PEMODELAN BERBASIS KONSEP UNTUK KATEGORISASI ARTIKEL BERITA BERBAHASA INDONESIA

Candra Triawati[1], M. Arif Bijaksana[2], Nur Indrawati[3], Widyanto Adi Saputro[4]
[1,2,3,4] Departemen Teknik Informatika, Institut Teknologi Telkom
Jl. Telekomunikasi No.1, Terusan Buah Batu, Bandung 40257
Telp. +62227564108, Fax. +62227562721
E-mail: candra_tria@yahoo.co.id, arifbijaksana@gmail.com, nurindrawati_jogja@yahoo.co.id, dyan_saputro@yahoo.com

ABSTRAK

Semakin pesatnya penggunaan Internet memacu pertumbuhan ketersediaan data, yang pada gilirannya memerlukan dukungan dari teknologi informasi untuk mengubah data tersebut menjadi suatu informasi, dan selanjutnya menjadi suatu pengetahuan yang bermanfaat. Text mining sebagai upaya pengelolaan dokumen yang berupa teks, timbul dari kebutuhan akan pemahaman dan pemanipulasian dokumen teks untuk mendapatkan informasi tertentu dari sekumpulan dokumen teks dengan memanfaatkan teknologi komputer. Sebagian besar teknik pengelompokkan dokumen dalam text mining bekerja berdasarkan analisis dari frekuensi kemunculan kata atau frasa pada dokumen tanpa menghiraukan sisi semantis, sehingga ada kemungkinan dua kata atau frasa mempunyai frekuensi yang sama padahal salah satu kata atau frasa tersebut memberikan kontribusi yang lebih besar dalam proses penentuan topik sebuah dokumen. Hal ini tentu saja menyebabkan hasil yang diperoleh kurang akurat. Untuk memperoleh hasil yang lebih akurat, diperlukan pendekatan baru yang melibatkan peran semantis suatu kata dalam kalimat. Salah satu pendekatan baru tersebut adalah pemodelan berbasis konsep. Tulisan ini membahas teori pemodelan berbasis konsep dan aplikasi pemodelan berbasis konsep dalam pemrosesan dokumen teks, khususnya pada kategorisasi artikel berita berbahasa Indonesia.

Kata Kunci: text mining, kategorisasi, pemodelan berbasis konsep

1. PENDAHULUAN

Salah satu hal yang umum dalam penelitian terkait *text mining* adalah dengan merepresentasikan teks sebagai kumpulan kata-kata atau lebih dikenal dengan istilah pendekatan *Bag-of-Words* atau *BoW* (Feldman dan Dagan, 1995). Kumpulan teks dalam suatu dokumen secara sederhana dianggap sebagai kumpulan kata-kata tanpa memperhatikan aspek sintaksis. Hal ini berarti pada pendekatan *BoW* tidak mempedulikan urutan kata dalam dokumen karena hanya memperhatikan suatu kata sebagai entitas tersendiri yang tidak dipengaruhi oleh struktur kalimat dimana kata itu berada. Pendekatan *BoW* juga tidak memperhatikan unsur semantis yang merupakan cabang linguistik yang membahas tentang arti dan makna, yang artinya sebuah kata diabaikan maknanya dan juga kedudukannya dalam sebuah dokumen secara semantis apakah berperan besar dalam mewakili isi dari suatu dokumen atau tidak.

Pendekatan *BoW* mengabaikan sintaksis dan semantis. Dalam *BoW,* dokumen direpresentasikan sebagai vektor yang elemennya merupakan hasil pembobotan berdasarkan frekuensi kemunculan suatu kata dalam dokumen saja (Feldman dan Dagan, 1995). Pendekatan *BoW* menggunakan analisa statistik dari frekuensi kemunculan *term* untuk mengetahui peran suatu kata dalam mewakili isi dokumen. Cara tersebut mempunyai kekurangan, salah satunya adalah bisa jadi dua buah kata mempunyai jumlah kemunculan yang sama dalam satu dokumen padahal kata yang satu lebih besar kontribusinya dalam mewakili isi dokumen dibandingkan dengan kata yang lain.

