KLASIFIKASI TEKS MENGGUNAKAN *CHI SQUARE FEATURE SELECTION* UNTUK MENENTUKAN KOMIK BERDASARKAN PERIODE, MATERI DAN FISIKDENGAN ALGORITMA *NAIVEBAYES*

Siti Anisah, Anton Setiawan H, Asih Pujiastuti
Program Studi Informatika
Sekolah Tinggi Teknologi Adisutjipto Yogyakarta
informatika@stta.ac.id

Abstract

A comic has its own characteristics compared the other types of books. The difference between comic and other books can be seen from the category of period, material and physical. Comicand other booksneeded an application of classification system. Looking for the problem, classification system was made using Chi Square Feature Selection and Naive Bayes algorithm to determine the comic based on the period, material and physical. Delphi programming language and Oracle Database are used to build the Classification System. Chi Square Feature Selection acquired trait a comic is in 0.10347 and which not comic is in 1.9531. Furthermore, data is classified by the Naive Bayes algorithm. From 120 titles of comic that consists 60 titles of comic and non comic used to build classesfor trainand 60 titles of comic and non comic used to test. The results of Naive Bayesalgorithm for comic is 96,67%with 3.33% error rate, and non comic is 90% with 10% error rate. The classification to determine comic is good.

Keywords : Chi Square Feature Selection, Naive Bayes Algorithm, Comic, Period, Content, Physic.

1. Latar Belakang

Membaca merupakan suatu proses yang membangun pemahaman isi bacaan yang tertulis. Berbagai macam jenis buku bacaan dibuat dan dapat dikelompokkan berdasarkan isi buku, seperti majalah, buku cerita, koran, maupun komik. Komik didefinisikan sebagai tatanan gambar dan balon kata yang berurutan, dalam sebuah buku komik. Teknis dan struktur komik didefinisikan sebagai *sequential art*, susunan gambar dan kata-kata untuk menceritakan sesuatu atau mendramatisasi suatu ide.

Sebuah komik memiliki ciri tersendiri dibandingkan dengan jenis buku lain. Perbedaan komik dengan buku lain dapat dilihat dari kategori periode penerbitan, fisik ataupun *interface* komik, serta materi atau isi cerita komik. Penggemar komik dapat membeli komik di toko buku atau meminjam komik di rental komik maupun sebuah taman bacaan. Sebuah taman bacaan hendaknya dapat membedakan setiap jenis buku yang ada dan mengelompokkan buku-buku tersebut ke dalam kelasnya. Misalnya apakah sebuah buku merupakan komik atau bukan komik.

Dibutuhkan suatu aplikasi sistem klasifikasi yang dapat membantu *admin* suatu taman bacaan untuk mengetahui apakah sebuah buku termasuk jenis komik ataupun bukan komik berdasarkan periode, materi, fisik menggunakan *Chi SquareFeature Selection* dan algoritma *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan teknik prediksi berbasis probabilistik sederhana dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Secara umum, dasar ide *feature selection* yaitu mencari semua kemungkinan kombinasi dari atribut dalam data untuk menemukan subset dari feature yang terbaik untuk prediksi.

2. Landasan Teori

2.1 Algoritma *Naïve Bayes Classifier*

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasarkan pada penerapan teorema *Bayes* (atau aturan *Bayes*) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, *Naïve Bayes*, model yang digunakan adalah "model fitur independen".

Dalam *Bayes* (terutama *Naive Bayes*), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama.

Prediksi *Bayes* didasarkan pada teorema *Bayes* dengan formula umum sebagai berikut :

$$P(H|E) = \frac{P(E|H)xP(H)}{P(E)} \quad \text{..... (2.1)}$$

Penjelasan dari formula tersebut terdapat pada Tabel 1.

Tabel 1. Keterangan Formula *Naive Bayes*

Parameter	Keterangan
P(H|E)	Probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
P(E|H)	Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H.
P(H)	Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun.
P(E)	Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati.

