Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government

Naïve Bayes Classification on Document Classification to Identify E-Government Content

Akhmad Pandhu Wijaya[1], Heru Agus Santoso[2]
[1,2]Jurusan Teknik Informatika, Universitas Dian Nuswantoro Semarang
Jl. Imam Bonjol 205-207 Semarang 50131
e-mail: [1] 111201106345@mhs.dinus.ac.id , [2]heru.agus.santoso@dsn.dinus.ac.id

Abstrak

Kebutuhan informasi adalah aspek penting yang harus dipertimbangkan, tidak semua informasi adalah informasi yang dibutuhkan. Karena banyaknya informasi digital dalam bahasa Indonesia, perlu untuk clustering dokumen berdasarkan apa yang dicari sehingga untuk mendapatkan beberapa informasi dapat dilakukan dengan sesuai, ringkas, menyeluruh, dan sesuai kebutuhan. Banyak penelitian tentang klasifikasi dokumen telah dibuat dan dikembangkan untuk mendapatkan hasil yang baik, penelitian pada beberapa website yang memiliki sumber informasi skala besar dan membutuhkan klasifikasi untuk mendapatkan informasi yang berkualitas dari situs yang diperiksa. Klasifikasi ini teknik dari data mining dan pertambangan teks juga digunakan untuk mencari atau mengatur kelas dibedakan dengan menggunakan beberapa fungsi dengan tujuan memungkinkan model untuk digunakan untuk data pengujian. Pada penelitian ini, objeknya adalah Situs Web Jawa Tengah dan diklasifikasikan oleh Naïve Bayes Classification (NBC). Dengan menggunakan metode ini diharapkan memfasilitasi klasifikasi dokumen bahasa Indonesia untuk identifikasi konten e-government.

Kata kunci— klasifikasi, dokumen, Naïve Bayes, e-government.

Abstract

Information's requirement is an important aspect that must be considered, not all available information is needed information. It's because bigger digital information in Indonesian language, need for document clustering based on what you're looking for so as to get some information can be done with a concise, thorough, and in accordance based on requirements. A lot of research on the documents classification has been made and developed in order to get good results, the research on some websites that have a large-scale source of information and requires classification to get quality information from the websites that are examined. The classification is technique from data mining and text mining is also used to locate or set of classes are distinguished by using several functions with purpose allowing the model to use for testing data. At this research, the object is Central Java Website and classified by NAIVE BAYES CLASSIFICATION (NBC), by use this methode are expected to facilitate the classification of Indonesian Language documents to Identify E-Government Content.

Keywords— classification, document, Naïve Bayes, e-government.

1. PENDAHULUAN

Informasi menjadi kebutuhan pokok bagi setiap orang, namun tidak semua informasi yang ada dapat menjadi kebutuhan. Dipengaruhi oleh kemajuan teknologi internet sehingga

informasi mengalami pelonjakan yang besar, sementara volume berita elektronik berbahasa Indonesia yang semakin besar adalahsumber informasi yang berharga, dan memungkinkan banyak pengguna informasi untuk merubah, memperbanyak, dan menghasilkan informasi baru. Sehingga dewasa ini perlu pencermatan lebih agar mendapatkan informasi yang relevan dan sesuai dengan apa yang diinginkan oleh pengguna informasi, pengelompokan berita dibutuhkan untuk mempermudah pencarian informasi mengenai suatu *event* tertentu [1].

Berbagai penelitian yang dilakukan oleh peneliti terdahulu mengenai text mining merupakan bukti banyaknya informasi media elektronik yang mengharuskan adanya pengembangan tentang proses penyaringan informasi secara berkala untuk menghasilkan informasi yang baik, serta dipengaruhi oleh permasalahan klasifikasi dokumen yang mendasar dan sangat penting. Dalam dokumen teks, tulisan yang terkandung adalah bahasa alami manusia, yang merupakan bahasa dengan struktur kompleks dan jumlah kata yang sangat banyak [1].

