Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain

Indah Maulida[1], Addy Suyatno[2], Heliza Rahmania Hatta[3]
Universitas Mulawarman, FKTI
[1,2,3]Jurusan Ilmu Komputer, Universitas Mulawarman
[1]maulida.ineff@gmail.com, [2]addysuyatno@yahoo.com, [3]heliza_rahmania@yahoo.com

Abstrak

Klasifikasi dapat diterapkan di semua bidang kehidupan termasuk dalam teks. Algoritma klasifikasi menggunakan semua fitur yang terdapat pada data untuk membangun sebuah model, padahal tidak semua fitur tersebut sesuai terhadap hasil klasifikasi. Seleksi fitur adalah teknik untuk memilih fitur penting dan relevan terhadap data dan mengurangi fitur yang tidak relevan. Seleksi fitur bertujuan untuk memilih fitur terbaik dari suatu kumpulan data fitur. Tujuan dari penelitian ini adalah menerapkan metode Information Gain dalam sistem seleksi fitur untuk dokumen teks berbahasa Indonesia. Metode Information Gain adalah metode yang menggunakan teknik scoring untuk pembobotan sebuah fitur dengan menggunakan maksimal entropy. Fitur yang dipilih adalah fitur dengan nilai Information Gain yang lebih besar atau sama dengan nilai threshold tertentu. Nilai threshold yang digunakan yaitu 0,02; 0,05 dan 0,07. Data yang digunakan adalah sekumpulan dokumen abstrak skripsi. Dari pengujian menggunakan 21 data, sistem dapat mereduksi fitur sebanyak 89% menggunakan threshold 0,07. Penelitian ini menghasilkan aplikasi yang dapat mengurangi dimensi fitur dan memilih fitur terbaik di dalam dokumen teks bahasa Indonesia.

Kata kunci*— seleksi fitur, fitur, information gain, entropy*

Abstract

Classification can be applied in all areas of life included in the text. The classification algorithm using all the features found on the data to build a model, though not all of these features are relevant to the classification results. Feature selection techniques is a technique to choose which feature is important and relevant to the data and reducing the irrelevant features. Feature Selection aims to select the best features of a data set of features. The purpose of this research is to apply the Information Gain method in the feature selection system for Indonesian language text documents. Information Gain method is a method that using scoring techniques for weighting a feature by using a maximum entropy. Selected features is featured with Information Gain values greater than or equal to a certain threshold value. The threshold value used is 0.02; 0.05 and 0.07. The data used is the thesis abstract collection of documents. From testing using 21 data, the system can reduce the features as much as 89% using a threshold of 0.07. This research resulted in an application that can reduce the dimensions of feature and choose the best features in the Indonesian language text documents.

Keywords*— feature selection, selection, features, information gain, entropy*

1. PENDAHULUAN

Kemudahan pencarian informasi merupakan dampak dari kemajuan teknologi. Data diolah sedemikian rupa agar lebih mudah ditemukan. Pengklasifikasian dilakukan agar data terorganisir sehingga menghasilkan informasi yang lebih baik. Algoritma klasifikasi menggunakan semua fitur yang terdapat pada data untuk membangun sebuah model [1]. Pada uraian ini, fitur yang dimaksud yaitu kata yang telah diekstrak menjadi kata dasar. Semakin banyak fitur, semakin besar kemungkinan terdapat

banyak fitur tidak relevan di dalam dokumen, sehingga mempengaruhi kinerja algoritma klasifikasi dari segi waktu dan keakuratan hasil klasifikasi.

