ISBN 979-978-96964-9-8

Fakultas Teknologi Industri
UNIVERSITAS ISLAM INDONESIA

SEMINAR NASIONAL

TEKNOIN

2012

Pengembangan Teknologi Manufaktur untuk Menunjang Penguatan Daya Saing Bangsa

TEKNIK INFORMATIKA

YOGYAKARTA, 10 NOVEMBER 2012

ISBN No. 978-979-96964-3-9

Prosiding

Seminar Nasional Teknoin 2012

"Pengembangan Teknologi Manufaktur untuk Menunjang Penguatan Daya Saing Bangsa"

Yogyakarta, 10 November 2012

Bidang Teknik Informatika

diselenggarakan oleh:

Fakultas Teknologi Industri
Universitas Islam Indonesia
Yogyakarta

ISBN No. 978-979-96964-3-9

Daftar Isi

ISBN No. 978-979-96964-3-9

Meningkatkan Kinerja Naïve Bayes Classifier (NBC) Untuk Klasifikasi Teks dengan Menggunakan Clustering untuk Pemilihan Feature Kata

Amir Hamzah
Jurusan Teknik Informatika, Fakultas Teknologi Industri, Institut Sains dan Teknologi AKPRIND Yogyakarta
Jl. Kalisahak 28 Komplek Balapan Yogyakarta
Telepon (0274) 563029
e-mail : miramzah@yahoo.co.id

Abstract

The rapid expansion of digital information, especially text information has prompted an intensive study of the techniques in text mining. One technique that is important because of the extensive applications in text mining is text categorization. A variety of text categorization algorithms have been proposed such as Naive Bayes Classifier (NBC), Support Vector Machine (SVM), K-Nearest Neighbour, artificial neural network (ANN) or decision tree. Of the various existing algorithms NBC algorithm is an algorithm that is relatively good and simple in computational processes. In some cases the performance is still slightly lower than the performance of the SVM. One possibility is the poor performance of the feature selection of words as the representation of the document. Selection of feature word is often done using frequency selection documents containing the word. This study applied a feature selection method using the clustering method. In the training phase the selected sample set of documents is clustered with cluster number equal to the number of categories. Furthermore, using cluster centers feature words that represent the clusters are selected. The data used is a text document word with 1000 document. Performance parameter used is the accuracy of categorization. Experiments showed that the use of clustering in selection of feature will increase the accuracy of about 5% to 10% for a variety of sample documents and test documents.

Keywords: text mining, feature words, NBC, categorization, clustering

Pendahuluan

Perkembangan informasi digital saat ini memiliki format informasi yang sangat beragam. Tetapi dari format yang beragam tersebut informasi teks menempati posisi yang paling dominan, yang menduduki sekitar 80% [14]. Menurut [1] dari 80% informasi yang didominasi teks tersebut sebagian besar bentuk informasi ada dalam bentuk yang tidak terstruktur. Melimpahnya informasi teks tidak terstruktur telah mendorongnya munculnya disiplin baru dalam analisis teks, yaitu *text mining* yang mencoba menemukan pola-pola informasi yang dapat digali dari suatu teks yang tidak terstruktur tersebut. Menurut [13], saat ini *text mining* telah mendapat perhatian dalam berbagai bidang, antara lain keamanan, biomedis, aplikasi on line, pemasaran dan aplikasi akademis. Salah satu kegiatan penting dalam text mining adalah klasifikasi atau ketegorisasi teks. Kategorisasi teks sendiri saat ini memiliki berbagai cara pendekatan antara lain misalnya pendekatan *probabilistic*, *support vector machine* , dan *artificial neural network.* Salah satu pendekatan berbasis probabilistic *Naïve Bayes Classifier* (NBC) memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi. Metode NBC untuk klasifikasi atau kategorisasi teks menggunakan atribut kata yang muncul dalam suatu dokumen sebagai dasar klasifikasinya. Penelitian [12] menunjukkan bahwa meskipun asumsi independensi antar kata dalam dokumen tidak sepenuhnya dapat dipenuhi, tetapi kinerja NBC dalam klasifikasi relatif sangat bagus.