Untuk mengatasi kekurangan pendekatan *BoW*, maka muncul sebuah pendekatan baru yang dikenal dengan *Bag-of-Concepts* (*BoC*). Pendekatan *BoC* menangkap kata-kata yang merupakan konsep dari sebuah dokumen, yang nantinya dapat mewakili isi dokumen. Pendekatan atau pemodelan berbasis konsep (disebut juga dengan istilah *concept-based mining model*) merupakan terobosan baru dalam *text mining*.

Tujuan dari penelitian ini adalah untuk mengimplementasikan *concept-based mining model* dalam rangka menghasilkan *dataset* bagi proses kategorisasi dokumen berbahasa Indonesia yang kemudian akan dianalisis keakuratan dan performansinya.

Adapun lingkup masalahnya adalah dokumen yang digunakan hanyalah dokumen berbahasa Indonesia yang terdiri dari empat topik yaitu: ekonomi, nasional, dan politik. Semua artikel tersebut telah melalui proses preprosesing secara manual berupa pelabelan dokumen berdasarkan peran semantiknya.

2. PEMODELAN BERBASIS KONSEP

Kategorisasi pada *text mining* menjadi hal yang penting ketika akan dilakukan pencarian dokumen

tertentu pada sejumlah data yang sangat besar. Penerapan kategorisasi dokumen telah dilakukan pada penentuan topik secara otomatis, pembuatan direktori topik, pengidentifikasian gaya penulisan dokumen, dan mengklasifikasikan kegunaan *hyperlink* berkaitan dengan sejumlah dokumen.

Kategorisasi dapat dilakukan dengan menggunakan beberapa metode, salah satunya adalah dengan metode atau pemodelan berbasis konsep. Pemodelan berbasis konsep terdiri atas 3 (tiga) bagian utama, yaitu pelabelan peran semantis, *graphical concept-based*, dan *statistical concept-based*. Masing-masing bagian memiliki fungsi dan peranan yang berbeda.

2.1 Pelabelan Peran Semantis

Mengidentifikasi peran semantis dapat memberikan level analisis semantis yang bermanfaat dalam menyelesaikan *task* pemrosesan bahasa alami. Peran semantis merepresentasikan partisipan dalam aksi atau keterhubungan yang digambarkan dengan kerangka semantis (Gildea dan Palmer, 2002).

Pelabelan peran semantis didefinisikan sebagai proses pengidentifikasian argumen dari predikat dalam suatu kalimat, dan menentukan (Gildea dan Jurafsky, 2002). Ilustrasi dari pelabelan peran semantis, misalnya untuk kalimat "Ibu memotong roti." adalah sebagai berikut. Pada kalimat "Ibu memotong roti." terdapat sebuah verba yaitu memotong dan argumen-argumennya yaitu "Ibu" dan "roti" . Pemberian label berdasarkan peran semantis dilakukan terhadap argumen dimana "Ibu" mendapatkan label ARG0 dan memiliki peran semantis sebagai pelaku (*agent*) serta "roti" mendapat label ARG1 dan memiliki peran semantis sebagai penderita (*patient*), sehingga struktur *verb-argument* yang terbentuk dari kalimat tersebut adalah sebagai berikut:

"[ARG0 Ibu] [TARGET **memotong**] [ARG1 roti]"

Label ARG0 dan ARG1, serta ARG3 sampai dengan ARG5 dan ARG-M mengacu pada label yang ada pada basis data PropBank (Gildea dan Jurafsky, 2002).

Salah satu manfaat pelabelan peran semantis adalah pada tahap *preprocessing* dari *concept-based mining system*, yang memanfaatkan representasi data berbasis grafikal dan statistikal untuk mencari konsep dari suatu kalimat dalam dokumen teks. Pencarian konsep pada tahap *preprocessing* dari sistem berbasis konsep merupakan salah satu cara untuk mendapatkan informasi yang berharga dari suatu kalimat dengan memperhatikan makna dari setiap kata dalam suatu kalimat.

