
plot	reason			nonsens
point	s why			nonsens
s &	I didn			_idea
nons	t like			idea
ensic	this			idea_rea
al	book			son
ideas				reason
as				reason_
well				whi whi
as				whi_i i
reas				i_t t
ons				t_book
why I				book
didn'				
t like				
this				
book				

3. Klasifikasi

Proses klasifikasi di sini adalah untuk menentukan sebuah kalimat sebagai anggota *class* positif atau *class* negatif berdasarkan nilai perhitungan probabilitas dari rumus *Bayes* yang lebih besar. Jika hasil probabilitas kalimat tersebut untuk *class* positif lebih besar dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* positif. Jika probabilitas untuk *class* positif lebih kecil dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* negatif. Penulis mendapatkan 4 kata dan 1 gabungan kata yang berhubungan dengan sentimen yang paling sering muncul, yaitu *amaz* yang berasal dari kata *amazing*, *favorit*, *great*, *love* dan *i_love*.

4.1.2. Pengujian Model dengan 10 Fold Cross Validation

Pada penelitian ini, penulis melakukan pengujian model dengan menggunakan teknik 10 cross validation, di mana proses ini membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Setelah itu proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar.

4.1.3. Optimasi Model dengan Gabungan Metode Pemilihan Fitur

Dengan menggabungkan metode pemilihan fitur filter dan wrapper, di mana dalam penelitian ini metode yang digunakan adalah Information gain dari filter dan Genetic algorithm dari wrapper. Data yang akan diolah diberikan bobot dari Information gain untuk meningkatkan akurasi pengklasifikasi Naive Bayes.

4.1.4. Eksperimen Terhadap Indikator Model

Untuk mendapatkan model yang baik, beberapa indikator disesuaikan nilainya agar didapatkan hasil akurasi yang tinggi. Dalam penyesuaian indikator pada Genetic algorithm, akurasi paling tinggi diperoleh dengan kombinasi population size=50, p initialize=0.8, p crossover=0.8, dan p generate=1.0. Hasil akurasi mencapai 84.50%. Jika indikator lainnya turut diubah nilainya, dapat menyebabkan proses pengolahan data menjadi semakin lama.

4.2. Pembahasan

Dengan memiliki model klasifikasi teks pada review, pembaca dapat dengan mudah mengidentifikasi mana review yang positif maupun yang negatif. Dari data review yang sudah ada, dipisahkan menjadi kata-kata, lalu diberikan bobot pada masing-masing kata tersebut. Dapat dilihat kata mana saja yang berhubungan dengan sentimen yang sering muncul dan mempunyai bobot paling tinggi. Dengan demikian dapat diketahui review tersebut positif atau negatif.

Dalam penelitian ini, hasil pengujian model akan dibahas melalui confusion matrix untuk menunjukkan seberapa baik model yang terbentuk. Tanpa menggunakan metode pemilihan fitur, algoritma Naive Bayes sendiri sudah menghasilkan akurasi sebesar 75.50%. Akurasi tersebut masih kurang akurat,

sehingga perlu ditingkatkan lagi menggunakan metode pemilihan fitur. Setelah menggunakan metode pemilihan fitur dari filter dan wrapper yang digabungkan, akurasi algoritma Naive Bayes meningkat menjadi 84.50% seperti yang bisa dilihat pada tabel 3.

Tabel 3. Model algoritma Naive Bayes sebelum dan sesudah menggunakan metode pemilihan fitur

	Algoritma Naive Bayes	Algoritma Naive Bayes + Information Gain & Genetic Algorithm
Sukses klasifikasi review positif	89	77
Sukses klasifikasi review negatif	62	92
Akurasi model	75.50%	84.50%

4.2.1. Pengukuran dengan Confusion Matrix

Pengukuran dengan confusion matrix di sini akan menampilkan perbandingan dari hasil akurasi model Naive Bayes sebelum ditambahkan metode pemilihan fitur yang bisa dilihat pada tabel 4 dan setelah ditambahkan metode pemilihan fitur, yaitu penggabungan Information gain dan Genetic algorithm yang bisa dilihat pada tabel 5.

Tabel 4. Confusion matrix model Naive Bayes sebelum penambahan metode pemilihan fitur

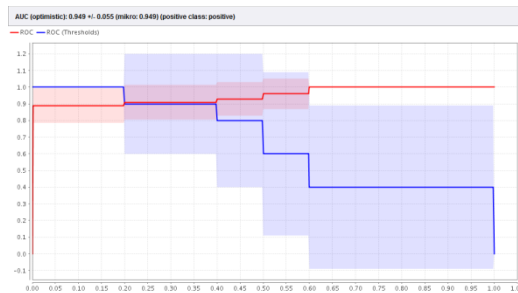
Akurasi Naive Bayes: 75.50% +/- 11.93% (mikro: 75.50%)			
	True negative	True positive	Class precision
Pred. negative	62	11	84.93%
Pred. positif	38	89	70.08%
Class recall	62.00%	89.00%	

