KLASIFIKASI ARTIKEL BERITA SECARA OTOMATIS MENGGUNAKAN METODE *NAIVE BAYES CLASSIFIER* YANG DIMODIFIKASI

Wayan Firdaus Mahmudy, Agus Wahyu Widodo

Abstrak: Klasifikasi dokumen berita digital menurut kategori tertentu diperlukan untuk mempermudah pencarian oleh pembaca. Peningkatan jumlah dokumen berita yang cukup besar tidak sebanding dengan ketersediaan editor ahli sehinggal diperlukan klasifikasi secara otomatis. Salah satu metode yang cukup akurat untuk klasifikasi artikel berita adalah *Naïve Bayes Classifier.* Makalah ini memaparkan modifikasi metode *Naïve Bayes Classifier* dengan melakukan pembobotan kata berdasarkan posisinya dalam berita. Percobaan dilakukan pada 900 dokumen berita. Sembilan ratus dokumen tersebut dibagi menjadi 9 kategori, sehingga masing-masing kategori diujikan 100 dokumen. Untuk mengetahui pengaruh jumlah data latih terhadap efektifitas naïve bayes classifier maka diambil beberapa kombinasi banyaknya dokumen latih dan dokumen uji. Secara berurutan kombinasi dokumen latih dan uji tersebut antara lain 5:95, 10:90, 15:85, 20:80, 25:75 dan 30:70. Dengan menggunakan metode *Naïve Bayes Classifier* didapatkan akurasi hasil klasifikasi berturut-turut 54%, 65%, 65%, 69%, 71% dan 76%. Dengan menggunakan metode *Naïve Bayes Classifier* yang dimodifikasi didapatkan akurasi hasil klasifikasi berturut-turut 57%, 68%, 69%, 70%, 72% dan 78%. Hasil ini menunjukkan bahwa pembobotan berdasarkan posisi kata meningkatan akurasi hasil klasifikasi.

Kata-kata Kunci : *Naïve Bayes Classifier,* klasifikasi otomotis, berita

Ketersedian dokumen digital di Internet yang berlimpah akan menyulitkan masyarakat untuk mengaksesnya jika dokumen tersebut tidak diatur secara layak. Pengaturan berita yang umum adalah dengan melakukan klasifikasi pada masing-masing artikel berita tersebut. Klasifikasi tersebut dapat didasarkan pada kondisi yang ada dalam masyarakat ataupun menurut standar khusus. Sebagai contoh, klasifikasi yang umum adalah politik, pendidikan, hiburan, ekonomi, olah raga, ilmu pengetahuan dan sebagainya. Jumlah klasifikasi tersebut sifatnya selalu berkembang. Proses klasifikasi dilakukan dengan melibatkan tenaga khusus yang memahami proses klasifikasi suatu artikel berita.

Pentingnya klasifikasi berita/dokumen ditunjukkan oleh sejumlah penelitian mengenai topik ini. Yong-feng & Yanping (2004) memaparkan beberapa metode pengkasifikasi dokumen yang umum dipakai serta menganalisis kelebihan dan kekurangannya. Miao & Kamel (2011) menggunakan algoritma *pairwise optimized Rocchio* untuk mengkasifikasikan teks. Hasil percobaan menggunakan beberapa data benchmark menunjukkan bahwa modifikasi mereka memberikan hasil yang lebih akurat dibanding algoritma asal. Suresh et al. (2011) membangun sistem pengklasifikasi dokumen yang efisien menggunakan reinforcement learning, cabang dari machine learning yang biasa digunakan dalam sistem pengambilan keputusan. Penelitian mengenai klasifikasi berita berbahasa Indonesia sebelumnya juga pernah dilakukan. Penelitian tersebut antara lain pengelompokan berita dengan algoritma *K-Means clustering*, dokumen berita dimasukkan kedalam cluster yang paling cocok berdasarkan ukuran kedekatan dengan *centroid*. *Centroid* adalah *vektor term* yang dianggap sebagai titik te-

Wayan Firdaus Mahmudy, Agus Wahyu Widodo adalah Dosen Program Studi Ilmu Komputer Universitas Brawijaya Malang

ngah cluster (Wibisono and Khodra, 2005). Dan juga penelitian dengan algoritma *Single Pass* Clustering, yaitu dengan menggunakan penghitungan tingkat kemiripan (*similarity*) dengan *Standard Cosine Similarity*. Similarity yang telah dihasilkan selanjutnya dievaluasi untuk menentukan pasangan-pasangan dokumen yang dinyatakan mirip berdasarkan nilai treshold tertentu (Arifin and Setiono, 2001).

