KLASIFIKASI DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN NAIVE BAYES CLASSIFIER

Rusdi Efendi, Reza Firsandaya Malik[1)], Jeni MilaSari U
[1]rezafm@unsri.ac.id

ABSTRACT

Document classification is a research field in information retrieval, by developing the method for categorize a document into one or more categories based on the content. Document classification aims for classify the unstructure document into category which describes the content of document. The document used can be a text document like news article. To solve the problem, the software is developed by using Naive Bayes Classifier with simple computation and quite high accuration level. There are three levels in doing document classification, preprocessing, training and classification. This research uses secondary data from online news portal with various categories such as economy, health, sport, technology, politics and education. The classification accuration level from the software is 86,67%, by using 60 documents, includes 30 train documents and 30 test documents. The error of classification is 13,33% lies on the politics and economy documents, because there are words dominant to both of those categories.

Keywords: Document Classification, Naive Bayes Classifier

I. PENDAHULUAN

Teknologi informasi telah berkembang sangat pesat hingga saat ini. Transmisi data dalam jaringan komputer dan internet semakin berkembang. Dokumen yang beredar di dunia maya terus tumbuh dan lebih mungkin menjadi kurang efektif dalam pencarian dan penyajian informasi [1]. Oleh karena itu, klasifikasi dokumen ke dalam kategori yang sesuai diperlukan untuk meningkatkan efektivitas dan efisiensi manajemen dan organisasi dokumen teks. Hal ini, mendukung kebutuhan untuk proses klasifikasi informasi – informasi [2].

Salah satu alternatif dalam memecahkan masalah adalah dengan mengelompokkan dokumen. Klasifikasi dokumen adalah bidang penelitian dalam perolehan informasi dengan mengembangkan metode untuk menentukan atau mengkategorikan dokumen ke dalam satu atau lebih kelompok yang sebelumnya telah diakui secara otomatis berdasarkan isi dokumen [3]. Klasifikasi dokumen bertujuan untuk mengklasifikasikan dokumen tidak terstruktur ke dalam kelompok yang menggambarkan isi dari dokumen. Dokumen dapat berupa teks dokumen seperti artikel berita.

Naive Bayes Classifier algoritma untuk mengklasifikasikan dokumen dalam berita itu, metode ini bekerja berdasarkan probabilitas dari sebuah dokumen di hadapan kata-kata yang sama dalam dokumen lain yang dalam kategori tersebut. Algoritma ini cocok untuk digunakan dalam mengklasifikasikan dokumen karena memiliki komputasi yang sangat sederhana dan memiliki akurasi yang tinggi [4]. Oleh karena itu, dalam penelitian ini akan menerapkan metode Naive Bayes dalam klasifikasi dokumen ditekankan dalam dokumen berita dan berita dalam bahasa Indonesia.

II. METODOLOGI PENELITIAN

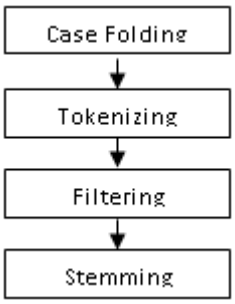
II.1. Teks Preprocessing

Cara yang digunakan dalam mempelajari suatu data teks, adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur-fitur yang mewakili, diperlukan tahapan ekstraksi yang dilakukan secara umum pada dokumen, yaitu tokenizing, filtering, stemming [5]. Proses ekstraksi dokumen ditunjukkan pada Gambar 1. Berikut penjelasan dari masing-masing komponen.

a. Tahap case folding adalah proses mengubah huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang

diterima. Karakter selain huruf dihilangkan dan dianggap sebagai delimiter.

b. Tahap tokenizing / parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
c. Tahap filtering adalah tahap mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma stoplist (membuang kata yang kurang penting). Stoplist / stopword adalah kata - kata yang tidak deskriptif yang dapat dibuang. Contoh stopwords adalah "yang", "dan", "di", "dari" dan seterusnya.
d. Tahap stemming adalah tahap mencari root kata dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki persamaan bentuk baku yang permanen.



Gambar 1.Text Preprocessing

II.2. Algoritma Naive Bayes Classifier

Metode ini menggunakan perhitungan probabilitas, tidak memperhatikan urutan kemunculan kata pada dokumen teks dan menganggap sebuah dokumen teks sebagai kumpulan dari kata-kata yang menyusun dokumen teks tersebut [6]. Naive Bayes merupakan salah satu contoh dari metode supervised document classification yang berarti membutuhkan data latih dalam melakukan klasifikasi.