2.2 *Chi Square Feature Selection*

Seleksi fitur (*feature selection)* dilakukan untuk mereduksi fitur-fitur yang tidak relevan dalam proses klasifikasi oleh *Naive Bayes*. Seleksi fitur *Chi Square* menggunakan teori statistikauntuk menguji independensi sebuah *term*dengan kategorinya. Salah satu tujuan penggunaan seleksi fitur adalah untuk menghilangkan fitur pengganggu dalam klasifikasi. Dalam seleksi fitur *Chi Square* berdasarkan teori statistika, dua peristiwa di antaranya adalah, kemunculan dari fitur dan kemunculan dari kategori, yang kemudian setiap nilai *term*diurutkan dari yang tertinggi. Uji *Chi Square* dalam statistika diterapkan untuk menguji independensi dari dua peristiwa.

Rumus :

$$X^2 = \Sigma \frac{(O-E)^2}{E} \quad \text{..... (2.2)}$$

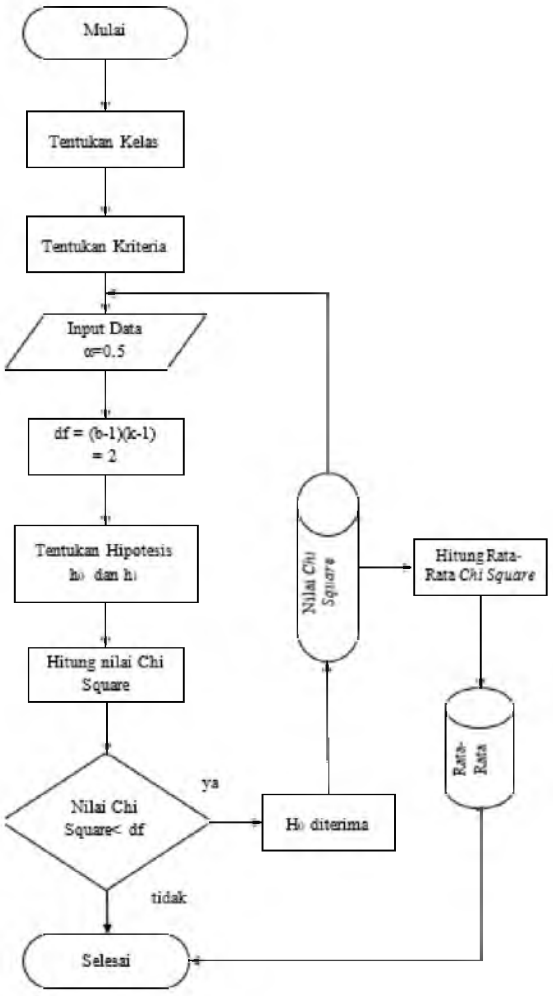
O : nilai Observasi (pengamatan)
E : nilai Expected (harapan)
Df = (b-1) (k-1)..... (2.3)
b : jumlah baris
k : jumlah kolom

3. Perancangan Sistem

3.1 *Flowchart* Seleksi Fitur

Flowchart untuk seleksi fitur yang menggunakan *Chi Square Feature Selection* ditunjukkan pada Gambar 1.

Langkah awal pada seleksi fitur menggunakan *Chi Square Feature Selection* sesuai Gambar 1 adalah menentukan kelas dan menentukan atribut yang nantinya akan membentuk sebuah tabel kontingensi. Tabel kontingensi berguna untuk menentukan nilai *Chi Square* yang berguna untuk penentuan kelas buku selanjutnya. Setelah tabel kontingensi terbentuk, *admin* dapat melakukan *input* data pada tabel kemudian menentukan nilai derajat kebebasan (df) dari tabel. Derajat kebebasan diperoleh dari jumlah baris kurang 1 dan jumlah kolom kurang 1 kemudian hasilnya dikalikan. Derajat kebebasan digunakan untuk membandingkan hasil *Chi Square* yang dihitung dengan tabel *Chi Square*.



