Optimasi Seleksi Fitur dengan Teknik Reduksi Dimensi pada Klasifikasi Abstrak Jurnal

Syukriyanto Latif[*1], Indrabayu[2], Intan Sari Areni[1]
[1]Departemen Teknik Elektro, Fakultas Teknik, Universitas Hasanuddin
Jl. Poros Malino Km. 6, Bontomarannu, Kabupaten Gowa, Sulawesi Selatan, 92171
[2]Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin
Jl. Poros Malino Km. 6, Bontomarannu, Kabupaten Gowa, Sulawesi Selatan, 92171
[*]Email: syukriyanto24@gmail.com

DOI: 10.25042/jpe.052018.08

Abstrak

Tujuan dari penelitian ini adalah untuk mengetahui nilai parameter reduksi dimensi pada seleksi fitur sehingga dapat meningkatkan akurasi dan mengurangi waktu komputasi. Sistem ini menggunakan teknologi text mining yang mengekstraksi data berupa teks untuk mencari informasi dari sekumpulan dokumen. Pembobotan kata *(Term Weighting)* dan Teknik Reduksi Dimensi *Term Frequency Thresholding* digunakan pada proses seleksi fitur, sedangkan pada proses klasifikasi menggunakan algoritma *Naive Bayes*. abstrak jurnal dikategorikan menjadi 3 yaitu *Data Mining (DM), Intelligent Transport System (ITS)* dan *Multimedia(MM)*. Jumlah seluruh data uji dan data latih sebesar 150 data. Hasil klasifikasi terbaik diperoleh saat nilai parameter reduksi dimensi 30%. Pada kondisi tersebut diperoleh nilai akurasi rata-rata sebesar 87.33% dengan waktu komputasi 4 menit 12 detik.

Abstract

Feature Selection Optimization with Dimensional Reduction Techniques in Abstract Journal Classification. The purpose of this research is to know dimension reduction parameter value at feature selection so as to improve accuracy and reduce computation time. This system uses text mining technology that extracts text data to find information from a set of documents. Word weighting and Term Reduction Technique The term Frequency Thresholding is used in the feature selection process, while in the classification process using the Naive Bayes algorithm. the abstract of the journal is categorized into 3 namely Data Mining (DM), Intelligent Transport System (ITS) and Multimedia (MM). The total number of test data and training data is 150 data. The best classification results are obtained when the dimension reduction parameter value is 30%. At that condition obtained an average accuracy of 87.33% with a computation time of 4 minutes 12 seconds.

Kata Kunci: *Klasifikasi, naive bayes, reduksi dimensi, term weighting*

1. Pendahuluan

Text mining merupakan suatu proses mencari informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen. Namun kebanyakan dokumen tidak diklasifikasi atau dikelompokkan sesuai dengan kelompoknya sehingga dokumen-dokumen yang berhubungan sulit untuk ditemukan.

Pemilihan fitur sangatlah penting dalam penelitian klasifikasi text. Karena pada tahap ini seluruh kata pada dokumen akan di seleksi berdasarkan kemunculan kata yang terdapat pada suatu dokumen. Namun dengan melimpahnya data yang akan diolah sering mengalami *over looping* pada proses klasifikasinya. Oleh karena itu penerapan *term weighting* dan teknik reduksi dimensi dapat mengurangi waktu komputasi dan mereduksi fitur-fitur yang non-informatif. Sehingga proses klasifikasi mendapatkan hasil yang baik.

K. K. Bharti dkk pada tahun 2015 melakukan penelitian mengenai teknik reduksi dimensi pada klasifikasi dokumen menggunakan teknik hybrid reduksi dimensi *Principal Component Analysis* (PCA). Dalam penelitian ini digunakan sebuah teknik reduksi dimensi dengan menggunakan *Term Varians* (TV) dan dokumen frekuensi (DF) untuk fitur perhitungan nilai relevansi. Kemudian menggunakan metode k-means dalam proses

pengelompokannya. Hasil dari ujicoba yang dilakukan pada parameter *C*1 dan *C*2 mencapai akurasi 20% dan 80% pada kondisi jumlah fitur kata 3131 dengan reduksi dimensi 50% [1].

