CLUSTERING ARTIKEL BERITA BERBAHASA INDONESIA MENGGUNAKAN *UNSUPERVISED FEATURE SELECTION*

Diah Pudi Langgeni[1], ZK. Abdurahman Baizal[2], Yanuar Firdaus A.W.[3]
Telp (022)7564108 ext 2298 Fax (022)7565934
[1,3]Program Studi Teknik Informatika, Fakultas Teknik Informatika Institut Teknologi Telkom, Bandung
[2]Program Studi Ilmu Komputasi, Fakultas Sains Intitut Teknologi Telkom, Bandung
Jl Telekomunikasi, Terusan Buah Batu, Bandung
e-mail : diah_pudi@yahoo.com[1], baizal@yahoo.com[2], yanuar@ittelkom.ac.id[3]

Abstrak
Meningkatnya penggunaan internet telah memicu pertumbuhan dan pertukaran informasi menjadi jauh lebih pesat dibandingkan era sebelumnya. Volume berita elektronik berbahasa Indonesia semakin bertambah besar dan menyimpan informasi yang berharga di dalamnya. Pengelompokkan berita berbahasa Indonesia merupakan salah satu solusi yang dapat digunakan untuk mempermudah mencerna informasi penting yang ada di dalamnya. Clustering dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan. Pada text clustering terdapat suatu permasalahan yaitu adanya fitur – fitur yang berdimensi tinggi. Diperlukan metode Feature selection untuk mengurangi dimensi fitur ini. Feature selection memiliki kemampuan mengurangi dimensionalitas suatu data sehingga dapat meningkatkan performansi clustering. Ada beberapa pendekatan sebagai teknik dari implementasi feature selection, salah satunya adalah filter based feature selection. Pada penelitian ini, dilakukan analisis perbandingan metode feature selection antara Term contribution dan Document Frequency. Metode-metode feature selection tersebut diterapkan secara filter feature selection. Pada akhir pengujian, dapat dibuktikan bahwa metode Term contribution lebih baik daripada Document Frequency karena memperhitungkan frekuensi kemunculan term pada suatu dokumen dan jumlah dokumen yang dimiliki term tersebut, sehingga term yang terpilih adalah term yang khas atau bersifat diskriminator. Hal ini dapat meningkatkan performansi clustering dokumen berdasarkan precision dan entropy.

Kata Kunci : *clustering, filter feature selection, Term contribution, Document Frequency*

1. PENDAHULUAN

Pesatnya pengunaan dan adopsi Internet telah memacu pertumbuhan dan pertukaran informasi yang sangat pesat dibandingkan era sebelumnya. Sebagai akibatnya, jumlah informasi terus meningkat secara exponensial, lebih dari 550 triliun dokumen saat ini. Sebanyak 7.3 juta Internet page baru tiap hari nya. Walau perkembangan ini memungkinan informasi untuk di akses penguna dengan mudah, jumlah yang terkendalikan ini telah menimbulkan isu and tantangan yang besar [1].

Demikian pula halnya dengan berita elektronik berbahasa Indonesia yang volumenya semakin bertambah besar. Berita yang disampaikan melalui media elektronik ini tentu merupakan sumber informasi yang berharga. Oleh karena itu dibutuhkan sebuah metode khusus untuk dapat mengelompokkan berita – berita tersebut sehingga dapat mempermudah pengambilan informasi penting yang ada di dalamnya. *Clustering* dokumen teks adalah salah satu operasi pada text mining untuk mengelompokkan dokumen yang memiliki kesamaan isi. *Clustering* dapat diaplikasikan untuk menemukan keterkaitan antar berita[15]. *Clustering* dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan.

Pada *clustering* teks terdapat suatu permasalahan yaitu adanya fitur – fitur yang berdimensi tinggi. Hal ini bisa disebabkan karena adanya data yang tidak relevan dan redundan. Kerja dari *Clustering* tidak akan optimal apabila di dalamnya terdapat fitur yang tidak relevan dan redundan. Oleh karena itu diperlukan metode untuk mengurangi dimensi fitur ini. Dalam hal ini ada 2 metode yang sering digunakan, yaitu *feature extraction* dan *feature selection*.

Feature extraction adalah proses mengekstrak fitur baru dari fitur asli melalui pemetaan fungsional. Sedangkan *feature selection* adalah sebuah proses pemilihan subset fitur dari fitur asli[8]. Kelebihan *feature selection* dibandingkan dengan *Feature extraction* adalah pada seleksi fitur memberikan pemahaman yang lebih baik mengenai data sedangkan *Feature extraction* tidak demikian.

Berdasarkan ada atau tidaknya informasi label, feature selection dapat dibedakan menjadi 2 jenis yaitu *supervised feature selection* dan *unsupervised feature selection.* Pada *supervised feature selection* dibutuhkan label kelas sedangkan pada *unsupervised feature selection* tidak. Metode *feature selection* telah banyak diaplikasikan pada classification teks tetapi jarang dilakukan pada clustering teks.