Metode NBC untuk klasifikasi berita telah dilakukan oleh beberapa peneliti. [15] meneliti metode NBC untuk kategorisasi berita yang mendapatkan hasil semakin banyak dokumen contoh maka akurasi akan semakin tinggi. Penelitian [16] membandingkan metode klasifikasi teks NBC dengan method *Support Vector machine* (SVM), C4.5 dan *K-Nearest Neighbour* (K-NN). Hasil penelitian menunjukkan SVM memiliki akurasi yang paling tinggi sedangkan NBC memiliki akurasi kedua tertinggi dibandingkan dua method yang lain. Penelitian [8] yang menerapkan method NBC untuk kategorisasi teks berita dan teks akademis menemukan bahwa baik dalam dokumen teks berita maupun dokumen teks akademis seleksi *feature* kata dengan menggunakan frekuensi kemunculan kata 4 atau 5, dalam arti kata akan digunakan sebagai *feature* jika ia muncul pada minimal 4 atau 5 dokumen, akan memberikan akurasi yang optimal

dalam klasifikasi. Tetapi penelitian tersebut belum memberikan rekomendasi yang jelas apabila koleksi dokumen memiliki skala yang lebih besar dari skala dokumen contoh yang digunakan dalam bahan penelitian. Pada penelitian sebelumnya [6] menemukan bahwa *clustering* dokumen menggunakan konsep sebagai *feature* memberikan hasil yang jauh lebih baik dibandingkan menggunakan *feature* kata. Dalam penelitian tersebut yang dimaksudkan dengan konsep adalah sejumlah kata yang memiliki bobot tertinggi dalam vector rata-rata cluster dokumen. Mengacu pada penelitian ini dapat diharapkan apabila seleksi kata dalam metode NBC menerapkan metode pemilihan kata berdasarkan hasil clustering dokumen maka akan dihasilkan hasil kategorisasi yang lebih baik jika dibandingkan dengan seleksi kata semata-mata berdasarkan frekuensi kemunculan kata.

Penelitian ini bertujuan untuk meneliti apakah teknik *clustering* dokumen dapat menghasilkan suatu metode pemilihan kata yang lebih baik jika digunakan sebagai *feature* dalam proses kategorisasi dokumen.

Klasifikasi dengan Naïve Bayes Classifier (NBC)

Jika sebuah koleksi dokumen dapat diwakili sebagai himpunan D, dimana $D=\{d_i \mid i=1,2,\ldots|D|\}=\{d_1,d_2,\ldots,d_{|D|}\}$ dan dokumen tersebut dapat dikategorikan dengan koleksi kategori $V=\{v_j|j=1,2,\ldots|V|\}=\{v_1,v_2,\ldots,v_{|V|}\}$. Klasifikasi NBC dilakukan dengan cara mencari probabilitas $P(V=v_j \mid D=di)$, yaitu probabilitas kategory v_j jika diketahui dokumen d_i. Dokumen d_i dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu $<a_1, a_2,\ldots,a_n>$. Selanjutnya klasifikasi dokumen adalah mencari nilai maksimum dari :

$$V_{MAP} = \underset{vj\in V}{\operatorname{argmax}} P(v_j \mid a_1, a_2, ..., a_n) \quad (1)$$

Dengan menerapkan teorema Bayes pada persamaan (1) dan dengan mengasumsikan nilai $P(a_1, a_2, ..., a_n)$ bernilai sama untuk semua v_j maka persamaan (1) akan menjadi :

$$V_{MAP} = \underset{vj\in V}{\operatorname{argmax}} P(a_1, a_2, ..., a_n \mid v_j) P(v_j) \quad (2)$$

Selanjutnya dengan menganggap bahwa setiap kata dalam $<a_1, a_2,\ldots,a_n>$ adalah independent, persamaan (2) dapat ditulis sebagai :

$$V_{MAP} = \underset{vj\in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i \mid v_j) \quad (3)$$

Nilai $P(v_j)$ ditentukan pada saat pelatihan, yang nilainya didekati dengan :

$$P(v_j) = \frac{|doc_J|}{|Contoh|} \quad (4)$$

dimana $|doc_J|$ adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan, sedangkan $|Contoh|$ banyaknya dokumen dalam contoh yang digunakan untuk pelatihan.