2.2 *Graphical Concept-Based*

Graphical concept-based adalah salah satu bagian dari *concept-based mining model* yang merepresentasikan dokumen ke dalam bentuk graf konseptual yang disebut dengan *conceptual ontological graph* (*COG*). Representasi *COG* dapat menunjukkan *term-term* yang mempunyai peran terhadap semantis kalimat. Setiap *term* dipilih berdasarkan posisinya pada representasi *COG*. Pada akhirnya, *term* yang terpilih tersebut dihubungkan dengan dokumen dimana *term* tersebut berada dan digunakan sebagai fitur untuk keperluan proses pengelompokkan pada *text mining*.

Representasi dari *COG* merupakan graf konseptual *G = (C, R)* dimana konsep dari kalimat digambarkan sebagai puncak (*C*). Relasi di antara konsep-konsep seperti subjek, objek, dan aksi digambarkan sebagai (*R*). *C* adalah sekumpulan dari simpul $\{c_1, c_2, \ldots, c_n\}$, dimana setiap simpul *c* mewakili sebuah konsep dalam kalimat atau sebuah graf konseptual bersarang *G*; dan *R* adalah sekumpulan dari sisi $\{r_1, r_2, \ldots, r_m\}$, dimana setiap sisi *r* adalah hubungan antara sepasang simpul yang terurut (c_i, c_j).

Hasil dari proses pelabelan secara semantis, yaitu verba dan argumen-argumennya digambarkan sebagai konsep beserta relasinya dalam representasi *COG*. Pertama, representasi *COG* menangkap konsep-konsep dan relasi diantara konsep-konsep tersebut. Kemudian, konsep dan relasinya digambarkan ke dalam suatu representasi graf konseptual yang berdasarkan kalimat. Setiap simpul pada representasi *COG* dapat berupa sebuah simpul konsep maupun sebuah graf konseptual. Jika sebuah simpul merujuk ke graf konseptual lainnya, hal ini berarti masih ada informasi yang lebih rinci mengenai topik yang ada pada simpul (konsep) tersebut, dan simpul tersebut digambarkan sebagai konsep dan relasinya pada tingkat yang berada di atas graf konseptual yang dirujuknya. Pada representasi *COG* konsep-konsep dari kalimat diurutkan secara menurun (*descending*), yang berarti simpul tertinggi mewakili konsep yang paling umum dari kalimat dan simpul terendah mewakili konsep yang paling rinci dari kalimat. Hal ini menghasilkan perbandingan konsep yang lebih informatif pada level kalimat dan level dokumen daripada jika hanya menggunakan perbandingan kata saja.

Representasi *COG* menyediakan berberapa macam level konsep yang bersarang dengan cara bertingkat. Level ini dibangun berdasarkan tingkat kepentingan dari konsep dalam kalimat, yang akan digunakan untuk menganalisa konsep kunci dalam kalimat. Representasi yang bertingkat dari *COG*, memberikan pemisahan antar konsep yang berkontribusi pada arti kalimat dengan jelas. Pemisahan ini diperlukan untuk membedakan antara konsep yang kurang berarti dengan konsep kunci pada kalimat. Konsep-konsep diletakkan pada representasi *COG* berdasarkan banyaknya *overlapping* antara *term*.

Langkah-langkah pembuatan *COG* dapat dilihat pada gambar 1 berikut (Shehata, Karray, dan Kamel, 2006):

1. Tentukan jumlah struktur verb-argument untuk masing-masing kalimat.
2. Nyatakan setiap term yang telah terlabeli, baik berupa verb atau argumen, sebagai konsep atau relasi dalam graf konseptual.
3. Hitung jumlah overlapping dari kata-kata di setiap term.
4. Tentukan tingkatan secara ontologi diantara graf konseptual berdasarkan jumlah overlapping term.
5. Bangun representasi COG dengan menggabungkan graf-graf konseptual yang dihasilkan dari langkah kedua ke dalam representasi COG.