Penelitian ini bertujuan untuk mengimplementasikan unsupervised feature selection yaitu *Document Frequency* (DF) dan *Term Contribution* (TC) pada clustering berita berbahasa Indonesia, Dua macam analisa yang dilkukan adalah sebagai berikut :

1. Analisis pengaruh metode *unsupervised feature selection* pada performansi *clustering* teks berdasarkan *entropy* dan *precision Measure*
2. Analisis perbandingan *Term Contribution (TC)* dengan *Document Frequency (DF)* pada *clustering* teks berdasarkan *entropy* dan *precision Measure*.

2. TINJAUAN PUSTAKA

2.1. *Text Mining*

Text mining sudah banyak didefinisikan oleh ahli riset dan praktisi [1,2,3,6]. *Text mining* memiliki definisi menambang data yang berupa teks di mana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata - kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Sistem *text mining* terdiri dari komponen *text preprocessing*, *feature selection,* dan komponen *data mining*. Komponen *text preprocessing* berfungsi untuk mengubah data tekstual yang tidak terstruktur seperti dokumen, kedalam data terstruktur dan disimpan ke dalam basis data. *Feature selection* akan memilih kata yang tepat dan berpengaruh pada proses klasifikasi. Komponen terakhir akan menjalankan teknik data mining pada output dari komponen sebelumnya.

2.2. *Text Prepocessing*

Teks yang akan dilakukan proses text mining, pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat noise pada data, dan terdapat struktur teks yang tidak baik[1]. 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 tahap *preprocessing* yang dilakukan secara umum dalam *text mining* pada dokumen, yaitu case folding, *tokenizing*, *filtering*, *stemming*, *tagging* dan *analyzing*.

Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

Tahap *tokenizing* / *parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.

Tahap *filtering* adalah tahap mengambil kata – kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). Stoplist / stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah "yang", "dan", "di", "dari" dan seterusnya.

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 rumus bentuk baku yang permanen.

2.3. Feature Selection

Terdapat dua pendekatan *feature selection* yang digunakan pada *machine learning*, yaitu *filtering* dan *wrapper*[2].

2.3.1. Pendekatan Feature Selection

2.3.1.1. Filter Feature Selection

Salah satu pendekatan Feature Selection dalam pemilihan feature adalah filter feature Selection. Pemilihan feature dengan filter model ini lebih murah dalam komputasi karena tidak melibatkan induksi algoritma dalam prosesnya[18]. Oleh karena itu, penerapan pendekatan filter ini cocok untuk data yang berdimensi tinggi seperti text mining.

Gambar 1 Filter Based *Feature Selection*

2.3.1.2. Wrapper Feature Selection

Pada pendekatan wrapper, pemilihan feature subset menggunakan fungsi evaluasi berdasarkan algoritma learning yang sama yang akan digunakan untuk proses clustering. Dalam algoritma pemilihan feature, terdapat dua komponen utama yaitu pembangkitan prosedur dan fungsi evaluasi . Komponen pertama dilakukan setelah feature space terbentuk, dilakukan pencarian prosedur yang menghasilkan subset dari fitur untuk dievaluasi. Komponen kedua melakukan fungsi evaluasi sebagai ukuran pemilihan subset.

Ide dasar dari pendekatan wrapper ini bahwa setiap fitur dari feature subset dievaluasi oleh inductive learning algorithm yang dibungkus didalam prosedur feature selection sebagai "black box".

2.3.2. Metode Unsupervised Feature Selection

2.3.2.1. *Document Frequency*

Document Frequency adalah jumlah dokumen yang mengandung suatu term tertentu. Tiap term akan dihitung nilai *Document Frequency*-nya (DF). Lalu term tersebut diseleksi berdasarkan jumlah nilai DF. Jika nilai DF berada di bawah threshold yang telah ditentukan, maka term tersebut akan dibuang.

Asumsi awalnya adalah bahwa term yang lebih jarang muncul tidak memiliki pengaruh yang besar dalam proses pengelompokkan dokumen. Pembuangan term yang jarang ini dapat mengurangi dimensi fitur yang besar pada text mining.

Perbaikan dalam pengelompokkan dokumen ini juga dapat terjadi jika term yang dibuang tersebut juga merupakan noise term. Document Frequency merupakan metode future selection yang paling sederhana dengan waktu komputasi yang rendah [17].

Ilustrasi dari metode *Document Frequency* ini adalah sebagai berikut. Jika terdapat data berjumlah 5000 dokumen, dan jumlah dokumen yang mengandung term "teknologi" adalah 150 dokumen. Maka nilai DF(teknologi) adalah 150.