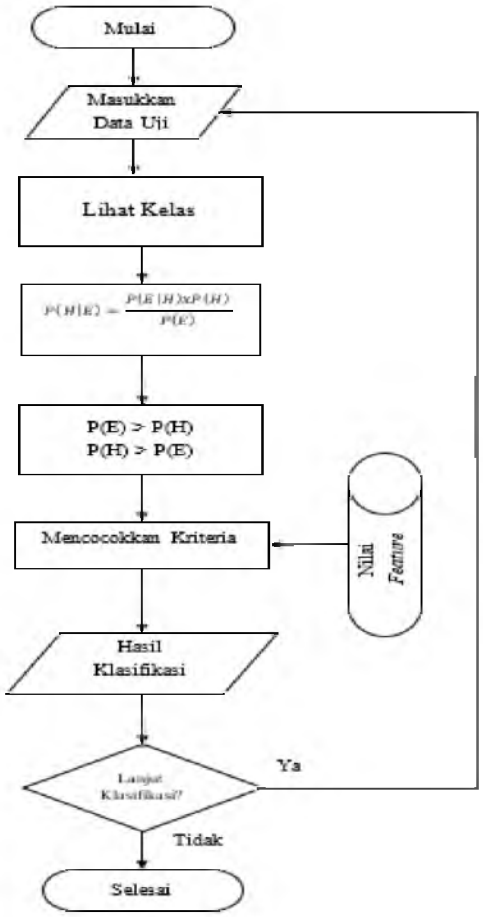
Gambar 1 *Flowchart* Seleksi Fitur dengan *Chi Square Feature Selection*

3.2 *Flowchart* Klasifikasi dengan *Naive Bayes*

Flowchart untuk klasifikasi dengan algoritma *Naive Bayes* ditunjukkan pada Gambar 2 berikut.

Langkah pertama pada proses klasifikasi dengan algoritma *Naive* Bayes sesuai Gambar 2 yaitu meng*input*kan data yang akan diklasifikasikan seperti judul buku beserta jumlah dari setiap kategori yaitu periode, materi dan fisik. Selanjutnya akan dicari kelas sesuaidata yang di*input*kan. Sistem akan menghitung nilai probabilitas setiap kelas. Probabilitas setiap kelas akan menentukan klasifikasi data dimana probabilitas terbesar akan diambil sebagai penentuan kelas apakah sebuah

data merupakan komik atau bukan. Apabila ingin memasukkan data buku lagi, maka kembali ke proses awal yaitu *input* data yang akan diuji.

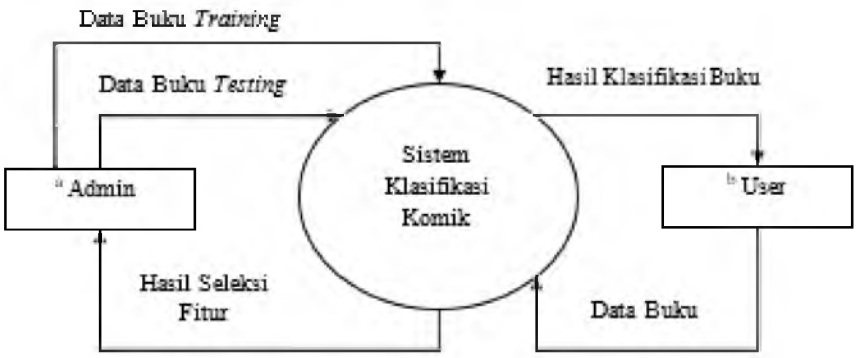


Gambar 2 *Flowchart* Proses Klasifikasi dengan Algoritma *Naive Bayes*

3.3 Diagram Alir Data

3.3.1 Diagram Konteks

Diagram konteks pada sistem klasifikasi komik berdasarkan periode, materi dan fisik menggunakan *Chi Square Feature Selection* dengan Algoritma *Naive Bayes* ditunjukkan pada Gambar 3.