Gambar 1. Langkah-langkah pembuatan *COG*

Dari algoritma pada gambar 1 dapat diketahui bahwa untuk menggambarkan *level* dari tingkatan *COG*, ada lima tipe struktur *verb-argument* yang digunakan untuk menentukan berada di level mana sebuah struktur *verb-argument* dalam *COG* (Shehata, Karray, dan Kamel, 2007):

1. *One*: Jika hanya ada satu struktur *verb-argument* yang dihasilkan.
2. *Main*: Jika ada lebih dari satu struktur *verb-argument*, dan struktur utama (*main structure*) mempunyai jumlah *term* (argumen) yang merujuk pada *term* di struktur *verb-argument* lainnya yang maksimal.
3. *Container*: Jika ada lebih dari satu struktur *verb-argument* yang dihasilkan, dan struktur *container* merujuk pada argumen yang lain, dan, pada saat yang sama struktur *container* tidak mempunyai jumlah rujukan *term* yang maksimal.
4. *Referenced*: Struktur yang ditunjuk (*referenced structure*) mempunyai *term* (baik *verb* atau argumen), yang dirujuk oleh struktur *main* maupun *container*.
5. *Unreferenced*: *Term* yang terdapat pada struktur *unreferenced* tidak dirujuk oleh satupun *term* yang lain.

Skema ini membuat sebuah graf konseptual untuk setiap struktur *verb-argument*. Setiap tipe dari struktur *verb-argument* ditempatkan dalam graf konseptual sesuai dengan posisinya. *COG* menggambarkan graf konseptual sebagai tingkatan, yang ditentukan berdasarkan tipe masing-masing.

Sebuah pengukuran L_{COG} digunakan untuk mengurutkan konsep dengan tetap memperhatikan semantis kalimat dalam representasi *COG*. L_{COG} digunakan pada tingkat *One*, *Unreferenced*, *Main*, *Container*, dan *Referenced* pada representasi *COG* dengan nilai 1, 2, 3, 4, dan 5 tergantung tipenya. Sebagai pengganti dari memilih konsep hanya pada satu level dalam representasi *COG*, konsep dari semua level dalam representasi *COG* diperhatikan dan di beri bobot.

Pembobotan untuk tiap konsep dilakukan oleh $weight_{COG}$ yang dalam representasi *COG* dirumuskan dengan:

$$weight_{COGi} = tf\ weight_i * L_{COGi} \quad (1)$$

Pada persamaan di atas, $tfweight_i$ merepresentasikan bobot dari konsep *i* dalam dokumen *d* pada *level* dokumen. Nilai L_{COGi} merepresentasikan tingkat kepentingan dari konsep *i* dalam dokumen *d* pada *level* kalimat berdasarkan pada besarnya kontribusi dari konsep *i* terhadap semantis kalimat yang digambarkan oleh level dari representasi *COG*. Hasil perkalian dari kedua nilai di atas digunakan untuk mengurutkan konsep-konsep dalam dokumen *d* dengan tetap memperhatikan kontribusi dari setiap konsep terhadap arti kalimat dan topik dalam dokumen.

Selain mengaplikasikan algoritma *graphical concept-based mining model*, dalam penelitian ini juga diupayakan untuk memodifikasi algoritma tersebut untuk kemudian dibandingkan hasilnya dengan algoritma aslinya dalam hal performansi. Beberapa hal telah diganti dan ditambah ke dalam algoritma *graphical concept-based mining model* awal. Perbedaan pertama adalah dari jumlah status untuk mendapatkan nilai L_{COG}.

Pada algoritma awal terdapat 5 macam nilai L_{COG} yaitu: 1 untuk *one*, 2 untuk *unreferenced*, 3 untuk *main*, 4 untuk *container*, dan 5 untuk *referenced*. Jika suatu konep mempunyai dua nilai, misal sebagai *container* dan *referenced*, maka yang diambil adalah status dengan nilai tertinggi yaitu *referenced*. Pada algoritma yang *graphical concept-based mining model* telah dimodifikasi terdapat 3 status tambahan yang nilainya merupakan rataan dari kedua nilai status yang disandang oleh konsep. Ketiga status tambahan itu adalah: *main-container* yang bernilai 3,5, *main-referenced* yang bernilai 4, *dan container-referenced* yang bernilai 4,5.

