KLASIFIKASI TEKS DENGAN NAÏVE BAYES CLASSIFIER (NBC) UNTUK PENGELOMPOKAN TEKS BERITA DAN ABSTRACT AKADEMIS

Amir Hamzah

Jurusaan Teknik Informatika, Fakultas Teknologi Industri,
Institut Sains dan Teknologi AKPRIND, Jalan Kalisahak 28, Yogyakarta 55222
Telepon: (0274)-563029; Fax: (0274)-563847
Email: miramzah@yahoo.co.id

ABSTRACT

The development of digital text information has been growing rapidly. Currently, an estimated 80% of digital text is in unstructured form. The high volume of text documents is triggered by the activity of a variety of news sources and academic activities of research, conferences and scientific meetings. The need for text mining analysis is indispensable in dealing with such unstructured text. An important task in text mining is text classification or categorization. Text categorization itself currently has a variety of approaches such as probabilistic approaches, support vector machines, and artificial neural network or decision tree classification. Probabilistic methods Naive Bayes Classifier (NBC) has some advantages in computational simplicity. But this method has a disadvantage in that the assumptions are difficult to fulfill, namely the independence of feature words. This study examines the performance of NBC for text categorization news and academic texts. The study used data of 1000 documents and 450 abstract of academic papers. The results showed on the news documents maximum accuracy reached 91% while 82% of academic documents. Selection appears in the word with a minimum of 4 or 5 documents provide the highest accuracy.
Keywords: naïve bayes classifier, classification, accurate, word selection

INTISARI

Perkembangan informasi teks digital telah tumbuh sangat cepat. Saat ini diperkirakan 80% teks digital dalam bentuk tidak terstruktur. Tingginya volume dokumen teks ini dipicu oleh aktivitas dari berbagai sumber berita dan aktivitas akademis dari kegiatan riset, konferensi dan pertemuan ilmiah yang makin meningkat. Kebutuhan analisis *text mining* sangat diperlukan dalam menangani teks yang tidak terstruktur tersebut. Salah satu kegiatan penting dalam text mining adalah klasifikasi atau kategorisasi teks. Kategorisasi teks sendiri saat ini memiliki berbagai cara pendekatan antara lain pendekatan *probabilistic*, *support vector machine* , dan *artificial neural network*, atau *decision tree classification*. Metode probabilistic *Naïve Bayes Classifier* (NBC) memiliki beberapa kelebihan kesederhanaan dalam komputasinya. Namun metode ini memiliki kelemahan dalam asumsi yang sulit dipenuhi, yaitu independensi feature kata. Penelitian ini mengkaji kinerja NBC untuk kategorisasi teks berita dan teks akademis. Penelitian menggunakan data 1000 dokumen berita dan 450 dokumen abstrak akademik. Hasil penelitian menunjukkan pada dokumen berita akurasi maksimal dicapai 91% sedangkan pada dokumen akademik 82%. Seleksi kata dengan minimal muncul pada 4 atau 5 dokumen memberikan akurasi yang paling tinggi.
Kata Kunci: naïve bayes classifier , klasifikasi, akurat, seleksi kata

PENDAHULUAN

Perkembangan yang pesat dalam informasi digital telah menyebabkan semakin meningkat pula volume informasi yang berbentuk teks. Diantara berbagai bentuk informasi digital, diperkirakan 80% dokumen digital adalah *dalam* bentuk teks (Tan, 1999). Tingginya volume dokumen teks ini misalnya dengan aktivitas yang terus meningkat dari berbagai sumber berita dan aktivitas penulisan dokumen akademis dari kegiatan riset, konferensi dan pertemuan-pertemuan ilmiah. Kondisi "kebanjiran informasi" ini telah menimbulkan kesulitan manusia dalam mencerna informasi. Menurut Bridge (2011), hal yang lebih menyulitkan dalam analisis adalah bahwa sekitar 80% sampai 85% bentuk informasi tersebut dalam format tidak terstruktur (*unstructured data*). 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. Dengan pengertian tersebut *text mining* mengacu juga kepada istilah *text data mining* (Hearst, 1997) atau penemuan pengetahuan dari basis data teks (Friedman and Dagan, 1995).