2.3.2.2. TERM CONTRIBUTION

Term Contribution diperkenalkan pertama kali oleh Tao Liu dan kawan-kawannya pada tahun 2003 [9]. Ide dasarnya adalah bahwa hasil dari clustering teks sangat bergantung pada kesamaan dokumen. Jadi, kontribusi dari sebuah term dapat dipandang sebagai kontribusinya terhadap kesamaan dokumen. Kesamaan antar dokumen **di** dan **di** dapat dihitung menggunakan dot product :

$$sim(d_{i,}d_j) = \sum_t f(t,d_i).f(t,d_j) \qquad (1)$$

$$TC(t) = \sum_{i,j \cap i \neq j} f(t,d_i).f(t,d_j) \qquad (2)$$

Di mana, f(t,d) merupakan bobot tf*idf dari term t di dokumen d.

Jadi kontribusi dari sebuah term pada dataset, sama dengan kontribusinya secara keseluruhan pada kesamaan dokumen. Persamaannya yaitu :

Metode "TC" digunakan untuk menghitung nilai tf*idf tiap term dengan cara menggunakan Term Frequency (TF) dan mengalikannya dengan bobot **Inverse** Document Frequency (IDF) dari term tersebut, dan akhirnya menormalisasikan panjang dokumen. Persamaannya yaitu :

$$w_{ik} = \frac{tf_{ik} x \log\left(N/n_i\right)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ik} x \log\left(N/n_i\right)\right]^2}} \qquad (3)$$

Jika bobot semua term sama, maka nilai f(t,d) = 1 ketika term t muncul di dokumen d. Sehingga nilai TC(t) bisa ditulis dalam persamaan berikut:

$$TC(t) = DF(t)(DF(t) - 1) \qquad (4)$$

2.3.2.3. Term Frequency (Tf)

Term Frequency merupakan salah satu metode untuk menghitung bobot tiap *term* dalam text. Dalam metode ini, tiap *term* diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan *term* tersebut pada text [10]. Bobot sebuah *term* t pada sebuah text d dirumuskan dalam persamaan berikut:

$$W(d,t) = TF(d,t) \qquad (5)$$

Dimana TF (d,t) adalah *term frequency* dari *term* t di text d. *Term frequency* dapat memperbaiki nilai *recall* pada *information retrieval*, tetapi tidak selalu memperbaiki nilai *precision.* Hal ini disebabkan *term* yang *frequent* cenderung muncul di banyak text, sehingga *term – term* tersebut memiliki kekuatan diskriminatif / keunikan yang kecil. Untuk memperbaiki permasalahan ini, *term* dengan nilai frekuensi yang tinggi sebaiknya dibuang dari set *term*. Menemukan *threshold* yang optimal merupakan fokus dari metode ini

2.3.2.4. Inverse Document Frequency (Idf)

Jika *Term Frequency* fokus pada kemunculan *term* dalam sebuah text, *Inverse Document Frequency (IDF)* fokus pada kemunculan *term* pada keseluruhan koleksi text. Pada IDF, *term* yang jarang muncul pada keseluruhan koleksi *term* dinilai lebih berharga. Nilai kepentingan tiap *term* diasumsikan berbanding terbalik dengan jumlah text yang mengandung *term* tersebut [10]. Nilai IDF sebuah *term* t dirumuskan dalam persamaan berikut:

$$IDF(t) = \log(N/df(t)) \qquad (6)$$

Di mana N adalah total jumlah text / dokumen pada koleksi dan df(t) adalah jumlah dokumen yang mengandung *term* t. Persamaan ini mengacu pada definisi Salton [11]. IDF dapat memperbaiki nilai *precision*, karena mengkhususkan fokus pada sebuah *term* dalam keseluruhan dokumen. Penelitian belakangan ini [11] telah mengkombinasikan TF dan IDF untuk menghitung bobot *term* dan menunjukkan bahwa gabungan keduanya menghasilkan performansi yang lebih baik. Kombinasi bobot dari sebuah *term* t pada text d didefinisikan sebagai berikut:

$$TFIDF(d,t) = TF(d,t) \cdot IDF(t) \qquad (7)$$

Faktor TF dan IDF dapat berkontribusi untuk memperbaiki nilai recall dan precision [11].

2.4. Clustering

Pesatnya pertambahan jumlah dan keanekaragaman dokumen dapat berdampak besar pada saat pencarian suatu dokumen. *Clustering* dokumen merupakan tool yang penting pada pengorganisasian dokumen yang efisien [13]. Subactivitas *clustering term*asuk representasi dokumen, penurunan dimensi, penggunaan *cluster* algorithm dan evaluasi [1]. Riset di text *clustering* ini sudah banyak dikerjakan, *term*asuk oleh [2, 4, 5, 7, 9, 10, 13, 16, 17].

Metode *clustering* tidak memiliki pendefinisian target *class*, karenanya disebut sebagai juga sebagai *unsupervised learning* [12]. Analisis *cluster* membagi data menjadi beberapa *cluster – cluster* (kelompok) yang memiliki arti, berguna atau keduanya. Untuk mendapatkan pengkategorian dari hasil pencarian yang baik, maka dapat diterap beberapa algoritma yang ada pada *clustering*, salah satu diantaranya yaitu Suffix Tree *Clustering* (S*TC*), K-Means dan algoritma lainnya.