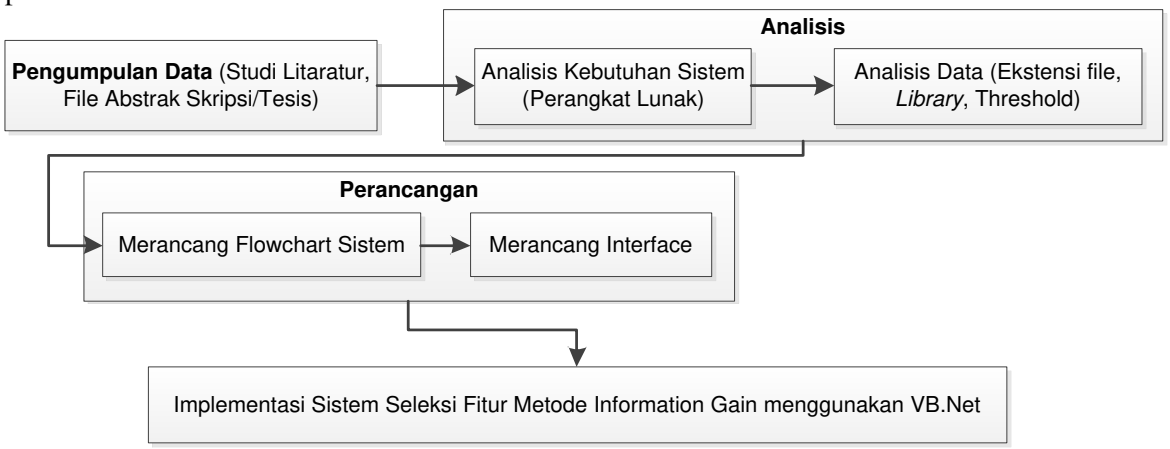
Untuk mencegah situasi ini, fitur yang diekstrak harus di*filter* sebelum fase klasifikasi. Karena itu diperlukan suatu metode untuk memilih fitur penting yang mewakili dokumen dan dapat mengurangi dimensi fitur karena dapat meningkatkan kinerja klasifikasi [2]. *Feature selection* merupakan teknik reduksi dimensi yang digunakan untuk memperkecil matriks data dengan memperhatikan informasi kata penting yang perlu diproses [3]. *Information Gain* merupakan salah satu algoritme seleksi fitur yang digunakan untuk memilih fitur terbaik. *Information Gain* dimanfaatkan untuk merangking kata-kata penting dari hasil reduksi fitur. Hasil dari proses *Information Gain* adalah kata penting yang bersifat informatif [3]. Metode IG dapat melihat setiap fitur untuk memprediksi label kelas yang benar yang karena memilih nilai yang tertinggi dan lebih efektif untuk mengoptimalkan hasil klasifikasi [2]. Yang dan Pedersen menggunakan teknik reduksi dimensi *Document Frequency* (DF) *thresholding*, *Information Gain* (IG), *Mutual Information* (MI), *Chi square,* dan *Term Strengt* (TS). Hasil percobaan menunjukkan bahwa akurasi IG paling bagus [3].

Dari hasil penelitian yang telah dilakukan oleh Sari (2013) [3] diperoleh matriks awal sebanyak 6.136 kata dengan hasil reduksi mengguakan *Information Gain* mencapai 1.219 kata [3]. Maka disimpulkan penggunaan seleksi fitur dapat mengurangi banyak fitur yang tidak mewakili dokumen dan menghasilkan fitur yang bersifat penting. Oleh sebab itu, seleksi fitur ini perlu dilakukan karena berfungsi untuk memilih fitur yang bersifat penting sehingga dapat meningkatkan kinerja algoritma klasifikasi. Penulis menggunakan *Information Gain* sebagai metode seleksi fitur karena memilih fitur dengan nilai yang tertinggi dan bersifat informatif. Berdasarkan uraian tersebut, dapat dirumuskan masalah yaitu bagaimana menerapkan metode *Information Gain* ke dalam sebuah sistem seleksi fitur pada dokumen teks Bahasa Indonesia.

Sistem ini dapat memberikan kemudahan kepada pengguna yang akan melakukan seleksi fitur pada dokumen teks berbahasa Indonesia, dapat membantu mengurangi fitur untuk meningkatkan kinerja algoritma klasifikasi serta dapat mengetahui fitur terbaik yang diperoleh dari *Information Gain* dan jumlah pengurangan fitur.

2. METODOLOGI PENELITIAN

Metode penelitian yang dilakukan seperti yang terlihat pada Gambar 1. Penulis melakukan metode studi literatur untuk memperoleh data dan informasi yang berhubungan dengan penelitian. Penulis menggunakan perangkat lunak VB.Net untuk pembuatan *interface* sekaligus *source code* sistem, Ms. Access untuk pembuatan *database* dan Ms. Word sebagai pengolah dan media dokumen untuk data input sistem.



Gambar 1. Kerangka Penelitian

2.1. Studi Literatur

Mengumpulkan data dan informasi yang berhubungan dengan penelitian dari beberapa buku, prosiding dan jurnal sebagai penunjang pada proses penelitian agar tidak menyimpang dari ketentuan yang telah ada sebelumnya.

2.1.1 Seleksi Fitur

Seleksi fitur adalah salah satu teknik terpenting dan sering digunakan dalam *pre-processing*. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi [5]. Tujuan utama dari seleksi fitur ialah memilih fitur terbaik dari suatu kumpulan fitur data [6].

2.1.2 *Preprocessing*

Preprocessing dapat didefinisikan sebagai proses pengontrolan ukuran daftar kata-kata yang dalam hal ini berupa jumlah kata-kata berbeda yang digunakan sebagai indeks *term.* Pengaturan ukuran daftar kata diharapkan dapat meningkatkan performa penemuan kembali informasi. *Preprocessing* juga bertujuan untuk menyaring kata-kata yang dianggap paling menonjol dari sebuah dokumen. Kata-kata konten seperti kata benda, kata kerja dan kata sifat merupakan sebagian besar pembawa semantik dari sebuah dokumen. Sementara itu, kata-kata fungsi seperti preposisi, kata ganti dan konjungsi dapat ditemukan di seluruh dokumen dan memiliki peran kecil dalam menentukan konten dari sebuah dokumen [7]. Penelitian ini menggunakan tiga tahap untuk *preprocessing*, yaitu *tokenization*, *stopword*, dan *stemming* [8]. *Tokenizing* adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri masing-masing [9]. Tahap *filtering* atau *stopword*, kata yang tidak relevan dalam penentuan topik sebuah dokumen akan dihilangkan, misal kata "adalah", "dari", "atau" [8]. *Stemming* adalah proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen [9].