Menurut Saraswati (2011), saat ini *text mining* telah mendapat perhatian dalam berbagai bidang, antara lain :

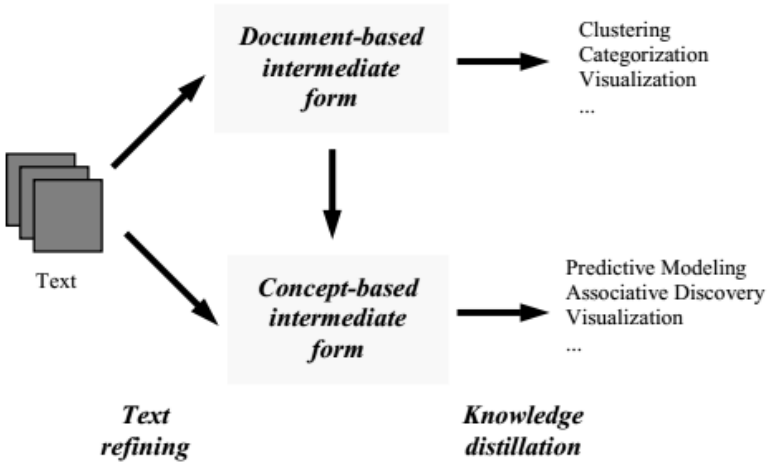
1. Aplikasi keamanan
   Banyak paket perangkat lunak *text mining* dipasarkan terhadap aplikasi keamanan, khususnya analisis plain text seperti berita internet.
2. Aplikasi biomedis
   Berbagai aplikasi *text mining* dalam literatur biomedis telah disusun. Salah satu contohnya adalah PubGene yang mengkombinasikan *text mining* biomedis dengan visualisasi jaringan sebagai sebuah layanan Internet.
3. Perangkat Lunak dan Aplikasi
   Departemen riset dan pengembangan perusahaan besar, termasuk IBM dan Microsoft, sedang meneliti teknik *text mining* dan mengembangkan program untuk lebih mengotomatisasi proses pertambangan dan analisis. Perangkat lunak *text mining* juga sedang diteliti oleh perusahaan yang berbeda yang bekerja di bidang pencarian dan pengindeksan secara umum sebagai cara untuk meningkatkan performansinya.
4. Aplikasi Media Online
   Text mining sedang digunakan oleh perusahaan media besar, seperti perusahaan Tribune, untuk menghilangkan ambigu informasi dan untuk memberikan pembaca dengan pengalaman pencarian yang lebih baik, yang meningkatkan loyalitas pada site dan pendapatan. Selain itu, editor diuntungkan dengan mampu berbagi, mengasosiasi dan properti paket berita, secara signifikan meningkatkan peluang untuk menguangkan konten.
5. Aplikasi Pemasaran
   Text mining juga mulai digunakan dalam pemasaran, lebih spesifik dalam analisis manajemen hubungan pelanggan.
6. Aplikasi Akademik
   Masalah *text mining* penting bagi penerbit yang memiliki database besar untuk mendapatkan informasi yang memerlukan pengindeksan untuk pencarian. Hal ini terutama berlaku dalam ilmu sains, di mana informasi yang sangat spesifik sering terkandung dalam teks tertulis.

Tan (1999) memberikan kerangka kerja dari *text mining* seperti pada Gambar 1. Pada tahap awal ditempuh langkah *text refining* yaitu pengubahan bentuk dari teks asli menjadi bentuk *intermediate* (*intermediate form*), yang dapat berbasis pada bentuk dokumen (*document-based intermediate form*) atau berbasis paad konsep (*concept-based intermediate form*). Tahap berikutnya adalah tahap *knowledge distillation*. Pada tahap ini jika bentuk *intermediate* berupa dokumen maka kegiatan distilasi pengetahuan dapat berupa kegiatan *clustering* dokumen, kategorisasi dokumen, visualisasi dan sebagainya. Untuk bentuk *intermediate* berupa konsep kegiatan distilasi dapat berupa *predictive modeling, assosiative discovery* dan visualisasi.

