Makara Journal of Technology

Volume 14 | Number 1 Article 5

4-1-2010

Ontology-Based Automatic Classification for News Articles in Indonesian Language

Prajna Wira Basnur
Fakultas Ilmu Komputer, Universitas Indonesia, Depok 16424, Indonesia

Dana Indra Sensuse
Fakultas Ilmu Komputer, Universitas Indonesia, Depok 16424, Indonesia, dana@cs.ui.ac.id

PENGKLASIFIKASIAN OTOMATIS BERBASIS ONTOLOGI UNTUK ARTIKEL BERITA BERBAHASA INDONESIA

Prajna Wira Basnur dan Dana Indra Sensuse[*)]

Fakultas Ilmu Komputer, Universitas Indonesia, Depok 16424, Indonesia

[*)]*E-mail: dana@cs.ui.ac.id*

Abstrak

Pencarian informasi tertentu akan sulit dilakukan bila mengandalkan *query* saja. Pemilihan *query* yang kurang spesifik akan berakibat banyaknya informasi yang tidak relevan ikut terambil oleh sistem. Salah satu cara yang paling berhasil untuk mengatasi permasalahan ini adalah dengan melakukan klasifikasi dokumen berdasarkan topiknya. Ada banyak metode digunakan untuk melakukan klasifikasi dokumen seperti menggunakan pendekatan statistik dan *machine learning*. Namun, metode klasifikasi dokumen tersebut membutuhkan *training* data atau dokumen pembelajaran. Pada penelitian ini penulis berusaha untuk melakukan klasifikasi dokumen menggunakan sebuah metode yang tidak memerlukan dokumen pembelajaran. Metode klasifikasi ini menggunakan ontologi untuk melakukan klasifikasi dokumen. Klasifikasi dokumen dengan menggunakan ontologi dilakukan dengan membandingkan nilai kemiripan di antara dokumen dan sebuah *node* yang ada di ontologi. Sebuah dokumen diklasifikasikan dalam sebuah kategori atau *node* jika memiliki nilai kemiripan paling tinggi di salah satu *node* di ontologi. Hasil penelitian menunjukkan bahwa ontologi dapat digunakan untuk melakukan klasifikasi dokumen. Nilai *recall* adalah 97,03%, *precision* 91,63%, dan *f-measure* 94,02%.

Abstract

Ontology-Based Automatic Classification for News Articles in Indonesian Language. Searching specific information will be difficult if relying only on query. Choosing less specific queries will result in a lot of irrelevant information fetched by the system. One of the most successful ways to overcome this problem is to perform document classification based on the topic. There are many methods that can be used to conduct such a classification, such as using statistical and machine learning approaches. However, those document classification methods require training the data or learning the documents. In this study, the authors attempted to classify documents using a method that doesn't require learning the documents. This classification method uses ontology to classify documents. Document classification using ontology is done by comparing the value of similarity among documents and existing node in the ontology. A document is classified into a category or a node if it has the highest similarity value in one of the nodes in the ontology. The results show that ontology can be used to perform document classification. The recall value is 97.03%, the precision is 91.63%, and the f-measure is 94.02%.

Keywords: Ontology, Naïve-Bayes, stopwords, stemming

1. Pendahuluan

Teknologi informasi telah berkembang sangat pesat sampai saat ini. Salah satu teknologi informasi dan komunikasi yang berkembang pesat adalah internet. Perkembangan internet sudah merambah ke berbagai lapisan masyarakat, mulai dari anak kecil hingga orang dewasa bahkan yang tua sekalipun. Internet sudah menjadi kebutuhan bagi banyak orang karena dengan internet kita dapat menemukan dan mengakses informasi dari seluruh dunia dengan cepat dan mudah.

Informasi yang diperoleh dari internet dapat berupa dokumen teks seperti dokumen berita, suara, video, maupun objek multimedia lainnya. Informasi tersebut dapat diakses melalui halaman web. Web memuat banyak informasi yang dihasilkan dari waktu ke waktu secara kontinu dari berbagai sumber. Jumlah informasi yang terus bertambah dari waktu ke waktu dapat menyulitkan para pencari informasi dalam menemukan informasi yang relevan. Salah satu cara yang paling berhasil untuk mengorganisasikan informasi dalam jumlah banyak dan dapat dipahami oleh para pencari

informasi adalah dengan melakukan klasifikasi dokumen berdasarkan topiknya [1].