Secara umum, metode pengklasifikasi teks dapat dikelompokkan dalam dua kelas. Yang pertama adalah metode berbasis statistik seperti Naive Bayes, maximum Shannon entropy model, K-Nearest Neighbor (KNN), dan Support Vector Machine. Yang kedua adalah metode berbasis pengetahuan seperti productive rules dan neural network. Metode berbasis statistik seperti Naive Bayes memiliki kelebihan hanya memerlukan komputasi matematika yang tidak terlalu komplek sehingga sangat efisien dalam aplikasi praktis (Yong-feng and Yan-ping, 2004). Metode ini juga terbukti handal dengan tingkat akurasi cukup tinggi (Wibisono, 2006). *Naïve Bayes Classifier* menggunakan teori probabilitas sebagai dasar teori. Dalam makalah ini digunakan metode klasifikasi *Naïve Bayes* yang sering disebut sebagai *Naïve Bayes Classifier*. Dalam penelitian sebelumnya, metode ini telah sukses digunakan untuk klasifikasi artikel berita (Widodo et al., 2007). Untuk meningkatkan tingkat tingkat akurasi dilakukan modifikasi rumus dengan melakukan pembobotan posisi kata di dalam dokumen artikel berita. Selanjutnya metode ini disebut *Modified Naïve Bayes Classifier.*

Berita

Berita adalah laporan mengenai suatu peristiwa atau kejadian yang terbaru; laporan mengenai fakta-fakta yang aktual, menarik perhatian, dinilai penting atau luar biasa (Budiman, 2005). Sedangkan menurut Dyoti (2003), berita adalah semua peristiwa yang sudah atau akan terjadi yang perlu diketahui oleh manusia pembacanya/pendengarnya/pemirsanya.

Secara umum, penulisan berita menggunakan model deduktif, artinya pembeberan fakta dimulai dari hal-hal yang bersifat umum ke hal yang khusus. Dari yang paling penting ke hal yang kurang penting. Struktur tulisannya dikenal dengan bentuk piramida terbalik. Karena struktur demikian ini cukup menarik untuk diteliti pengaruh pembobotan berdasarkan posisi kata terhadap keakuratan hasil klasifikasi.

Tahapan Text Mining

Pengklasifikasian artikel berita secara otomatis bisa dikategorikan sebagai *text mining*. Proses *text mining* dibagi menjadi 3 tahap utama, yaitu proses awal terhadap teks (*text preprocessing*), transformasi teks ke dalam bentuk antara (*text transformation*/feature *generation*), dan penemuan pola (*pattern discovery*) (Even and Zohar, 2002). Masukan awal dari proses ini adalah suatu data teks dan keluarannya berupa pola sebagai hasil interpretasi.

Text Preprocessing

Tahapan ini bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya. Tindakan yang dilakukan meliputi tindakan kompleks dan tindakan sederhana. Contoh tindakan yang bersifat kompleks pada tahap ini adalah *part-of-speech* (*pos*) *tagging*, membangkitkan *parse tree*. Contoh tindakan yang bersifat sederhana adalah proses parsing sederhana terhadap teks, yaitu memecah suatu kalimat menjadi sekumpulan kata. Selain itu pada tahapan ini biasanya juga dilakukan *case folding*, yaitu pengubahan karakter huruf menjadi huruf kecil (Garcia, 2005).

Text Transformation/Feature Generation

Pada tahap ini hasil yang diperoleh dari tahap *text preprocessing* akan melalui proses transformasi. Proses transformasi ini dilakukan dengan mengurangi jumlah kata-kata yang ada dengan penghilangan kata-kata yang dianggap tidak penting atau disebut *stopword*. Contoh kata-kata yang masuk dalam *stopword* adalah kata sambung dan kata kepunyaan. Dalam proses ini juga dilakukan pengubahan kata-kata ke dalam bentuk dasarnya atau disebut *stemming*.