2.2.1. Proses Pelatihan

Pada saat pelatihan, dokumen yang akan diinputkan terlebih dahulu telah diketahui kategorinya. Dokumen yang digunakan pada saat pelatihan disebut dokumen contoh. Dokumen contoh terlebih dahulu mengalami praproses untuk masuk ke proses ini. Proses pelatihan berguna untuk membentuk pengetahuan berupa nilai probabilitas kata. Pada proses pelatihan ini, memiliki modul yang hampir sama dengan tahap klasifikasi. Bedanya, hanya pada proses pelatihan tidak menjalankan modul klasifikasi, tetapi hanya menghasilkan dokumen yang mengandung kata – kata untuk mengkarakteristik suatu kategori.

Pada setiap kata yang muncul pada dokumen latih, hitung nilai probabilitas masing – masing katanya dengan persamaan 1. Setelah itu, hitung probabilitas kategori dengan menggunakan persamaan 2.

$$P\ (w_k\ |v_j\ )\ = \frac{n_k+1}{n+\ |kosakata|} \qquad (1)$$

$P\ (w_k\ |v_j\ )$: Probabilitas kata - kata pada sebuah kategori
n_k: Frekuensi kemunculan kata pada sebuah kategori
n: Jumlah seluruh kata pada dokumen dalam suatu kategori
$|kosakata|$: jumlah total kata (*distinc*) pada semua data latihan.

$$P(v_j) = \frac{|docs_j|}{|\text{contoh}|} \qquad (2)$$

Keterangan :
$P(v_j)$: Probabilitas dokumen kategori
$|docs_j|$: Jumlah seluruh dokumen pada sebuah kategori
|Sample|: Jumlah keseluruhan data yang dilatih

2.2.2. Proses Klasifikasi

Pada proses klasifikasi, dokumen yang diinputkan belum diketahui kategorinya. Sama seperti pada saat proses pelatihan, proses klasifikasi ini pun harus melewati praproses terlebih dahulu. Pada proses ini, dilakukan tahap

mencari kata – kata yang ada pada dokumen input sesuai dengan pengetahuan kata di data latih yang disebut P($a_i|v_j$). Lalu gunakan nilai P(v_j) yaitu probabilitas dokumen yang telah diperoleh pada proses pelatihan yang terlebih dahulu telah disimpan di pengetahuan pada saat pelatihan, maka untuk setiap kategori hitung $P(v_j)\prod_i P(a_i|v_j)$. Untuk mendapatkan nilai tersebut, terlebih dahulu menghitung nilai $\prod_i$ P($a_i|v_j$) yaitu kumulatif perkalian probabilitas kemunculan kata sama pada data latih lalu dikalikan lagi dengan nilai probabilitas dokumen kategorinya masing - masing P(v_j) .

Setelah hasil perkalian didapatkan dan dilakukan pada masing – masing kategori maka akan dibandingkan dan dicari nilai terbesar untuk mengklasifikasikan data uji pada dokumen berita yang akan masuk ke dalam kategori tesebut, sesuai dengan persamaan 3 berikut.

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j)\prod_i P(a_i|v_j) \qquad (3)$$

III. HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian yang telah dilakukan pada tahap sebelumnya, dapat disimpulkan bahwa implementasi unit dan antar muka perangkat lunak berjalan dengan baik. Hal ini ditandai dengan kesimpulan hasil skenario pada kasus uji semuanya memberikan kesimpulan yang sama, yaitu diterima.

Pada subbab ini, akan dilakukan analisis hasil pengujian dari perangkat lunak yang telah dibangun. Pengujian keakuratan dalam melakukan klasifikasi dokumen berita menggunakan artikel dalam bentuk file teks. Di dalam pengujian ini menggunakan 60 dokumen berita dengan berbagai kategori dari media elektronik yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan. Dokumen berita yang digunakan terdiri dari dua bagian, dokumen pelatihan dan dokumen pengujian. Dokumen pelatihan berperan sebagai data contoh yang akan digunakan dalam proses pelatihan. Sedangkan dokumen pengujian digunakan sebagai data pengujian untuk melihat tingkat akurasi.

Dalam pengujian ini, digunakan dokumen pelatihan sebanyak 30 dokumen dan dokumen pengujian 30 dokumen. Rata – rata ukuran dokumen pelatihan sebesar 2 kb dengan jumlah kata yang terkandung didalamnya sebanyak ±200 kata. Pada setiap proses klasifikasi ataupun pelatihan, semua dokumen yang digunakan harus melewati proses text mining terlebih dahulu, yaitu proses tokenizing (pemecahan kata), filtering (penyaringan kata), stemming (penghilangan imbuhan). Pada proses pelatihan yang dilakukan terbentuklah 1272 token atau kosakata.