Perbedaan kedua terletak pada perhitungan nilai *tf*. Pada algoritma hasil modifikasi, nilai *tf* suatu konsep tidak dihitung apa adanya. Jika suatu konsep lebih dari satu kata maka akan diberi suatu nilai tambah yang berasal dari nilai subset dari konsep tersebut. Perhitungan *tf* ini juga menerapkan metode *stemming*.

Perbedaan ketiga adalah dari jumlah konsep yang diambil. Pada algoritma yang sudah dimodifikasi, semua konsep yang mempunyai bobot diambil sebagai perwakilan dari suatu dokumen. Sedangkan pada algoritma awalnya, konsep yang diambil sebagai wakil dari dokumen adalah konsep dengan bobot tertinggi saja.

2.3 *Statistical Concept-Based*

Tujuan dari pemodelan ini adalah untuk menganalisis kata dan frasa dalam level kalimat dan dokumen, tidak hanya analisis *single term* dalam dokumen saja. Analisis dalam level kalimat muncul akibat dari penggunaan konsep sehingga hasilnya akan lebih akurat jika dibandingkan analisis single term saja sebab memperhitungkan peran semantik kata dalam suatu kalimat. Untuk menganalisis masing-masing konsep pada level kalimat digunakan istilah *ctf* (*conceptual term frequency*). *Ctf* adalah jumlah kemunculan konsep *c* dalam struktur *verb-argument* dari kalimat *s*. Untuk menganalisis masing-masing konsep pada level dokumen digunakan istilah frekuensi *tf* (*term frequency*), yaitu jumlah kemunculan sebuah konsep *c* dalam dokumen asli dihitung.

Nilai *tf* dan *ctf* dihitung dengan algoritma *statistical analyzer*. Algoritmanya seperti pada gambar 2 berikut (Shehata, Karray dan Kamel, 2007):

```
1. d is a new Document. L is an empty
   List (L is a top  concept list).
2. for each labeled sentence s in d do
3.     c_i is a new concept in s
4.     for each concept c_i in s do
5.           compute tf_i in d
6.           compute ctf_i in d
7.           compute weightstat_i of
             concept c_i
8.           add concept c_i to L
9.           end for
10. end for
11. sort L descendingly based on max
    (weight_stat)
12. output the max (weight_stat) from L
```

Gambar 2. Algoritma *statistical analyzer*

Algoritma *statistical concept-based* atau *statistical analyzer* seperti pada gambar 2 di atas mendeskripsikan proses penghitungan *tf* dan *ctf* konsep pada suatu dokumen. Prosedur dimulai dengan pemrosesan dokumen baru, dimana kalimat – kalimatnya sudah dilabeli sesuai dengan peran semantis masing-masing kata dalam kalimat-kalimat tersebut.

Untuk setiap kalimat yang terlabeli, konsep pada struktur *verb-argument* yang merepresentasikan semantis kalimat dibobotkan dengan menggunakan weightstat berdasarkan dari nilai *ctf* dan *tf* . Konsep yang bobotnya paling tinggi di outputkan.

Statistical concept-based mining model merupakan faktor utama yang memperhatikan penting atau tidaknya suatu konsept pada sebuah dokumen dan pada kalimat. Konsep dengan nilai tertinggi merupakan keluaran dari sistem.

