MODEL KLASIFIKASI ABSTRAK SKRIPSI MENGGUNAKAN TEXT MINING UNTUK PENGKATEGORIAN SKRIPSI SESUAI BIDANG KAJIAN

Angga Cahyo Pradikdo
Fakultas Teknik, Program Studi Sistem Informasi
Universitas Nusantara PGRI Kediri
Email: anggacahyo567@gmail.com

Aidina Ristyawan
Fakultas Teknik, Program Studi Sistem Informasi
Universitas Nusantara PGRI Kediri
Email: ristykdr@gmail.com

ABSTRAK

Dengan melakukan observasi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, penulis mendapati bahwa dokumen skripsi pada Program Studi tersebut selalu bertambah setiap tahun, sehingga dapat dijadikan referensi pemilihan bidang penelitian yang sesuai untuk Mahasiswa Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri. Selain itu penulis juga pernah melakukan penelitian tentang pemodelan klasifikasi abstrak prosiding yang bisa digunakan untuk penyusunan letak skripsi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri. Dari hasil penelitian tersebut penulis mendapatkan saran tentang data yang digunakan. Saran tersebut berupa penggunaan data penelitian mahasiswa sebelumnya pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, supaya lebih tepat dan sesuai dengan studi kasusnya. Maka dari itu penulis terinspirasi untuk melakukan penelitian dengan menggunakan data penelitian mahasiswa Program Studi Sistem Informasi yang tersimpan di SIMKI (Sistem Informasi Manajemen Karya Ilmiah) Universitas Nusantara PGRI Kediri. Dengan memanfaatkan data penelitian mahasiswa sebelumnya serta metode teknik *text mining* diantaranya *prepocesing* dan *trasformation* dengan didukung dengan algoritma naive bayes sebagai proses untuk menghitung nilai probabilitas tertinggi sebagai proses klasifikasi yang akan digunakan untuk menguji data tersebut. Dari hasil pengujian 9 siklus menghasilkan pengetahuan bahwa siklus ke 1 merupakan siklus terbaik dengan akurasi 82,76%, yang dapat digunakan sebagai model klasifikasi skripsi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, untuk dapat membantu memudahkan mahasiswa untuk mencari referensi karena sudah memuat bidang kajian yang sesuai dan Program Studi Informasi mendapatkan model klasifikasi dengan data hasil dari skripsi mahasiswa Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri.

Kata kunci: *text mining; naive bayes classifier*; *confusion* matrik; skripsi.

ABSTRACT

By observing the Information Systems Study Program at Universitas Nusantara PGRI Kediri, the authors find that the thesis documents in the Study Program are always increasing every year, so that they can be used as a reference for the selection of research fields that are suitable for Information Systems Students of Nusantara PGRI University of Kediri.In addition, the author has also conducted research on proceedings abstract classification modeling that can be used for the preparation of the thesis location in the Information Systems Program of Universitas Nusantara PGRI Kediri. From the results of the study the authors get advice about the data used. The suggestion is in the form of the use of previous student research data on the Information Systems Program of Universitas Nusantara PGRI Kediri, so that it is more appropriate and in accordance with the case study. Therefore the author was inspired to conduct research using research data of Information Systems Study Program students stored in the SIMKI (Scientific Work Management Information System) Universitas Nusantara PGRI Kediri. By utilizing previous student research data and text mining techniques, prepocesing and transformation are supported by the Naive Bayes algorithm as a process to calculate the highest probability value as a classification process that will be used to test the data. From the results of testing 9 cycles produce knowledge that the first cycle is the best cycle with an accuracy of 82.76%, which can be used as a thesis classification model in the Information Systems Program Universitas Nusantara PGRI Kediri, to be able to help facilitate students to search for references because they contain fields an appropriate study and Information Study Program get a classification model with data from the students' Information Systems Study Program Universitas Nusantara PGRI Kediri.

Keywords: *text mining; naive bayes classifier; confusion matrix*; *thesis*.

1. PENDAHULUAN

Dokumen teks telah berkembang sangat pesat salah satunya ialah skripsi, dapat diartikan skripsi sebagai karya tulis oleh mahasiswa yang berdasarkan hasil penelitian lapangan dan studi kepustakaan yang disusun mahasiswa sesuai dengan bidang studinya sebagai tugas akhir dalam studi formalnya di perguruan tinggi.