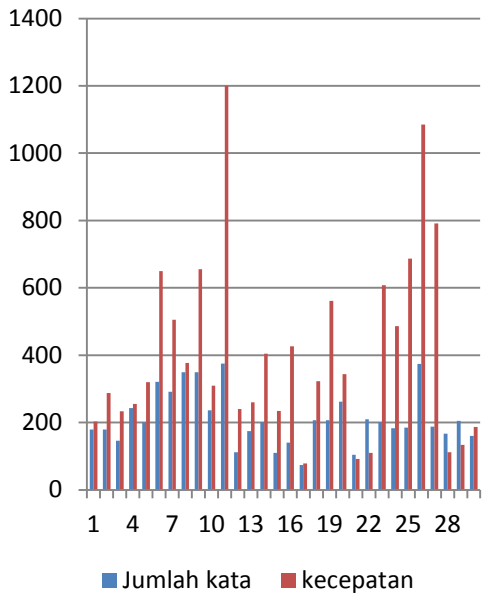
Pada beberapa kasus, jika contoh dokumen berita yang dimasukkan memiliki kata – kata yang dominan ke arah kategori yang lain maka dokumen berita akan salah diklasifikasikan dan tidak sesuai dengan kategorinya. Sebagai contoh dokumen berita ekonomi berikut dengan judul berita "Defisit anggaran dan rasio utang terhadap PDB yang rendah". Pada kasus ini, dokumen ekonomi tersebut dilakukan klasifikasi berita dengan menggunakan program aplikasi yang dibuat dan menunjukkan bahwa dokumen berita ekonomi tersebut salah diklasifikasikan ke berita politik. Adapun hasil pengujian ditampilkan dalam bentuk Tabel 1. Untuk menghitung persentase keakuratan dilakukan dengan cara:

$$\frac{\text{Jumlah dokumen yang benar dikenali}}{\text{Jumlah dokumen pengujian}} \text{x}100\% \qquad (4)$$

$$= \frac{26}{30} \text{ x } 100$$

$$= 86{,}67\ \%$$

Dari hasil pengujian, keakuratan yang didapat oleh aplikasi ini mencapai 86,67 %. Hasil 86,67 % didapat dari tabel 1, dimana jumlah pengujian 30 contoh dokumen uji dan hanya 26 contoh dokumen uji yang memiliki nilai benar. Hal ini dikarenakan dokumen pelatihan yang sedikit menyebabkan kurangnya kata – kata yang penting yang mencirikan suatu dokumen dan juga terdapat kata – kata yang dominan ke kategori lain yang bukan kategorinya sehingga dapat menimbulkan kesalahan dalam pengklasifikasian dokumen. Sedangkan, dalam hasil pengujian terhadap banyaknya kata dan waktu yang diperlukan dalam praproses sebuah dokumen berita dapat dilihat pada Gambar 2.



Gambar 2. Pengaruh Waktu Praproses dengan Jumlah Kata pada Dokumen

Berdasarkan Gambar 2 di atas, batang biru menggambarkan jumlah kata dan batang merah menggambarkan kecepatan yang menunjukan menunjukkan bahwa, dokumen yang memiliki kata – kata yang banyak memiliki waktu praproses yang lama terlihat pada dokumen 11 dengan judul berita "Dua Atlet Andalan Kembali Perkuat Jabar" memiliki kata – kata yang berjumlah 375 dengan lamanya waktu yang diperlukan dalam praproses yaitu 1200 ms. Namun frekuensi kata yang dibutuhkan saat proses stemming pada tiap dokumen juga berpengaruh terhadap waktu praproses, sehingga waktu praproses pada tiap dokumen dapat lebih lama dibandingkan dokumen dengan frekuensi kata yang lebih banyak namun memiliki frekuensi kata saat stemming lebih sedikit.

Misalnya, pada dokumen ke-27 memiliki frekuensi kata 188 dengan waktu proses stemming 791 ms, sedangkan pada dokumen ke-28 memiliki frekuensi kata 167 dengan waktu proses stemming 112 ms. Pada dokumen ke-27 memiliki frekuensi kata yang diperlukan proses stemming lebih banyak dibandingkan dengan dokumen ke-28, dokumen ke-27 memiliki frekuensi kata yang dibutuhkan saat proses stemming sebanyak 107 sedangkan dokumen ke-28 memiliki frekuensi kata yang dibutuhkan saat proses stemming sebanyak 92, sehingga diperlukan waktu yang lebih lama untuk melakukan proses stemming.

IV. KESIMPULAN

Berikut ini beberapa kesimpulan yang dapat dibuat dari penelitian ini:

1. Pengembangan perangkat lunak pada penelitian ini telah dapat mengklasifikasikan dokumen berita. Berdasarkan penelitian yang diperoleh tingkat akurasi perangkat lunak ini mencapai 86,67% dengan menggunakan 60 dokumen yang terdiri dari 30 dokumen latih dan 30 dokumen uji untuk enam kategori berita yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan.
2. Dalam penelitian ini, kesalahan klasifikasi sebesar 13,33 % terletak pada dokumen yang berkategori politik dan ekonomi, dikarenakan cukup banyaknya kata-kata yang dominan ke arah kedua kategori tersebut.
3. Pada penelitian ini, menunjukkan bahwa pengaruh kecepatan/waktu yang diperlukan pada praproses tidak selalu berdasarkan pada banyaknya kata pada dokumen tetapi berdasarkan banyaknya kata – kata yang diproses pada tahap stemming.