Permasalahan mendasar pada *clustering* dokumen adalah tingginya dimensi data. Beberapa metode untuk mengurangi dimensi tersebut telah dilakukan. Ada 2 cara untuk mengurangi dimensi data, yaitu *feature selection* dan feature transformation.

2.5. Algortima K-Means

Eksperimen ini menggunakan algoritma yang paling umum digunakan dalam *clustering* yaitu algoritma *K-Means*. Algoritma ini populer karena mudah diimplementasikan dan kompleksitas waktunya linear. Kelemahannya adalah algoritma ini sensitif terhadap inisialisasi *cluster*

Dasar algoritmanya adalah sebagai berikut:

1) Inisialisasi *cluster*
2) Masukkan setiap dokumen ke *cluster* yang paling cocok berdasarkan ukuran kedekatan dengan centroid. Centroid adalah vektor *term* yang dianggap sebagai titik tengah *cluster*.
3) Setelah semua dokumen masuk ke *cluster*. Hitung ulang centroid *cluster* berdasarkan dokumen yang berada di dalam *cluster* tersebut.
4) Jika *centroid* tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2.

Ukuran kedekatan antara dua vector *term* t1,t2 yang digunakan pada paper ini adalah cosinus sudut antara kedua vektor tersebut

$$Cos(t_1,t_2) = \frac{t_1 t_2}{|t_1||t_2|} \qquad (8)$$

2.6. *Evaluation Measures*

Untuk mengevaluasi apakah hasil clustering yang diperoleh baik atau tidak maka perlu dilakukan validasi *clustering* yang bertujuan untuk membandingkan hasil *clustering* dengan informasi class sesungguhnya dan membandingkan antara dua hasil *clustering* untuk mengetahui hasil mana yang lebih baik [14].

2.6.1. Precision

Precision adalah rasio penempatan cluster yang benar oleh sistem dibagi keseluruhan penempatan oleh sistem. Semakin besar nilai *Precision*, maka semakin bagus *cluster* yang dihasilkan [9]. Rumusnya:

$$\Pr ecision(A) = \frac{1}{|A|}\max\left(\left\{d_i \middle| label(d_i) = c_j\right\}\right) \quad (9)$$

$$\Pr ecision = \sum_{k=1}^{G'} \frac{|A_k|}{N} \Pr ecision(A_k) \quad (10)$$

Penjelasan untuk semua notasi diatas :

- A= jumlah dokumen yang diklasterkan dalam satu kategori
- d_j= kategori yang diberikan
- c_j = *cluster* yang dibentuk
- L = jumlah *class*
- A_k =jumlah dokumen yang diklasterkan pada semua kategori
- N =total jumlah dokumen

2.6.2. Entropy

Entropy mengukur kemurnian dari klaster yang dihasilkan dengan memperhatikan pada kategori yang ada. Nilai *Entropy* yang lebih kecil menghasilkan klaster yang lebih bagus kualitasnya [9].

$$P_{jk} = \frac{1}{|A_k|}\left|\left\{d_i \middle| label\left(d_i = c_j\right)\right\}\right| \quad (11)$$

$$Entropy = -\sum_{k=1}^{G'} \frac{|A_k|}{N} \sum_{j=1}^{G} p_{jk} x \log\left(p_{jk}\right) \quad (12)$$

Penjelasan untuk semua notasi diatas :

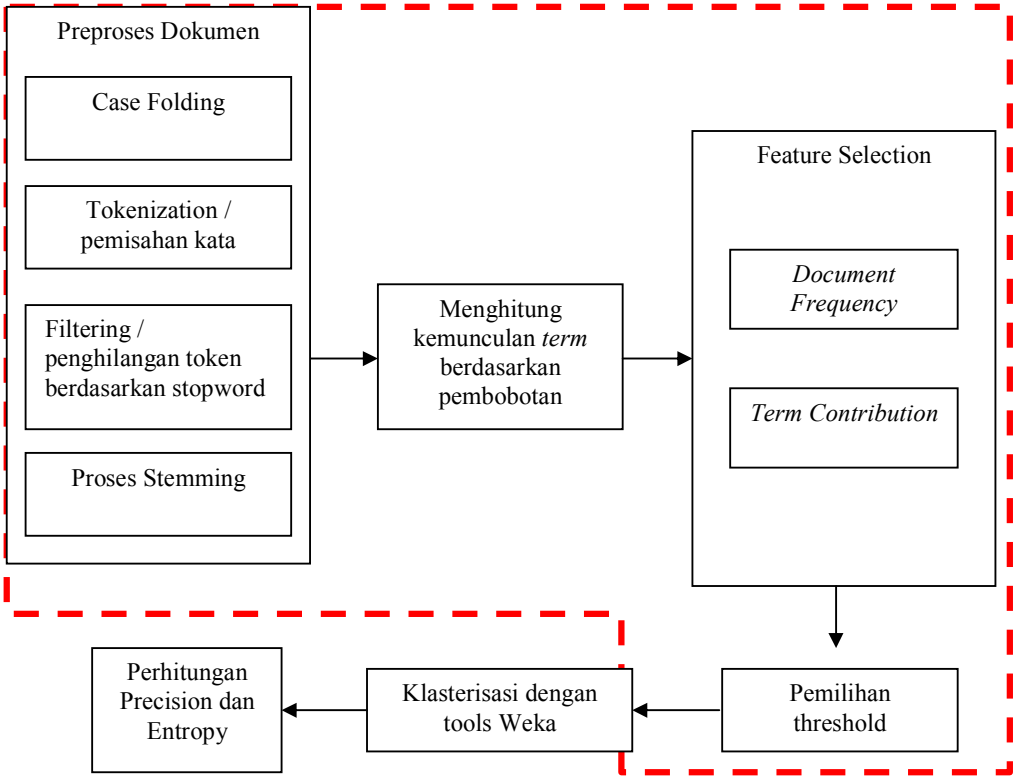
- A= jumlah dokumen yang diklasterkan dalam satu kategori
- d_j= kategori yang diberikan
- c_j = *cluster* yang dibentuk
- L = jumlah *class*
- A_k =jumlah dokumen yang diklasterkan pada semua kategori
- N =total jumlah dokumen
- p_{jk}= peluang dokumen kategori k masuk ke klaster j