Salah satu kegiatan penting dalam distilasi pengetahuan adalah klasifikasi atau kategorisasi teks dengan pendekatan *supervised learning*. Kategorisasi teks sendiri saat ini memiliki berbagai cara pendekatan antara lain berbasis numeris, misalnya pendekatan *probabilistic*, *support vector machine* , dan *artificial neural network*, serta berbasis non numeris seperti *decision tree classification.* Dari kelompok pendekatan berbasis numeris, 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 Rish (2001) 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. Wibisono (2005) meneliti metode NBC untuk kategorisasi berita menggunakan 291 dokumen sampel dan 291 dokumen uji mendapatkan hasil akurasi 86.9%, sedangkan jika 70% dokumen contoh dan 30% dokumen uji akurasinya naik jadi 90,23%. Penelitian Wulandini dan Nugroho (2009) membandingkan method klasifikasi teks NBC dengan method *Support Vector machine* (SVM), C4.5 dan *K-Nearest Neighbour* (K-NN).

Hasil penelitian menunjukkan akurasi masing-masing methode urut dari yang terbaik adalah SVM akurasi 92%, NBC akurasi 90% C4.5 akurasi 77.5%% dan yang terendah K-NN akurasi 50%. Sebelumnya perbandingan NBC dengan algoritma *decision tree* juga telah dilakukan oleh Lewis and Renguette (1994), yang memberikan hasil keduanya memiliki keunggulan dan kelemahan.



Gambar 1 Kerangka Kerja *Text Mining* (Tan,1999)

Keunggulan *decision tree* adalah dalam seleksi *feature* kata yang dapat disiapkan lebih baik dari pada seleksi *feature* kata dalam NBC yang cenderung random. Sedangkan keungulan NBC adalah dalam kesederhanaan komputasinya.

Penelitian ini bertujuan untuk meneliti sejauh mana kinerja algoritma NBC dalam kategorisasi teks yang berupa teks berita dan teks akademis berupa abstrak akademis dari berbagai disiplin ilmu. Fokus kajian dilakukan dalam hal pemilihan feature kata yang akan dijadikan sebagai representasi dokumen dan sejauh mana efek banyaknya kata dalam vocabulary mempengaruhi kinerja algoritma klasifikasinya.

Metode Naïve Bayes Classification (NBC)

Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan *term* yang muncul dalam dokumen yang diklasifikasi.

Lebih konkritnya jika diasumsikan dimiliki koleksi dokumen $D=\{d_i \mid i=1,2,\ldots|D|\}=\{d_1,d_2,\ldots,d_{|D|}\}$ dan 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 category v_j jika diketahui dokumen d_i. Dokumen d_i dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu $<a_1, a_2,\ldots,a_n>$, yang frekuensi kemunculannya diasumsikan sebagai variable random dengan distribusi probabilitas Bernoulli (McCallum and Nigam, 1998). Selanjutnya klasifikasi dokumen adalah mencari nilai maksimum dari :

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

Teorema Bayes menyatakan tentang probabilitas bersyarat menyatakan :

$$P(B|A)=\frac{P(A \mid B)P(B)}{P(A)} \quad (2)$$

Dengan menerapkan teorema Bayes persamaan (1) dapat ditulis :

$$V_{MAP} = \arg\max_{vj \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j)P(v_j)}{P(a_1, a_2, ..., a_n)} \quad (3)$$

Karena nilai $P(a_1, a_2, ..., a_n)$ untuk semua v_j besarnya sama maka nilainya dapat diabaikan, sehingga persamaan (3) menjadi :

$$V_{MAP} = \arg\max_{vj \in V} P(a_1, a_2, ..., a_n \mid v_j) P(v_j) \quad (4)$$

Dengan mengasumsikan bahwa setiap kata dalam $<a_1, a_2, \ldots, a_n>$ adalah independent, maka $P(a_1, a_2, ..., a_n \mid v_j)$ dalam persamaan (4) dapat ditulis sebagai :

$$P(a_1, a_2, ..., a_n \mid v_j) = \prod_i P(a_i \mid v_j) \quad (5)$$

Sehingga persamaan (4) dapat ditulis :

$$V_{MAP} = \arg\max_{vj \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (6)$$

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

$$P(v_j) = \frac{|doc_J|}{|Contoh|} \quad (7)$$

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 :

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |vocabulary|} \quad (8)$$

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,

METODE

Dalam penelitian ini digunakan dokumen teks yang terdiri dari dokumen berita dan dan dokumen akademik. 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.