Untuk nilai $P(w_k \mid v_j)$, yaitu probabilitas kata w_k dalam kategori j ditentukan dengan :

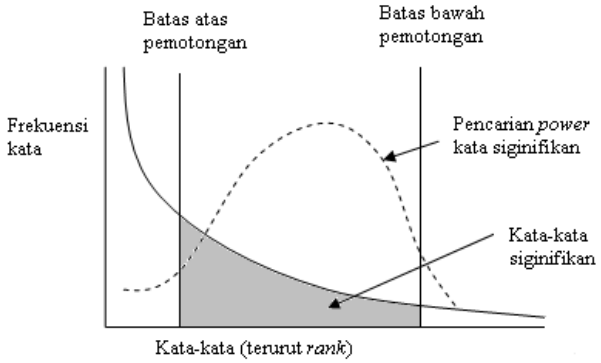
$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |vocabulary|} \quad (5)$$

Dimana n_k adalah frekuensi munculnya kata w_k dalam dokumen yang ber kategori v_j , sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori v_j, dan $|vocabulary|$ adalah banyaknya kata dalam contoh pelatihan. Nilai yang terakhir inilah, yaitu banyaknya kata dalam contoh, yang selama ini berpengaruh besar dalam proses komputasi dan terpenuhinya asumsi independensi. Dari persamaan (2) dan (3) dapat diduga bahwa jika vocabulary memiliki dimensi yang cukup besar maka proses komputasi dalam tahap klasifikasi akan lama. Ketepatan kata yang dipilih dalam vocabulary juga diduga akan menentukan dan mempengaruhi akurasi hasil klasifikasi.

Seleksi Feature Kata

Masalah krusial dalam klasifikasi dokumen adalah menentukan kata-kata yang akan digunakan sebagai *feature* atau atribut dalam proses klasifikasi. Pada umumnya sebaran frekuensi kemunculan kata dalam dokumen adalah mengikuti pola seperi pada gambar 1. Menurut [10], ada kata-kata dengan frekuensi sangat tinggi yang harus dibuang, karena ia

tidak dapat dijadikan pembeda dokumen. Ada juga kata-kata dengan frekuensi sangat jarang yang juga harus dihilangkan. Seleksi awal adalah dengan cara membuang kata-kata yang bukan pembeda dokumen, yaitu yang dikenal dengan STOP-WORD filtering, yaitu kata-kata yang hampir selalu muncul dalam setiap dokumen, misalnya 'ini', 'itu', 'yang', 'dari', dan lain-lain. Penurunan jumlah kata juga dapat dilakukan dengan melakukan *stemming*, yaitu mencari akar kata sedemikian sehingga semua kata-kata yang memiliki akar kata yang sama dianggap sebagai satu kata yang sama. *Stemming* ini dapat menurunkan dimensi kata sampai 25% [4]. Pemilihan kata dengan melihat variasi kemunculan kata juga membuktikan bahwa dengan 20% kata dengan variasi tertinggi mampu memberikan feature kata yang dapat membedakan dokumen dengan sangat baik [2][3][5]. Selanjutnya metode lain untuk seleksi kata adalah dengan melihat frekuensi kemunculan kata dalam dokumen. Penelitian [7] menghasilkan bahwa untuk proses clustering dokumen, seleksi kata dengan menyisakan 5% sampai 10% dari total kata dalam koleksi dokumen. Berbagai metode lain untuk penurunan dimensi (jumlah kata) dalam clustering dan klasifikasi dokumen telah banyak dicoba antara lain penggunaan SVD [11], atau menggunakan ontology [9].