Nilai yang digunakan untuk membedakan konsep yang penting atau tidak penting adalah nilai $weight_{stat}$, dengan memperhatikan semantis kalimat. Berikut adalah persamaannya (Shehata, Karray dan Kamel, 2007):

$$weightstat_i = tfweight_i + ctfweight_i \quad (2)$$

Pada penghitungan $weigh_{stat,}$ dihitung juga nilai *tfweight*, dimana *tf* $weight_i$ ini merepresentasikan bobot konsep *i* dalam dokumen *d* pada dokumen. Sedangkan nilai $ctfweighth_i$ merepresentasikan bobot konsep *i* terhadap semantis kalimat dalam dokumen *d*. Hasil penjumlahan antara $ctfweight_i$ dan $tfweight_i$ merepresentasikan penghitungan yang akurat kontribusi konsep terhadap makna kalimat dan topik pada dokumen.

Pada persamaan (3), nilai tf_{ij} dinormalisasi dengan panjang vektor dokumen tf_{ij} dalam dokumen d, dimana j=1,2,.., c_n (Shehata, S., Karray, F., Kamel, M., 2007).

$$tfweight_i = \frac{tf_{ij}}{\sum_{j=1}^{cn} (tf_{ij})^2} \quad (3)$$

Cn diperoleh dari jumlah konsep yang mempunyai nilai *tf* pada dokumen.

Pada persamaan (4), nilai ctf_{ij} dinormalisasi dengan panjang vektor dokumen dari ctf_{ij} di dokumen d, dimana j=1,2,3,…,c_n (Shehata, Karray, Kamel, 2007):

$$ctfweight_i = \frac{ctf_{ij}}{\sum_{j=1}^{cn} (ctf_{ij})^2} \quad (4)$$

C_n diperoleh dari jumlah konsep yang mempunyai nilai *tf* pada dokumen.

Algoritma *statistical concept-based mining model* ini bisa dikembangkan menjadi beberapa varian. Pengembangan ini dilakukan dengan cara menganalisis kekurangan – kekurangan pada metode *statistical analyzer* pada *concept-based mining model* kemudian diusulkan cara-cara untuk memperkecil atau menghilangkan kekurangan yang ada pada algoritma *statistical analyzer* pada *concept based mining model*. Berikut adalah pengembangan algoritma *statistical analyzer:*

1. Penghitungan subset *term* dengan memperhatikan urutan *term* yang menjadi subset pada suatu kata. Penghitungan subset *term* ini dilakukan dengan tujuan untuk mencari kata – kata yang sering muncul pada konsep atau frasa dokumen lain. Konsep yang memiliki subset yang banyak mengindikasikan bahwa konsep tersebut penting. Konsep yang memiliki subset terhadap konsep lain diberi bobot lebih. Persamaan yang digunakan adalah:
   tf_modified=nilai_subset + tf_lama (5)
   Pengecekan konsep suatu kalimat pada dokumen yang tidak sama, dilakukan sesuai langkah-langkah penghitungan seperti pada gambar 3.

```
1. Hitung jumlah kata dalam konsep yang
   akan dihitung bobotnya.
2. Cari dalam dokumen subset kata
   tersebut, lalu beri bobotnya sesuai
   persamaan berikut:
   Kata subset = (probabilitas kemunculan
   kata) x (nilai kata subset) x (jumlah
   kemunculan kata subset dalam dokumen.
```

Gambar 3. Algoritma penghitungan konsep

Setelah diperoleh nilai *tf_subset*, maka bobot konsep dihitung dengan menggunakan persamaan berikut:

$$weight\_concept=(ctfweight+tf\_modifiedWeight)*idf \quad (6)$$

2. Melakukan *pseudo stemming* pada kata kerja untuk peningkatan performansi. Kata kerja dicari bentuk dasarnya, kemudian bentuk dasar kata kerja tersebut dicocokkan dengan bentuk dasar argumen kata kerja tersebut. Jika kata dasarnya sama, maka bobot argumen diberi bobot tambahan sebesar 1.