2.1.3 *Information Gain*

Information Gain merupakan teknik seleksi fitur yang memakai metode *scoring* untuk nominal ataupun pembobotan atribut kontinu yang didiskretkan menggunakan maksimal entropy. Suatu entropy digunakan untuk mendefinisikan nilai *Information Gain*. Entropy menggambarkan banyaknya informasi yang dibutuhkan untuk mengkodekan suatu kelas [6]. *Information Gain* (IG) dari suatu *term* diukur dengan menghitung jumlah bit informasi yang diambil dari prediksi kategori dengan ada atau tidaknya *term* dalam suatu dokumen. Secara matematis dituliskan dengan [10] :

$$InfoGain\ (S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{S}\ Entropy(S_v) \quad (1)$$

$$Entropy(S) = - \sum \frac{|S_i|}{S} \log \frac{S_i}{S} \quad (2)$$

Dimana S adalah jumlah seluruh fitur, A adalah kategori, S_v adalah jumlah sampel untuk nilai v, v adalah nilai yang mungkin untuk kategori A, S_i adalah fitur ke i dan $Value(A)$ adalah himpunan nilai-nilai yang mungkin untuk kategori A.

Fitur yang dipilih adalah fitur dengan nilai *Information Gain* yang tidak sama dengan nol dan lebih besar dari suatu nilai *threshold* tertentu. Ide dibalik *Information Gain* untuk memilih fitur adalah menyatakan fitur dengan informasi yang paling signifikan terhadap kategori [10].

2.2. Tahap Analisis

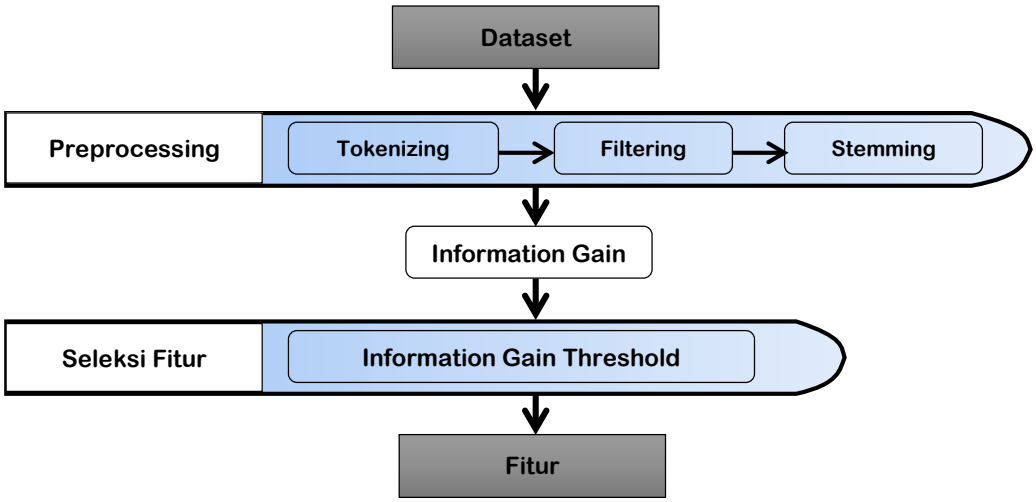
Penulis melakukan analisis data yang diperlukan oleh sistem. Hasil analisis untuk menunjang pembangunan sistem, yaitu:

1. Dokumen teks abstrak skripsi atau tesis sebagai data *input* yang berekstensi *.doc dan *.docx.
2. Sistem menggunakan 2 prediksi kategori yaitu Ilmu Komputer dan Biologi.

3. Data yang digunakan sebanyak 70 dokumen yang terdiri dari dokumen *training* sebanyak 49 (Ilmu Komputer 25 dokumen, Biologi 24 dokumen) dan dokumen *testing* sebanyak 21 (Ilmu Komputer 11 dokumen, Biologi 10 dokumen).
4. *Threshold Information Gain* yang digunakan yaitu: 0,02; 0,05 dan 0,07.
5. Dibutuhkan *library* dalam sistem untuk menampilkan isi dokumen *input*. *Library* tersebut menggunakan *spiredoc*.
6. Dibutuhkan data *stopword* di dalam sistem untuk mem*filter* fitur. Data *stopword* diperoleh dari penelitian Sarawati, dkk (2015) [11]. Contoh kata yang di*filter* dalam sistem yaitu: adanya, akan, bagaimana, hanya, jadi, karena, lanjut, maka, namun, oleh, pada, rata, sebab, tanpa, untuk, yang.

