REDUKSI FITUR UNTUK KATEGORISASI TEXT DENGAN KLASIFIKASI MENGGUNAKAN NEURAL NETWORK

Eva Yulia Puspaningrum[1], Shofiya Syidada[2]
[1] Teknik Informatika, UPN Veteran Jawa Timur
[2] Teknik Informatika, Universitas Wijaya Kusuma Surabaya
email : evayulia@ymail.com [1], cpya12@gmail.com [2]

Abstrak. Data memiliki ruang fitur yang sangat tinggi dimensi. Dalam kategorisasi teks menggunakan jaringan syaraf tiruan sebagai klasifier teks, data dilatih menggunakan ruang fitur. Untuk itu diusulkan teknik pengurangan dimensi untuk mengurangi ruang fitur. Untuk menguji keefektifan dari model yang diusulkan, percobaan dilakukan menggunakan dataset dari Routers-21578 untuk uji kategorisasi teks. Pada paper ini diusulkan dengan membandingkan hasil kategorisasi teks dengan cara mereduksi fitur dengan TF/DF Thresholding dan TF/DF Thresholding ditambah dengan PCA. Hasil yang didapatkan menunjukkan dengan reduksi fitur menggunakan TF/DF thresholding mampu mereduksi kata hingga 45,37 % dan TF/DF Thresholding ditambah dengan PCA mampu mereduksi dokumen asli menjadi 98,5%. Pada saat klasifikasi akurasi yang didapat setelah reduksi dimensi dengan TF/DF Thresholding mempunyai nilai akurasi yang lebih baik dibandingkan dengan hasil akurasi setelah reduksi dimensi dengan PCA.

Keyword: reduksi dimensi, kategorisasi text, neural network.

1. PENDAHULUAN

Pergeseran informasi berlangsung cepat dari waktu kewaktu, dimulai dari media cetak hingga saat initeknologi penyebaran informasi telah menggunakandigital. Perkembangan internet menyebabkan membanjirnya informasi digital. Berbagai informasi bisa didapatkan dengan mudah, cukup meng-*klik* atau menekan 'enter'. Banyak informasi digital yang tersebar di dunia maya yang tidak semua kita perlukan. Oleh karena itu, perlu adanya pengelompokan informasi berdasarkan konten nya. Selain itu, dimensi data yang besar juga menyebabkan proses komputasi yang besar. Oleh karena itu teknik reduksi dimensi digunakan agar dapat mengurangi waktu komputasi dan dapat mereduksi fitur-fitur yang non-informatif. Penelitian Lailil dan Baharuddin [1] menggunakan *Document Frequency* (DF) *thresholding* dalam melakukan pengelompokan dokumen menggunakan SVD dan FCM.

Sementara pada Penelitian Savio L.Y. Lam&Dik Lun Lee [2] menggunakan teknik reduksi dimensi *Principal Componen Analisis.* Teknik reduksi dimensi dengan menggunakan PCA digunakan untuk mengurangi dimensi data dengan mempertahan kanvariasi data yang ada. Untuk itu dalam penelitian ini digunakan sebuah teknik reduksi dimensi dengan menggunakan TF/DF thresholding yang akan di bandingkan dengan reduksi fitur yang menggabungkan antara TF/DF thresholding dan PCA untuk mereduksi fitur-fitur yang dianggap non-informatif pada saat kategorisasi teks.

2. TINJAUAN PUSTAKA

2.1 Text Mining

Text mining adalah salah satu bidang khusus dari data mining. Text mining dapat didefinisikan sebagai suatu proses menggali informasi dari data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen.

Agar dapat dikomputasikan, maka sebuah kumpulan dokumen teks harus diolah agar menjadi numerik. Teknik tersebut dinamakan preprocessing.Teknik yang terdapat dalam preprocessing yaitu tokenizing, filtering dan stemming. Tokenizing digunakan untuk memisahkan kata-kata yang ada dalam dokumen serta membuang tanda baca. Filtering digunakan membuang kata-kata yang kurang penting (*Stopword*), seperti kata penghubung, kata sambung, dll. Dan selanjutnya mengubah kata-kata sesuai dengan kata dasarnya. Teknik tersebut dinamakan dengan Stemming.

2.2 Term Frequency Thresholding (TF)

Term Frekuensi mengukur relevansi sebuah kata dalam sebuah teks tunggal. TF thresholding merupakan teknik untuk reduksi kata. Jumlah

kejadian dari kata / jumlah kata dalam dokumen dihitung dan menghapus kata-kata yang kurang berpengaruh dan jarang muncul. Pada penelitian ini digunakan teknik TF thresholding dengan cara kata yang memiliki TF<2 akan dihapus [1].

2.3 Document Frequency Thresholding (DF)

DF *thresholding* adalah teknik untuk reduksi kata. Frekuensi dokumen adalah banyaknya kata yang muncul dalam dokumen. Pada DF *thresholding,* menghitung frekuensi dokumen untuk setiap kata dalam dokumen dan menghapus kata-kata yang kurang dari batas yang telah ditentukan. Asumsi dasarnya adalah menghapus semua kata yang frekuensi tinggi yang tidak relevan dan menghapus dokumen yang jarang atau yang tidak berpengaruh.