2.4 Analisis Pemodelan Berbasis Konsep

Concept-based mining model berhasil memberikan suatu sudut pandang baru dalam dunia *text mining*. Pelibatan peran semantik kata dalam suatu kalimat membuat kita mempunyai cara baru dalam memproses suatu dokumen sebelum dikategorisasikan. Meskipun demikian, ada harga yang harus dibayar dalam mengaplikasikan algoritma ini. Struktur algoritma yang lebih kompleks daripada TF-IDF berdampak pada semakin lamanya waktu yang diperlukan untuk memproses dokumen. Hal ini menyebabkan algoritma ini tidak efektif jika digunakan dalam dunia nyata, khususnya untuk aplikasi berbasis web. Kemudian penerapan dua algoritma sekaligus yaitu *graphical concept-based mining model* dan *statistical concept-based mining model* terasa mubazir sebab salah satunya saja sudah mencukupi untuk menghasilkan dataset yang diperlukan untuk proses kategorisasi. *graphical concept-based mining model* dan *statistical concept-based mining model* sebenarnya berjalan secara independen, sehingga sebaiknya bersifat opsional dalam arti salah satu saja yang perlu kita gunakan. Dan yang terakhir, *concept-based mining model* bukanlah algoritma yang fleksibel karena penerapannya tergantung dari bahasa yang digunakan dalam dokumen.

3. HASIL PENELITIAN

3.1 Data yang Digunakan

Data yang digunakan pada penelitian ini adalah artikel berita berbahasa Indonesia yang diambil dari website Harian Kompas periode bulan Januari 2009 dan bersifat *offline*, sejumlah 100 artikel. Artikel-artikel tersebut disimpan dalam file berekstensi .txt. yang disimpan dalam sebuah direktori dan dikelompokkan dalam folder sesuai dengan nama kategori. Kategori dokumen diambil berdasarkan kategori asli pada website Harian Kompas, yaitu: ekonomi, nasional dan politik dengan asumsi pengkategorian yang dilakukan dalam website tersebut benar.

Data yang diambil dari website tersebut kemudian diberi label peran semantis (melalui proses pelabelan peran semantis) dan kemudian disimpan dalam basis data. Untuk pengujian, digunakan data yang belum dilabeli dan data yang telah dilabeli secara semantis melalui proses pelabelan peran semantis.

3.2 Hasil Pengujian

Pengujian dalam penelitian ini dilakukan dengan mengukur performansi kategorisasi berdasarkan *f-measure*. Perhitungan performansi berdasarkan nilai *f-measure* yang mengkombinasikan nilai *precision* dan *recall* dengan bobot yang sama. Pengujian ini dilakukan pada masing-masing kategori dari dokumen uji. Pengujian untuk algoritma statistikal dilakukan untuk ketiga algoritma pembobotan yaitu *statistical concept-based mining model* yang standard dan modifikasi serta *tf-idf*. Output dari ketiga algoritma statistikal ini kemudian dijadikan sebagai data input pada proses kategorisasi dengan menggunakan tools WEKA.

Dari hasil pengujian terhadap *graphical concept-based mining model* didapatkan bahwa algoritma *tf-idf* relatif lebih bagus dibandingkan algoritma *graphical concept-based mining model* standar. Secara keseluruhan dapat dilihat nilai *f-measure tf-idf* lebih baik dari *graphical concept-based mining model* dimana nilai *f-measure* tertinggi untuk *tf-idf* adalah 0.909 dan untuk *graphical concept-based mining model* sebesar 0.842. Hasil yang terbaik didapatkan oleh pembobotan dengan algoritma *graphical concept-based mining model* yang telah dimodifikasi, dimana nilai *f-measure* di ketiga kategori yang berbeda mencapai nilai maksimal yaitu satu.

Meskipun terlihat bahwa nilai *f-measure* pada *graphical concept-based mining model* selalu lebih rendah jika dibandingkan *tf-idf*, hal ini bukan berarti pendekatan *graphical concept-based mining model* lebih buruk dari pendekatan tradisional. Hal ini dibuktikan dengan sedikit melakukan modifikasi pada algoritma *graphical concept-based mining model* maka akan didapatkan hasil yang sangat bagus dengan peningkatan kualitas *f-measure* yang signifikan.