Dengan melakukan observasi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, penulis mendapati bahwa dokumen skripsi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri selalu bertambah setiap tahun, sehingga dapat dijadikan referensi yang baik untuk Mahasiswa Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, karena keterbaruan data setiap tahun terjaga.

Pada penelitian sebelumnya dengan judul Model Klasifikasi Prosiding Berdasarkan Abstrak Untuk Penyusunan Letak Skripsi, yang telah penulis 1 presentasikan mendapatkan beberapa pertanyaan salah satunya, kenapa tidak menggunakan data dari Program Sistem Informasi Universitas Nusantara PGRI Kediri, supaya lebih tepat dan sesuai dengan studi kasusnya. Maka dari itu penulis akan melakukan penelitian dengan menggunakan data dari SIMKI (Sistem Informasi Manajemen Karya Ilmiah) Universitas Nusantara PGRI Kediri.

SIMKI (Sistem Informasi Manajemen Karya Ilmiah) Universitas Nusantara PGRI Kediri merupakan sistem publikasi untuk skripsi mahasiswa yang telah dijurnalkan pada Universitas Nusantara PGRI Kediri, salah satunya skripsi mahasiwa pada Prodi Sistem Informasi, tapi sangat disayangkan bahwa pada SIMKI Universitas Nusantara PGRI Kediri belum memuat bidang kajian yang telah diteliti oleh mahasiswa, hanya memuat nama Fakultas dan Program Studi peneliti yang ditampu tidak mengacu pada bidang kajian yang diteliti.

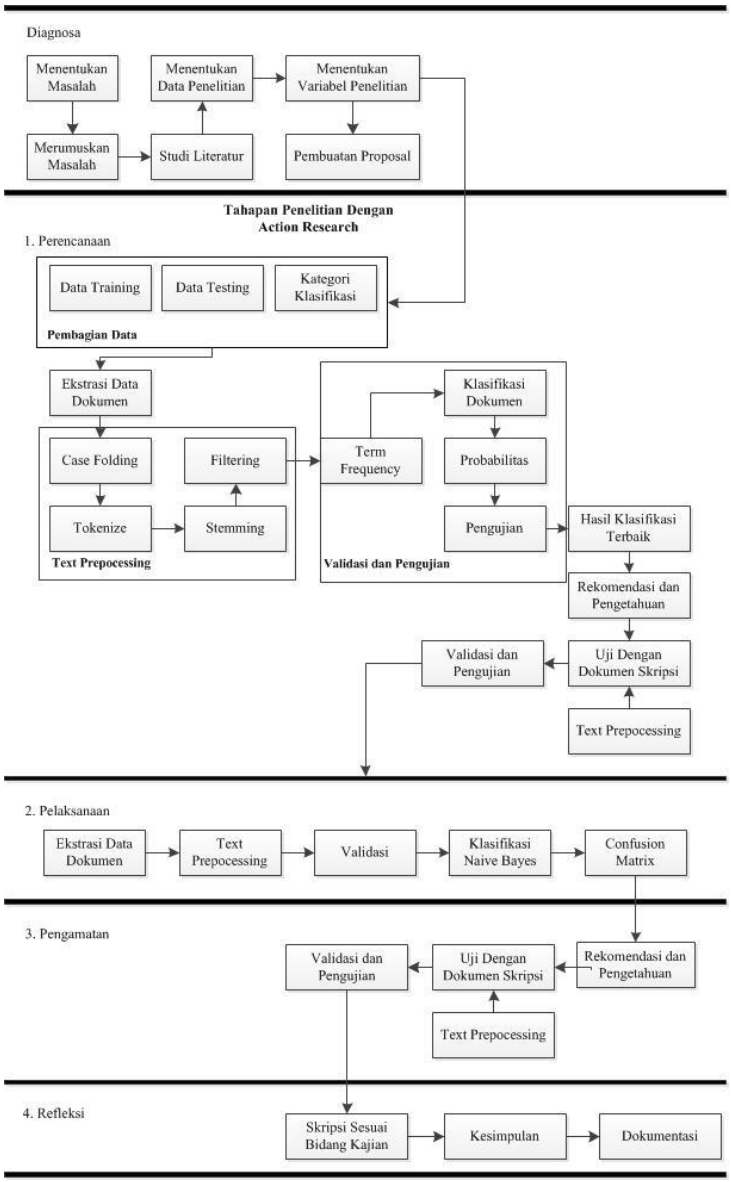
Maka dari itu penulis mendapatkan ide untuk membuat "Model Klasifikasi Abstrak Skripsi Menggunakan Text Mining Untuk Pengkategorian Skripsi Sesuai Bidang Kajian" dengan menggunakan data SIMKI Fakultas Teknik Program Studi Sistem Informasi pada tahun 2012-2013 yang telah dilabelkan menggunakan penelitian penulis sebelumnya dengan judul Model Klasifikasi Prosiding Berdasarkan Abstrak Untuk Penyusunan Letak Skripsi Menurut Bidang Kajian mendapatkan 3 model terbaik, salah satunya model terbaik berdasarkan pengujian *confusion matrik*, maka dari itu penulis menggunakan model ini sebagai teknik untuk membuat model klasifikasi untuk Program Studi Sistem Informasi Universitas Nusantara PGRI yang diharapkan dapat membantu memudahkan mahasiswa untuk mencari referensi karena sudah memuat bidang kajian yang sesuai dan Program Studi Informasi mendapatkan model klasifikasi dengan data hasil dari skripsi mahasiswa Program Studi Sistem Informasi Universitas Nusantara PGRI.