Kualitas pemilihan *stopword* dalam sistem *text mining* menentukan hasil dari *text mining*. Dengan kata lain sistem *text mining* bergantung kepada faktor bahasa. Selain itu, proses penghilangan stopword tetap digunakan karena proses karena sangat mengurangi waktu komputasi dan beban kerja sistem. Dengan menghilangkan *stopword* dari suatu teks maka sistem hanya akan memperhitungkan kata-kata yang dianggap penting.

Stemming Bahasa Indonesia

Dalam bahasa Indonesia, afiks/imbuhan terdiri dari sufiks (akhiran), infiks (sisipan) dan prefiks (awalan). Pada penelitian ini proses *stemming* yang dibangun hanya menangani kata yang mengalami penambahan prefiks dan sufiks. Hal ini dilakukan karena proses penambahan infiks dalam bahasa Indonesia jarang terjadi sehingga tidak ada pengaruh yang signifikan terhadap akurasi sistem. Selain itu, penanganan kata yang mengandung infiks relative sulit dan membebani waktu komputasi sistem.

Terdapat 5 aturan tahap dalam proses *stemming* pada bahasa Indonesia sebagai berikut:

1. Penanganan terhadap partikel *inflek-sional*. yaitu : lah, kah dan tah. Contoh : duduklah, apakah.
2. Penanganan terhadap kata ganti *inflek-sional*, yaitu : ku, mu, nya. Contuh : sepedamu, mobilnya.
3. Penanganan terhadap prefiks *deriva-sional* pertama, yaitu : meng dan semua variasinya, peng dan semua variasinya, di, ter, dan ke. Contoh : membakar, pegukur, kekasih
4. Penanganan terhadap prefiks *deriva-sional* kedua, yaitu : ber dan semua variasinya serta per dan semua variasinya. Contoh : berlari, belajar, perjelas.
5. Penanganan terhadap sufiks *derivasio-nal*, yaitu: kan, an, i. Contoh : makanan, gantikan, tandai.

Karena struktur morfologi dalam bahasa Indonesia yang rumit, maka kelima tahap aturan diatas tidak cukup untuk menangani proses *stemming* bahasa Indonesia. Kesulitan dalam membedakan suatu kata yang mengandung imbuhan baik prefiks maupun sufiks dengan suatu kata dasar yang salah satu suku katanya merupakan bagian dari imbuhan, terutama dengan kata dasar yang mempunyai suku kata lebih besar dari dua.

Contoh :

- sekolah → sekolah
  (kata dasar,tidak dilakukan stemming)
- duduklah → duduk
  (dilakukan proses stemming)

Berdasarkan urutan tahapan pada penanganan kata berimbuhan, maka terdapat beberapa kemungkinan dalam kesulitan membedakan suatu suku kata merupakan imbuhan atau bagian kata dasar :

1. kata dasar mempunyai suku kata lebih besar dari dua dan suku kata terakhir Є { partikel infleksional}. Serta kata dasar tersebut tidak mendapatkan imbuhan apapun.
   Contoh : sekolah, istilah.
2. kata dasar mempunyai suku kata terakhir Є { partikel *infleksional* }dan mempunyai prefiks.
   Contoh : bersalah, pemakalah.

3. kata dasar mempunyai suku kata lebih besar dari dua dan suku kata terakhir Є { kata ganti *infleksional*}. Serta kata dasar tersebut tidak mendapatkan imbuhan apapun.
   Contoh : maluku, terungku.
4. kata dasar mempunyai suku kata terakhir Є { kata ganti *infleksional*} dan mempunyai prefiks.
   Contoh : pelaku, bertanya.
5. kata dasar mempunyai suku kata lebih besar dari dua dan suku kata pertama Є { prefiks *derivasional* pertama }. Serta kata dasar tersebut tidak mendapatkan imbuhan apapun.
   Contoh : melati, kereta, diagram.
6. kata dasar mempunyai suku kata pertama Є { prefiks *derivasional* pertama } dan mempunyai sufiks derivasional.
   Contoh : kejutan,terusan.
7. kata dasar mempunyai suku kata lebih besar dari dua dan suku kata pertama Є { prefiks *derivasional* kedua}. Serta kata dasar tersebut tidak mendapatkan imbuhan apapun.
   Contoh : perawan, berlian, berita.
8. kata dasar mempunyai suku kata pertama Є { prefiks *derivasional* kedua} dan mempunyai sufiks derivasional.
   Contoh : peranan, bereskan.
9. kata dasar mempunyai suku kata lebih besar dari dua dan suku kata terakhir Є { sufiks *derivasional*}.
   Contoh : adegan, pantai.