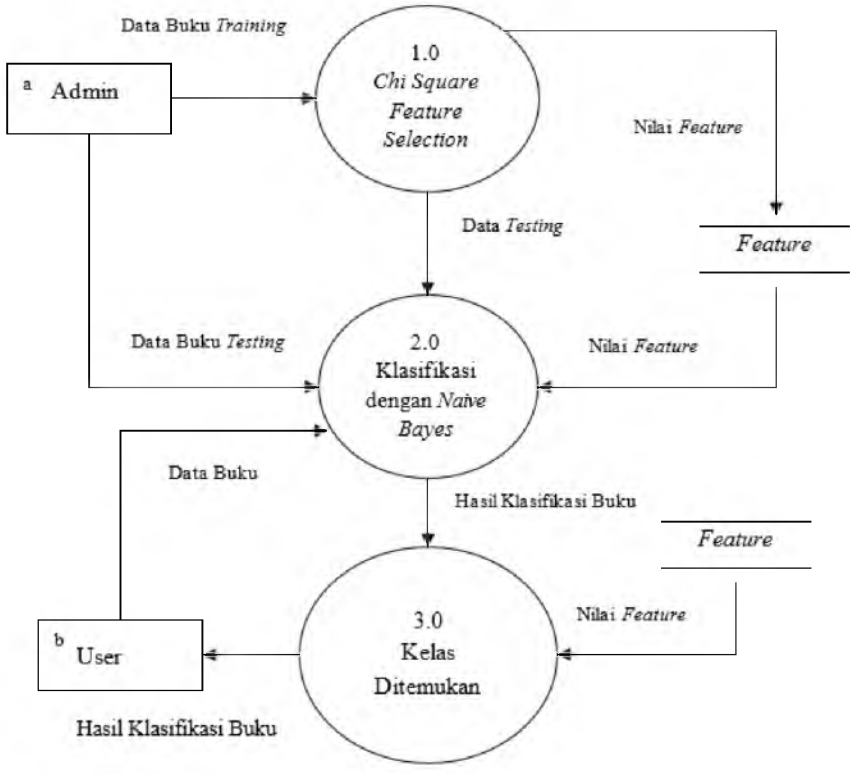
Gambar 3 Diagram Konteks Klasifikasi Komik

Berdasarkan diagram konteks pada Gambar 3 terdapat dua entitas yaitu *Admin* dan *User*. *Admin* meng*input*kan periode, materi dan fisik sebuah buku kemudian sistem akan melakukan perhitungan dengan *Chi Square* yang menentukan nilai masing-masing kategori.Dengan nilai tersebut*User* dapat mengklasifikasikan sebuah buku menjadi komik ataupun bukan komik berdasarkan nilai seleksi fitur dengan menggunakan algoritma *Naive Bayes.*

3.3.2 DAD Level 0

DAD Level 0 klasifikasi komik menggambarkan rincian proses-proses yang telah digambarkan pada Diagram konteks. Adapun DAD Level 0 klasifikasi komik dapat dilihat pada Gambar 4.

Tahap yang dilakukan oleh *Admin* untuk menentukan kelas yaitu melakukan *input* nilai periode, materi dan fisik kemudian menghitung nilai *Chi Square Feature Selection* masing-masing kelas dan mendapatkan ciri kelas (*feature*). Selanjutnya, *admin* dapat melakukan klasifikasi dengan *Naive Bayes* untuk menentukan kelas apakah sebuah buku merupakan komik atau bukan dengan melakukan *input* data buku. Hasil klasifikasi ditentukan oleh beberapa data *training* melalui proses *Feature Selection* sebelumnya. Dengan meng*input*kan data sebuah buku, maka *user* dapat menentukan apakah buku tersebut merupakan komik atau bukan komik.