Pada penelitian ini digunakan teknik DF *thresholding* dengan cara DF yang panjangnya lebih dari sama dengan setengah dokumen akan dihapus [3].

2.4 Principal Component Analysis (PCA)

Analisis komponen utama (PCA) adalah statistikteknik untuk mereduksi dimensi yang bertujuan untuk meminimalkan kerugian dalam varians dalam data asli[4]. Hal ini dapat dipandang sebagai suatu teknik independen domain untuk ekstraksi fitur,yang berlaku untuk berbagai macam data.

Dalam rangka untuk melakukan analisis komponen utama padaset dokumen pelatihan, akan diwakili set fitur vektor acak oleh vektor berdimensi *N* (x). Dataset matrik *x* berukuran (*M* × *N*) yang dimisalkan x_1, x_2, ..., x_M adalah *N* ×1 vectors.Algoritma dari analisis komponen utama adalah sebagai berikut :

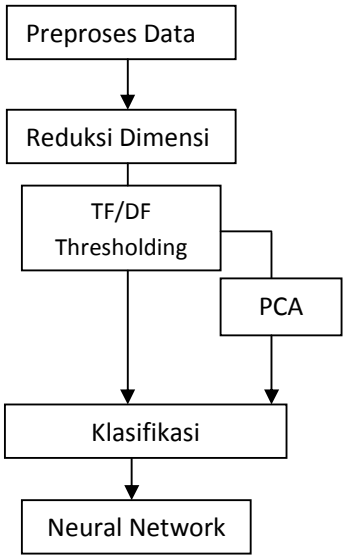
1. Hitung vektor rata-rata
$$\bar{x} = \frac{\sum_{i=1}^{M} xi}{M}$$
2. Hitung Mean $\Phi_i = x_i - \bar{x}$
3. Dari Matrik $A = [\Phi_1, \Phi_2, \dots \Phi_M]$
Hitung matrik kovariansi **C** dengan:
$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n{}^T = A A^T$$
4. Hitung eigenvalue dari
$$C: \lambda_1 > \lambda_2 > \cdots > \lambda_N$$
5. Htung eigenvectors dari
$C: u_1, u_2, \dots, u_N$
6. Reduksi dimensi.
Hanya berhubungan dengan *K eigenvalue* terbesar.
$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \ \ where\ K \ll N$$
7. Transformasi $R^N \rightarrow R^K$ dengan melakukan reduksi dimensi
$$(x_i - \bar{x}) = U^T (x_i - \bar{x})$$
8. Sehingga banyak principal Component
$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > threshold$ (e.g 0.9 or 0.95)

2.5 Neural Network

Dengan menggunakan teknik reduksi dimensi, set dokumen dikategorikan berubah menjadi set vektor fitur dalam fitur dimensi yang relatif rendah. Set vektor fitur dikurangi kemudian diumpankanke classifier teks sebagai masukan. Dalam penelitian ini, jaringan saraf yang digunakan dalam penelitian adalah backpropagation yang terdiri dari lapisan masukan, lapisan tersembunyi, dan lapisan output.

3. METODOLOGI

Dataset yang digunakan adalah dataset dari UCI *Reuters-21578.* Karena banyaknya dokumen pada dataset tersebut maka di perlukan adanya reduksi dimensi pada kategorisasi teks. Tahap pertama dokumen tersebut di preproses telebih dahulu untuk menghilangkan kata penghubung, mengubah kata dasar. Presproses ini dilakukan di *Weka.* Setelah preproses maka diteruskan dengan reduksi dimensi dengan TF/DF Thresholding yang kemudian direduksi lagi dengan PCA setelah proses reduksi dimensi maka dilakukan kategorisasi teks dengan menggunakan *neural network*. Metodologi tersebut dapat dilihat pada Gambar 1.



Gambar 1. Metodologi

4. HASIL DAN PEMBAHASAN

4.1 Dataset

Dataset yang digunakan dalam uji coba diambil dari UCI *Reuters-2157* yang terdiri dari 3 kelas yaitu kelas *acq, crude, earn.* Seperti pada Tabel 1.

Tabel 1. Dataset

Jumlah Dokumen	Jumlah Fitur Awal	
	Allterm	Stemmed
90	2008	1546
120	2311	1770
300	3202	2460

4.2 Hasil Uji Coba

Reduksi Fitur dilakukan sebelum kategorisasi text, hali ini dilakukan untuk mengurangi fitur dokumen. Tabel 2 menunjukkan hasil reduksi yang telah dilakukan.

Reduksi dimensi yang menggunakan df/tf thresholding mengurangi jumlah fitur sampai 45.3% pada dokumen allterm dan 40.5% pada dokumen stemmed.