2.3. Tahap Perancangan

Ditahap ini penulis merancang alur proses sistem yang akan dibuat. Perancangan sistem dapat dilihat pada gambar 2.



Gambar 2. Rancangan Proses Sistem

Dari gambar 2, dataset adalah dokumen abstrak yang di*input* ke dalam sistem. Dataset memasuki sistem *preprocessing* yang terdiri dari *tokenizing*, *filtering* dan *stemming* untuk menghasilkan sekumpulan fitur. Fitur hasil *preprocessing* kemudian masuk ke sistem pembobotan menggunakan *Information Gain*. Setelah memiliki nilai bobot, fitur akan diseleksi dengan menggunakan *threshold* sehingga menghasilkan output berupa fitur terbaik.

Setelah merancang alur sistem, penulis merancang basis data yang akan digunakan kemudian merancang antarmuka sistem yang akan penulis buat. Antarmuka dirancang secara sederhana untuk memudahkan implementasi sistem. Pada tahap ini penulis merancang antarmuka sistem menjadi 9 bagian yaitu form Halaman Utama, form *Learning Process*, form *Feature Selection*, form Data Fitur, form Dokumen, form Hasil Seleksi, form Kamus dan form Info.

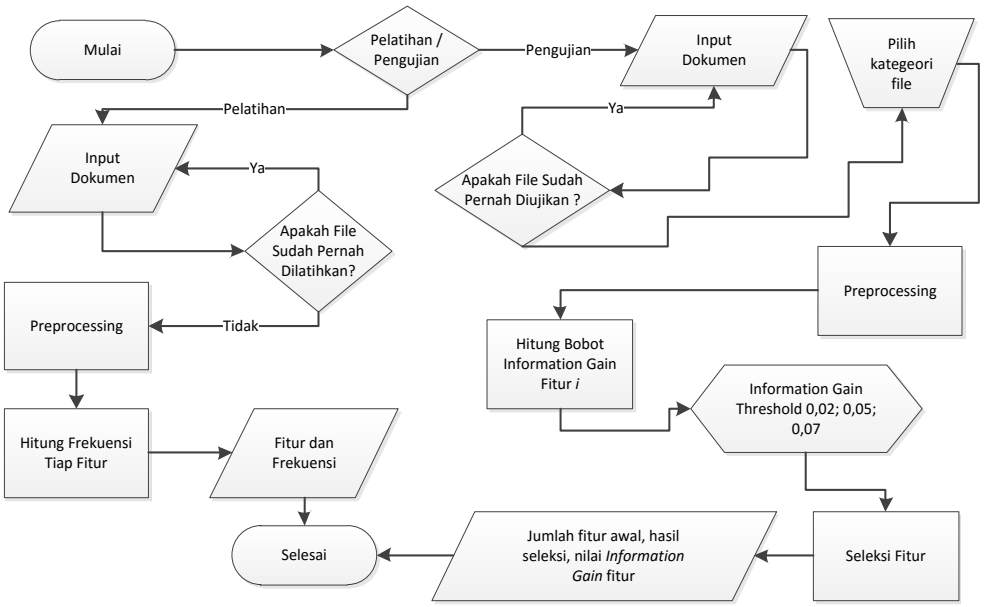
2.4. Tahap Implementasi

Tahap dimana peneliti merealisasikan rancangan sistem dan design sistem ke dalam baris-baris kode assembly serta membuat design form sesuai dengan perancangan. Peneliti membangun sistem dengan menggunakan Visual Basic 2010 dan Access 2007 sebagai database.

3. HASIL DAN PEMBAHASAN

3.1. Perancangan Sistem

Penulis merancang sistem seleksi fitur menjadi 2 sistem, yaitu sistem pelatihan dan sistem pengujian. Pada sistem pelatihan penulis hanya membuat proses *preprocessing*, proses menyimpan data dokumen dan proses menyimpan data hasil dari *preprocessing* berupa fitur dan frekuensi. Pada sistem pengujian penulis membuat proses *preprocessing*, proses perhitungan nilai bobot *Information Gain*, proses seleksi fitur dan proses menyimpan data. Gambar 3 menunjukkan alur sistem. Data yang sudah pernah diinputkan ke dalam sistem tidak akan bisa diinputkan kembali.