Gambar 1. Hubungan frekuensi kata dan kata ter-*ranking* frekuensi (Luhn, 1958)

Seleksi Feature Kata dengan Clustering

Koleksi dokumen dalam *clustering* model ruang vector dianggap sebagai koleksi vektor dokumen dengan elemen vaktor adalah bobot kepentingan kata terhadap dokumen tersebut. Permasalahan yang biasa muncul dalam *clustering* dokumen adalah tidak adanya informasi berapa buah kluster harus dibuat. Dalam kategorisasi dokumen, pada tahap pelatihan sudah diketahui secara pasti berapa jumlah kategori dari dokumen yang dimaksudkan. Dengan demikian jika dilakukan *clustering* dokumen pada dokumen sampel dengan jumlah kategori dokumen sebagai banyaknya cluster maka hasil *clustering* akan mencerminkan pemusatan kelompok dokumen pada kategorisasi sampel. Penelitian [6] menunjukkan bahwa dengan mengambil kata-kata dengan bobot tertinggi dari pusat kluster akan didapatkan sejumlah kata yang mencerminkan isi atau "konsep" apa sebenarnya cluster tersebut. Demikian dapat diharapkan kata-kata yang diambil dari pusat cluster dengan bobot tertinggi akan dapat mewakili secara lebih independen kategorisasi dokumen yang dimaksudkan. Dengan demikian teknis pemilihan kata dilakukan semata-mata dengan melakukan *clustering* dengan jumlah kategori sebagai cacah kluster.

Metodologi

Dalam penelitian ini digunakan dokumen teks yang terdiri dari dokumen berita yang terdiri dari 1000 dokumen dalam kategori 14 macam kategori.

Pada setiap koleksi, dokumen diformat dalam bentuk seperti Gambar 2. Keterangan tag adalah <DOC></DOC> untuk membedakan dokumen yang satu dengan dokumen yang lain, tag <DOCNO>..</DOCNO> untuk identifikasi nomor dokumen dan tag <CATNO>..</CATNO> untuk identifikasi dari nomor kategori dokumen. Predefine kategori dalam <CATNO> diperlukan untuk menguji kemampuan algoritma dalam proses kategorisasi teks jika dokumen tersebut terpilah sebagai dokumen uji. Format dokumen yang dipisahkan dalam <DOCNO> diperlukan untuk mengidentifikasi bahwa dokumen dalam tag tersebut adalah dokumen yang berbeda dari dokumen lain dalam tag <DOCNO> yang lain. Seluruh dokumen disimpan dalam satu file dalam format koleksi dokumen dengan kode tag <DOC>..</DOC>.

<DOC>
<DOCNO>news035-html</DOCNO>
<CATNO>01</CATNO>
…. perjalanan jamaah haji tahun ini tidak menemukan kendala yang berarti. Pemerintah berharap bahwa kondisi jamaah harus tetap diupayakan selalu dalam stamina tinggi sehingga akan meminimalkan terjadinya korban jiwa karena pelayanan yang tidak optimal.
….
….
</DOC>

Gambar 2. Format Dokumen dalam Koleksi

Pembuatan program dilakukan dengan menggunakan komputer PC Intel Pentium IV 2.8GHz, RAM 1GB, Hard Disk 80 GB, dan sistem operasi Windows XP Professional. Bahasa pemrograman yang dipergunakan adalah java jdk1.6.4.
Tahapan percobaan dibuat kode program untuk melakukan proses Pre Prosesing dokumen, Proses Pelatihan dan Proses Klasifikasi.