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.

Tabel 1 Koleksi dokumen teks yang dijadikan uji coba

Koleksi	Cacah dok	Tipe dokumen	Cacah Kategori
Nws500.dok	500	Berita	10
Nws1000.dok	1000	Berita	14
Abs.dok	450	Akademik	11

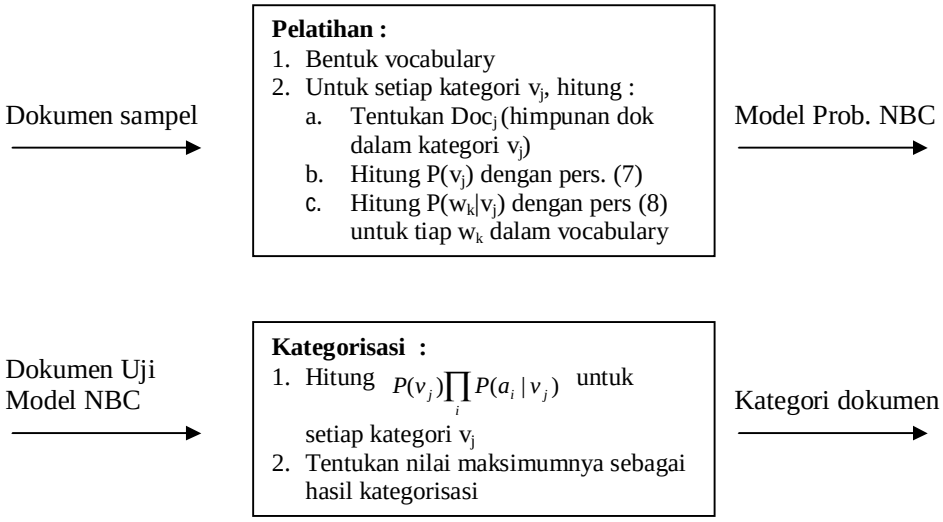
<DOC>
<DOCNO>news035-html</DOCNO>
<CATNO>01</CATNO>
mayjen syafrie samsuddin akan jadi kapuspen tni jakarta media mantan pangdam jaya mayjen syafrie samsuddin akan menjadi kapuspen tni menggantikan marsekal muda graito husodo menurut informasi yang diperoleh antara jakarta kamis syafrie samsuddin menjadi
……
</DOC>

Gambar 2 Format dokumen dalam koleksi

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.

Tahapan **Pelatihan** dan **Klasifikasi** tersaji seperti pada *flowchart* Gambar 3 berikut ini :



Gambar 3 Algoritma Pelatihan dan Pengujian

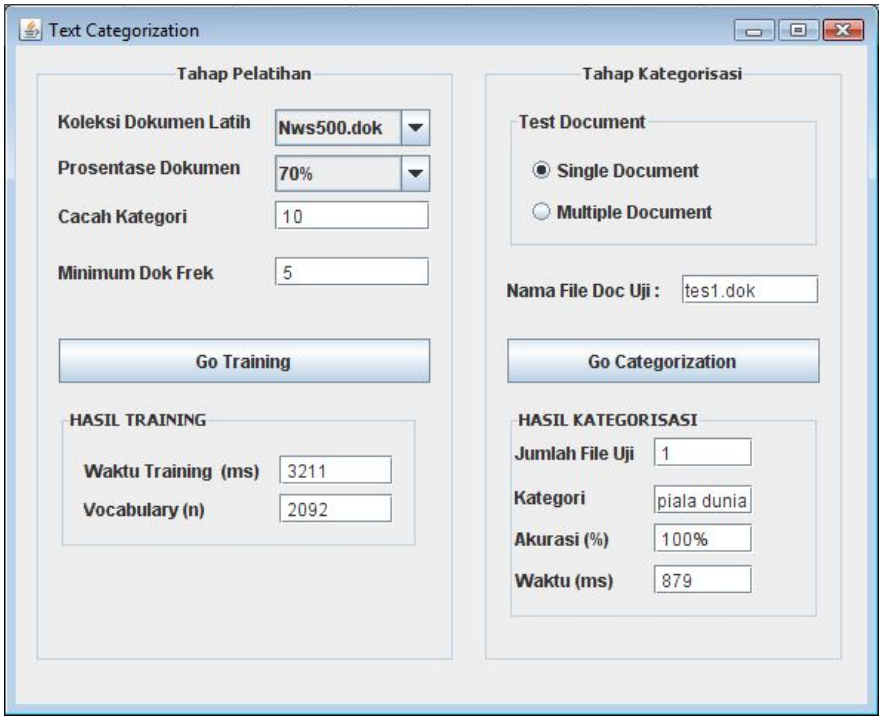
Untuk evaluasi kinerja algoritma dalam melakukan kategorisasi ditetapkan berdasarkan nilai akurasi, yaitu :