Gambar 3. Alur Proses Sistem

3.2. Analisis Hasil Simulasi

3.2.1 Form Learning Process

Halaman *learning process* merupakan halaman untuk sistem pelatihan dokumen. Pelatihan dokumen menggunakan abstrak skripsi Ilmu Komputer sebanyak 25 data dan Biologi 24 data. Dokumen abstrak tersebut akan diproses sistem *preprocessing* yang terdiri dari 3 tahap yaitu *tokenizing*, *stopword* dan *stemming*. Setelah proses stemming selesai, sistem akan menghitung frekuensi dari tiap fitur yang diperoleh. Daftar fitur dan frekuensinya akan ditampilkan setelah proses pelatihan selesai. Pelatihan dokumen ini berfungsi untuk memperoleh fitur beserta frekuensinya. Hasil proses pelatihan kemudian disimpan ke *database* untuk digunakan pada sistem pengujian. Implementasi form *learning process* dapat dilihat pada Gambar 4. Langkah penggunaannya yaitu pengguna mengklik tombol bergambar folder untuk meng*input* dokumen yang akan dilatihkan, lalu mengklik tombol proses agar sistem melakukan proses *preprocessing*. Setelah proses selesai, sistem akan menampilkan hasil berupa daftar fitur beserta frekuensinya seperti yang terlihat pada Gambar 5. Selanjutnya pengguna memilih kategori sesuai sumber dokumen lalu mengklik tombol simpan.

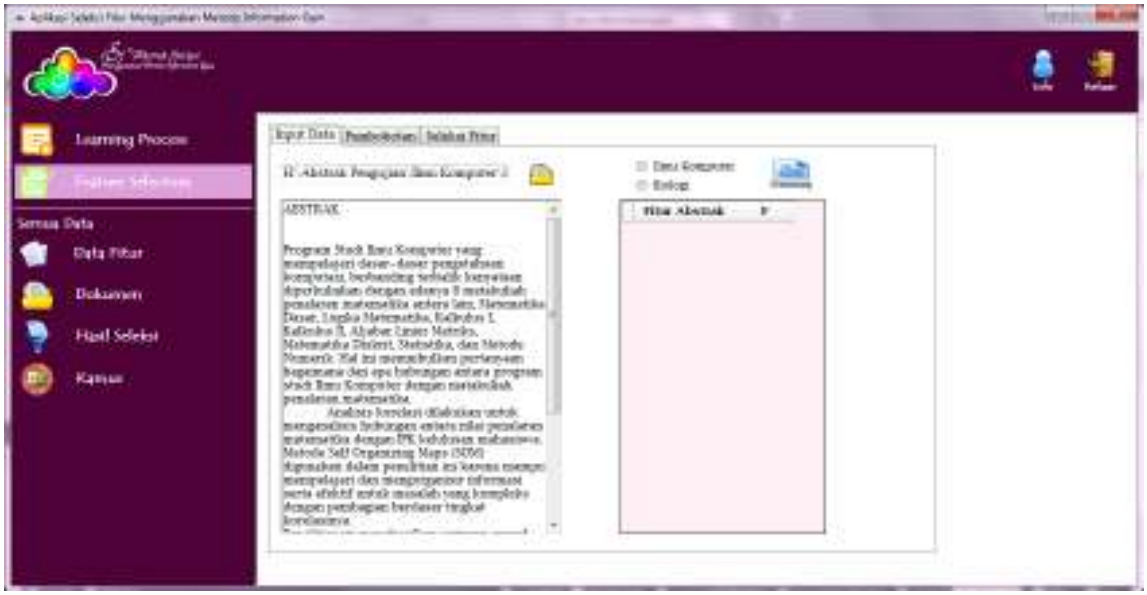
Gambar 4. Implementasi Learning Process

Fitur	Frekuensi
individu	6
jenis	5
konservasi	5
hutan	4
kelelawar	4
kaltim	4
famili	4
rea	4
identifikasi	4
plantations	3
pt	3
ve	vespertilionidae
rivoula	2