Pre Processing dilakukan dengan mengidentifikasikan kata unik dalam seluruh dokumen yang dipilih sebagai contoh, membuang *stop word* seperti kata 'ini', 'itu','yang', dan lain-lain dan memilih *feature* kata kunci sebagai *vocabulary* berdasarkan frekuensi kemunculan kata dalam dokumen. Selanjutnya disusun matriks kata-dokumen yang mencatat frekuensi kemunculan kata dalam tiap dokumen. Koleksi dokumen contoh dalam format vector dokumen selanjutnya dilakukan clustering dengan metode *hierarchical* dengan model UPGMA dan *partitional* K-Means *Clustering*. Dari pusat-pusat kluster selanjutnya dipilih sejumlah kata tertentu yang memiliki ranking pembobotan paling tinggi

Tahapan **Pelatihan** dan **Klasifikasi** tersaji seperti pada algoritma berikut ini :

Pelatihan :
1. Bentuk vocabulary dari kata yang terpilih dalam proses clustering, yaitu kata yang mewakili pusat-pusat kluster.
2. Untuk setiap kategori v_j, hitung :
   a. Tentukan Doc_j (himpunan dok dalam kategori v_j)
   b. Hitung $P(v_j)$ dengan persamaan (4)
   c. Hitung $P(w_k|v_j)$ dengan persamaan (5) untuk tiap w_k dalam vocabulary

Kategorisasi :
1. Tetapkan < a1,a2,…,ai> dari dokumen uji
2. Hitung $P(v_j)\prod_i P(a_i \mid v_j)$ untuk setiap kategori v_j
3. Tentukan nilai maksimumnya sebagai hasil kategorisasi dokumen uji

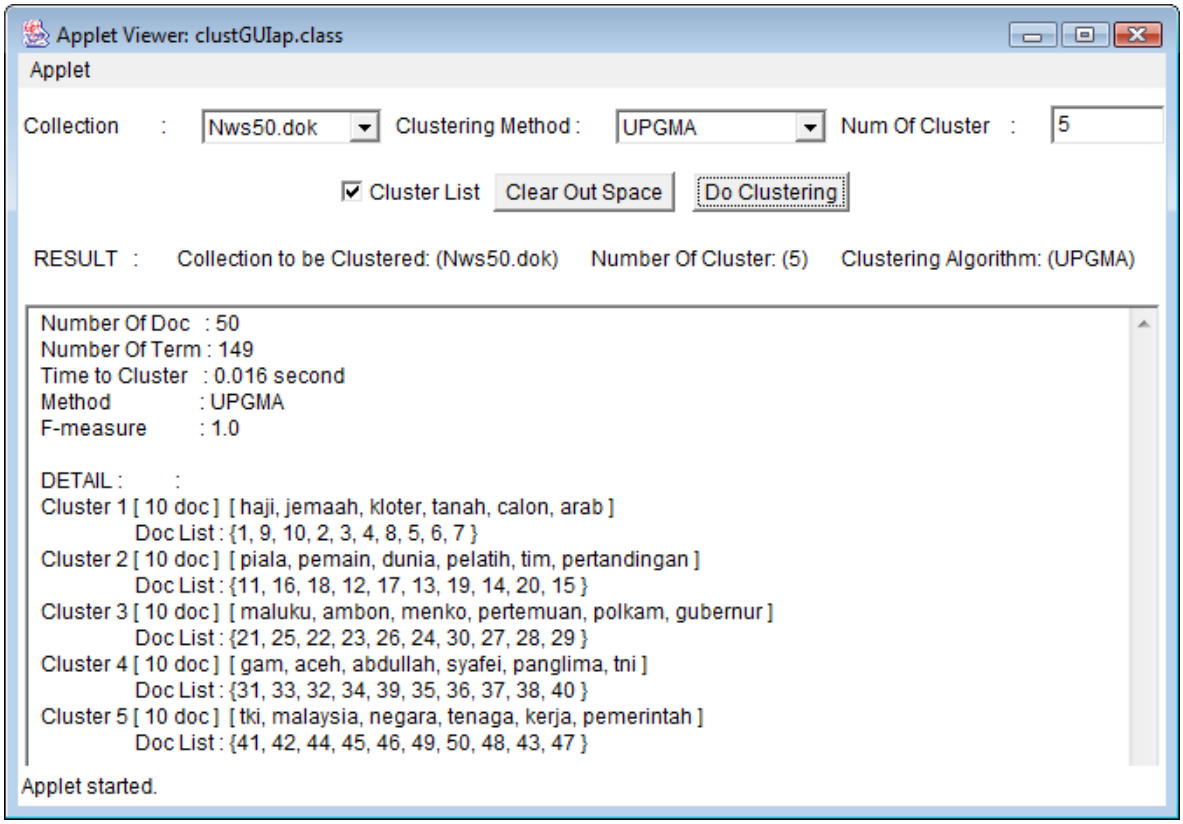
Untuk evaluasi kinerja algoritma dalam melakukan kategorisasi ditetapkan berdasarkan nilai akurasi, yaitu :