Selanjutnya, penelitian yang dilakukan Fajar Rohman Hariri dkk pada tahun 2015 juga melakukan penelitian klasifikasi abstrak tesis menggunakan metode *learning vector quantization* dengan membagi 3 kategori bidang yaitu SIRPL(Sistem Informasi Rekayasa Perangkat Lunak), CAI (*Computation – Artificial Intellegence*) dan Multimedia, dalam penelitian *feature selection* yang digunakan yaitu reduksi dimensi sehingga dapat menambah akurasi dalam proses validasinya. Akurasi 100% yang didapat pada bidang SIRPL dan CAI, 70% untuk bidang minat Multimedia, sehingga rata-rata keseluruhan akurasi yang di dapat 90%, kondisi terbaik di dapatkan dengan parameter reduksi dimensi 20% [2].

Amalia Anjani dkk pada tahun 2015 juga menerapkan Algoritma NBC *(Naive Bayes Classifier)* dengan *Confix Stripping Stemmer* sebagai klasifikasi artikel berita bahasa indonesia. Berdasarkan hasil validasi menggunakan 10 *Cross validation* NBC mencapai akurasi terbaik sebesar 86,97% [3]. I Gusti A Socrates dkk pada tahun 2016 melakukan penelitian menggunakan metode *NBC* dengan menerapkan pemilihan fitur pembobotan *Gain Ratio* dalam mengklasifikasi teks bahasa indonesia, tingkat akurasi NBC mencapai 91% [4].

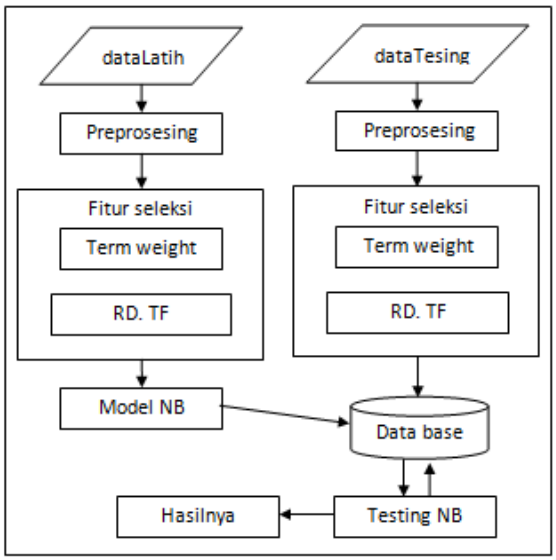
Penelitian selanjutnya oleh Mohamed K. Elhadad dkk tahun 2017 juga melakukan penelitian klasifikasi dokumen web dengan menerapkan teknik reduksi dimensi dengan *Term Frequency-Inverse Documen Frequency* (TF-IDF) dan metode ekstraksi *Principal Component Analysis* (PCA) pada proses seleksi fiturnya. Beberapa metode klasifikasi yang diuji dalam penelitian ini yaitu *Naive Bayes* (NB), J48, JRip dan *Support Vector Machine* (SVM), SVM mencapai akurasi tertinggi dari yang lainya dengan 85,15%, J48 81%,JRip 75.70% dan NB 75.30% [5].

Dari beberapa penelitian tersebut maka penulis mengusulkan untuk menggunakan *Term Weighting* dan teknik reduksi dimensi pada seleksi fitur sehingga akurasi sistem klasifikasi dapat ditingkatkan.