Gambar 5. Hasil Pelatihan Dokumen

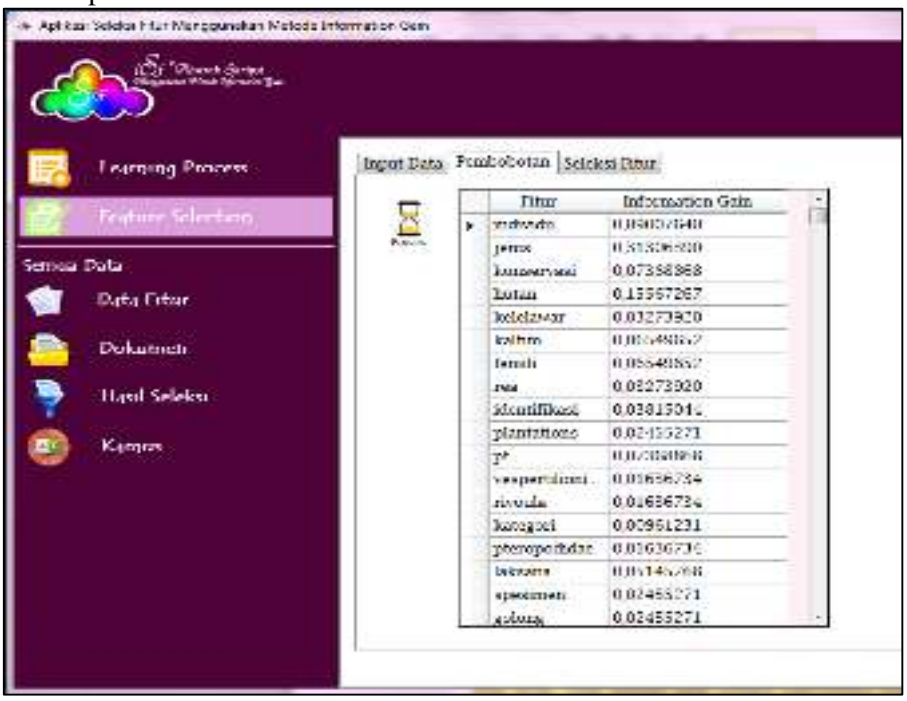
3.2.2 Form Feature Selection

Halaman *feature selection* merupakan halaman untuk sistem pengujian dokumen dimana proses seleksi fitur dilakukan. Pengguna dapat memilih dokumen yang akan diujikan dengan mengklik tombol bergambar folder, memilih kategori dokumen sesuai dari sumbernya lalu mengklik tombol "stemming" maka sistem akan melakukan proses *preprocessing*. Setelah *preprocessing* selesai, sistem akan menampilkan daftar fitur dokumen serta frekuensinya dan meng*update* data fitur di *database*. Implementasi rancangan form *feature selection* tab *page input* data dapat dilihat pada Gambar 6.



Gambar 6. Implementasi Form Feature Selection Tab Input Data

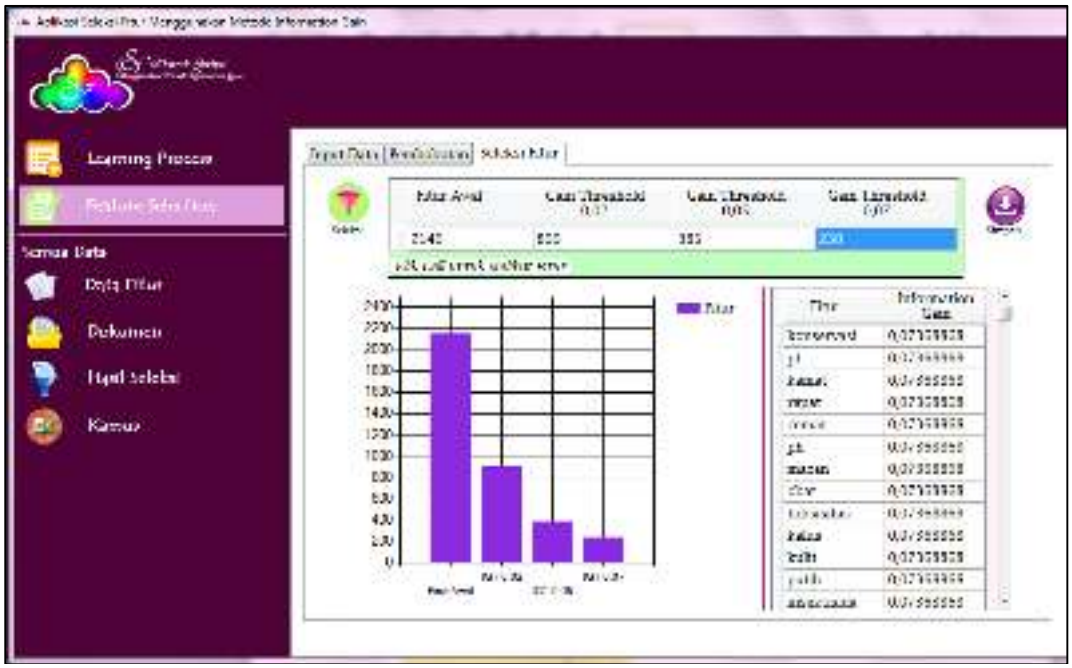
Pengguna dapat memilih *tab page* pembobotan kemudian mengklik tombol proses agar sistem dapat memulai proses menghitung nilai bobot setiap fitur dengan menggunakan *Information Gain*. Setelah proses selesai, sistem akan menampilkan hasil nilai bobot *Information Gain* untuk setiap fitur seperti yang terlihat pada Gambar 7.



Gambar 7. Sistem Menampilkan Fitur Beserta Nilai Bobotnya

Pengguna masuk ke *tab page* seleksi fitur untuk proses seleksi fitur. Fitur dengan nilai bobot ≥ *threshold* merupakan fitur terbaik yang akan dijumlahkan oleh sistem. Proses pertama, sistem menyeleksi fitur dengan *Information Gain Threshold* 0,02. Proses kedua, sistem menggunakan *Threshold* 0,05. Proses ketiga, sistem menyeleksi fitur dengan *Threshold* 0,07. Fitur yang memiliki nilai bobot di bawah ketiga nilai *threshold* tersebut tidak akan digunakan.