Salah satunya penelitian terhadap situs e-Government yang penulis beserta tim lakukan khususnya pada dokumen politik dan ekonomi, bertujuan untuk mengetahui sejauh mana perkembangan konten-konten politik dan ekonomi yang disediakan pada situs tersebut dan diharapkan mampu membantu menjadikan acuan bagi developer portal agar dapat memenajemen konten-konten yang terdapat di dalamnya dengan baik, serta menjadikan situs e-Goevernment lebih informatif. Pada situs e-Government banyak sekali informasi-informasi yang disertakan, untuk mengetahui tingkat efektivitas konten diperlukan pengolah informasi yang ada pada teks tersebut, pada penelitian ini penulis menggunakan metode *Naïve Bayes Classification*, penelitian ini berusaha untuk mengklasifikasikan dokumen dengan metode tersebut. Klasifikasi ini ditekankan untuk dokumen berbahasa Indonesia, sementara keterkaitan antar dokumen diukur berdasarkan probabilitas.

Erfian Junianto [2], dengan judul "Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan Naive Bayes Classifier". Penelitian terkait selanjutnya oleh Amir Hamzah [3], melakukan penelitian dengan judul "Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan *Abstract* Akademis*"*. Melihat hasl dari penelitian tersebut menjadikan NBC sebagai metode yang dipilih pada penelitian ini.

Penggunaan NBC pada penelitian ini diharapkan mampu menghasilkan data akurat agar dapat dijadikan bahan penelitian lebih lanjut, kelebihan NBC dibandingkan algoritma lain adalah pada kemampuannya mengklasifikasi dokumen dengan kesederhanaan dan kecepatan komputasinya namun memiliki komputasi tinggi, metode NBC juga memiliki kinerja yang baik terhadap pengklasifikasian data dokumen yang mengandung angka maupun teks. Sebelum tahap klasifikasi, dokumen harus dipresentasikan menjadi vektor [2].

Pengujian algoritma ini menggunakan dataset berupa dokumen dengan format HTML yang kemudian dilakukan perubahan ekstensi menjadi TXT dengan tujuan mempermudah pemrosesan. Sehingga penulis melakukan penelitian dengan judul "KLASIFIKASI DOKUMEN WEB DENGAN *NAIVE BAYES CLASSIFICATION* (NBC*)* UNTUK MENGETAHUI JENIS KONTEN *E-GOVERNMENT* ".

2. METODE PENELITIAN

Tekhnik pengolahan teks atau *Text minning* adalah cara yang digunakan untuk ekstraksi informasi yang lebih berkualitas dari dataset yang tersedia. Penelitian ini mengusulkan metode klasifikasi dengan algotitma *Naïve Bayes Classification* (NBC)

2.1 Teks mining

Teks mining secara umum adalah teori tentang pengolahan koleksi dokumen dalam jumlah besar yang ada dari waktu ke waktu dengan menggunakan beberapa analisis, tujuan pengolahan teks tersebut adalah mengetahui dan mengekstrak informasi yang berguna dari

sumber data dengan identifikasi dan eksplorasi pola menarik dalam kasus text mining, sumber data yang dipergunakan adalah kumpulan atau koleksi dokumen tidak terstuktur dan memerlukan adanya pengelompokan untuk diketahui informasi sejenis.

Text mining terdiri dari 3 proses yang biasa dilakukan [1], ketiga proses tersebut adalah sebagai berikut

1. *Characterization of data*

   Seluruh teks yang akan diproses distrukturkan terlebih dahulu dikarenakan tardapat tag HTML yang tidak dibutuhkan, proses tersubut menggunakan parsing dan dimasukan ke dalam sebuah database.
2. *Data mining*

   Dari data yang ada kemudian dilakukan pencarian dengan algoritma tertentu untuk mendapatkan pola dari data tersebut.
3. *Data visualization*

   Hasil pencarian yang ada akan menghasilkan output dalam bentuk teks yang dapat dipahami dengan mudah.

Text mining adalah bidang khusus dari data mining, hanya saja yang menbedakan adalah jenis datasetnya, Pada data *mining* terdapat *dataset* dipergunakan seperti data terstruktur, sementara pada *text mining* data yang dipergunakan adalah dataset yang tidak terstruktur berupa teks.