3. GAMBARAN UMUM SISTEM

3.1. Analisis Sistem

Dalam sistem yang akan dikerjakan pada Tugas Akhir ini, masukkan ke sistem adalah artikel berita yang berasal dari website media cetak yang diambil secara offline. Awalnya, artikel tersebut akan dimasukkan ke dalam file dokumen berita berekstensi .txt. Awalnya, dilakukan preproses dokumen untuk mengubah struktur dokumen menjadi data dengan struktur yang sesuai dengan cara melakukan *case folding, tokenization, filtering dan stemming-*.

Setelah preproses berakhir, dilakukan proses *Feature Selection*. Akan tetapi sebelum proses *Feature Selection* dilakukan, perlu dilakukan perhitungan kemunculan *term* dalam dokumen. Hasil perhitungan inilah yang akan menjadi inputan bagi proses *Feature Selection*. Kemudian dilakukan *clustering* menggunakan tools. Kemudian proses klasifikasi pun berakhir dengan perhitungan precision dan entropy sebagai pengukuran performansi yang dihasilkan. Proses ini ditunjukkan pada gambar 2.



Gambar 2. Gambaran proses sistem secara umum

Secara umum, hasil akhir dari system yang dibuat pada Tugas Akhir ini adalah record hasil *Feature Selection* yang siap untuk diolah ke dalam tools sesuai bagian yang diberi garis putus-putus.

4. PENGUJIAN DAN ANALISIS

Dalam sistem yang dibuat, masukkan ke sistem adalah artikel berita yang berasal dari website media cetak yang diambil secara offline. Awalnya, artikel tersebut akan dimasukkan ke dalam file dokumen berita berekstensi .txt. Awalnya, dilakukan preproses dokumen untuk mengubah struktur dokumen menjadi data dengan struktur yang sesuai dengan cara melakukan *case folding, tokenization, filtering dan stemming*-.

Setelah preproses berakhir, dilakukan proses *Feature Selection*. Akan tetapi sebelum proses Feature Selection dilakukan, perlu dilakukan perhitungan kemunculan term dalam dokumen. Hasil perhitungan inilah yang akan menjadi masukan bagi proses *Feature Selection*. Kemudian dilakukan clustering menggunakan tools. Kemudian proses klasifikasi pun berakhir dengan perhitungan precision dan *entropy* sebagai pengukuran performansi yang dihasilkan.

Dataset yang digunakan yaitu *Indonesian TREC-like Corpus* yaitu dataset yang berisikan kumpulan artikel-artikel yang berasal dari media surat kabar kompas (www.kompas.com), dataset ini menggunakan 120 artikel dengan 6 kategori dan 225 artikel dengan 6 kategori.

4.1. Analisis Precision

Dalam hal ini, diukur nilai *Precision* dengan clustering *K-Means* terhadap dataset 120 dokumen yang telah dilakukan *Feature Selection* dengan metode *Document Frequency* dan *Term Contribution*. Hal ini ditunjukkan pada gambar 1 pada lampiran.

Pada *Term Contribution* dapat dilihat bahwa, nilai precision cenderung meningkat, mulai dari pembuangan *term* 60% hingga 96%, kemudian sedikit menurun ketika term dibuang hingga 96 %, dan nilai *precision* terus stabil ketika term dibuang 97% hingga 99%. Nilai *precision* yang semakin meningkat membuktikan bahwa performansi clustering menjadi lebih baik ketika fiturnya dikurangi sebanyak 96%.