Hasil seleksi ditampilkan bersama dengan jumlah fitur awal seperti pada Gambar 8. Nilai Information Gain fitur dapat dilihat dengan mengklik *cell* data hasil dan sistem akan menampilkan data fitur berdasarkan *cell* yang diklik. Daftar fitur yang ditampilkan merupakan fitur terbaik yang diperoleh.



Gambar 8. Hasil Proses Seleksi Fitur

Tabel 1. Perubahan Jumlah Fitur Sebelum dan Sesudah Seleksi Menggunakan Information Gain Threshold (IGT)

Banyak Data	Atribut	Banyak Fitur
21 data	Fitur Awal	2146
	IGT 0,02	899
	IGT 0,05	385
	IGT 0,07	230

Threshold 0,02 $= \frac{Jumlah\ Fitur\ Awal - Jumlah\ Fitur\ Terseleksi}{Jumlah\ Fitur\ Awal} \times 100\%$

$= \frac{2146 - 899}{2146} \times 100\%$

$= 58\%$

Threshold 0,05 $= \frac{Jumlah\ Fitur\ Awal - Jumlah\ Fitur\ Terseleksi}{Jumlah\ Fitur\ Awal} \times 100\%$

$= \frac{2146 - 385}{2146} \times 100\%$

$= 82\%$

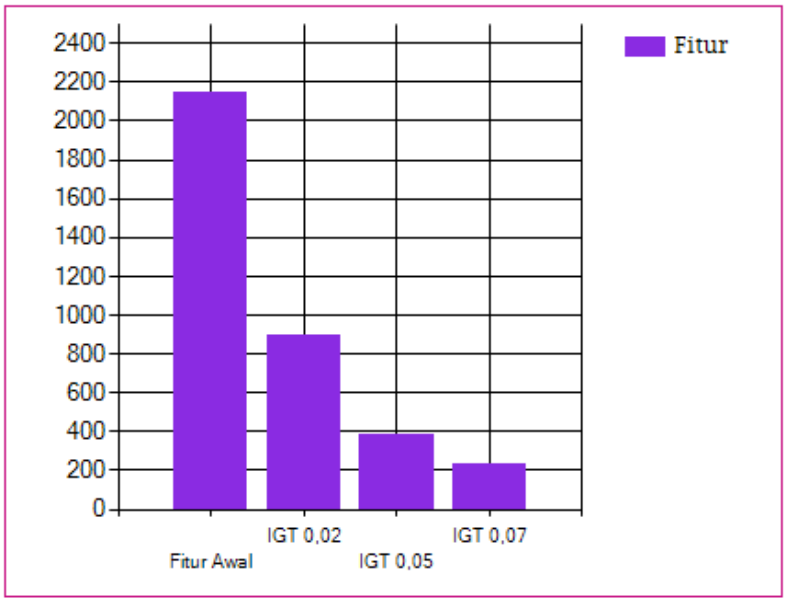
Threshold 0,07 $= \frac{Jumlah\ Fitur\ Awal - Jumlah\ Fitur\ Terseleksi}{Jumlah\ Fitur\ Awal} \times 100\%$

$= \frac{2146 - 230}{2146} \times 100\%$

$= 89\%$

Dari tabel 1 dapat dilihat bahwa sistem menghasilkan sebanyak 2146 fitur awal. Dari pengujian menggunakan 21 dokumen, reduksi *Information Gain* dengan *threshold* 0,02 menghasilkan 899 fitur terbaik dengan reduksi sebesar 58% dari fitur awal. Reduksi dengan *threshold* 0,05 sebesar 82% menghasilkan 385 fitur terbaik. Persentase reduksi fitur terbesar dihasilkan oleh *Information Gain*

dengan *threshold* 0,07 yang mencapai 89% dari fitur awal dengan hasil 230 fitur terbaik. Pengujian tersebut menunjukkan hasil bahwa reduksi dengan *threshold* 0,07 dapat mengurangi lebih banyak fitur yang kurang bersifat penting. Perbedaan jumlah hasil fitur tiap *threshold Information Gain* dapat dilihat melalui grafik pada gambar 9.



Gambar 9. Grafik Perbedaan Jumlah Fitur yang Dihasilkan Metode Information Gain

Semakin besar nilai *Information Gain threshold* yang digunakan, semakin tinggi persentase reduksi atau pengurangan fitur. Apabila *Information Gain threshold* bernilai kecil, mengakibatkan informasi penting terlalu banyak sehingga dapat menimbulkan *noise* karena masih luasnya lingkup fitur penting. Fitur yang terpilih oleh sistem memiliki nilai *Information Gain* tertinggi seperti yang terlihat pada gambar 10.