2.2 Bahasa Indonesia

Bahasa Indonesia adalah bahasa nasional yang digunakan di Indonesia, bahasa ini adalah dasar dari bahasa Melayu dan telah dimoderenkan dan dikembangkan sekian lamanya. Bahasa Indonesia atau diartikan sebagai Bahasa, standar untuk penulisan dan pengucapan yang ditulis pada (Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan) [4].Panduan tersebut menjelaskan bagaimana penulsan yang benar, penggunaan tanda baca, huruf capital dan cetak miring, juga penulisan kata sebaik *adaptive words.*

Pada Bahasa Indonesia, imbuhan kata (awalan, akhiran) dapat ditemukan hamper di setiap kata, imbuhan digunakan untuk kata jadi dan dapat memiliki arti yang berbeda tergantung apa dan bagaimana imbuhan diletakkan.Imbuhan dalam Bahasa Indonesia dibagi menjadi 3 yaitu : imbuhan sederhana, imbuhan terkombinasi, imbuhan khusus [4].

2.3 Text preprocessing

Pada text preprocessing, terdapat beberapa langkah seperti tokenizing, stopword, filtering, stemming, word frequency counting, computation ofTF-IDFfeature, dan normalization.[5].

1. *Tokenizing*

Pada *tokenizing* terdapat bebearapa proses yang harus dilakukan adalah merubah semua huruf besar menjadi kecil (*text to lowercase*). Proses selanjutnya adalah penguraian*,* proses penguraian yang dimaksud adalah membagi teks menjadi kumpulan kata tanpa memperhatikan keterhubungan dantara kata satu dengan yang lain serta peran dan posisinya pada kalimat, karakter diterima dalam kumpulan kata menurut abjad. Sedangkan untuk perulangan kata dalam Bahasa Indonesia akan terbagi menjadi dua kata.

2. *Stopword filtering*

Proses selanjutnya adalah memeriksa *stop word list, stopeord list* adalah daftar kata-kata yang semestinya dihilangkan, jika kata pada dataset terdapat pada *stop word list* maka kata akan dihilangkan. Tetapi jika tidak terdapat di dalamnya maka proses akan beranjut tanpa menghilangkan kata pada dokumen.

3. *Word frequency training*

Kata-kata yang telah selesai dilakukan proses stemming kemudian disimpan sebagai data percobaan, setiap kata pada data percobaan dirubah menjadi format yang tidak diketahui

oleh metode untuk selanjutnya dijadikan sebagai data masukan untuk prosses pembelajaran dengan metode terkait. Proses tersebut mencari 3 frekuensi kata pada setiap dokumen.

4. TF-IDF *features*

Setiap dokumen diwakili oleh vektor dengan pengenalan elemen-elemen yang dikenali dari tahap ekstraksi dari dokumen.Vektor yang terdiri dari bobot setiap pemberhentian yang menggunakan dasar perhitungan pada metode TF-IDF.TF-IDF adalah metode pembobotan yang mengaitkan antara *term frequency* (TF) dan *inverse document frequensi* (IDF). Langkah awal pada pembobotan TF-IDF adalah menemukan nomor kata yang diketahui sebagai bobot atau *frequency term* di setiap dikumen setelah dilakukan pengalian oleh *inverse deocument frequency*. Adapun rumus untuk menemukan bobot dari kata menggunakan TF-IDF adalah :

a. TF (*Term Frequency*)

Term Frequency adalah cara pembobotan *term* (kata) yang paling sederhana [1]. Bobot kata t pada dokumen diberikan dengan :

$$w_{ij} = tf_{ij} . idf \quad (1)$$

b. IDF (*Inverse Document Frequency*)

Jika TF memperhatikan kemunculan kata dalam dokumen, IDF memperhatikan kemunculan kata pada kumpulan dokumen [1]. Faktor IDF pada suatu kata t diberikan oleh :

$$idf = log \frac{N}{df_j} \quad (2)$$

Dimana $w_{i\,j}$ adalah bobot kata *i* pada dokumen *j,* semantara N adalah jumlah dokumen, dan *term frequency* adalah $tf_{i\,j}$ adalah jumlah dari kemunculan kata *i* pada dokumen *j*, df_j *(document frequency)* adalah jumlah dokumen *j* yang berisi kata *i*.