2. Metodologi

Penelitian ini merupakan penelitian eksperimental yang menerapkan *term weighting* dan teknik reduksi dimensi pada proses seleksi fitur juga algortimaNB dalam klasifikasinya.. Data konten Abstrak didapat pada laman www.computer.org. dari 3 kategori, yaitu *Data mining* (DM), *Intelligent Transport System* (ITS) dan Multimedia (MM). Data file abstrak jurnal yang digunakan sebanyak 150 dengan ekstensi pdf. Untuk analisis data dari dokumen abstrak yang diperoleh akan dilakukan analisis percobaan dengan 9 ujicoba reduksi dimensi yang berbeda-beda.

Perancangan sistem yang diusulkan ditujukkan pada Gambar 1.



Gambar 1. Desain sistem

2.1. *Preprocessing*

Ada beberapa tahap yang perlu dilakukan dalam proses ini, yaitu:

a. *Case folding*
   Pada bagian ini mengubah seluruh huruf yang terdapat di dalam dokumen menjadi huruf kecil (*lowercase*).

b. *Tokenizing*
   Tahap ini pemotongan *string input* berdasarkan tiap kata yang menyusunnya.

c. *Filtering*
Tahap ini mengambil kata-kata penting dari hasil token dengan menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting).

d. *Stemming*
Tahap ini mencari *root* kata atau kata dasar dari tiap kata hasil filtering. *Stemming* merupakan suatu proses yang terdapat dalam sistem IR yang mentransformasikan kata-kata yang terdapat dalam suatu dokumen ke kata-kata akar atau dasarnya *(root word)* dengan menggunakan aturan-aturan tertentu.

2.2. *Feature Selection*

Dalam tahap ini ada 2 proses yang akan dilakukan, yaitu:

a. *Term Weighting*
Pemberian bobot kata *(term)* yang muncul pada setiap abstrak jurnal

b. Teknik reduksi dimensi
Mengecilkan dimensi data sehingga waktu komputasi dibutuhkan lebih sedikit. Namun proses reduksi dimensi harus memperhatikan karakteristik data, karena dimensi yang hilang bisa jadi juga menghilangkan karakteristik data. Jumlah kejadian dari kata / jumlah kata dalam dokumen dihitung dan menghapus kata-kata yang kurang berpengaruh dan jarang muncul. Pada penelitian sebelumnya digunakan teknik *Term Frequency (TF) thresholding* dengan cara kata yang memiliki TF<2 akan dihapus [6]. Namun, pada penelitian ini menggunakan *TF thresholding* dengan cara kata yang memiliki TF < Nilai persentase kemunculan kata yang ditentukan maka akan dihapus. Dengan memberikan nilai persentase pada reduksi dimensi diharapkan dapat memberikan peningkatan akurasi dan kecepatan komputasi yang lebih baik.

Dari 150 data abstrak jurnal setelah melewati proses *preprocessing* menghasilkan 6.640 *term* (kata). Hasil jumlah term yang dihasilkan setelah dilakukan proses reduksi dimensi ditunjukan pada Tabel 1.

Tabel 1. Hasil jumlah kata reduksi dimensi *Term Frequency Thresholding* pada tiap kategori

Nilai Reduksi Dimensi	Jumlah kata DM	Jumlah kata ITS	Jumlah kata MM
10%	168	156	130
20%	136	120	116
30%	117	115	97
40%	87	81	62
50%	69	59	50
60%	42	38	31
70%	30	29	19
80%	20	18	11
90%	13	10	8

2.3. Klasifikasi

Klasifikasi merupakan salah satu tugas penting dalam *data mining*. Sebuah data akan masuk ke dalam kelompok tertentu yang sebelumnya telah ditentukan. Setiap hari, jumlah dokumen semakin bertambah. Diantara berbagai bentuk informasi digital, diperkirakan 80% dokumen digital adalah dalam bentuk teks. Tingginya volume dokumen teks ini dikarenakan aktivitas yang terus meningkat dari berbagai sumber berita dan aktivitas penulisan dokumen akademis dari kegiatan riset, konferensi dan pertemuan-pertemuan ilmiah [7].