Tabel 5. Confusion matrix model Naive Bayes sesudah penambahan metode pemilihan fitur

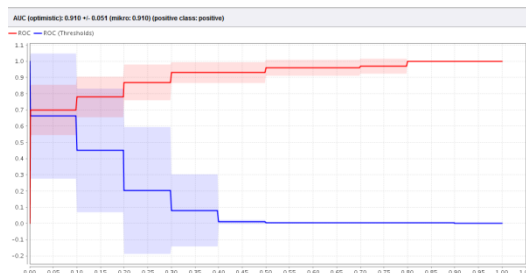
Akurasi Naive Bayes: 84.50% +/- 5.22% (mikro: 84.50%)			
	True negative	True positive	Class precision
Pred. negative	92	23	80.00%
Pred.	8	77	90.59%

<i>positif</i>			
<i>Class recall</i>	92.00%	77.00%	

Berikut adalah tampilan kurva ROC dari hasil uji data. Gambar 3 adalah kurva ROC untuk model *Naïve Bayes* sebelum menggunakan metode pemilihan fitur dan gambar 4 adalah kurva ROC untuk model *Naïve Bayes* setelah menggunakan metode pemilihan fitur.



Gambar 3. Kurva ROC model *Naïve Bayes* sebelum menggunakan metode pemilihan fitur



Gambar 4. Kurva ROC model *Naïve Bayes* setelah menggunakan metode pemilihan fitur

4.3. Implikasi Penelitian

Implikasi penelitian ini mencakup beberapa aspek, di antaranya:

1. Implikasi terhadap aspek sistem
Hasil evaluasi menunjukkan penerapan *Information gain* dan *Genetic algorithm* untuk seleksi fitur dapat meningkatkan akurasi *Naïve Bayes* dan merupakan metode yang cukup baik dalam mengklasifikasi teks review buku. Dengan demikian penerapan metode tersebut dapat membantu para calon pembaca dalam mengambil keputusan saat ingin membaca suatu buku, maupun calon penonton film yang kisahnya diangkat dari buku-buku tersebut, agar bisa mengurangi waktu dalam membaca review dan komentar dari buku tersebut.
2. Implikasi terhadap aspek manajerial
Membantu para pengembang sistem yang berkaitan dengan review buku, baik dari sumber www.goodreads.com maupun dari sosial media lainnya seperti Twitter, Blog,

dan lain-lain agar menggunakan aplikasi RapidMiner dalam membangun suatu sistem.

3. Implikasi terhadap aspek penelitian lanjutan
Penelitian selanjutnya bisa menggunakan metode pemilihan fitur ataupun dataset dari domain yang berbeda, seperti review produk, review restoran, dan sebagainya.

5. KESIMPULAN

Dari pengolahan data yang sudah dilakukan, penggabungan metode pemilihan fitur, yaitu *filter* dan *wrapper*, terbukti dapat meningkatkan akurasi pengklasifikasi *Naïve Bayes*. Data review buku dapat diklasifikasi dengan baik ke dalam bentuk positif dan negatif. Akurasi *Naïve Bayes* sebelum menggunakan penggabungan metode pemilihan fitur mencapai 75.50% Sedangkan setelah menggunakan penggabungan metode pemilihan fitur, yaitu *Information gain* dan *Genetic algorithm*, akurasinya meningkat hingga mencapai 84.50%. Peningkatan akurasi mencapai 9%.

Model yang terbentuk dapat diterapkan pada seluruh data review buku, sehingga dapat dilihat secara langsung hasilnya dalam bentuk positif dan negatif. Hal ini dapat membantu seseorang untuk menghemat waktu saat mencari review suatu buku yang khususnya buku-buku *best-seller* yang akan diangkat ke layar lebar menjadi sebuah film tanpa harus mengkhawatirkan adanya *spoiler* dan pemberian rating yang tidak sesuai dengan reviewnya.

6. DAFTAR PUSTAKA

- Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, 53, 453–462.
- Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with *Naïve Bayes*. *Expert Systems with Applications*, 36(3), 5432–5435.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82.
- Gorunescu, F. (2011). *Data Mining Concept Model Technique*.
- Gunal, S. (2012). Hybrid feature selection for text classification, 20.
- Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment

- Analysis. *Procedia Computer Science*, 17, 26–32.
- Han, J., & Kamber, M. (2007). *Data Mining Concepts and Techniques*.
- Markov, Z., & Daniel, T. (2007). *Uncovering Patterns in*.
- Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633.
- Santoso, Budi. 2007. *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
- Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36, 226–235.
- V, S. R. R., Somayajulu, D. V. L. N., & Dani, A. R. (2010). Classification of Movie Reviews Using Complemented Naive Bayesian Classifier, 1(4), 162–167.
- Ye, Q., Zhang, Z., & Law, R. (2009). Expert Systems with Applications Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems With Applications*, 36(3), 6527–6535.
- Yessenov, K. (2009). Sentiment Analysis of Movie Review Comments 6.863, 1–17.
- Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, 38(6), 7674–7682.