Karena alasan yang telah diuraikan ini, maka pada sistem terdapat 9 kamus kecil untuk melengkapi proses *stemming*.

Pattern Discovery

Tahap penemuan pola atau pattern discovery adalah tahap terpenting dari seluruh proses text mining. Tahap ini berusaha menemukan pola atau pengetahuan dari keseluruhan teks. Terdapat dua teknik pembelajaran pada tahap *pattern discovery* ini, yaitu *unsupervised* dan *supervised learning*. *Supervised learning* melakukan klasifikasi suatu data baru berdasarkan data latih. *Unsupervised learning* data latih dikelompokkan berdasarkan ukuran kemiripan pada suatu kelas (Luz, 2006).

Naïve Bayes Classifier

Naïve bayes classifier termasuk ke dalam algoritma pembelajaran bayes. Algoritma pembelajaran bayes menghitung probabilitas eksplisit untuk menggambarkan hipotesa yang dicari. Suatu data pada *naïve bayes classifier* direpresentasikan dengan konjungsi dari nilai-nilai atribut dan sebuah fungsi target f(x) yang dapat memiliki nilai apapun dari himpunan set domain V (Dumais et al., 2002). Sistem dilatih menggunakan data latih lengkap berupa pasangan nilai-nilai atribut dan nilai target kemudian sistem akan diberikan sebuah data baru dalam bentuk <a1,a2,a3,…an> dan sistem diberi tugas untuk menebak nilai fungsi target dari data tersebut (Mitchell, 1997).

Naïve bayes classifier memberi nilai target kepada data baru menggunakan nilai V_{MAP}, yaitu nilai kemungkinan tertinggi dari seluruh anggota himpunan set domain V.

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j \mid a_1, a_2, a_3, \ldots a_n) \quad \ldots (1)$$

Terorema bayes kemudian digunakan untuk menulis ulang Persamaan 1 menjadi Persamaan 2.

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, a_3, \ldots a_n \mid v_j)\, P(v_j)}{P(a_1, a_2, a_3, \ldots a_n)} \quad \ldots (2)$$

Karena $P(a_1,a_2,a_3,\ldots a_n)$ nilainya konstan untuk semua v_j sehingga Persamaan 2 dapat ditulis menjadi Persamaan 3.

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, a_3, \ldots a_n \mid v_j)\, P(v_j) \quad \ldots (3)$$

Tingkat kesulitan menghitung nilai $P(a_1,a_2,a_3,\ldots an \mid v_j)$ menjadi tinggi karena jumlah *term* $P(a_1,a_2,a_3,\ldots a_n \mid v_j)$ bisa jadi akan sangat besar. Ini disebabkan jumlah *term* tersebut sama dengan jumlah kombinasi posisi kata dikali dengan jumlah kategori. Naïve bayes classifier menyederhanakan hal ini dan bekerja dengan dasar asumsi bahwa atribut-atribut yang digunakan bersifat *conditionally independent* antara satu dan yang lainnya, dengan kata lain dalam setiap kategori, setiap kata independen satu sama lain. Sehingga :

$$P(a_1,a_2,a_3,\ldots a_n \mid v_j) = \mathbf{\Pi}_i \; P(a_i \mid v_j) \quad \ldots(4)$$

Subsitusi Persamaan 4 dengan Persamaan 3 menjadi Persamaan 5.

$$V_{NB} = \underset{v_j \epsilon V}{\operatorname{argmax}} \; P(v_j) \; \mathbf{\Pi}_i \; P(a_i \mid v_j) \quad \ldots(5)$$