Dari hasil pembobotan *Term Weighting* dan teknik reduksi dimensi yang dilakukan selanjutnya akan digunakan sebagai data latih dan data uji pada proses *Naive Bayes*. Proses naive bayes meliputi 2 tahapan yaitu tahapan latih dan tahapan testing, *Naive Bayes Classifier* merupakan model penyederhanaan dari algoritma bayes yang cocok dalam pengklasifikasian text atau dokumen seperti pada persamaan berikut.

$$P\left(C_i \left| W_k \right.\right) = \frac{P\left(C_i \left| W_k \right.\right) \times P(C_i)}{P\left(W_k\right)} \qquad (1)$$

Dimana:

- $P\left(C_i \left| W_k \right.\right)$ adalah probabilitas kemunculan kategori C_i dengan kata W_k
- $P\left(W_k\right)$ ”konstan” untuk semua kategori sehingga hanya terbentuk $P\left(W_k \left| C_i \right.\right) \times P(C_i)$ yang perlu dimaksimumkan

- C_i adalah kategori yang tersedia (C_1, C_2, ...Ci)
- $P(C_i)$ adalah probabilitas kemunculan kategori C_i
- $P(W_k|C_i)$ adalah probabilitas kemunculan kata W_k pada ketegori C_i

Sehingga persamaan untuk menyelesaikan permasalahan ini sebagai berikut:

$$P(kata|kategori) = \frac{P(kata|kategori) \times P(kategori)}{P(kategori)} \quad (2)$$

a. Proses latih

Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin merepresentasikan dokumen, juga pembentukan kelas dan sebagai acuan bagaimana konten abstrak akan diklasifikasikan. untuk memodelkan probabilitas proses latih digunakan persamaan berikut.

$$P(c_i)\frac{fd(c_i)}{|D|}$$

$$P(W_k|C_i) = \frac{n_k + 1}{n + |vocabulary|} \quad (3)$$

Dimana:

- $P(c_i)$ adalah jumlah konten abstrak yang memiliki kategori c_i
- $|D|$ adalah jumlah seluruh training konten abstrak.
- n_k nilai kemunculan kata pada tiap kategori c_i
- n adalah jumlah keseluruhan kata pada kategori c_i
- $|vocabulary|$ adalah jumlah keseluruhan kata.

Sehingga Persamaan untuk memodelkan probabilitas proses latih digunakan persamaan berikut.

$$P(kategori) = \frac{J \cdot Konten\ abstrak\ dalam\ kategori}{j \cdot seluruh\ data\ latih\ konten\ abstrak} \quad (4)$$

$$P(kata|kategori) = \frac{Nilai\ kemunculan\ kata\ pada\ tiap\ Kategori + 1}{J \cdot semua\ kata\ pada\ kategori + j \cdot semua\ kata\ latih\ data} \quad (5)$$

b. Proses testing

Tahapan ini sebagai inti dalam *naive bayes* yaitu untuk mengetahui keakuratan model yang dibangun pada proses *training* untuk memprediksi label kelas yang belum diketahui.

- Hasilkan probabilitas untuk masing-masing kelas sesuai dengan persamaan (1) dengan menggunakan *P(kategori)* dan *P(kata | kategori)* yang telah diperoleh dari pelatihan.
- Nilai probabilitas kelas maksimum adalah kategori kelas terpilih sebagai hasil klasifikasi.

3. Hasil

Setelah dilakukan proses reduksi dimensi, selanjutnya dilakukan proses pelatihan menggunakan *Naive Bayes* (NB). Dari bobot akhir yang dihasilkan, kemudian diprediksi dengan cara menghitung probabilitas kemunculan kata ditahap testing pada metode NB. Dengan 100% data digunakan sebagai data testing dan data training dengan 9 kali uji coba reduksi dimensi yang berbeda menghasilkan akurasi dan waktu komputasi seperti ditunjukkan pada Tabel 2 berikut.