Kebutuhan akan dokumen pembelajaran untuk melakukan klasifikasi dokumen merupakan salah satu permasalahan yang sering muncul dalam topik klasifikasi dokumen [2]. Permasalahan lain yang muncul adalah seberapa banyak dokumen pembelajaran yang dibutuhkan agar klasifikasi dokumen memberikan akurasi yang maksimal. Apabila jumlah dokumen pembelajaran yang digunakan terlalu sedikit, maka tidak akan menghasilkan tingkat akurasi yang maksimal. Permasalahan dokumen pembelajaran untuk melakukan klasifikasi dokumen ini dapat diatasi dengan pendekatan baru yang tidak memerlukan dokumen pembelajaran. Pendekatan ini dikenal dengan nama pendekatan ontologi [3].

Klasifikasi dokumen adalah bidang penelitian dalam perolehan informasi yang mengembangkan metode untuk menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok yang telah dikenal sebelumnya secara otomatis berdasarkan isi dokumen [4]. Klasifikasi dokumen bertujuan untuk mengelompokkan dokumen yang tidak terstruktur ke dalam kelompok-kelompok yang menggambarkan isi dari dokumen. Dokumen dapat berupa dokumen teks seperti artikel berita. Pada bagian ini membahas tentang penelitian dalam bidang klasifikasi artikel berita berbahasa Indonesia.

Penelitian dilakukan oleh Slyvia Susanto yaitu pengklasifikasian dokumen berita berbahasa Indonesia dengan menggunakan Naïve Bayes *classifier* (*stemming* atau *non-stemming*) [2]. Eksperimen yang dilakukan dalam penelitian ini dengan menggunakan *stemming* dan *non-stemming*. Hasil eksperimen dalam penelitian ini menunjukkan bahwa jumlah dokumen pembelajaran 90% dan jumlah dokumen pengujian 10% (*stemming*) menghasilkan akurasi yang paling tinggi yaitu dengan *recall* 93,5%, *precision* 90,36%, dan *f-measure* 93,81%. Kesimpulan yang diperoleh dari penelitian ini adalah kinerja Naïve Bayes *classifier* yang menggunakan *stemming* lebih baik dari pada *non-stemming*.

Penelitian dalam bidang klasifikasi dokumen berbahasa Indonesia yang dilakukan oleh Yudi Wibisono dan Slyvia Susanto membutuhkan dokumen pembelajaran untuk melakukan klasifikasi dokumen baru [2,5]. Pada penelitian ini mengajukan sebuah metode baru untuk melakukan klasifikasi dokumen, yaitu dengan menggunakan ontologi. Metode klasifikasi dokumen dengan menggunakan ontologi tidak memerlukan dokumen pembelajaran.

Ontologi adalah sebuah deskripsi formal tentang sebuah konsep secara eksplisit dalam sebuah domain, properti dari setiap konsep beserta dengan batasannya [6]. Sebuah konsep di ontologi dapat memiliki objek (*instances*). Secara teknis, ontologi direpresentasikan dalam bentuk *class*, *property*, *facet*, dan *instances*. *Class* menerangkan konsep atau makna dari suatu domain. *Class* adalah kumpulan dari elemen dengan sifat yang sama. Sebuah *class* bisa memiliki sub *class* yang menerangkan konsep yang lebih spesifik.

Property merepresentasikan hubungan di antara dua individu. *Property* menghubungkan individu dari domain tertentu dengan individu dari *range* tertentu. Ada tiga jenis *property*, yaitu *object property*, *data type property,* dan *annotation property*. *Object property* menghubungkan suatu individu dengan individu lain. *Object property* terdiri dari empat tipe, yaitu *inverse property*, *functional property*, *transitive property*, dan *symmetric property*. *Data type property* menghubungkan sebuah individu ke sebuah tipe data pada *resource description framework* (RDF) *literal* atau pada *extensible markup language* (XML). *Annotation property* digunakan untuk menambah informasi (*metadata*) ke kelas, individu, dan *object/data type property*.

Facet digunakan untuk merepresentasikan informasi atau batasan tentang *property*. Ada dua jenis *facet*, yaitu *cardinality* dan *value type*. *Cardinality facet* merepresentasikan nilai eksak yang bisa digunakan untuk *slot* pada suatu kelas tertentu. *Cardinality facet* dapat bernilai *single* dan *multiple cardinality*. *Value type* menggambarkan tipe nilai yang dapat memenuhi *property*, seperti *string*, *number*, *boolean*, dan *enumerated*.