Reduksi dimensi dengan menggunakan df/tf thresholding dan PCA dengan threshold 0.8 pada dokumen stemmed dan allterm adalah 98.1% dan 98.5% dari jumlah fitur awal

Tabel 2. Hasil Reduksi Dokumen

Jumlah Dokumen	*Allterm*			*Stemmed*		
	threshold tf/df	t PCA>0.8	t PCA>0.9	threshold tf/df	t PCA>0.8	t PCA>0.9
90	52.59%	98.85%	98.31%	46.83%	98.51%	97.80%
120	52.01%	98.70%	98.05%	46.89%	98.42%	97.57%
300	31.51%	98.03%	96.72%	28.01%	97.64%	96.02%
Rata-rata Fitur yang tereduksi	45.37%	98.53%	97.69%	40.58%	98.19%	97.13%

Dengan threshold PCA 0.9 jumlah fitur pada dokumen stemmed dan allterm akan berkurang menjadi 97.1% dan 97.6%

Pada uji coba dilakukan dengan beberapa skenario yang dilakukan dengan 30 iterasi pada masing-masing dataset, kemudian dari hasil uji coba tersebut diambil 3 hasil terbaik sehinggan diperoleh rata-rata akurasi dan nilai MSE.

Berikut adalah hasil skenario uji coba yang dilakukan:

a) Skenario 1. Klasifikasi dokumen *allterm* dan *stemmed* setelah reduksi fitur dengan TF/DF Thresholding
   Hasil Uji Coba Skenario1 dapat dilihat pada Tabel 3.

Tabel 3. Hasil Uji Coba Skenario 1

Jml dokumen	All term		Stemmed	
	Akurasi	MSE	akurasi	MSE
90	83.33333	0.1341333	92.5926	0.1022919
120	94.4447	0.1147	83.33333	0.18038
300	-		-	

b) Skenario 2. Klasifikasi dokumen *allterm* dan *stemmed* setelah reduksi fitur dengan TF/DF Thresholding dan reduksi PCA dengan threshold 0.8 dan 0.9.
   Hasil Uji Coba Skenario 2 dapat dilihat pada Tabel 4.

Tabel 4. Hasil Uji Coba Skenario 2

Jumlah Dokumen	All Term				Stemmed			
	Akurasi		MSE		Akurasi		MSE	
	threshold = 0.8	threshold = 0.9	threshold = 0.8	threshold = 0.9	threshold = 0.8	threshold = 0.9	threshold = 0.8	threshold = 0.9
90	55.1852	56.2963	0.457123	0.4749867	58.5185	59.62963	0.45601	0.407007
120	57.22223	63.8889	0.3939	0.3982433	61.94443	66.66663	0.425847	0.441017
300	58.8889	61.333333	0.40437	0.39635	67.33333	63.66667	0.400987	0.347287

4.3 Pembahasan

Dari hasil uji coba skenario pertama terlihat bahwa hasil klasifikasi dengan menggunakan dokumen *allterm* vs *stemmed* tidak jauh berbeda. Akurasi klasifikasi dokumen *allterm* dengan jumlah dokumen 90 dan 120 adalah 83.33% dan 94.44%. Sedangkan akurasi dokumen *stemmed* dengan jumlah dokumen 90 dan 120 adalah 92.59% dan 83.33%.

Dari hasil uji coba skenario kedua terlihat bahwa hasil klasifikasi dengan menggunakan dokumen *stemmed* lebih baik dari hasil klasifikasi dengan menggunakan dokumen *allterm*. Akurasi klasifikasi dokumen *stemmed* setelah reduksi fitur menggunakan TF/DF Thresholding dan PCA dengan threshold 0.9 lebih baik dari pada menggunakan threshold 0.8 karena jumlah fitur yang digunakan untuk klasifikasi lebih banyak.

Rata-rata hasil akurasi yang diperoleh dengan menggunakan skenario pertama lebih baik daripada hasil akurasi dengan menggunakan skenario kedua.

5. KESIMPULAN

Reduksi fitur dengan menggunakan TF/DF Thresholding mengurangi jumlah fitur sampai 45.37%. Sedangkan Reduksi fitur dengan menggunakan PCA akan mengurangi jumlah fitur sampai 98.5%.

Klasifikasi dengan setelah reduksi dimensi dengan TF/DF Thresholding mempunyai akurasi yang lebih baik dibandingkan dengan hasil akurasi setelah reduksi dimensi dengan PCA tetapi karena menggunakan jumlah fitur yang lebih banyak maka waktu komputasionalnya relatif lama dan memerlukan memory yang besar.

6. DAFTAR PUSTAKA

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

[2]. Lam, Savio L dan Lee, Dik Lun. 1997, Feature Reduction for Neural Network Based Text Categorization, epartment of Computer Science Honh Kong University of Science and Technoly Clear Dwater Bay, Hong Kong

[3] Yang, Yaming dan Pedersen, 1997, J.O, A Comparative study on Feature Selection in Text Categorization, Schoolof Computer Science, Carnegie Mellon University, USA

[4] I. T. Jolliffe. Principal Component Analysis. Springer- Verlag, New York, 1986.