2. METODOLOGI PENELITIAN

2.1 Tahapan Membangun Model

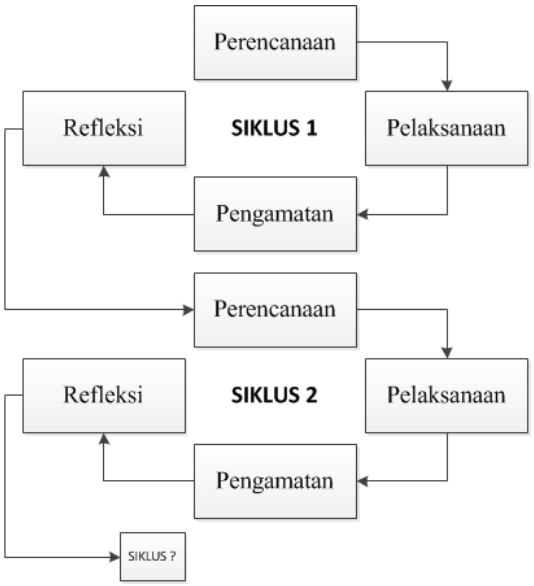
Jenis penelitian yang dikerjakan adalah penelitian *Action Research* dengan memahami dan mencatat pola yang ada. Secara metodologis tidak kuat tapi ada *knowladge* (pengetahuan) yang bisa digali dari situ, dalam penelitian menggunakan *action research* adapun siklus *action research* bisa di lihat di Gambar 2. *Siklus Action Research* : [1]

Adapun alur penelitian yang dikerjakan menggunakan model *action research* dengan 5 tahapan pengerjaan yaitu : Diagnosa, Perencanaan, Pelaksanaan, Pengamatan dan Refleksi.



Gambar 1. Alur Penelitian

Pada Gambar 1. Alur Penelitian, tahap pertama melakukan diagnosa terhadap masalah untuk menentukan data serta variabel yang digunakan dalam penelitian penulis menetapkan data dari SIMKI pada tahun 2012-2013 dan abstrak sebagai variabel penelitian, tahap kedua perencanaan untuk menentukan teknik yang dapat menyelesaikan masalah yang telah dirumuskan agar dapat terselesaikan dan mendapatkan suatu hasil atau pengetahuan penulis menetapkan teknik *sampling purposive* untuk membagi antara *data training* dan *data testing*, teknik *text prepocessing* meliputi *case folding*, *tokenize, Stemming, filtering* dan teknik validasi dan pengujian *confusion matrik*, tahap ketiga pelaksanaan, menguji data dari diagnosa dan teknik yang telah direncanakan, tahap keempat pengamata, mengamati setiap pengujian atau siklus yang penulis uji dan tahap kelima refleksi mengumpulkan semua hasil dari pengujian untuk penetapan hasil yang terbaik yang dapat diterapkan sesuai tujuan penulis.