Ontologi dapat digunakan untuk melakukan klasifikasi dokumen teks dalam penelitian ini karena ontologi bersifat unik dan memiliki struktur hierarkis. Selain itu, sebuah model ontologi dapat menghilangkan makna ambigu, sehingga dapat menanggulangi masalah yang muncul pada bahasa alami dimana sebuah kata memiliki lebih dari satu makna atau arti bergantung pada konteks kalimatnya.

Ontologi dalam penelitian ini dibuat dengan menggunakan konsep hierarki. Ontologi terdiri dari konsep, relasi antar konsep, fitur konsep, dan batasan untuk klasifikasi dokumen. Konsep atau *class* merepresentasikan *term* atau kata dalam domain yang spesifik. Fitur atau *instance* merepresentasikan individu dari sebuah kelas. Relasi atau *property* merepresentasikan hubungan di antara konsep. Ada dua relasi yang digunakan dalam penelitian ini, yaitu relasi "*is-a*" dan "*has-a*". *Constraint* merepresentasikan kondisi yang harus dipenuhi di sebuah konsep.

2. Eksperimental

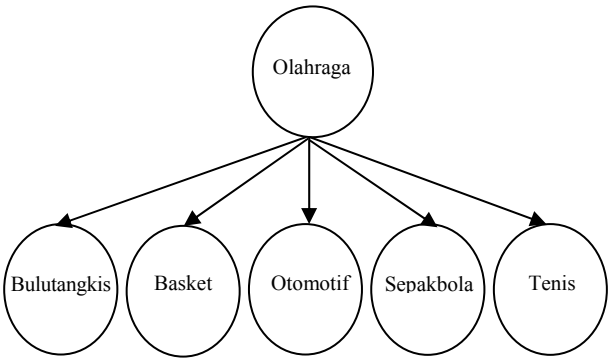
Pemodelan ontologi dalam penelitian ini diawali dengan mendefinisikan *root* dari ontologi. *Root* ini diberi nama "olahraga". *Root* direpresentasikan sebagai sebuah kelas.

Kelas *root* memiliki lima subkelas, yaitu bulutangkis, basket, otomotif, sepakbola, dan tenis (Gambar 1).

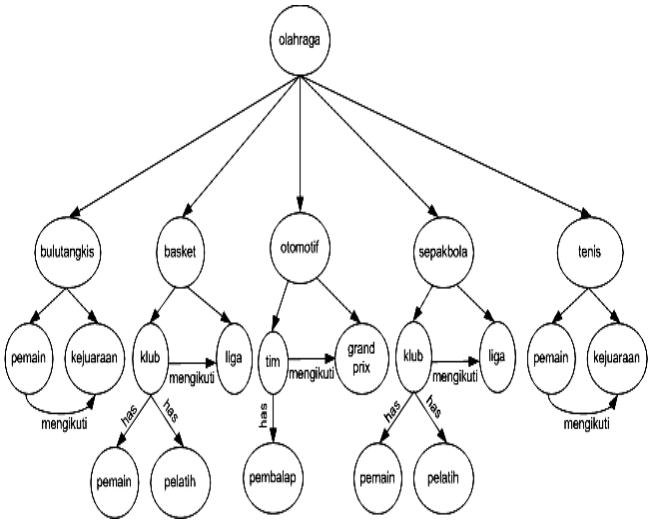
Kelima subkelas tersebut (bulutangkis, basket, otomotif, sepakbola, dan tenis) juga terdiri dari beberapa subkelas. Setiap subkelas dari kelas memiliki *property* dan *instance. Property* digunakan untuk mendefinisikan atribut dari subkelas. Selain itu, *property* juga digunakan untuk mendefinisikan relasi antara satu subkelas dengan subkelas lain. *Instance* digunakan untuk mendefinisikan objek dari sebuah *property*. Representasi subkelas masing-masing kategori dapat dilihat pada Gambar 2.

Proses pengklasifikasian artikel berita berbahasa Indonesia terdiri atas dua langkah. Pertama, proses penemuan kosakata kunci dalam dokumen. Kedua, pemetaan kosakata tersebut ke sebuah *node* dalam konsep hierarki (ontologi). Proses pemetaan dilakukan setelah melakukan proses persiapan dokumen dan pembobotan kata.