$$\text{Akurasi} = \frac{jumlah\_kategorisasi\_benar}{jumlah\_dokumen\_uji} x100\% \quad (6)$$

Hasil dan Pembahasan

Proses *clustering* dokumen untuk mendapatkan daftar feature kata didapatkan seperti pada tampilan Gambar 3. Terlihat pada setiap kluster ada 6 kata terpilih yang mewakili pusat kluster tersebut. Misalnya daftar kata dalam kluster 1, yaitu 'haji', 'jemaah', 'kloter', 'tanah', 'calon' dan 'arab' bisa memberikan gambaran bahwa kluster tersebut adalah koleksi dokumen tentang 'perjalanan haji'. Demikian juga dari kata-kata yang terpilih pada kluster 2 yaitu 'piala', 'pemain', 'dunia', 'pelatih', 'tim', dan 'pertandingan' dapat disimpulkan bahwa kluster kedua adalah kelompok dokumen tentang pertandingan piala dunia. Dengan cara seperti ini ada kemungkinan besar independensi dari *feature* kata yang mewakili dokumen akan lebih terjamin. Hal ini seperti terlihat dalam pilihan 10 kata terpenting dalam tiga buah kluster pada Tabel 1. Dalam tampilan Tabel 1 terdapat angka di depan kata yang menunjukkan bobot elemen kata tersebut dalam vektor pusat klusternya.

ISBN No. 978-979-96964-3-9



Gambar 3. Seleksi 6 Kata Terpenting dari Pusat Cluster

Tabel 1. Contoh 10 kata Terpenting dalam Suatu Kluster

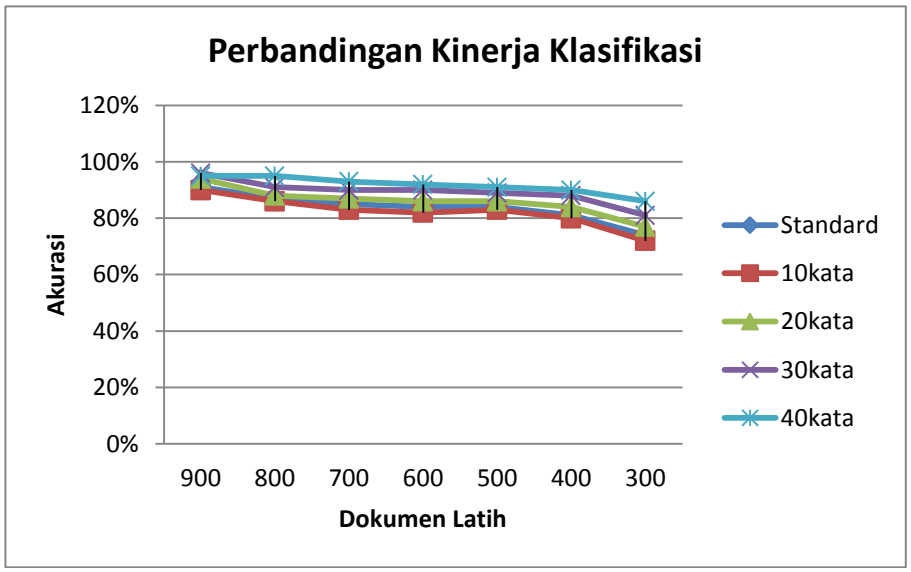
0.2428	kloter	0.0937	prancis	0.1289	akbar
0.1168	menag	0.0818	italia	0.1037	jaya
0.1049	medan	0.0724	bermain	0.0978	sidang
0.0974	makassar	0.0682	cina	0.0776	hakim
0.0972	biaya	0.0508	striker	0.0762	ketua
0.0959	embarkasi	0.0506	kapten	0.0657	hukum
0.0712	barat	0.0488	kartu	0.0652	keterangan
0.0676	asrama	0.0447	unggul	0.0629	jakarta
0.0658	perjanjian	0.0446	milan	0.0532	tandjung
0.0595	mendarat	0.0419	sapporo	0.0516	syarief