Gambar 2. Siklus *Action Research*

Pada Gambar 2. terdapat beberapa tahapan yaitu : Perencanaan, Pelaksanaan, Pengamatan dan Refleksi yang akan digunakan untuk menguji setiap siklus yang ditentukan dalam penelitian.

2.2 Naive Bayes

Klasifikasi NBC dilakukan dengan cara mencari probabilitas dari kategori Vj dan kata-kata dalam dokumen dengan rumus : [2]

$$P(Vj) = \frac{docsj}{Contoh} \quad (1)$$

$$P(Wk \mid Vj) = \frac{nk + 1}{n + |vocabulary|} \quad (2)$$

$$Vmap = \text{argmax } vj \in vP(Vj) \prod i \, P(ai \mid Vj) \quad (3)$$

Keterangan rumus (1), (2), (3)

Vmap	: Nilai kemungkinan tertinggi dari seluruh anggota himpunan V
V_j	: Nilai kategori dari himpunan V
a_1-a_n	: Frekuensi kemunculan kata
P	: Peluang
P(B|A)	: Peluang B jika diketahui keadaan A
P(Vj)	: Peluang dari nilai kategori dari himpunan V
Docsj	: Jumlah dokumen yang memiliki kategori j dalam pelatihan
Contoh	: Jumlah dokumen yang digunakan
nk	: frekuensi munculnya kata wk dalam dokumen yang berkategori vj
n	: banyaknya keseluruh kata dalam dokumen berkategori vj
vocabulary	: banyaknya kata di dalam contoh pelatihan. (Hamzah, 2012)

2.3 Text Mining

Menurut [3] *text mining* adalah penemuan dan ekstraksi pengetahuan yang menarik dan tidak sepele dari teks bebas atau tidak terstruktur. Ini mencakup segala sesuatu mulai dari pengambilan informasi yaitu pengambilan dokumen atau pengambilan situs web) untuk klasifikasi dan pengelompokkan teks, untuk (agak lebih baru) entitas, relasi, dan ekstraksi peristiwa. Dengan beberapa tahapan sebagai berikut : [4]

1) *Pre processing*
   Tahapan dari pemrosesan awal dokumen yang dilakukan ialah :

a. *Case Folding* (Merubah huruf menjadi *lowercase*)
b. *Tokenize* (Merubah kalimat menjadi kata-kata serta menghilangkan tanda baca dan angka)

2) Tranformation
Tahapan setelah *Pre processing* menggubah data menjadi model yang dapat digunakan dalam penelitian.

a. *Stemming* (Merubah kata menjadi kata dasar sesuai Kamus Besar Bahasa Indonesia)
b. *Filtering* (Menyaring kata-kata penting hasil dari tokenize) tahap filtering dapat menggunakan teknik *Stopword* (Penghapusan kata penghubung seperti di, ke, dll) dan Token Filtering (Mengambil kata berdasarkan kecocokan antara keluaran kata dengan list yang digunakan).
c. Pembobotan kata dengan *Term Frequency*

2.4 Confusion Matrik

Confusion matrix melakukan pengujian untuk memperkirakan obyek yang benar dan salah.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Sensitivitas dan spesifitas tidak memberikan informasi untuk nilai diagnosa yang benar. Maka perlu adanya Precision untuk menghitung ketepatan antara informasi yang diminta dengan jawaban yang di berikan sistem dengan rumus :

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

Dan membutuhkan Recall untuk menilai tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi dengan rumus :