V_{NB} adalah nilai probabilitas hasil perhitungan *naïve bayes classifier* untuk nilai fungsi target yang bersangkutan. Frekuensi kemunculan kata menjadi dasar perhitungan nilai dari $P(v_j)$ dan $P(a_i|v_j)$. Himpunan dari nilai-nilai probabilitas ini berkorespondensi *dengan* hipotesa yang ingin dipelajari. Hipotesa kemudian digunakan untuk mengklasifikasikan data-data baru. Pada pengklasifikasian teks, perhitungan Persamaan 4 dapat didefinisikan dalam Persamaan 6 dan 7.

$$P(v_j) = \frac{docs_j}{examples} \quad \ldots(6)$$

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |kata|} \quad \ldots(7)$$

Keterangan :

1. $docs_j$: kumpulan dokumen yang memiliki nilai target vj.
2. examples : adalah jumlah dokumen yang digunakan dalam pelatihan (kumpulan data latih).
3. n : adalah jumlah total kata yang terdapat di dalam data tekstual yang memiliki nilai fungsi target yang sesuai.
4. nk : adalah jumlah kemunculan kata wk pada semua data tekstual yang memiliki nilai fungsi target yang sesuai.
5. |kata| : adalah jumlah kata yang berbeda yang muncul dalam seluruh data tekstual yang digunakan.

Perbedaan antara *naïve bayes classifier* dengan metoda pembelajaran lainnya terletak pada proses pembangunan hipotesa. Pada *naïve bayes classifier*, hipotesa langsung dibentuk tanpa proses pencarian (*searching*), hanya dengan menghitung frekuensi kemunculan suatu kata dalam data latih, sedangkan pada metoda pembelajaran lainnya biasanya dilakukan pencarian hipotesa yang sesuai dari ruang hipotesa.

METODE

Pada makalah ini koleksi berita bahasa Indonesia yang diambil sebagai objek penelitian bersumber dari www.kompas.com, yang merupakan salah satu situs berita berbahasa Indonesia yang banyak diakses oleh pencari berita ditanah air. Koleksi berita ini terdiri atas 900 dokumen yang dikumpulkan dari berita yang diterbitkan dari bulan November 2006 sampai dengan Januari 2007. Pada URL-nya, kompas telah memberikan kategori/ label pada sebagaian berita yang diterbitkannya. Adapun kategori tersebut adalah: internasional, nasional, metropolitan, kesehatan, dikbud, IPTEK, olah raga, hiburan dan ekonomi.

Untuk memperoleh data latih (training set) yang tepat dan untuk mempermudah pengujian kebenaran dan keakuratan pada data uji (testing set) maka dokumen berita yang digunakan dalam penelitian ini adalah berita-berita yang telah dikelompokkan/dikategorikan oleh Kompas.

Deskripsi Umum Sistem

Pengklasifikasian berita yang dibuat terdapat dua tahap. Tahap pertama adalah proses pembelajaran atau pelatihan terhadap sekumpulan dokumen berita (*training set*) dan tahap selanjutnya adalah proses pengklasifikasian berita yang belum diketahui kategorinya (*testing set*) berdasarkan pengetahuan yang telah terbentuk dari *training set*.

Pada tahap pembelajaran, proses-proses yang dilakukan adalah:

1. *User* memasukkan teks berita yang akan dijadikan objek pembelajaran
2. *User* menentukan kategori dari teks berita yang diinputkan
3. Kemudian oleh sistem, teks berita tersebut diurai(*parsing)*, dilakukan proses pemilihan dan *stemming* kata (*prepocessing* dan *tranformation*)
4. Untuk setiap kata yang dihasilkan, dihitung frekuensi kemunculan kata tersebut pada dokumen
5. Semua kata beserta frekuensi yang dihasilkan dari dokumen tersebut digabungkan dengan kumpulan kata yang sudah tersimpan dalam pengetahuan, untuk membentuk pengetahuan yang baru.
6. Dihitung probabilitas setiap kata pada pengetahuan untuk setiap kategori berita yang ada. Probabilitas inilah yang digunakan untuk proses testing set.
7. Sistem juga menyimpan perubahan jumlah dokumen data latih.