2.4 Naïve Bayes Classification (NBC)

Klasifikasi adalah proses untuk menentukan model atau fungsi yang membedakan konsep atau kelas data [1], dengan tujuan untuk memperkirakan kelas yang tidak tersedia pada objek, dalam pengklasifikasian terdapat 2 proses yang dilakukan yaitu :

1. Proses *training*

   Pada proses ini dilakukan *training set* yang sudah diketahui label-labelnya untuk membangun model.
2. Proses *testing*

   Proses ini intuk mengetahui keakuratan model yang dibangun pada proses *training*, umumnya digunakan data yang disebut *test set* untuk memprediksi label.

Metode NBC terdiri dari dua tahap dalam proses klasifikasi teks, tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap *sample* dokumen berupa pemilihan *vocabulary* yaitu kata yang dimungkinkan muncul dalam koleksi dokumen *sample* yang menjadi representasi dokumen. Langkah selanjutnya adalah menentukan probabilitas bagi tiap kategori berdasarkan sampel dokumen. *Naïve Bayes* membangun model probabilistik dari *term documents matrix* data *labeled.*

Klasifikasi dokumen dilakukan dengan terlebih dahulu menentukan kategori *c* $\varepsilon C = \{ c_1, c_2, c_3, \dots, c_n \}$dari suatu dokumen *d* $\varepsilon D = \{ d_1, d_2, d_3, \dots, d_n \}$ berdasarkan kata – kata yang ada pada dokumen. Proses penentuan ketegori dari sebuah dokumen dilakukan dengan melakukan perhitungan menggunakan persamaan sebagai berikut :

$$c^* = \arg\max_{ci \in C} p\left(c_i \middle| d_j\right)$$

$$= \arg\max_{ci \in C} \prod_k p\left(w_{kj} \middle| c_i\right) x\, p(c_i)$$

dimana w_{kj} adalah fitur atau kata dari dokumen d_j yang ingin diketahui kategorinya. Nilai $p\left(w_{kj}|c_i\right)$ diketahui dari data *training* yang dilimiliki.

2.5 Accuracy

Metode evaluasidipergunakan untuk mengukur keakuratan hasil klasifikasi, digunakan perhitungan accuracy. Mengevaluasi banyaknya label prediksi yang sesuai dengan label actual . Semakin besar nilai accuracy, maka perormansi ckassifier semakin bagus.

$$\text{Accuracy} = \frac{Jumlah\ dokumen\ terklasifikasi\ dengan\ benar}{Jumlah\ dokumen\ keseluruhan} \text{ x } 100$$

3. HASIL DAN PEMBAHASAN

Pada bab ini akan dibahas mengenai penjelasan langkah-langkah dalam persiapan dokumen, tahap ini meliputi *converting* dan *filtering* kemudian pemrosesan file seperti proses pengenalan pola klasifikasi, metode pengukuran dan hasil pengukuran, kualitas informasi pada klasifikasi dokumen menggunakan metode *Naïve Bayes Classification*.

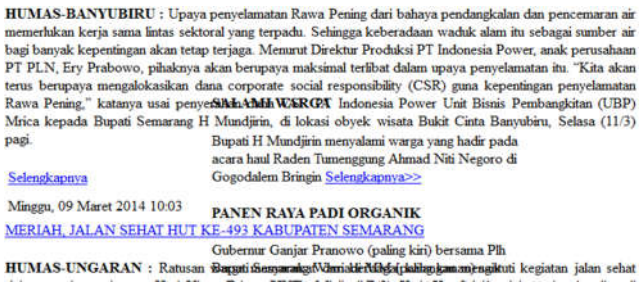
3.1 Dokumen yang Digunakan

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 merepresentasikan dokumen, pada tahap pelatihan terdapat dokumen *training* yang menjadi acuan untuk proses *testing*.