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 Refleksi

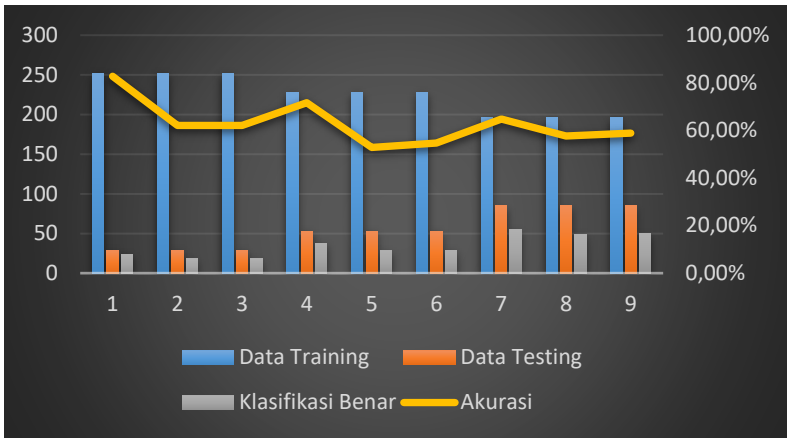
Tabel 1. Rincian hasil penelitian

Siklus ke	*Data Training*	*Data Testing*	*Klasifikasi Benar*	*Akurasi*	*Keterangan*
1	252	29	24	82,76%	Pada Penggunaan Token case menggunakan Kata Kunci dari data penelitian sebelumnya
2	252	29	18	62,07%	Pada Penggunaan Token case menggunakan Kata Kunci dari keseluruhan data sampel dari SIMKI tahun 2012-2013
3	252	29	18	62,07%	Pada Penggunaan Token case menggunakan Kata Kunci dari Gabungan penelitian sebelumnya keseluruhan data sampel dari SIMKI tahun 2012-2013
4	228	53	38	71,70%	teknik dan Filtering token case sama dengan siklus 1 hanya berbeda pada jumlah persentase yang digunakan
5	228	53	28	52,83%	teknik dan Filtering token case sama dengan siklus 2 hanya berbeda pada jumlah persentase yang digunakan

Siklus ke	*Data Training*	*Data Testing*	*Klasifikasi Benar*	*Akurasi*	*Keterangan*
6	228	53	29	54,72%	teknik dan Filtering token case sama dengan siklus 3 hanya berbeda pada jumlah persentase yang digunakan
7	196	85	55	64,71%	teknik dan Filtering token case sama dengan siklus 1 hanya berbeda pada jumlah persentase yang digunakan
8	196	85	49	57,65%	teknik dan Filtering token case sama dengan siklus 2 hanya berbeda pada jumlah persentase yang digunakan
9	196	85	50	58,82%	teknik dan Filtering token case sama dengan siklus 3 hanya berbeda pada jumlah persentase yang digunakan

Pada tabel 1. Rincian Hasil Penelitian, melakukan beberapa kali pengujian berdasarkan jumlah *data training* dan *data testing* yang telah di tentukan dengan memanfaatkan teknik *textmining* lebih jelasnya bisa dilihat pada keterangan pada tabel 1. Rincian Hasil Penelitian.

Berdasarkan pengujian dengan 9 siklus, peneliti menetapkan model klasifikasi terbaik dengan melihat beberapa kondisi pengetahuan yang dihasilkan dalam penelitian ini. Dengan melihat *barchart* sebagai berikut :