Proses persiapan dokumen meliputi proses *case folding*, tokenisasi, pembuangan *stopwords*, dan pemotongan imbuhan [7]. Tujuan dari proses persiapan dokumen adalah untuk menghilangkan karakter-karakter selain huruf, menyeragamkan kata, dan mengurangi volume kosakata.



Gambar 1. Representasi Ontologi Olahraga



Gambar 2. Representasi Subkelas Masing-masing Kategori

Proses pembobotan kata adalah proses memberikan nilai atau bobot ke sebuah kata berdasarkan kemunculannya pada suatu dokumen teks [7]. Proses persiapan dokumen teks dalam penelitian ini menghasilkan kumpulan kata atau *term* yang kemudian direpresentasikan dalam sebuah *terms vector*. *Terms vector* dari suatu dokumen teks d adalah *tuple* bobot semua *term* pada d. Nilai bobot sebuah *term* menyatakan tingkat kepentingan *term* tersebut dalam merepresentasikan dokumen teks. Pada penelitian ini, proses pembobotan kata menggunakan metode *term frequency-inverse document frequency* (TF-IDF).

Term frequency–inverse document frequency atau biasa sering disebut TF-IDF adalah metode pembobotan kata dengan menghitung nilai TF dan juga menghitung kemunculan sebuah kata pada koleksi dokumen teks secara keseluruhan [7]. Pada pembobotan ini, jika kemunculan *term pada* sebuah dokumen teks tinggi dan kemunculan *term* tersebut pada dokumen teks lain rendah, maka bobotnya akan semakin besar. Akan tetapi, jika kemunculan *term* tersebut pada dokumen teks lain tinggi, maka bobotnya akan semakin kecil. Tujuan penghitungan IDF adalah untuk mencari kata-kata yang benar-benar merepresentasikan suatu dokumen teks pada suatu koleksi. Metode pembobotan kata yang digunakan dalam penelitian ini adalah metode TF-IDF. Metode ini digunakan karena metode ini paling baik dalam perolehan informasi [8]. Rumus TF-IDF dapat dilihat pada Persamaan (1) [9].

$$tfidf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right) \qquad (1)$$

dengan tf(i,j) adalah frekuensi kemunculan *term* j pada dokumen teks $d_i \in D^*$, dimana i = 1,2,3,...,N, df(j) adalah frekuensi dokumen yang mengandung *term* j dari semua koleksi dokumen, dan N adalah jumlah seluruh dokumen yang ada di koleksi dokumen.

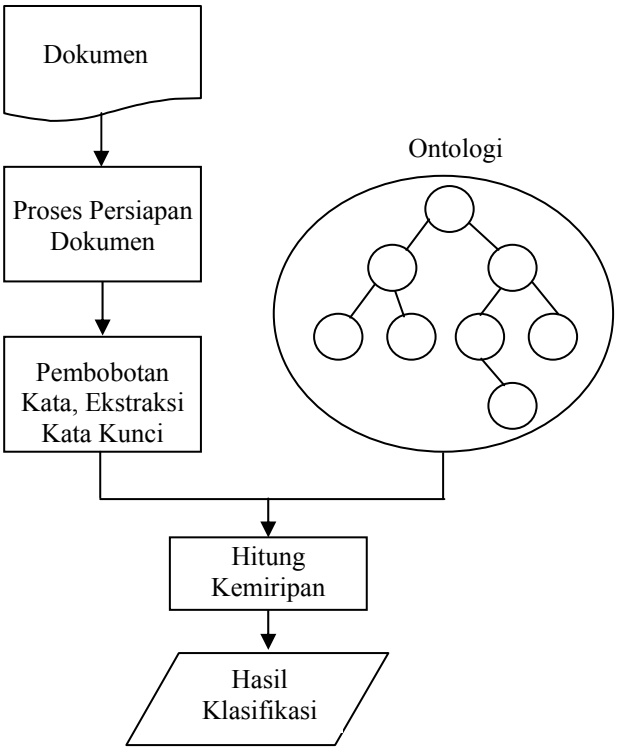
Proses klasifikasi dokumen dengan menggunakan ontologi dilakukan setelah melakukan pembobotan kata. Proses klasifikasi dilakukan dengan memetakan dokumen teks ke sebuah *node* dengan nilai kemiripan paling tinggi dan dokumen teks tersebut diklasifikasikan tepat ke satu *class*. Rumus untuk menghitung nilai kemiripan dapat dilihat pada Persamaan (2).