Sedangkan pada tahap pengklasifikasian berita, proses-proses yang dilakukan adalah :

1. *User* memasukkan teks berita yang ingin diklasifikasi/diketahui kategori-nya
2. Seperti pada tahap pembelajaran, sistem akan melakukan *prepocessing* dan *tranformation* pada teks berita untuk menghasilkan sekumpulan kata yang akan diproses
3. Untuk setiap kata yang dihasilkan, cari probabilitasnya pada pengetahuan yang tersimpan untuk setiap kategori yang ada
4. Setelah semua kata dicari dan dihitung probabilitasnya, maka bandingkan nilai probalitas yang didapatkan antara kategori
5. Kemudian sistem akan mengkategorikan berita tersebut berdasarkan nilai probabilitas yang paling tinggi.

Batasan sistem

Batasan dari sistem yang akan dibuat adalah :

1. Pada proses *stemming* tidak memperhitungkan adanya infiks(sisipan). Proses *stemming* yang dibangun hanya melakukan penghilangan prefiks dan sufiks
2. Suatu kata dianggap berdiri sendiri dan lepas dari kata disekitarnya. Hal ini dimaksudkan untuk menyederhanakan proses pengolahan. Jika kata diproses secara berpasangan tentu saja akan muncul berbagai faktor yang harus diperhitungkan
3. Kata yang dihasilkan dari *headline*/judul berita, teras berita dan tubuh berita dianggap sama. Probabilitas yang dibangun hanya berdasarkan frekuensi kemunculan kata tidak memperdulikan posisi kata tersebut pada teks berita

Learn naive bayes

Pada tahap *learn naive bayes*, untuk membentuk pengetahuan maka sistem akan belajar dari sekumpulan data latih. Pengetahuan inilah yang akan digunakan sebagai dasar pada tahap *classify naive bayes*. Pengetahuan terdiri dari pengetahuan kata dan pengetahuan dokumen. Pengetahuan kata berisi semua jenis kata pada seluruh data latih, frekuensi kemunculan kata tersebut untuk setiap kategori dan nilai probabilitas $P(w_k \mid v_j)$ sedangkan pengetahuan dokumen berisi jumlah dokumen data latih pada setiap kategori dan

nilai probabilitas $P(v_j)$. Adapun proses pada tahapan *learn naive bayes*:

1. Untuk setiap jenis kata yang muncul pada data latih cari kedalam pengetahuan kata yang sudah ada.
   - Jika ada, maka tambahkan angka jumlah kemunculan kata tersebut pada pengetahuan kata untuk kategori yang bersesuaian.
   - Jika tidak ada, maka tambahkan kata baru dan juga jumlah kemunculan kata tersebut pada pengetahuan kata untuk kategori yang bersesuaian.
2. Setelah semua kata dan frekuensi kemunculannya ditambahkan pada pengetahuan kata maka hitung ulang probabilitas $P(w_k \mid v_j)$ pada pengeta-huan sesuai dengan Persamaan 7
3. Tambahkan jumlah dokumen yang bersesuaian pada pengetahuan dokumen
4. Hitung ulang $P(v_j)$ sesuai dengan Persamaan 6

Classify naive bayes

Classify naive bayes berusaha mencari nilai probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat. Tahapan pada *classify naive bayes* adalah :

1. Untuk setiap kata yang muncul cari kedalam pengetahuan kata
   - Jika ada, maka cari nilai probabilitas $P(w_k \mid v_j)$ untuk setiap kategori
   - Jika tidak ada, abaikan kata tersebut. Dengan kata lain, nilai probabilitas $P(w_k \mid v_j)$ dianggap nol.
2. Jumlahkan semua nilai probabilitas $P(w_k \mid v_j)$ yang didapat pada setiap kategori.
3. Untuk setiap kategori hitung $P(v_j)$ **Π** $P(w_k \mid v_j)$
4. Setelah hasil perkalian didapatkan maka hasil dari semua kategori akan dibandingkan dan untuk nilai yang terbesar maka dokumen berita ter-masuk ke dalam kategori tersebut.