$$\text{Akurasi} = \frac{jumlah\_kategorisasi\_benar}{jumlah\_dokumen\_uji} x100\% \quad (9)$$

PEMBAHASAN

Antar muka grafis untuk proses training dan kategorisasi adalah seperti Gambar 4. Pengguna dapat memilih koleksi dokumen uji. Proses dimulai dari training dan dilanjutkan dengan proses kategorisasi yang dapat dilakukan per dokumen atau sekelompok dokumen.

Untuk hasil kategorisasi dokumen pada teks dokumen berita dari koelksi 1000 dokumen dengan berbagai variasi banyaknya dokumen latih dan dokumen uji disajikan dalam Tabel 2.



Gambar 4 Antar Muka Kategorisasi

Pada tabel 2 terlihat untuk bahwa untuk 1000 dokumen, penggunaan 50% dokumen sebagai dokumen latih memberikan akurasi yang cukup tinggi, yaitu 84%. Akurasi menurun dengan semakin sedikitnya dokumen yang digunakan dalam pelatihan.

Tabel 2. Akurasi Uji dokumen berita

Dokumen Latih	Dokumen Uji	Akurasi (%)
900	100	91%
800	200	87%
700	300	85%
600	400	84%
500	500	84%
400	600	81%
300	700	74%

Pada dokumen akademik hasil kategorisasi dari koleksi dkoumen sejumlah 450 dokumen dengan berbagai variasi dokumen latih dan dokumen uji memberikan hasil seperti pada tabel 3.
Dari tabel 3 terlihat bahwa seperti pada dokumen berita, terjadi tingkat penurunan akurasi jika jumlah dokumen yang dijadikan pelatihan berkurang. Dengan jumlah dokumen pelatihan sebanyak 50% atau lebih akan memberikan akurasi 75% atau lebih.

Jika dibandingkan tingkat akurasi kategorisasi pada koleksi dokumen berita, maka koleksi dokumen akademik memberikan hasil dengan akurasi yang lebih rendah. Kemungkinan menurunnya akurasi ini adalah terkait dengan *feature* kata yang terpilih dalam dokumen akademik memberikan ambiguitas yang lebih tinggi dari pada feature kata dalam dokumen berita. Seperti kita ketahui tidak semua kata unik dapat digunakan sebagai feature dalam kategorisasi teks, tetapi kata-kata yang dipilih untuk dapat mewakili suatu kategori dari suatu teks.

Tabel 3. Akurasi Uji dokumen akademik

Dokumen Latih	Dokumen Uji	Akurasi (%)
405	45	82%
360	90	81%
315	135	78%
270	180	75%
225	225	75%
180	270	68%
135	315	65%

Dalam penelitian ini penulis menggunakan filter frekuensi dokumen sebagai kriteria apakah suatu kata dapat dimasukkan dalam vocabulary. Filter frekuensi dokumen mempertimbangkan frekuensi dokumen yang memuat kata tersebut. Asumsinya semakin banyak dokumen dalam suatu kategori memuat suatu kata, maka kata tersebut dianggap semakin baik untuk dijadikan sebagai feature pembeda kategori dokumen tersebut dengan kategori dokumen yang lain.

Berikut ini efek dari banyaknya feature kata berdasarkan filter frekuensi dokumen yang memuat kata tersebut. Untuk koleksi dokumen berita sebanyak 1000 dokumen tersedia feature kata unik adalah 18.139 kata. Untuk memilih dilakukan seleksi berdasarkan frekuensi dengan minimal dokumen memuat kata tersebut sebagai dasar pemilihan kata. Berikut ini efek pemilihan feature kata pada jumlah kata yang tepilih dan hasil akurasinya. Dokumen latih dipilih sebanyak 90%.