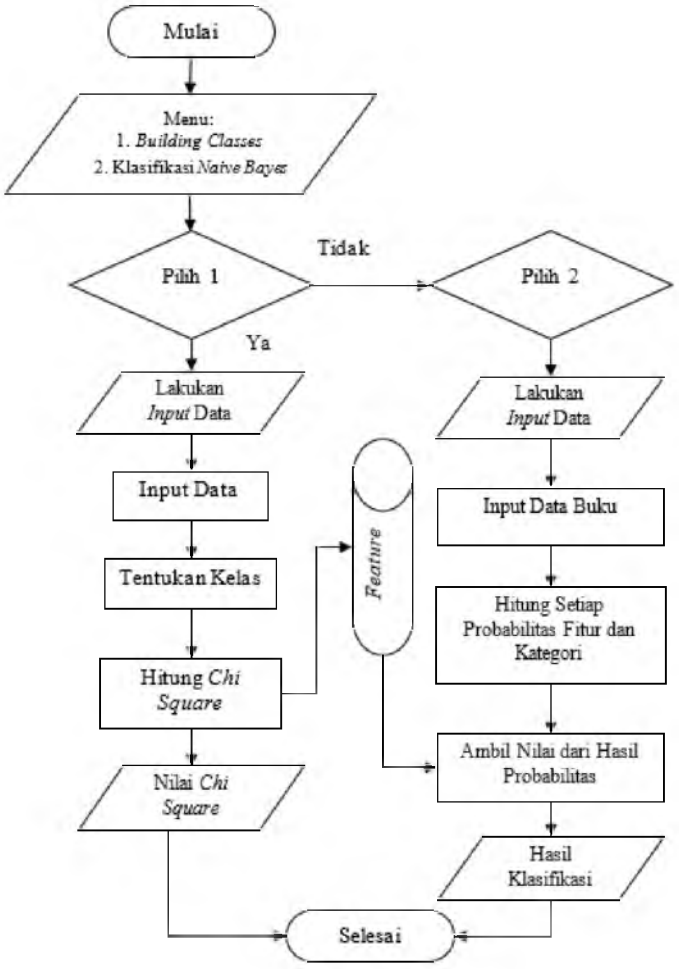


Gambar 4 DAD Level 0 Klasifikasi Komik

3.4 *Flowchart* Program

Untuk *flowchart* program kasifikasi berdasarkan periode, materi dan fisik menggunakan *Chi Square Feature Selection* dengan Algoritma *Naive Bayes* ditunjukkan pada Gambar 5. Dari Gambar 5 diketahui bahwa program terdiri dari dua menu utama yaitu *Building Classes* dan proses klasifikasi *Naive Bayes*. Apabila memilih menu *Feature Selection* maka akan masuk ke proses seleksi fitur. Data yang telah di*input*kan selanjutnya dihitung dengan memilih kelasnya dan

menghitung *Chi Square* dari data tersebut. Kemudian data yang telah dihitung akan disimpan untuk proses klasifikasi.



Gambar 3.8 *Flowchart* Program Klasifikasi

4. Implementasi dan Pengujian

4.1 Implementasi Sistem

Dalam implementasi suatu sistem dapat diketahui cara kerja suatu sistem yang dijalankan, apakah telah berjalan baik atau tidak. Untuk mengetahuinya, program ini dibangun dengan menggunakan bahasa pemrograman *Delphi* dan menggunakan *database oracle*. Ketika program dijalankan, tampilan pertama yaitu *form* Menu Utama yang memuat dua tombol, masing-masing tombol memiliki fungsi untuk ke *Form* selanjutnya. Tombol pertama yaitu tombol *Building Classes* yang digunakan apabila pengguna ingin membangun kelas dengan *Chi Square Feature Selection* dan tombol kedua yaitu tombol Klasifikasi*Naive Bayes* yang digunakan apabila pengguna ingin melakukan proses klasifikasi.

4.2 Pengujian Klasifikasi Teks Menggunakan *Chi Square Feature Selection* dan Algoritma *Naive Bayes*

Pengujian *Chi Square Feature* Selection dilakukan pada 60 judul buku dan pengujian Algortima *Naive Bayes* dilakukan pada 60 judul buku. Pengujian dilakukan berdasarkan nilai probabilitas dan nilai ekstraksi ciri. Hasil pengujian akan dilampirkan dengan tabel *confusion matrix.*

4.2.1 Tabel Ekstraksi Ciri

Berdasarkan perhitungan yang dilakukan oleh sistem, maka diperoleh nilai ekstraksi ciri masing-masing kelas yang disusun pada Tabel 2. Nilai ekstraksi ciri diperoleh dari nilai rata-rata (*average*) dari total nilai *Chi Square Feature Selection*.

Tabel 2 Tabel Ekstraksi Ciri

Komik	**Bukan Komik**
0,10347	1,95157

4.2.2 *Confusion Matrix* Hasil Pengujian

1) Kelas Komik

Pengujian terhadap 30 judul yang termasuk Komik dilampirkan menggunakan tabel *Confusion Matrix* seperti pada Tabel 3.