Modified naive bayes classifier

Modifikasi dilakukan dengan mempertimbangkan struktur berita berbahasa Indonesia. Pada perhitungan frekuensi kemunculan kata, kata yang muncul pada setengah bagian atas diberi bobot dua dan sisanya diberi bobot satu. Hal ini menunjukkan bahwa kata-kata yang muncul pada setengah bagian atas dianggap lebih penting dan lebih mewakili kategori berita.

HASIL DAN PEMBAHASAN

Pada pengujian sistem klasifikasi ini, digunakan 900 dokumen berita. Sembilan ratus dokumen tersebut dibagi menjadi 9 kategori, sehingga masing-masing kategori akan diujikan 100 dokumen.

Untuk mengetahui pengaruh jumlah data latih terhadap efektifitas naïve bayes classifier maka diambil beberapa kombinasi jumlah dokumen latih dan dokumen uji. Secara berurutan kombinasi dokumen latih dan uji tersebut antara lain 5:95, 10:90, 15:85, 20:80, 25:75 dan 30:70. Dengan kata lain persentase dokumen latih terhadap keseluruahan dokumen adalah 5%, 10%, 15%, 20%, 25% dan 30%.

Akurasi dihitung dengan membandingkan jumlah klasifikasi yang benar dan jumlah dokumen uji yang dapat dinyatakan dengan Persamaan 8.

$$Akurasi = \frac{banyaknya\ klasifikasi\ benar}{banyaknya\ dokumen\ uji} x\ 100 \quad \text{... (8)}$$

Evaluasi pada metode naive bayes

Banyaknya klasifikasi benar pada rumus naive bayes dan prosentase akurasinya disajikan pada Tabel 1 dan Tabel 2.

Tabel 1. Banyaknya Klasifikasi Benar Pada Rumus Asal

No	Kategori	Banyaknya dokumen latih					
		5	10	15	20	25	30
1	DikBud	71	59	61	59	57	58
2	Ekonomi	61	73	73	69	66	63
3	Hiburan	29	40	41	42	41	44
4	Internasional	75	72	69	66	64	62
5	IPTEK	32	38	38	38	38	39
6	Kesehatan	59	79	72	70	67	65
7	Metropolitan	36	53	41	46	45	48
8	Nasional	37	34	34	34	35	39
9	Olah raga	62	77	72	70	67	63

Tabel 2. Persentase Akurasi Pada Rumus Asal

No	Kategori	Banyaknya dokumen latih					
		5	10	15	20	25	30
1	DikBud	75	66	72	74	76	83
2	Ekonomi	64	81	86	86	88	90
3	Hiburan	31	44	48	52	54	63
4	Internasional	79	80	81	82	85	89
5	IPTEK	34	42	45	48	50	55
6	Kesehatan	62	88	85	88	89	93
7	Metropolitan	38	59	48	58	60	69
8	Nasional	39	38	40	42	46	55
9	Olah raga	65	86	85	88	89	90
	Rata-Rata	**54**	**65**	**65**	**69**	**71**	**76**

Dari Tabel 2 terlihat bahwa nilai persentase akurasi NBC memiliki kecenderungan yang tinggi sebanding dengan jumlah dokumen latih yang digunakan. Jika diamati masing-masing kategori dokumen maka persentase yang cukup rendah adalah kategori Hiburan dan Nasional. Hal ini diasumsikan bahwa penyebabnya adalah luasnya cakupan berita hiburan dan nasional.

Evaluasi pada metode naive bayes yang dimodifikasi

Banyaknya klasifikasi benar pada rumus naive bayes yang dimodifikasi dan prosentase akurasinya disajikan pada Tabel 3 dan Tabel 4.

Tabel 3. Banyaknya Klasifikasi Benar Pada Rumus Dimodifikasi

No	Kategori	Banyaknya dokumen latih					
		5	10	15	20	25	30
1	DikBud	72	73	68	69	62	65
2	Ekonomi	60	74	75	70	67	68
3	Hiburan	29	36	42	42	40	42
4	Internasional	66	72	70	65	62	60
5	IPTEK	40	47	42	34	34	30
6	Kesehatan	60	77	73	72	70	69
7	Metropolitan	42	54	46	52	52	55
8	Nasional	50	40	36	35	33	35
9	Olah raga	70	77	73	68	68	65

