**CLUSTERING ARTIKEL BERITA BERBAHASA INDONESIA
MENGUNAKAN UNSUPERVISED FEATURE SELECTION****Catur tyas surya wigati
175110012**Universitas Mitra Indonesia, Tehnnik informatika
Caturtyassuryawigati.student@umitra.ac.id**ABSTRAK**

Meningkatnya penggunaan internet telah memicu pertumbuhan dan pertukaran informasi menjadi jauh lebih pesat dibandingkan era sebelumnya. Volume berita elektronik berbahasa Indonesia semakin bertambah besaran menyimpan informasi yang berharga di dalamnya. Pengelompokan 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 penggunaan dan adopsi Internet telah memacu pertumbuhan dan pertukaran informasi yang sangat pesat dibandingkan era sebelumnya. Sebagai akibatnya, jumlah informasi terus meningkat secara eksponensial, lebih dari 550 triliun dokumen saat ini. Sebanyak 7.3 juta Internet page baru tiap hari nya. Walau perkembangan ini memungkinkan informasi untuk di akses pengguna dengan mudah, jumlah yang terkendali ini telah menimbulkan isu dan 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 dilakukan 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 analisis 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, ke dalam 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 Preprocessing

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 / stopwords 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 pengelompokan dokumen. Pembuangan term yang jarang ini dapat mengurangi dimensi fitur yang besar pada text mining. Perbaikan dalam pengelompokan dokumen ini juga dapat terjadi jika term yang dibuang tersebut juga merupakan noise term. Document Frequency merupakan metode feature 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]. Idedarnya 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 dapat dihitung menggunakan dot product :

$$=_{\text{t}} \left(\begin{array}{c} \text{ij ij sim(d d) ft, d . ft, d , (1)} \\ -()() \\ =- \\ = \\ \text{ij ij} \\ \text{ij TC t f t d f t d} \\ , \\ () , \dots (2) \end{array} \right)$$

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

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

$$- = -$$

-

-

-

-

-

-

- ---

-

=

M

j

i

ik

i

ik

ik

n

tf x N

n

tf x N

w

1

2

log

log

(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:

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

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:

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

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 di buang dari set term. Menemukan threshold yang optimal merupakan fokus dari metode ini Seminar Nasional Informatika 2010 (semnasIF 2010) ISSN: 1979-2328

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:

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

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 gabungannya menghasilkan performansi yang lebih baik. Kombinasi bobot dari sebuah term t pada text d didefinisikan sebagai berikut:

$TFIDF(d,t) = TF(d,t) \times IDF(t)$ (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]. Subaktivitas clustering termasuk representasi dokumen, penurunan dimensi, penggunaan cluster algorithm dan evaluasi [1]. Riset di text clustering ini sudah banyak dikerjakan, termasuk 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 diterapkan beberapa algoritma yang ada pada clustering, salah satu diantaranya yaitu Suffix Tree Clustering (STC), 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. Algoritma 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 didalam cluster tersebut.

4) Jika centroid tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2. Ukuran kedekatan antara dua vector term t_1, t_2 yang digunakan pada paper ini adalah cosinus sudut antara kedua vektor tersebut

()

1 2

1 2

1 2 .

.

,

t t

t t

$\cos t t = (8)$

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:

$(\{ \}) i i j d \text{ label } d c$

A

recision A = max () =

1

Pr ()

(9)

Pr Pr ()

,

1

k

G

k

k ecision A

N

A

recision _

=

= (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

• k A = 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].

$\{ () \} i i j$

k

$j k d$ label $d c$

A

$P = = 1$

(11)

$-- ()$

$= =$

$= -$

G

j

$j k j k$

G

k

$k p \times p$

N

A

Entropy

$1 1$

$\log$

,

(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

• k A = jumlah dokumen yang diklasterkan pada semua kategori

- N = total jumlah dokumen

- p_{jk} = peluang dokumen kategori k masuk ke klaster j

3. STUDI KASUS

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. DISKUSI

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, 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 Menghitung kemunculan term berdasarkan pembobotan

Feature Selection

Document

Frequency

Term Contribution

Pemilihan

threshold

Klasterisasi dengan

tools Weka

Perhitungan

Precision dan

Entropy

Preproses Dokumen

Proses Stemming

Tokenization /

pemisahan kata

Filtering /

penghilangan token

berdasarkan stopword

Case Folding

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] PUTRA, Arie Setya; FEBRIANI, Ochi Marshella. Knowledge Management Online Application in PDAM Lampung Province. In: *Prosiding International conference on Information Technology and Business (ICITB)*. 2018. P. 181-187.
- [2] FEBRIANI, Ochi Marshella; PUTRA, Arie Setya. Sistem Informasi Monitoring Inventori Barang Pada Balai Riset Standardisasi Industri Bandar Lampung. *Jurnal Informatika*, 2014, 13.1: 90-98
- [3] Putra, A. S. (2018, July 9). 2018 Artikel Struktur Data , Audit dan Jaringan Komputer. Retrieved from osf.io/3uq8w

- [4] Adiwijaya, Igg. (2006). Texi Mining dan Knowledge Discovery. Komunitas Data mining Indonesia & Soft-computing Indonesia.
- [5] Chen Jinxiu, Ji, Tan, Unsupervised Feature Selection for Relation Extraction, National University of Singapore, 2005
- [6] Dash Manojan, Liu. Dimensionality Reduction. National University of Singapore. 1997