$$Sim(node,d) = \frac{\sum_{i=0}^{n} freq_{i,d} / \max_{i,d}}{N} \times \frac{V_d}{V} \qquad (2)$$

dengan N adalah frekuensi fitur dari sebuah *node*, $freq_{i,d}$ merepresentasikan frekuensi fitur dari fitur i yang cocok di dokumen d, $max_{i,d}$ merepresentasikan frekuensi fitur

yang paling cocok di dokumen d, V adalah jumlah *constraint*, dan V_d adalah jumlah *constraint* yang terpenuhi di dokumen d. Proses klasifikasi dokumen hanya dilakukan ketika menggunakan relasi "*is-a*" dan "*has-a*". Ketika *node* lain cocok dengan fitur di dokumen, maka *node* tersebut juga dimasukkan ke dalam proses klasifikasi dokumen untuk menghitung nilai kemiripannya. Perancangan klasifikasi dokumen teks dengan menggunakan ontologi dapat dilihat pada Gambar 3.

Data (dokumen teks) yang digunakan dalam penelitian ini berjumlah 1300 dokumen berita berbahasa Indonesia yang diperoleh dari situs *http://www.kompas.com*. Dokumen uji coba tersebut merupakan kumpulan berita dari kategori olahraga. Jenis dokumen uji coba terdiri dari dokumen bulutangkis, basket, sepakbola, tenis, dan otomotif. Rincian daftar kategori dan jumlah dokumen teks yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.



Gambar 3. Perancangan Klasifikasi Dokumen Menggunakan Ontologi

Tabel 1. Daftar Kategori dan Jumlah Dokumen Teks

Kategori	Jumlah Dokumen
Bulutangkis	155
Basket	59
Otomotif	160
Sepakbola	767
Tenis	159

Pada penelitian ini, nilai akurasi klasifikasi dokumen menggunakan ontologi dibandingkan dengan klasifikasi dokumen dengan menggunakan metode Naïve Bayes *classifier*.

Naïve Bayes Classifier. Naïve Bayes merupakan salah satu contoh dari metode *supervised document classification.* Metode ini menggunakan perhitungan probabilitas. Naïve Bayes tidak memperhatikan urutan kemunculan kata pada dokumen teks dan menganggap sebuah dokumen teks sebagai kumpulan dari kata-kata yang menyusun dokumen teks tersebut. Metode ini memiliki tingkat akurasi yang tinggi dengan penghitungan sederhana [10].

Naïve Bayes menggunakan teorema dasar yang dikenal dengan nama Teorema Bayes (Persamaan (3)) [11].

$$P(C = c_a \mid D = d_b) = \frac{P(C = c_a \cap D = d_b)}{P(D = d_b)} \quad (3)$$

dengan $P(C = c_a \mid D = d_b)$, probabilitas kategori c_a jika diketahui dokumen d_b. Kemudian dari Persamaan (3) kita dapat membuat Persamaan (4).

$$P(C=c_a \cap D=d_b)=P(D=d_b|C=c_a)xP(C=c_a) \quad (4)$$

sehingga didapatkan Teorema Bayes seperti pada Persamaan (5).

$$P(C=c_a \mid D=d_b)=\frac{P(D=d_b \mid C=c_a)\times P(C=c_a)}{P(D=d_b)} \quad (5)$$

dengan $P(D = d_b \mid C = c_a)$ merupakan nilai probabilitas dari kemunculan dokumen d_b jika diketahui dokumen tersebut memiliki kategori c_a, $p(C = c_a)$ adalah nilai probabilitas kemunculan kategori c_a, dan $p(D = d_b)$ adalah nilai probabilitas kemunculan dokumen d_b. Klasifikasi dokumen teks dilakukan dengan terlebih dahulu menentukan kategori $c \in C$ dari suatu dokumen $d \in D$ dimana $C = \{c_1, c_2, c_3, \ldots, c_m\}$ dan $D = \{d_1, d_2, d_3, \ldots, d_n\}$ dan $P(C = c_a \mid D = d_b)$ memiliki nilai maksimum dari suatu distribusi probabilitas $P=\{ P(C = c_a \mid D = d_b) \mid c \in C \text{ dan } d \in D \}$.