Pada *Document Frequency*, dapat disimpulkan bahwa nilai precision pada Document Frequency, cenderung mengalami peningkatan hingga term dibuang sebesar 94%, berbeda dengan *Term Contribution* yang baru menurun stabil ketika term dibuang sebesar 96%. Hal ini berarti jumlah maksimal term yang dapat dibuang

untuk meningkatkan performansi *clustering*, lebih sedikit 2% jika menggunakan metode Term Contribution daripada metode *Document Frequency*.

Dari pengujian ini juga dapat dilihat bahwa adanya kedua metode *feature selection* ini dapat memperbaiki performansi *clustering*. Karena ketika jumlah term masih 100% atau tidak dikurangi fiturnya, maka nilai precisionnya lebih rendah jika dibandingkan dengan nilai *precision* setelah pengurangan term yaitu 4-975, kecuali jika *term* tersebut dibuang 40%. Selain itu jika dibandingkan dengan seksama, maka dapat disimpulkan nilai precision *Term Contribution* rata – rata lebih baik untuk setiap tahap pembuangan term. Hal ini berarti metode *Term Contribution* lebih baik daripada *Document Frequency* dalam memperbaiki performansi *clustering*. Hal ini dapat dihubungkan dengan perbedaan sifat kedua metode tersebut, yaitu bahwa *Term Contribution* lebih handal daripada Document Frequency, karena *Term Contribution* memberikan bobot pada sebuah *term* dengan mempertimbangkan kontribusi atau peran term tersebut yaitu adanya nilai *term frequency* dan inverse *Document Frequency* pada setiap pasangan dokumen. Jika term tersebut terjadi di banyak pasangan dokumen, berarti *term* tersebut memiliki kemiripan atau kesamaan topic.

4.2. Analisis Entropy

Pada percobaan dengan 120 dokumen ini dapat dilihat bahwa unsupervised *feature selection* yaitu *Document Frequency* dan *Term Contribution* dapat memperbaiki performansi *clustering* yaitu nilai *Entropy*. *Entropy* mengukur kemurnian dari klaster yang dihasilkan dengan memperhatikan pada kategori yang ada. Nilai *Entropy* yang lebih kecil menghasilkan klaster yang lebih bagus kualitasnya [9].

Dalam hal ini, diukur nilai Entropy dengan *clustering K-Means* terhadap dataset 120 dokumen yang telah dilakukan *Feature Selection* dengan metode *Document Frequency* dan *Term Contribution*. Hal ini ditunjukkan pada gambar 1 pada lampiran.

Nilai maksimum *entropy* pada *Document Frequency* dan *Term Contribution* adalah sama yaitu 0.49 ketika term dibuang 40%. Hal ini berarti jumlah term yang dibuang saat itu sama dan menghasilkan performansi *clustering* yang paling buruk.

Dari hasil percobaan ini dapat dilihat bahwa adanya kedua metode *feature selection* ini dapat memperbaiki performansi *clustering*. Karena ketika jumlah *term* masih 100% atau tidak dikurangi fiturnya, maka nilai *entropy*-nya lebih tinggi jika dibandingkan dengan nilai *entropy* setelah pengurangan *term* yaitu 0.43, kecuali jika term tersebut dibuang 40%. Selain itu jika dibandingkan dengan seksama, maka dapat disimpulkan nilai entropy *Term Contribution* rata – rata lebih baik untuk beberapa tahap pembuangan term. Hal ini berarti metode *Term Contribution* lebih baik daripada *Document Frequency* dalam memperbaiki performansi *clustering*.

4.3. ANALISIS PERSAMAAN *TERM CONTRIBUTION*

Term Contribution melihat pentingnya pengaruh kesamaan dokumen – dokumen yang ada terhadap *clustering* teks. Jadi, kontribusi atau pengaruh suatu term, dapat dipandang sebagai kontribusi atau pengaruhnya terhadap kesamaan seluruh dokumen yang ada

Document Frequency menganggap setiap *term* memiliki tingkat kepentingan yang sama walaupun terdapat di berbagai dokumen. Hal ini berarti semakin banyak term tersebut terdapat di dalam dokumen yang berbeda, maka nilainya semakin besar dan memiliki pengaruh yang semakin besar pula pada clustering dokumen

Jika diperhatikan dari kedua percobaan sebelumnya, maka dapat diketahui nilai *precision* dan *entropy* kedua percobaan di atas sama hingga term dibuang sebanyak 40%. Hal ini terjadi karena kedua metode tersebut membuang term yang sama. Untuk mempermudah proses analisa, maka dilakukan percobaan dengan dataset yang lebih sedikit yaitu berjumlah 4 dokumen dan 621 term. Ketika term dibuang 40% maka terlihat, term yang dibuang adalah term yang sama yaitu yang memiliki nilai *Document Frequency* yang kecil. Tabel 1 dan 2 pada lampiran menunjukkan hasil percobaan ini.