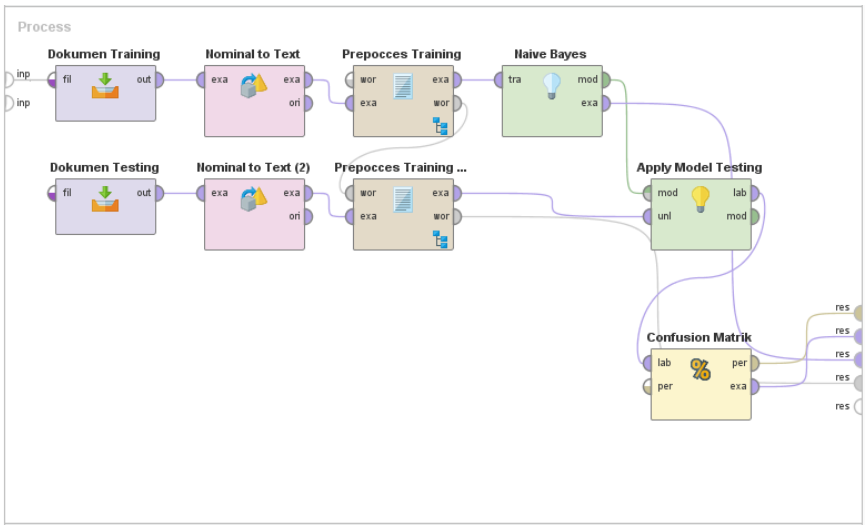


Gambar 3. *Chart* Hasil Penelitian

Hasil dari analisa pada Gambar 3. *Chart* Hasil Penelitian menggunakan teknik penelitian *action research* dengan menggunakan 9 siklus pengujian dan metode *textmining* dengan algoritma naive bayes mendapatkan pengetahuan bahwa siklus ke 1 merupakan siklus terbaik dengan akurasi 82,76% dan juga pada siklus ke 1 jarak antara *data testing* dan Jumlah klasifikasi benar tidak terlalu jauh.

3.2 Model Untuk pengkategorian Dokumen Skripsi

Pada pengujian dalam penelitian ini, untuk mencari model terbaik menggunakan software RapidMiner dengan desain model tampilan sebagai berikut :



Gambar 4. Model Rapidminer

Pada Gambar 4 penulis menjelaskan bahwa dalam penelitian ini menggunakan *Software* RapidMiner untuk memperoleh suatu pengetahuan, dengan memanfaatkan beberapa operator dari RapidMiner untuk menghasilkan model pengujian yang sesuai pada setiap siklus yang telah direncanakan, seperti *Read* Excel *(*dokumen *training)* untuk membaca data yang akan di *normalize* menggunakan *operator nominal to text*, untuk dapat di *prepocessing* menggunakan operator *process document from data* (*prepocessing training*) dengan menggunakan algoritma naive bayes untuk membentuk sebuah model klasifikasi yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada. Selanjutnya *Read* Excel (dokumen *testing)* untuk membaca data yang akan di uji untuk mengetahui kelasnya dengan operator *apply model* sebagai *validation set* untuk mencari parameter yang paling baik untuk setiap *data testing* yang akan di uji tingkat akurasinya menggunakan operator *Confusion Matrik.*

Gambar 5. Model *Prepocessing*

Pada Gambar 5. Penulis menggunakan beberapa operator untuk melakukan *prepocessing* atau pemrosesan awal, antara lain *Tokenize* (Merubah kalimat menjadi kata-kata serta menghilangkan tanda baca dan angka), *Transform Cases* (merubah hasil dari *tokenize* ke *lowercase*, *Stem* indonesia (Merubah kata menjadi kata dasar sesuai Kamus Besar Bahasa Indonesia), *filter token* (Mengambil kata berdasarkan kecocokan antara keluaran kata dengan list yang digunakan).

4. KESIMPULAN

Dengan memanfaatkan data SIMKI Universitas Nusantara PGRI Kediri dan teknik *text mining* diantaranya *prepocesing* dan *trasformation* dengan didukung dengan algoritma naive bayes sebagai proses untuk menghitung nilai probabilitas tertinggi sebagai proses klasifikasi yang akan digunakan untuk menguji, dari hasil pengujian 9 siklus menghasilkan pengetahuan bahwa siklus ke 1 merupakan siklus terbaik dengan akurasi 82,76%, yang dapat digunakan sebagai model klasifikasi pada Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri, untuk dapat membantu memudahkan mahasiswa untuk mencari referensi karena sudah memuat bidang kajian yang sesuai dan Program Studi Informasi mendapatkan model klasifikasi dengan data hasil dari skripsi mahasiswa Program Studi Sistem Informasi Universitas Nusantara PGRI Kediri.

DAFTAR PUSTAKA

[1] Z. A. Hasibuan, *Metodologi Penelitian Pada Bidang Ilmu Komputer dan Teknologi Informasi*. Depok: Fasilkom Universitas Indonesia, 2007.
[2] A. Hamzah, "KLASIFIKASI TEKS DENGAN NAÏVE BAYES CLASSIFIER (NBC) UNTUK PENGELOMPOKAN TEKS BERITA DAN ABSTRACT AKADEMIS," *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, vol. Periode III, hlm. 269–277, 2012.
[3] A. Kao dan S. R. Poteet, *Natural Language Processing and Text Mining*. Washington: Springer, 2007.
[4] R. Feldman dan J. Sanger, *The Text Mining HandBook Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2007.