Tabel 3. *Confusion Matrix* kelas Komik

N=30	A	B
A	29	-
B	-	1

Keterangan :
- A : Komik
- B : Bukan Komik

Hasil Akurasi : $\frac{29}{30}\ x\ 100\% = 96{,}67\%$

Tingkat *Error* : $\frac{1}{30}\ x\ 100\% = 3{,}33\%$

2) Kelas Bukan Komik

Pengujian terhadap 30 judul yang termasuk Bukan Komik dilampirkan menggunakan tabel *Confusion Matrix* seperti pada Tabel 4.

Tabel 4. *Confusion Matrix* Kelas Bukan Komik

N=30	A	B
A	27	-
B	-	3

Keterangan :
- A : Komik
- B : Bukan Komik

Hasil Akurasi : $\frac{27}{30}\ x\ 100\% = 90\%$

Tingkat *Error* : $\frac{3}{30}\ x\ 100\% = 20\%$

5. Penutup

5.1 Kesimpulan

Dari hasil pengujian sistem klasifikasi teks menggunakan *Chi Square Feature Selection* dan Algoritma *Naive Bayes* untuk menentukan komik berdasarkan periode, materi dan fisik dapat diperoleh beberapa kesimpulan. Adapun kesimpulan yang diperoleh dari hasil pengujian yaitu :

1. Sistem klasifikasi dibangun menggunakan *tool* yang membantu dalam menganalisis sistem klasifikasi teks menggunakan *Chi Square Feature Selection* dan algoritma *Naive Bayes*. *Tool* dibangun menggunakan Bahasa pemrograman Delphi yang dapat membangun kelas-kelas untuk

proses klasifikasi, ekstraksi ciri dan melakukan klasifikasi (pengujian) menggunakan Algoritma *Naive Bayes*.

2. Berdasarkan hasil perhitungan dengan *Chi Square Feature Selection* diperoleh ciri buku yang merupakan komik yaitu dengan nilai 0,10347 dan yang bukan komik memiliki nilai 1,9531. Selanjutnya data diklasifikasi dengan algoritma *Naive Bayes*. Perhitungan sistem yang dilakukan pada 120 judul komik yang terdiri dari 60 judul komik dan bukan komik untuk membangun kelas yang merupakan tahap *training* serta 60 judul komik dan bukan komik untuk tahap *testing*, diperoleh hasil akurasi dari Algoritma *Naive Bayes* untuk kelas komik adalah 96,67% dengan tingkat *error* 3,33%, sedangkan untuk kelas bukan komik diperoleh hasil akurasi sebesar 90% dengan tingkat *error* 10% sehingga sistem klasifikasi untuk menentukan komik telah berjalan dengan baik.

5.2 Saran

Berdasarkan kesimpulan dan hasil peneltian yang didapatkan, maka penelitian ini dapat dikembangkan. Adapun saran untuk pengembangan selanjutnya, sebagai berikut :

1. Sistem klasifikasi teks menggunakan *Chi Square Feature Selection* dan Algoritma *Naive Bayes* dapat dikembangkan lagi dengan menggabungkan *Feature Selection* lain seperti *Mutual Information Gain* ataupun *Within Class Popularity* dan dikembangkan dengan algoritma selain *Naive Bayes* serta dengan membangun kelas yang lebih beragam, misalnya klasifikasi berdasarkan *Genre* dengan kelas romantis, misteri, atau petualangan.
2. Sistem dapat dikembangkan dalam penelitian selanjutnya untuk pengklasifikasian data yang lebih besar dan banyak seperti klasifikasi untuk data *mining*.

Daftar Pustaka

[1] Jogiyanto, Hartono. 2005. *Analisis dan Desain Sistem Informasi.* Yogyakarta : Penerbit Andi.

[2] Kadir, Abdul. 2009. *Dasar Perancangan dan Implementasi Database Relasional*. Yogyakarta : Andi Offset.

[3] Kadir, Abdul. 2005. *Pemrograman Database dengan Delphi 7 Menggunakan Access dan ADO*. Yogyakarta : Andi Offset.

[4] Nugroho, Bunafit. 2005. *Database Relational dengan MySQL.* Yogyakarta : Andi Offset.