Berikut ini adalah efek penggunaan clustering dalam pemilihan *feature* kata pada beberapa kemungkinan *feature* kata dalam kinerja klasifikasi dibandingkan dengan pemilihan *feature* kata hanya dengan menggunakan batasan minimal frekuensi saja. Dalam hal ini minimal frekuensi yang diambil adalah 4, yaitu kata akan diambil jika minimal muncul dalam 4 dokumen (dalam tabel adalah kolom ketiga). Beberapa pilihan kata yang diambil dari proses *clustering* adalah 10 kata, 20 kata, 30 kata dan 40 kata. Hasil kategorisasi pada berbagai dokumen latih dan dokumen uji dapat dicermati dalam Tabel 2 dan grafik pada Gambar 4. Dari Tabel 2 dan gambar 4 terlihat bahwa kinerja klasifikasi standard, yaitu tanpa pemilihan kata akan lebih tinggi jika dibandingkan dengan kategorisasi dengan kata diseleksi menggunakan *clustering* untuk pilihan 10 kata. Tetapi jika pilihan kata diambil 20 kata, 30 kata dan 40 kata akan terlihat kinerja kategorisasi akan meningkat secara rata-rata 2.29% untuk 20 kata, 5.57% untuk 30 kata dan 8.00% untuk 40 kata.

Kinerja ini sesuai dengan asumsi semula bahwa jika independensi antar kata dapat ditingkatkan maka kinerja kategorisasi diharapkan akan meningkat.

Tabel 2. Akurasi Kategorisasi pada Berbagai Pilihan Kata dan Akurasi Standard

Dokumen Latih	Dokumen Uji	Akurasi Standard	Akurasi dengan Seleksi Kata			
			10kata	20kata	30kata	40kata
900	100	91%	90%	94%	96%	95%
800	200	87%	86%	88%	91%	95%
700	300	85%	83%	87%	90%	93%
600	400	84%	82%	86%	90%	92%
500	500	84%	83%	86%	89%	91%
400	600	81%	80%	84%	88%	90%
300	700	74%	72%	77%	81%	86%
Rata-rata kenaikan			-1.43%	2.29%	5.57%	8.00%



Gambar 4. Perbandingan Kinerja Klasifikasi Dokumen

Kesimpulan dan Saran

Beberapa kesimpulan penting yang dapat ditarik dari penelitian ini adalah :