1. Dokumen *training*

   Berfungsi untuk pembentukan kelas dan sebagai acuan bagaimana dokumen akan diklasifikasikan, dalam penelitian ini penulis menggunakan sumber data yang telah diklasifikasikan menjadi dokumen politik dan ekonomi pada portal www.jawapos.com, acuan yang dituju adalah pelabelan dokumen berdasarkan *domain expert*.
2. Dokumen *testing*

   Dalam penelitian yang dilakukan, jenis dokumen yang digunakan dalam penelitian ini yaitu dokumen website dalam bentuk ekstensi html yang berisi tambang informasi dan didapatkan dengan isi yang tidak terstruktur dikarenakan terdapat tag-tag html yang menjadikan dokumen pada penelitian tidak bermakna, sementara untuk keakuratan klasifikasi dibutuhkan dokumen yang terstruktur dan dapat dipahami isinya. Dokumen website yang digunakan adalah dokumen asli yang tercantum pada website. Dokumen percobaan adalah dokumen website PemerintahKabupaten Semarang dengan jumlah dokumen sebanyak 2781 dokumen html.


HUMAS-BANYUBIRU : Upaya penyelamatan Rawa Pening dari bahaya pendangkalan dan pencemaran air memerlukan kerja sama lintas sektoral yang terpadu. Sehingga keberadaan waduk alam itu sebagai sumber air bagi banyak kepentingan akan tetap terjaga. Menurut Direktur Produksi PT Indonesia Power, anak perusahaan PT PLN, Ery Prabowo, pihaknya akan berupaya maksimal terlibat dalam upaya penyelamatan itu. "Kita akan terus berupaya mengalokasikan dana corporate social responsibility (CSR) guna kepentingan penyelamatan Rawa Pening," katanya usai penyerahan CSR PT Indonesia Power Unit Bisnis Pembangkitan (UBP) Mrica kepada Bupati Semarang H Mundjirin, di lokasi obyek wisata Bukit Cinta Banyubiru, Selasa (11/3) pagi.

Selengkapnya

Minggu, 09 Maret 2014 10:03

MERIAH, JALAN SEHAT HUT KE-493 KABUPATEN SEMARANG

HUMAS-UNGARAN : Ratusan

Bupati H Mundjirin menyalami warga yang hadir pada acara haul Raden Tumenggung Ahmad Niti Negoro di Gogodalem Bringin Selengkapnya>>

PANEN RAYA PADI ORGANIK

Gubernur Ganjar Pranowo (paling kiri) bersama Plh

Gambar 1. Isi dokumen

3.2 Preprocessing Dokumen

Tahap yang dilakukan sebelum proses klasifikasi adalah *preprocessing* untuk mencari makna pada dokumen *training* maupun *testing* dan mendukung proses klasifikasi, proses ini harus dilakukan karena pada data uji dokumen berupa paragraf beserta tag-tag yang menghilangkan arti dari dokumen tersebut. Penulis mengalami kesulitan dalam memahami isi dokumen uji sebelum dilakukan proses *preprocessing*. *Preprocessing* juga dapat mempengaruhi identifikasi teks yang bertujuan menentukan fitur. Hal pertama dalam pemrosesan dokumen adalah memecah kumpulan karakter ke dalam kata atau token, sering disebut sebagai tokenisasi. Tokenisasi adalah hal yang kompleks untuk program komputer karena beberapa karakter dapat ditemukan sebagai *token delimiters*. Delimiter adalah karakter spasi, tab, dan baris, sedangkan karakter ( ) < > ! ? " kadangkala dijadikan delimiter namun tergantung pada lingkungannya [6].

3.3 Proses Identifikasi

Proses identifikasi teks sangatlah penting untuk mengenali pola teks yang akan diklasifikasikan dan mengenali jenis – jenis teks yang akan digunakan sebagai *training*. Permasalahan yang timbul saat identifikasi adalah tidak teraturnya pola teks yang didapatkan meski telah diproses menggunakan *stopwords* pada langkah sebelumnya, hal ini mengakibatkan penulis sedikit kesuliatan dalam mengidentifikasi teks dan memerlukan ketelitian dalam pengamatan.