Apabila urutan kemunculan kata dalam dokumen teks tidak diperhatikan, maka perhitungan probabilitas $P(D = d_b \mid C = c_a)$ dapat dianggap sebagai hasil perkalian dari probabilitas kemunculan kata-kata dalam dokumen d_b. Sebuah dokumen d_b terdiri dari kata-kata, maka dapat dituliskan sebagai $d_b = \{w_{1b}, w_{2b}, w_{3b}, \ldots, w_{kb}\}$ sehingga probabilitas $P(C = c_a \mid D = d_b)$ dapat dituliskan seperti pada Persamaan (6).

$$P(C = c_a \mid D = d_b) = \frac{\prod_k P(w_{kb} \mid C = c_a) \times P(C = c_a)}{P(w_1, w_2, w_3, \ldots, w_n)} \quad (6)$$

dengan $\prod_k P(w_{kb} \mid C = c_a)$ adalah hasil perkalian probabilitas kemunculan semua kata pada dokumen teks d_b, jika diketahui dokumen kategorinya adalah c_a.

Tahap klasifikasi dilakukan dengan membuat model probabilistik dari dokumen pembelajaran, yaitu dengan menghitung nilai $P(w_{kb}|c_a)$. Probabilitas yang mungkin dapat dicari untuk seluruh nilai w_{kb} dengan menggunakan Persamaan (7) dan (8) [11].

$$P(w_{kb} \mid c_a) = \frac{f(w_{kb}, c_a)+1}{f(c_a)+|w|} \quad (7)$$

dengan $f(w_{kb}, c_a)$ adalah fungsi yang menghasilkan nilai kemunculan kata w_{kb} pada kategori c_a, $f(c_a)$ adalah fungsi yang menghasilkan jumlah keseluruhan kata pada kategori c_a, dan $|W|$ adalah jumlah kemungkinan nilai dari w_{kb} (jumlah keseluruhan kata yang digunakan). Persamaan $f(w_{kb}, c_a)$ sering kali dikombinasikan dengan *Laplacian smoothing* (tambah satu) untuk mencegah persamaan mendapatkan nilai 0. Hal ini dilakukan karena nilai 0 dapat mengganggu hasil klasifikasi secara keseluruhan.

$$P(c_a) = \frac{f_d(c_a)}{|D|} \quad (8)$$

dengan $fa(c_a)$ adalah fungsi yang menghasilkan jumlah dokumen teks yang memiliki kategori c_a, dan $|D|$ adalah jumlah seluruh dokumen pembelajaran.

Pemberian kategori dari sebuah dokumen teks dilakukan dengan memilih nilai c yang memiliki nilai probabilitas $P(C = c_a \mid D = d_b)$ maksimum, seperti pada Persamaan (9).

$$c_* = \arg\max_{ca \in C} P(c_a \mid d_b)$$
$$c^* = \arg\max_{ca \in C} \prod_k P(w_{kb} \mid c_a) \times P(c_a) \quad (9)$$

Kategori a^* merupakan kategori yang memiliki nilai probabilitas $p(C = c_a | D = d_b)$ maksimum.

Evaluasi. Evaluasi digunakan untuk mengukur kinerja suatu system, khusus dalam penelitian ini digunakan untuk mengukur keakuratan metode klasifikasi dokumen teks. Metode evaluasi yang digunakan, yaitu: *recall, precision,* dan *F-measure*. Rumus *recall, precision,* dan *F-measure* dapat dilihat pada Persamaan (10), (11), dan (12).

$$Recall = \frac{\text{Jumlah dokumen relevan terkenali}}{\text{Jumlah dokumen relevan}} \quad (10)$$

$$Precision = \frac{\text{Jumlah dokumen relevan terkenali}}{\text{Jumlah dokumen terkenali}} \quad (11)$$