Fitur	Information Gain
pegawai	0,18049300
alam	0,17197077
bayes	0,16758689
sms	0,16758689
tree	0,16758689
ponsel	0,16758689
jamur	0,16186436
konsentrasi	0,16186436
kelapa	0,16186436
image	0,15468290
anak	0,15468290
mining	0,15468290
decision	0,15468290
tanah	0,15376060

Gambar 10. Fitur dan Nilai Information Gain

4. KESIMPULAN

Kesimpulan yang dapat diambil dari hasil penelitian ini, yaitu:

1. Metode *Information Gain* yang diterapkan pada sistem seleksi fitur untuk dokumen teks bahasa Indonesia dapat mengurangi fitur hingga 89% dari fitur awal.

2. Pengujian sistem dengan menggunakan sebanyak 21 dokumen abstrak, metode *Information Gain* pada *Threshold* 0,02 dapat menghasilkan fitur terbaik berjumlah 899, *threshold* 0,05 menghasilkan sebanyak 385 dan *threshold* 0,07 menghasilkan 230 fitur terbaik.
3. Pengurangan fitur tertinggi dihasilkan oleh *threshold* 0,07 sebesar 89%.
4. Sistem yang dihasilkan hanya dapat meng*input*kan dokumen teks berekstensi *.doc dan *.docx, tanpa gambar ataupun simbol.

5. SARAN

Saran yang dapat penulis berikan dalam pengembangan sistem adalah aplikasi seleksi fitur dapat dilanjutkan ke tahap pengklasifikasian dokumen atau pencarian dokumen, membangun aplikasi seleksi fitur berbasis web, ekstensi file dokumen *input* ditambahkan *.pdf atau *.html.

DAFTAR PUSTAKA

[1] [1]Firqiani, Hida N., Kustiyo, Aziz & Giri, Endang P. 2008. *Seleksi Fitur Menggunakan Fast Correlation Based Filter Pada Algoritma Voting Feature Intervals 5*. Jurnal Ilmiah Ilmu Komputer, Institut Pertanian Bogor, Vol. 6 No. 2, Hal. 41-47, ISSN : 1693 -1629.

[2] Hatta, Heliza R., Arifin, A. Z. & Yuniarti, A. 2013. *Metode Hibridasi Ant Colony Optimization dan Information Gain Untuk Seleksi Fitur Pada Dokumen Teks Arab*. SCAN Vol. 8 No. 2, ISSN : 1978-0087.

[3] Sari, Yuita Arum & Puspaningrum, Eva Yulia. 2013. *Pencarian Semantik Dokumen Berita Menggunakan Essential Dimension of Latent Semantic Indexing dengan Memakai Reduksi Fitur Document Frequency dan Information Gain Thresholding*. Seminar Nasional Teknologi Informasi dan Multimedia, STIMIK AMIKOM Yogyakarta, 19 Januari 2013, ISSN : 2302-3805.

[4] Qonita, Khurina Azka. 2012. Penerapan Text Mining dengan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) Untuk Pencarian Tafsir Al Qur'an. *Skripsi*, S-1 Ilmu Komputer Fakultas MIPA, Universitas Mulawarman, Samarinda.

[5] Djatna, Taufik & Morimoto, Yasuhiko. 2008. *Pembandingan Stabilitas Algoritma Seleksi Fitur Menggunakan Transformasi Ranking Normal*. Jurnal Ilmiah Ilmu Komputer, Institut Pertanian Bogor, Vol. 6 No. 2, ISSN : 1693 -1629.

[6] Abadi, Delki. 2013. Perbandingan Algoritme Feature Selection Information Gain Dan Symmetrical Uncertainty Pada Data Ketahanan Pangan. *Skripsi*, Institut Pertanian Bogor, Bogor.

[7] Rachmania, Aini. 2011. Klasifikasi Kategori dan Identifikasi Topik Pada Artikel Berita Berbahasa Indonesia. *Skripsi*, Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember, Surabaya.

[8] Luthfiarta, A., Zeniarja, J. & Salam, A. 2013. *Algoritma Latent Semantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen*. Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2013, 16 November 2013, ISBN : 979-26-0266-6.

[9] Karyono, Giat & Utomo, Fandy S. 2012. *Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model*. Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012, 23 Juni 2013, ISBN : 979-26-0255-0.

[10] Sofiana, Ika, Atastina, Imelda & Ardiyanti, Arie. 2012. Analisis Pengaruh Feature Selection Menggunakan Information Gain dan Chi-Square Untuk Kategorisasi Teks Berbahasa Indonesia, https://openlibrary.telkomuniversity.ac.id/home/catalog/id/95720/slug/analisis-pengaruh-feature-selection-menggunakan-information-gain-dan-chi-square-untuk-kategorisasi-teks-berbahasa-indonesia.html, diakses tanggal 1 Oktober 2014.

[11] Saraswati, Ani, Suyatno, Addy & Hatta, Heliza R. 2015. Pengkategorian Dokumen Berita Bahasa Indonesia Menggunakan Algoritma *Term Frequency Inverse Document Frequency* (TF-IDF). *Skripsi*, S-1 Ilmu Komputer Fakultas MIPA, Universitas Mulawarman, Samarinda.