Pada proses identifikasi yang dilakukan penulis perlu membuka dokumen satu persatu untuk memahami pola yang ada pada teks tersebut, untuk pola sendiri didapatkan tidak beraturan dalam peletakan konten.

3.4 Proses training label

Proses penentuan label pada dokumen training dilakukan secara manual berdasarkan domain expert yang penulis ambil dari www.jawapos.com berdasarkan kategori yang telah ditentukan pada domain tersebut. Penentuan label digunakan untuk memberikan acuan pada proses klasifikasi dokumen atau mengelompokkan sesuai dengan kategori label yang telah ditentukan sebelumnya. Berdasarkan hasil identifikasi dokumen yang mengacu pada konten yang terdapat pada dokumen, data akan diklasifikasikan menjadi dua kelas, kelas politik dan kelas ekonomi.

3.5 Penentuan fitur

Penentuan fitur dapat dilakukan dengan term frekuensi, pembuktian dari eksperimental hanya 10% memilih kata-kata yang sering muncul, akantetapi hal tersebut tidak mempengaruhi proses klasifikasi [1].

Tabel 1. Fitur klasifikasi

Doc/fitur	Dok1	Dok2	Dok3	Dok4	Dok5
harga	$w_{a,1}$	$w_{a,2}$	$w_{a,3}$	$w_{a,4}$	$w_{a,5}$
pasar	$w_{b,1}$	$w_{b,2}$	$w_{b,2}$	$w_{b,4}$	$w_{b,5}$
ekonomi	$w_{c,1}$	$w_{c,2}$	$w_{c,3}$	$w_{c,4}$	$w_{c,5}$
daerah	$w_{d,1}$	$w_{d,2}$	$w_{d,3}$	$w_{d,4}$	$w_{d,5}$
partai	$w_{e,1}$	$w_{e,2}$	$w_{e,3}$	$w_{e,4}$	$w_{e,5}$
politik	$w_{f,1}$	$w_{f,2}$	$w_{f,3}$	$w_{f,4}$	$w_{f,5}$
dpr	$w_{g,1}$	$w_{g,2}$	$w_{g,3}$	$w_{g,4}$	$w_{g,5}$
pemilu	$w_{h,1}$	$w_{h,2}$	$w_{h,3}$	$w_{h,4}$	$w_{h,5}$

3.6 Pembobotan tf-idf

Pengambilan citra user juga dipengaruhi jarak pengambilan antar user dengan webcam di bawah ini merupakan hasil pengukuran jarak antar fitur wajah dengan jarak pengambilan 40 cm dari webcam.

3.7 Klasifikasi dokumen

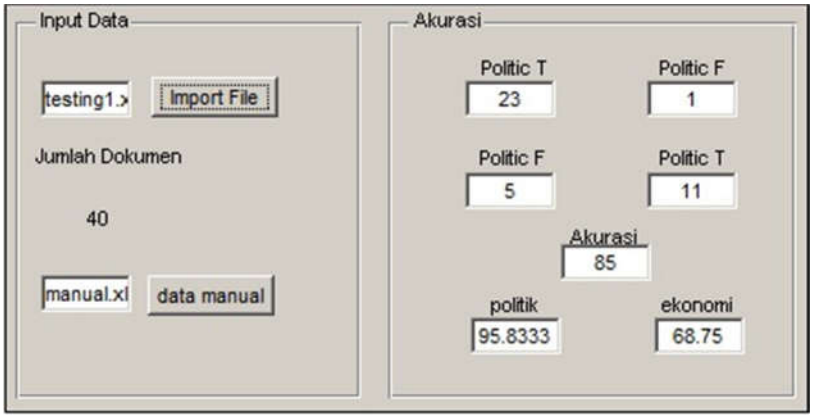
Proses klasifikasi dokumen membutuhkan perhitungan yang melibatkan jumlah dokumen label x, jumlah dokumen label y, serta jumlah keseluruhan dokumen *training*, yang disebut $p(c_i)$ yaitu pada kategori x pembagian jumlah dokumen berkategori x dengan jumlah seluruh data *training,* serupa dengan kategori y adalah pembagian jumlah dokumen berkategori y dengan jumlah seluruh data *training*.