$$F - Measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (12)$$

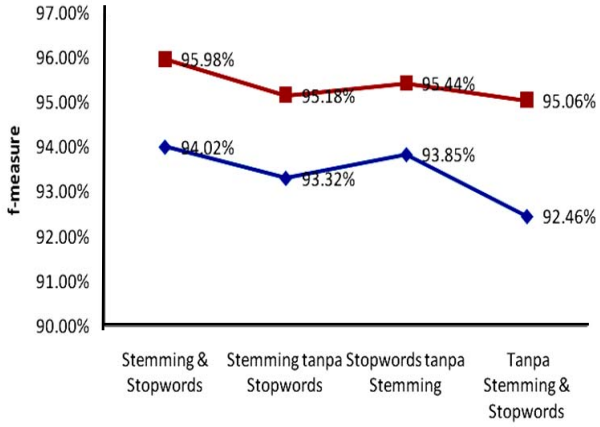
3. Hasil dan Pembahasan

Eksperimen dilakukan dengan memperhatikan dua aspek, yaitu penggunaan *stopwords* dan pemotongan imbuhan (*stemming*) serta jumlah kategori. Eksperimen penggunaan *stopwords* dan *stemming* dilakukan dengan 4 kombinasi perlakuan, yaitu penggunaan *stopwords* dan *stemming*, penggunaan *stopwords* tanpa *stemming*, penggunaan *stemming* tanpa *stopwords*, dan tanpa menggunakan *stopwords* dan *stemming*. Hasil eksperimen menunjukkan bahwa nilai akurasi klasifikasi dokumen yang menggunakan *stopwords* dan *stemming* lebih tinggi daripada nilai akurasi klasifikasi dokumen menggunakan *stopwords* tanpa *stemming*, menggunakan *stemming* tanpa *stopwords*, dan tanpa menggunakan *stemming* dan *stopwords* (Tabel 2).

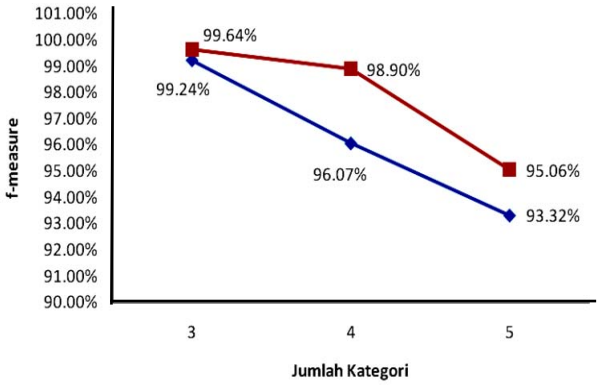
Nilai akurasi metode klasifikasi menggunakan Naïve Bayes lebih tinggi daripada nilai akurasi metode klasifikasi menggunakan ontologi jika ditinjau dari aspek penggunaan *stopwords* dan *stemming* (Gambar 4). Nilai akurasi metode klasifikasi menggunakan ontologi lebih rendah daripada menggunakan Naïve Bayes karena jumlah informasi yang digunakan sebagai data untuk memodelkan ontologi masih sedikit sehingga diperlukan lebih banyak data agar nilai akurasinya sama atau lebih tinggi dari nilai akurasi metode Naïve Bayes.

Eksperimen jumlah kategori bertujuan untuk mengetahui pengaruh jumlah kategori terhadap hasil akurasi klasifikasi dokumen teks. Jumlah kategori dalam eksperimen ini dimulai dari tiga, empat, dan lima buah kategori. Kategori bulutangkis, otomotif, dan sepakbola digunakan sebagai eksperimen untuk jumlah kategori sama dengan tiga. Kategori bulutangkis, basket, otomotif, dan sepakbola digunakan sebagai eksperimen untuk jumlah kategori sama dengan empat.

Semua kategori digunakan sebagai eksperimen untuk jumlah kategori sama dengan lima. Nilai akurasi untuk empat kategori menggunakan metode Naïve Bayes merupakan rata-rata dari nilai akurasi dengan menggunakan 100, 500, dan 1000 data pembelajaran. Nilai akurasi untuk tiga kategori menggunakan metode Naïve Bayes merupakan rata-rata dari nilai akurasi dengan menggunakan 100, 500, dan 800 data pembelajaran. Perbandingan nilai *f-measure* jumlah



Gambar 4. Pengaruh Penggunaan *Stopwords* dan *Stemming* pada Ontologi (-♦-) dan Naïve Bayes (-■-)



Gambar 5. Pengaruh Jumlah Kategori pada Ontologi (-♦-) dan Naïve Bayes (-■-)

kategori untuk metode Naïve Bayes dan ontologi dapat dilihat pada Gambar 5.

Hasil eksperimen pada data artikel media massa berbahasa Indonesia menunjukkan penurunan akurasi (Gambar 5). Pada metode Naïve Bayes, akurasi klasifikasi menunjukkan penurunan dari 99,64% pada penggunaan tiga kategori menjadi 98,90% pada penggunaan empat buah kategori, kemudian pada penggunaan lima buah kategori hasil akurasi kembali menurun menjadi 95,06%. Pada metode ontologi terjadi penurunan dari 99,24% pada penggunaan tiga buah kategori menjadi 96,07% pada penggunaan empat buah kategori, kemudian pada penggunaan lima buah kategori hasil akurasi kembali menurun menjadi 93,32%. Hasil ini sesuai dengan perkiraan bahwa penambahan jumlah kategori akan menurunkan akurasi klasifikasi dokumen teks.