Tabel 4. Persentase Akurasi Pada Rumus Dimodifikasi

No	Kategori	Jumlah dokumen latih					
		5	10	15	20	25	30
1	DikBud	76	81	80	86	83	93
2	Ekonomi	63	82	88	88	89	97
3	Hiburan	31	40	49	53	53	60
4	Internasional	69	80	82	81	83	86
5	IPTEK	42	52	49	43	45	43
6	Kesehatan	63	86	86	90	93	99
7	Metropolitan	44	60	54	65	69	79
8	Nasional	53	44	42	44	44	50
9	Olah raga	74	86	86	85	91	93
	Rata-Rata	**57**	**68**	**69**	**70**	**72**	**78**

Tabel 5. Perbandingan Persentase Akurasi Pada Rumus Sebelum dan Sesudah Dimodifikasi

Jumlah Dokumen Latih	Akurasi dengan rumus asal	Akurasi dengan rumus dimodifikasi	Kenaikan Akurasi
5	54	57	3
10	65	68	3
15	65	69	4
20	69	70	1
25	71	72	1
30	76	78	2
Rata-Rata			2.3

Tabel 5 menunjukkan modifikasi rumus NBC dengan pembobotan posisi kata meningkatkan akurasi hasil klasifikasi. Secara rata-rata didapatkan kenaikan akurasi sebesar 2,3%.

KESIMPULAN

Tulisan ini telah memaparkan pemodelan dan pembuatan sistem pengklasifikasi artikel otomatis dengan metode naive bayes yang telah dimodifikasi. Modifikasi dilakukan dengan melakukan pembobotan berasarkan posisi kata dalam berita. Akurasi sistem meningkat dengan meningkatnya jumlah data latih yang digunakan sebagai pembelajaran. Pembobotan posisi kata meningkatkan akurasi klasifikasi rata-rata sebesar 2,3%.

DAFTAR RUJUKAN

Arifin, A. Z. & Setiono, A. N. 2001. *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*. Surabaya: Institut Teknologi Sepuluh Nopember (ITS).

Budiman, K. 2005. *Dasar-Dasar Jurnalistik.* Available: www.infojawa.org/index.html [Accessed 20 Januari 2011].

Dumais, S., Platt, J., Heckerman, D. & Sahami, M. 2002. *Inductive Learning Algorithms and Representations for Text Categorization.*

Dyoti, K. 2003. *Jurnalisme in Action*. Available: http://www.deliveri.org/Guidelines [Accessed 20 Januari 2011].

Even, Y. & Zohar 2002. *Introduction to Text Mining*, University of Illionis.

Garcia, E. 2005. *Document Indexing Tutorial for Information Retrieval Students and Search Engine Marketers*. Available: http://www.miislita.com/information-retrieval-tutorial/indexing.html [Accessed 29 September 2006].

Luz, S. 2006. *Machine Learning of Text Categorization.* Trinity College, Department of Computer Science.

Miao, Y.-Q. & Kamel, M. 2011. *Pairwise optimized Rocchio algorithm for text categorization*. *Pattern Recognition Letters,* 32**,** 375-382.

Mitchell, T. M. 1997. *Machine Learning,* Singapore, McGraw –Hill.

S.R.Suresh, T.Karthikeyan, D.B.Shanmugam & J.Dhilipan 2011. *Text categorization using qlearning alogrithm*. *Indian Journal of Computer Science and Engineering,* 2**,** 315-324.

Wibisono, Y. 2006. *Klasifikasi Berita Berbahasa Indonesia Menggunakan Naive Bayes Classifier.* Bandung: FPMIPA Universitas Pendidikan Indonesia.

Wibisono, Y. & Khodra, M. 2005. *Clustering Berita Berbahasa Indonesia*. Bandung: FPMIPA Universitas Pendidikan Indonesia.

Widodo, A. W., Mahmudy, W. F. & Maisuroh, M. 2007. *Klasifikasi artikel otomatis, sebuah kajian eksperimen*. *FKP2T,* 2.

Yong-Feng, S. & Yan-Ping, Z. 2004. *Comparison of text categorization algorithms*. Wuhan University Journal of Natural Sciences*,* 9**,** 798-804.