DAFTAR PUSTAKA

[1] Hamzah, A., Adhi Susanto, F.Soesianto, Jazi Eko Istiyanto, 2007, "Studi Komparasi Algoritma *Hierarchical* Dan *Partitional* Untuk *Clustering* Dokumen Teks Berbahasa Indonesia", Academia Ista.

[2] Natalius,S. 2010. Metode *Naïve Bayes Classifier* dan Penggunaannya pada Klasifikasi Dokumen. Departemen Teknik Informatika, Sekolah Teknik Elektro Informatika, Institut Teknologi Bandung.

[3] Ramadan, R. 2006. Penerapan Pohon Untuk Klasifikasi Dokumen Teks Berbahasa Inggris. Departemen Teknik Informatika, Sekolah Teknik Elektro Informatika, Institut Teknologi Bandung.

[4] Kang,D., Honavar,V., and Silvescu,A. 2006. A Recursive Naïve Bayes Learner for Sequence Classifiication, Artificial Intelligence Research Laboratory, IOWA State University.
[5] Mooney, J. 2006. CS 391L : Machine Learning Text Categorization. Austin: University of Texas.
[6] Wibisono,Y. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan *Naïve Bayes Classifier*. Departemen Pendidikan Matematika, FMIPA UPI.

No.	Document	Classification Result	Truth Value	Number of Words	Time (ms)
1	Defisit anggaran dan rasio utang terhadap PDB yang rendah.txt	Politic	False	179	203
2	Harga Minyak Lebih Tinggi di Perdagangan Asia.txt	Economic	True	179	288
3	SERAPAN APBN 30% Anggaran di Kemenhub tak terserap.txt	Politic	False	146	233
4	Harga Minyak Lebih Tinggi di Perdagangan Asia.txt	Economic	True	243	255
5	Harga Gula di Solo Mulai Tembus Rp11 RibuKg.txt	Economic	False	200	320
6	Langkah Ini Bisa Bantu Membersihkan Paru-paru.txt	Health	True	321	650
7	Masyarakat Harus Peduli Obat Generik.txt	Health	True	291	505
8	Pasang Gigi Palsu di Tukang Gigi Seperti Memasang Ranjau dalam Mulut.txt	Health	True	349	377
9	Taman Bermain Tradisional dapat Mengurangi Obesitas pada Anak.txt	Health	True	349	655
10	Zat Pewarna Soda Memicu Kanker.txt	Health	True	236	309
11	Dua Atlet Andalan Kembali Perkuat Jabar.txt	Sport	True	375	1200
12	Gol Telat Toto Salvio Menangkan Atletico Madrid.txt	Sport	True	112	240
13	Kalahkan Lakers, Thunder Kuasai Wilayah Barat.txt	Sport	True	174	260
14	Lakers dan Mavericks Telan Kekalahan.txt	Sport	True	201	404
15	Messi Leg Kedua Kami Akan Menang.txt	Sport	True	110	234
16	Akses Siswa Miskin ke PTN Semakin Sempit.txt	Education	True	140	426
17	Beasiswa Education Belum Sentuh Siswa Mentawai.txt	Education	True	74	78
18	Belenggu Disiplin Education di	Education	True	207	323

	Papua.txt				
19	BOS SMA/SMK Ditambah.txt	Education	True	207	561
20	Guru Honorer Tuntut Gaji Sesuai Upah Minimum.txt	Education	True	262	344
21	Berkantor di Istana, Presiden Kumpulkan Menteri.txt	Politic	True	104	92
22	Menteri BUMN sempat mengamuk di pintu tol Semanggi.txt	Economic	True	210	110
23	Demokrat Tak Akan Berani Depak PKS dari Koalisi.txt	Politic	True	201	608
24	DPD RI Ajak Demonstran Aksi Damai.txt	Economic	False	183	486
25	Malam Ini SBY Harus Umumkan Pembatalan Kenaikan Harga BBM.txt	Politic	True	185	687
26	kehadiran ponsel terbarunya, Nokia N9.txt	Technology	True	374	1085
27	keluarga 'badai', BlackBerry Storm 3.txt	Technology	True	188	791
28	New iPhone Akan Pakai Layar 4,6 Inci.txt	Technology	True	167	112
29	Nokia Bawa Fitur Smartphone ke Ponsel Asha.txt	Technology	True	205	134
30	Opera Mini 7 Sudah Bisa Didownload di Android.txt	Technology	True	160	187