Nilai akurasi metode Naïve Bayes lebih tinggi daripada metode ontologi (Gambar 5) baik dengan jumlah kategori sama dengan tiga, empat, dan lima kategori. Perbedaan hasil akurasi klasifikasi mencapai 0,4% pada penggunaan tiga buah kategori, 2,83% pada penggunaan empat buah kategori, dan 1,74% pada penggunaan lima buah kategori.

4. Simpulan

Berdasarkan hasil eksperimen yang diperoleh, dapat ditarik beberapa simpulan. Pertama, jumlah kategori yang ada mempengaruhi kinerja klasifikasi dokumen menggunakan metode Naïve Bayes dan ontologi. Secara umum, penambahan jumlah kategori dapat menurunkan tingkat akurasi klasifikasi dokumen. Selain itu, tingkat kemiripan di antara kategori juga dapat mempengaruhi tingkat akurasi klasifikasi dokumen. Jika tingkat kemiripan di antara dua kategori tinggi, maka akan sulit untuk membedakan kedua kategori tersebut sehingga tingkat akurasi klasifikasi dokumen akan turun. Kedua, penggunaan *stopwords* dan *stemming* dapat meningkatkan tingkat akurasi klasifikasi dokumen. Klasifikasi dokumen dapat menggunakan ontologi dan memiliki nilai *f-measure* mencapai 94,02%. Meskipun nilai *f-measure* klasifikasi dokumen menggunakan ontologi tidak lebih tinggi daripada nilai akurasi klasifikasi dokumen menggunakan metode Naïve Bayes. Namun, metode klasifikasi dokumen dengan menggunakan ontologi memiliki kelebihan, yaitu tidak memerlukan proses pembelajaran atau data eksperimental sedangkan metode Naïve Bayes membutuhkan proses pembelajaran agar dapat mengklasifikasikan dokumen baru. Klasifikasi dokumen teks (metode Naïve Bayes dan ontologi) yang dilakukan pada penelitian ini masih memiliki kekurangan. Beberapa saran yang mungkin berguna untuk melakukan penelitian klasifikasi dokumen selanjutnya, antara lain: (1) Mencari lebih banyak dokumen, khususnya dokumen berbahasa Indonesia sehingga dapat menganalisis pengaruh jumlah dokumen terhadap tingkat akurasi klasifikasi dokumen, (2) Mengumpulkan lebih banyak data untuk memodelkan ontologi dan merancang metode klasifikasi yang lebih efisien dan akurat, (3) Pemodelan ontologi untuk sebuah konsep atau *class* dibuat seunik mungkin dari konsep atau *class* lain sehingga dapat meningkatkan nilai akurasi klasifikasi dokumen.

Daftar Acuan

[1] D. Anggraeni, Skripsi Sarjana, Fakultas Ilmu Komputer, Universitas Indonesia, Indonesia, 2009.
[2] S. Susanto, Skripsi Sarjana, Fakultas Ilmu Komputer, Universitas Indonesia, Indonesia, 2006
[3] S.Y. Lim, M.H. Song, S.J. Lee, Ontology-based Automatic Classification of Web Documents, ICIC 2006, LNAI 4114, Springer-Verlag, Berlin, 2006, p.690.
[4] L. Tenenboim, B. Shapira, P. Shoval, Proceedings of the Intelligent Information and Engineering Systems Conference, Varna, Bulgaria, 2008.
[5] Y. Wibisono, Seminar Nasional Matematika, Universitas Pendidikan Indonesia, Bandung, Jawa Barat, 2005.
[6] N.F. Noy, D.L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Knowledge Systems Laboratory (KSL) of Department of Computer Science Stanford, USA: Technical Report, KSL-01-05, 2001.
[7] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, New York, 1999, p.544.
[8] L.M. Khodra, Y. Wibisono, Konferensi Nasional Sistem Informasi, Universitas Pasundan Bandung, Indonesia, Februari 2006.
[9] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill, New York, 1986, p.400.

[10] D.K. Kang, A. Silvescu, V. Honavar, Proceedings of the Tenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2006, p.45.

[11] T. Mitchell, Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression, in Machine Learning, McGraw Hill, New York, 2006, p.414.