Tabel 2. Hasil akurasi dan waktu komputasi pada skenario reduksi dimensi

Nilai Reduksi Dimensi	Akurasi (%)	Waktu komputasi *(time)*
10%	74	0:05:30
20%	81.33	0:04:58
30%	87.33	0:04:12
40%	72	0:03:55
50%	56	0:03:20
60%	47.33	0:02:43
70%	47.33	0:02:43
80%	47.33	0:02:43
90%	47.33	0:02:43

Dari tabel 2 dapat diketahui pada nilai reduksi dimensi 60%-90% memperoleh waktu komputasi yang lebih cepat namun hanya menghasilkan akurasi 47.33%, dan yang menghasilkan nilai akurasi terbaik adalah saat dilakukan reduksi dimensi 30% dengan nilai akurasi mencapai 87.33% dengan waktu komputasi 4 menit 12 detik. Dengan rincian pengenalan untuk masing-masing kategori pada Tabel 3 berikut.

Tabel 3. Rincian akurasi pengenalan reduksi dimensi 30%

Kategori	Jumlah konten absrak	Jumlah berhasil diklasifikasi	akurasi
DM	50	45	90%
ITS	50	46	92%
MM	50	40	80%

Dari Tabel 3 diketahui bahwa metode NB paling baik dalam mengklasifikasikan abstrak untuk kategori DM dengan perhitungan akurasi mencapai 90% berhasil mengenali 45 abstrak jurnal, 46 abstrak jurnal pada kategori ITS dengan akurasi 92% dan MM hanya berhasil mengenali 40 dari 50 abstrak jurnal dan menghasilkan nilai akurasi sebesar 80%.

4. Kesimpulan

Optimasi seleksi fitur dengan teknik reduksi pada klasifikasi abstrak jurnal berbahasa Inggris telah dilakukan pada penelitian ini dengan jumlah konten sebesar 150 data. Tahapan klasifikasi abstrak jurnal menggunakan metode teknik reduksi dimensi untuk ekstraksi fitur dan *Naive Bayes* pada proses klasifikasi. Dari hasil penelitian terlihat bahwa teknik reduksi dimensi sebesar 30% dapat meningkatkan hasil akurasi NB dengan rata-rata 87.33% dengan waktu komputasi 4 menit 12 detik.

Referensi

[1] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Systems with Applications*, vol. 42, p. 3105–3114, 2015.

[2] Fajar hariri R. Ema Utami, and Armadyah Amborowati, 2015, *Learning Vector Quantization for abstract thesis classification.* (*Citec Journal*), vol. 2, No. 2, Februari 2015 – April 2015 ISSN: 2354-5771.

[3] Amalia Anjani A. Arif Djunaidy, Renny P. Kusumawardani, 2015. Classification Articles News On The Net Based Naive Bayes Classifier Using Confix-Stripping Stemmer, (ITS-Undergraduate)-38449-5210100106.

[4] I Gusti. A. Socrates, Afrizal L. Akbar, dan M. Sonhaji Akbar, 2016. Optimization of Naive Bayes with Feature Selection and Weighted Gain Ratio,,(Lontar komputer) vol 7, No.1, April 2016.

[5] Mohamed K. Elhadad, Khaled M. Badran, and Gouda I. Salama, 2017. A Novel Approach for Ontology-based Dimensionality Reduction for web Text Document Classification. 978-1-5090-5507-4/17/$31.00 ©2017 IEEE ICIS 2017, May 24-26, 2017, Wuhan, China.

[6] Muflikhah, Lailil, dan Baharudin, Baharum, 2009, *Document Clustering using Concept Space and Cosine Similarity Measurement*, 2009 IEEE InternationalConference on Computer Technology and Development.

[7] Yang, Y., Pedersen, J. O., 1997, A Comparative Study on Feature Selection in Text Categorization, (Proceedings of the Fourteenth International Conference on Machine Learning) Nashville, Tennessee, USA, 8-12 Juli 1997.