Pada pengujian terhadap *statistical concept-based mining model*, nilai *f-measure* dari hasil kategorisasi berita ekonomi menunjukkan hasil yang hampir sama pada algoritma *statistical concept-based mining model* standar maupun yang telah dimodifikasi yaitu antara 0,861 dan 0,882. Nilai *f-measure* pada kategori nasional menunjukkan bahwa *f-measure* untuk *tf-idf* mempunyai nilai yang

paling rendah dibandingkan pada *statistical concept-based mining model* yaitu sebesar 0.824. Sedangkan algoritma *statistical concept-based mining model* standar mempunyai nilai tertinggi yaitu 0.877. Pada kategori politik, nilai f-measure tertinggi didapatkan oleh algoritma *statistical concept-based mining model* yang telah dimodifikasi yaitu sebesar 0.913. Hasil pengujian dan perbandingan untuk masing-masing algoritma dapat dilihat pada tabel 1.

Tabel 1. Hasil *statistical concept-based mining model*, *graphical concept-based mining model*, modifikasi, dan perbandingannya dengan *tf-idf*

Metode	*f-measure*		
	Ekonomi	*Nasional*	*Politik*
tf-idf	0.909	0.824	0.909
Graphical	0.842	0.762	0.8
Modified graphical	1	1	1
Statistical	0.882	0.877	0.881
Modified statistical	0.861	0.866	0.913

Secara umum nilai *f-measure* menunjukkan bahwa performansi kategorisasi dengan menggunakan algoritma *concept-based*, baik *statistical concept-based mining model* standar maupun yang telah dimodifikasi, lebih baik di bandingkan dengan performansi kategorisasi menggunakan algoritma *single term*, yaitu *tf-idf*. Hal ini menunjukkan bahwa analisis kata atau *concept* pada tingkat kalimat dan dokumen mempengaruhi performansi kategorisasi suatu dokumen.

4. KESIMPULAN

Penerapan algoritma *statistical concept-based* dan *statistical concept based* yang sudah dimodifikasi *(modified statistical concept-based)* pada kategorisasi artikel berita berbahasa Indonesia tidak selalu menghasilkan performansi kategorisasi yang lebih baik daripada metode *tf-idf*, bergantung pada karakteristik dokumen, yaitu jumlah kata-kata yang sama pada dokumen yang berbeda. Metode ini mempunyai tingkat *f-measure* yang semakin bagus untuk dokumen yang antar dokumen satu dengan dokumen lainnya mempunyai isi yang hampir sama.

Penerapan algoritma *graphical concept-based mining model* pada kategorisasi artikel berita berbahasa Indonesia dapat mempengaruhi performansi proses kategorisasi dokumen, yaitu relatif dapat memperbaiki performansi kategorisasi tergantung penerapannya. Banyaknya atribut yang mewakili dokumen pada pendekatan *concept-based mining model* dapat mempengaruhi kinerja sistem, yang ditunjukkan dari nilai *f-measure*-nya. *Concept-based mining model* memberikan dasar bagi penelitian dan pengembangan varian-varian algoritmanya.

PUSTAKA

Feldman, R., Dagan, I. (1995). Knowledge Discovery in Textual Databases. *Proceedings of First International Conference on Knowledge Discovery and Data Mining*, 112–117.

Gildea, D., Jurafsky, D. (2002). *Automatic Labeling of Semantic Roles*. Diakses pada 16 Oktober 2008 dari http://www.cs.rochester.edu/~gildea/gildea-cl02.pdf.

Gildea, D., Palmer, M. (2002). *The Necessity of Parsing for Predicate Argument Recognition.* Diakses pada tanggal 15 Oktober 2008 dari http://acl.ldc.upenn.edu/P/P02/P02-1031.pdf.

Shehata, S., Karray, F., Kamel, M. (2006). Enhancing Text Retrieval Performance using Conceptual Ontological Graph. *Proceedings of ICDMW 2006*, San Jose, 39-44.

Shehata, S., Karray, F., Kamel, M. (2007). A Concept-based Model for Enhancing Text Categorization. *Proceedings of ICDMW 2007*, San Jose, 12–15.