Tabel 2. Klasifikasi dokumen

Kategori		politik	ekonomi
$p(c_i)$		0.53	0.46
$p(w_{kj}\|c_i)$	harga	12.95	533.14
	industri	13.56	245.40
	pasar	7.39	219.42
	ekonomi	22.17	171.10
	daerah	129.50	48.25
	partai	460.59	0.75
	politik	173.18	16.35
	dpr	299.56	12.92
	pemilu	144.48	4.95

3.8 Akurasi

Pada penelitian denggan dokumen *testing* sebanyak 40 dokumen yang mengacu pada 260 dokumen politik dan 222 dokumen ekonomi sebagai data *training* menghasilkan akurasi yang baik pada dokumen politik sebesar 95.8% sedangkan pada dokumen ekonomi hanya 68.75%.



Gambar 2. Akurasi

4. KESIMPULAN

Dari pembahasan seperti yang dikemukakan pada bab sebelumnya, maka penulis dapat mengambil kesimpulan sebagai berikut :

Tekhnik klasifikasi dokumen dengan NBC dan pembobotan fitur metode *tf-idf* menghasilkan nilai yang pasti dan akurasi yang baik karena bobot memperkecil kemungkinan kesalahan pada pengklasifikasian, fitur yang mempunyai frekuensi tertentu dapat mempengaruhi keakuratan dalam klasifikasi bergantung pada frekuensi fitur dan dokumen yang mengandung fitur tersebut.

Hasil dari klasifikasi dokumen menggunakan NBC pada penelitian ini dengan data *training* sebanyak 260 dokumen politik dan 222 dokumen ekonomi menggunakan 40 data *testing* menunjukkan nilai akurasi yang baik pada keseluruhan klasifikasi, dengan akurasi keseluruhan klasifikasi sebesar 85%.

5. SARAN

Adapun saran-saran yang dapat diberikan dalam penelitian ini untuk pengembangan lebih lanjut agar meningkatkan kualitas dan fungsionalitas dari metode pengklasifikasian dokumen ini, adalah sebagai berikut :

1. Memperbaiki pengolahan dan identifikasi dokumen serta mengembangkan tahap preprocessing dengan menyeleksi lebih banyak kata-kata yang dianggap tidak perlu ada pada dokumen untuk meningkatkan proses klasifikasi dokumen.
2. Penelitian ini mengklasifikasikan dokumen menggunakan Naive Bayes Classification dengan kombinasi pembobotan kata menggunakan metode tf-idf , dalam penelitian selanjutnya dapat dikembangkan dengan metode klasifikasi lainnya seperti Support Vector Machine, Neural Network.
3. Pada pemilihan fitur penelitian ini menggunakan metode term frequency, untuk penelitian selanjutnya dapat menggunakan metode lainnya seperti metode chi-square, expected cross entropy, odds ratio, the weight of evidence of text dan sebagainya.

DAFTAR PUSTAKA

[1] L. Novianti, A. Ardiyanti dan A. Prima, "Pengklasifikasian Dokumen Berita Berbahasa Indonesia Menggunakan Latent Semantic Indexing (LSI) dan Support Vector Machine (SVM)," *ISSN:1979-911X,* 2012.

[2] E. Junianto, "Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan Naive Bayes Classifier," Program Pascasarjana Magister Ilmu Komputer Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, 2014.

[3] A. Hamzah, "Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan Akademis," *ISSN:1979-911I,* vol. 3, 2012.

[4] F. Wulandari dan A. S. Nugroho, "Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases," International Conference on Rural Information and Communication Technology, 2009.

[5] D. Y. Liliana, A. Hardianto dan M.Ridok, "Indonesian News Classification using Support Vector Machine," *Worls Academy of Science,* vol. 5, 2011.

[6] D. Isa, L. H. Lee, V. P. Kallimani dan R. Rajkumar, "Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model," *Computer and Information Science,* vol. 1, no. 4, 2008.