1. Clustering dokumen dapat diterapkan untuk meningkatkan kinerja kategorisasi dokumen dengan cara mengambil kata-kata trpenting dalam pusat kluster sebagai feature kata dalam vocabulary dokumen latih.
2. Pemilihan feature kata meskipun demikian masih belum ditemukan patokan berapa banyak feature kata harus diambil, karena pengambilan 10kata, 20kata, 30kata dan 40kata semata-mata trial dan error, dan jika jumlah kata terlalu sedikit (misalnya 10kata) ternyata kinerja kategorisasi justru menurun, tetapi jika feature kata cukup banyak maka kinerja kategorisasi akan mengkiat sampai 8% rata-rata peningkatannya
3. Peningkatan kinerja klasifikasi belum dianalisis lebih jauh dari sisi kompleksitas waktunya, mengingat penggunaan clustering akan menambah waktu dalam proses preprocessing.
4. Perlu diteliti lebih jauh bagaimana mencari nilai optimal jumlah kata yang akan diambil sebagai feature dari pusat kluster.

Daftar Pustaka

[1] Bridge, C., 2011, *Unstructured Data and the 80 Percent Rule*. (Online di: http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551 ; di ases 29 September 2012)

[2] Dhillon, S. I. and D. S. Modha, "Concept Decomposition for Large Sparse Text data Using Clustering", *Machine Learning*, 42(I):143-175, January 2001

[3] Dhillon, I., J. Kogan, and C. Nicholas, *Feature Selection and Document Clustering*, www.csee.umbc.edu/cadip/koghan.pdf, 2002

[4] Hamzah, A, "Pengaruh Stemming Kata Dalam Peningkatan Unjuk Kerja Document Clustering Untuk Dokumen Berbahasa Indonesia" , *Proseding Seminar Nasional Riset Teknologi Informasi*, AKAKOM, Juli , 2006a

[5] Hamzah, A., F. Soesianto, A.Susanto, J.E.,Istyanto, "Seleksi Feature Kata Berdasarkan Variansi Kemunculan Kata dalam Peningkatan Unjuk Kerja Document Clustering untuk Dokumen Berbahasa Indonesia", *Pakar, Jurnal Teknologi Informasi dan Bisnis* , Vol.7,No.3. , pp. 181-190, 2006b

[6] Hamzah, A, A. Susanto, F. Soesianto, and J.E. Istiyanto, *"*Concept-Based Document Clustering*", International Conference On Electrical Engineering and Informatics*, ICEEI2007, ITB 17-19 June 2007

[7] Hamzah, A., Efek Penambahan Frasa dalam *Feature* Kata untuk *Clustering* Dokumen Teks, *Jurnal TECHNOSCIENTIA*, Vol 1. No.2, pp.140-147, 2009

[8] Hamzah, A., "Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis", to appear in *Seminar Nasional Sains dan Teknologi SNAST 2012*, 3 November 2012, Yogyakarta, 2012.

[9] Khan,L., R., "Ontology-Based Information Selection", ***PhD*** *Dissertation*, Faculty of the Graduate School, University of Southern California,2000.

[10] Luhn, H.P., "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, 2:159-165, 1958

[11] Osinki, S., 2004, "Dimensionality Reduction Techniques for Search Engine Results Clustering", ***Master Thesis***, University of Sheffield, UK

[12] Rish, Irina, 2001 , "An Empirical Study of the Naïve Bayes Classifier", *T.J. Watson Research Center* (on line di http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf ; di ases 29 September 2012).

[13] Saraswati, N.W.S., "Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machine untuk Sentimen Analysis", *Thesis Program Studi Teknik Elektro*, Program pasca Sarjana Universitas Udayana, Bali, 2011.

[14] Tan, Ah-Hwee, "Text Mining: The state of the art and the challenges", *Kent Ridge Digital Labs* 21 Heng Mui Keng Terrace Singapore 119613, 1999

[15] Wibisono, Y., "Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier". (Online di: http:// fpmipa.upi.edu/staff/yudi/yudi_0805.pdf ; diases 29 September 2012), 2005.

[16] Wulandini, F. & Nugroho, A. N., "Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases". *International Conference on Rural Information and Communication Technology*, 2009. (Online di: http://asnugroho.net/papers/rict2009_textclassification.pdf ; diases 28 September 2012).