Tabel 4. Efek Pemilihan Kata pada Akurasi Pada Dokumen Berita 1000

Min Dok	Kata Terpilih	Akurasi
1	5970	87%
2	3994	89%
3	3038	90%
4	2471	91%
5	2092	88%
6	1787	87%
7	1577	87%

Dari tabel 4 terlihat bahwa dengan menggunakan filter minimal kata muncul dalam 4 dokumen maka kata tersebut dijadikan sebagai feature kategorisasi yang layak masuk dalam vocabulary. Dengan menggunakan filter 4 memberikan hasil akurasi yang paling tinggi. Dengan filter tersebut juga jumlah kata hanya sebanyak 2471, hanya sebesar 13,6% dari seluruh kata unik diluar STOP WORD, yaitu 18.139 kata.

Untuk dokumen akademik, efek pemilihan feature kata berdasarkan filter frekuensi dokumen yang memuat kata tersebut disajikan seperti dalam tabel 5. Pada koleksi 450 dokumen akademik, cacah kata unik semuanya diluar stopword adalah 6.200 kata. Pada kategorisasi digunakan 90% dokumen sebagai dokumen latih dan 10% sebagai dokumen uji.

Jika diperhatikan perubahan akurasi pada tabel 4, terlihat pola yang sama dengan tabel 4, yaitu jika digunakan cacah kata yang menggunakan filter frekuensi dokumen 1, 2, atau 3 akan mendapat hasil kinerja kategorisasi yang lebih kecil jika dibandingkan jika menggunakan filter 4 atau 5. Nilai akurasi optimal dicapai pada penggunaan filter kata minimal muncul di 5 dokumen.

Tabel 5. Efek Pemilihan Kata pada Akurasi Pada Dokumen Berita 450

Min Dok	Kata Terpilih	Akurasi
1	3233	70%
2	2038	71%
3	1536	72%
4	1188	74%
5	973	75%
6	814	73%
7	691	70%

KESIMPULAN

Algoritma NBC memiliki kinerja yang cukup tinggi untuk klasifikasi dokumen teks, baik dokumen berita maupun dokumen akademik. Pada klasifikasi dokumen berita didapatkan akurasi yang lebih tinggi (maksimal 91%) dibandingkan dengan dokumen akademik (maksimal 82%).

Baik pada dokumen berita maupun dokumen akademik, penggunaan 50% dokumen sebagai dokumen pelatihan memberikan kinerja akurasi diatas 75%.

Penggunaan kata unik dalam koleksi dokumen latih tanpa filter memberikan kinerja yang kurang optimal. Telah dicoba melakukan filter kata menggunakan frekuensi dokumen. Ditemukan filter minimal kata muncul dalam 4 atau 5 dokumen memberikan hasil akurasi yang paling tinggi dibandingkan dengan filter yang lain. Meskipun demikian belum dapat diketahui acuan batasan nilai minimal ini jika jumlah dokumen lebih banyak lagi.

Disarankan untuk mencari teknik yang lebih baik dalam melakukan seleksi feature kata yang digunakan sebagai dasar klasifikasi, karena ditemukan jumlah kata yang banyak dengan mengunakan seluruh kata unik dalam koleksi dokumen tidak memberikan hasil klasifikasi yang terbaik.

DAFTAR PUSTAKA

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)

Feldman, R. & Dagan, I., 1995 *Knowledge discovery in textual databases (KDT).* In proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117.

Hearst, M. A., 1997, *Text data mining: Issues, techniques, and the relationship to information access*. Presentation notes for UW/MS workshop on data mining, July 1997

Lewis, D. D. & Ringuette, M. A *Comparison of Two Learning Algorithms for Text Categorization.* In Third Annual Symposium on Document Analysis and Information Retrieval, 1994, p. 81-93 (on line : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.860; diases 29 SEptember 2012)

McCallum, A. and Nigam, K., 1998, *A comparison of event models for Naive Bayes text classification*, online pada http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.1529 diases 29 September 2012-10-10

Saraswati, N.W.S., 2011, *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.

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

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

Wulandini, F. & Nugroho, A. N. 2009. *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).

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).