Pada *Term Contribution*, suatu term dianggap penting jika term tersebut terjadi di sedikit dokumen dan memiliki frekuensi term yang besar di sebuah dokumen. Dengan cara ini, maka hasil clustering yang dihasilkan akan lebih baik, karena term yang tersisa adalah term yang khas atau bersifat diskriminator. Sedangkan jika nilai *term frequency* tidak dilihat, atau frekuensi kemunculan term dalam sebuah dokumen hanya bersifat Boolean, maka metode TC akan mirip dengan metode DF yaitu hanya memperhatikan jumlah dokumen yang dimiliki term. Oleh karena itu, metode DF adalah bentuk khusus dari metode TC.

Kesimpulan lain, yang dapat diambil adalah bahwa nilai *precision* dan *entropy* pada kedua percobaan di atas, memiliki pola yang hampir sama, namun naik turun. Artinya, *feature selection* tidak selalu dapat memperbaiki kualitas atau performansi clustering. Dapat dilihat pada percobaan pertama, performansi clustering mengalami penurunan pada pembuangan term sebesar 40%. Oleh karena itu perlu dicari titik maksimal di mana feature selection menghasilkan nilai performansi yang terbaik. Pada kedua percobaan di atas, diperoleh hasil bahwa feature selection akan memperoleh hasil terbaik pada pembuangan term 96% untuk *Term Contribution* dan 94% untuk *Document Frequency*.

5. KESIMPULAN

Dari hasil analisis dan pengujian dalam penelitian ini, maka didapatkan kesimpulan :

1. *Feature Selection* dapat mempengaruhi performansi *clustering* berdasarkan pengukuran precision dan entropy. Tetapi *feature selection* tidak selalu dapat memperbaiki kualitas atau performansi *clustering*, ada beberapa titik di mana *feature selection* justru dapat mengurangi nilai performansi, oleh karena itu perlu dicari titik maksimal pembuangan *term* dari data set.
2. *Term Contribution* dan *Document Frequency* dapat mengurangi dimensi data serta memperbaiki performansi *clustering* berdasarkan pengukuran precision dan entropy dengan cara menghilangkan fitur yang tidak relevan, redundant dan noise.
3. Pada *Term Contribution*, *term* yang dianggap baik adalah *term* yang memiliki nilai *term* frequency yang besar dan terjadi di sedikit dokumen.
4. Pada *Document Frequency*, *term* yang dianggap baik adalah *term* yang terjadi di banyak dokumen.
5. *Term Contribution* lebih baik daripada *Document Frequency* yaitu dapat menghasilkan nilai precision dan entropy lebih baik dengan fitur yang lebih sedikit. Hal ini dikarenakan *Term Contribution* mempertimbangkan frekuensi kemunculan *term* dan frekuensi dokumen sebuah *term*, sehingga *term* yang tetap dipertahankan adalah *term* yang khas atau bersifat diskriminator, berbeda halnya dengan *Document Frequency* yang hanya mempertahankan *term – term* yang bersifat umum.

6. DAFTAR PUSTAKA

[1] Adiwijaya, Igg. (2006). *Texi Mining dan Knowledge Discovery.* Komunitas *Data mining* Indonesia & Soft-omputing Indonesia.

[2] Chen Jinxiu, Ji,Tan, *Unsupervised Feature Selection for Relation Extraction*, National University of Singapore,2005

[3] Dash Manorjan , Liu. Dimensionality Reduction. National University of Singapore. 1997

[4] Dash Manorjan, Liu, *Feature Selection for Clustering*, PAKDD, 2000.

[5] Devaney, M. & A. Ram. *Efficient feature selection in conceptual clustering*. In p*roceedings of the Fourteenth International Conference on Machine Learning*, pages 92–97, 1997.

[6] Franke J, Gholamreza Nakhaeizadeh, and Ingrid Renz. *Text mining*: Theoretical Aspects and Applications

[7] Lerman, Kristina. (1999). *Document Clustering in Reduced Dimension Vector Space.*

[8] Liu Huan & Lei Yu. (2005). Toward Integrating *Feature Selection* Algorithms for Classification and *Clustering*.

[9] Liu, Liu, Chen, Ma, *An Evaluation of feature selection for clustering*, ICML Conference, 2003

[10] Mark A. Hall and Llioyd A. Smith. Feature Subset Selection : A Correlation Based Filter Approach. University of Wakaito.

[11] Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-wesley, Reading, Pennsylvania.

[12] Tan, Pang-ning, Michael Steinbach, dan Vipin Kumar. 2006. *Introduction to Data mining*. Pearson education, Inc.

[13] Tien Dung Do, Hui, Fong, *Associative Feature Selection for Text mining,* Nanyang Technological University,2006

[14] Tokunaga, Takenobu. Iwayama, Makoto. (1994). *Text Categorization based on Weighted Inverse Document Frequency.*

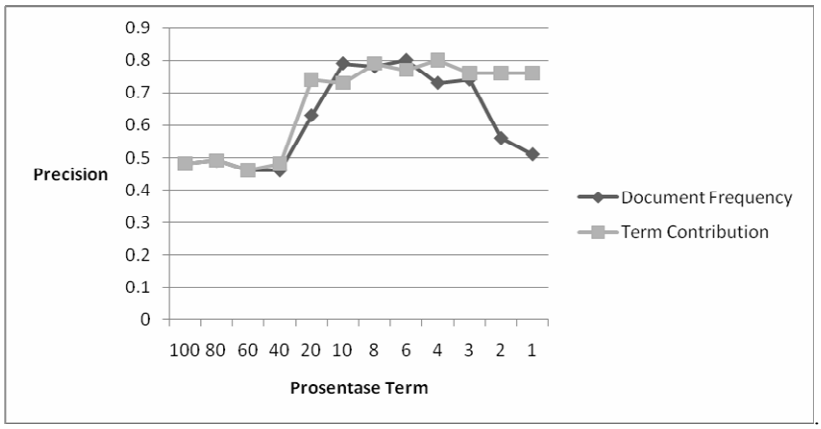
[15] Wibisono, Yudi., & Khodra, M. L. (2006). *Clustering Berita Berbahasa Indonesia*

[16] Wiratunga Nirmalie, Lothian, Massie, *Unsupervised Feature Selection for Text data* , Proceedings of the 8th European Conference on Case-Based Reasoning,2006

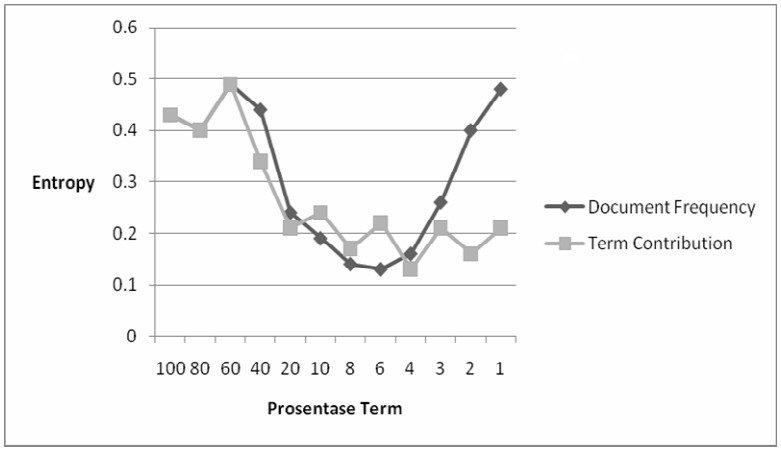
[17] Yang, Y., & Pedersen, J. O. (1997). A comparative study on *feature selection* in text categorization. *Proc. of ICML- 97* (pp. 412-420).

[18] Zexuan Zhu, Yew-Soon Ong, and Manoranjan Dash. Wrapper-Filter *Feature Selection* Algorithm Using A memetic Framework. Nanyang Technological University, Singapore

LAMPIRAN



***Gambar 1: Nilai Precision dengan* clustering *K-Means terhadap dataset 120 dokumen yang telah dilakukan* Feature Selection *dengan metode* Document Frequency *dan* Term Contribution.**



Gambar 2 Nilai Entropy dengan clustering K-Means terhadap dataset 120 dokumen yang telah dilakukan Feature Selection dengan metode Document Frequency dan Term Contribution.

Tabel 1. Tabel *Term* pada 1% *term* terakhir dengan metode DF

No	*Term*	No	*Term*
1	dua	26	sementara
2	jakarta	27	ri
3	media	28	akhirnya
4	as	29	indonesia
5	pukul	30	menyatakan
6	wib	31	terus
7	empat	32	rabu
8	sedang	33	bagian
9	kata	34	menghadapi
10	sebelum	35	tinggi
11	atas	36	ii
12	tenggara	37	dua

13	melakukan	38	bela
14	menurut	39	memerintah
15	terjadi	40	mata
16	dalam	41	pagi
17	sekitar	42	katak
18	keadaan	43	mencapai
19	mengenai	44	dolar
20	sehingga	45	pemasaran
21	akibat	46	bank
22	kembali	47	dibanding
23	hari	48	menguat
24	besar	49	uang
25	ditutup	50	nya

Tabel 2 Tabel *Term* pada 1% *term* terakhir dengan metode *TC*

No	*Term*	No	*Term*
1	tni	26	Megawati
2	as	27	Cm
3	pesawat	28	Bank
4	presiden	29	Naik
5	sisi	30	rp
6	air	31	Pedagang
7	berjalan	32	Disbanding
8	ditutup	33	Menguat
9	ri	34	Menghadapi
10	indonesia	35	Level
11	tim	36	Menutup
12	terus	37	Uang
13	menghadapi	38	Nya
14	tinggi	39	Poin
15	bela	40	Modal
16	mata	41	l
17	pagi	42	Transaksi
18	posisi	43	Dimainkan
19	mebel	44	pertandingan
20	dolar	45	Regional
21	pemasaran	46	United
22	banjir	47	rupiah
23	pintu	48	Yen
24	menggenang	49	Selian
25	cenderung		