REDUKSI DIMENSI FITUR MENGGUNAKAN ALGORITMA ALOFT UNTUK PENGELOMPOKAN DOKUMEN

Mamluatul Hani'ah[1)], Chastine Fatichah[2)], dan Diana Purwitasari[3)]
[1, 2, 3)] Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Kampus ITS, Sukolilo, Surabaya 60111
e-mail: mamluatul.haniah@gmail.com[1)], chastine.fatichah@gmail.com[2)], diana.purwitasari@gmail.com[3)]

ABSTRAK

Pengelompokan dokumen masih memiliki tantangan dimana semakin besar dokumen maka akan menghasilkan fitur yang semakin banyak. Sehingga berdampak pada tingginya dimensi dan dapat menyebabkan performa yang buruk terhadap algoritma clustering. Cara untuk mengatasi masalah ini adalah dengan reduksi dimensi. Metode reduksi dimensi seperti seleksi fitur dengan metode filter telah digunakan untuk pengelompokan dokumen. Akan tetapi metode filter sangat tergantung pada masukan pengguna untuk memilih sejumlah n fitur teratas dari keseluruhan dokumen. Algoritma ALOFT (At Least One FeaTure) dapat menghasilkan sejumlah set fitur secara otomatis tanpa adanya parameter masukan dari pengguna. Karena sebelumnya algoritma ALOFT digunakan pada klasifikasi dokumen, metode filter yang digunakan pada algoritma ALOFT membutuhkan adanya label pada kelas sehingga metode filter tersebut tidak dapat digunakan untuk pengelompokan dokumen. Pada penelitian ini diusulkan metode reduksi dimensi fitur dengan menggunakan variasi metode filter pada algoritma ALOFT untuk pengelompokan dokumen. Sebelum dilakukan proses reduksi dimensi langkah pertama yang harus dilakukan adalah tahap preprocessing kemudian dilakukan perhitungan bobot tfidf. Proses reduksi dimensi dilakukan dengan menggunakan metode filter seperti Document Frequency (DF), Term Contribution (TC), Term Variance Quality (TVQ), Term Variance (TV), Mean Absolute Difference (MAD), Mean Median (MM), dan Arithmetic Mean Geometric Mean (AMGM). Selanjutnya himpunan fitur akhir dipilih dengan algoritma ALOFT. Tahap terakhir adalah pengelompokan dokumen menggunakan dua metode clustering yang berbeda yaitu k-means dan Hierarchical Agglomerative Clustering (HAC). Dari hasil ujicoba didapatkan bahwa kualitas cluster yang dihasilkan oleh metode usulan dengan menggunakan algoritma k-means mampu memperbaiki hasil dari metode VR.

***Kata Kunci**: ALOFT, metode filter, pengelompokan dokumen, reduksi dimensi*

ABSTRACT

Document clustering still have a challenge when the volume of document increases, the dimensionality of term features increases as well. This contributes to the high dimensionality and may cause poor performance of clustering algorithm. A way to overcome this problem is dimension reduction. Dimension reduction methods such as feature selection using filter method has been used for document clustering. But filter method is highly dependent on user input to select number of n top features from the whole document. ALOFT (At Least One feature) Algorithm can generate a number of feature set automatically without user input. In previous research ALOFT algorithm used on classification documents so filter method require labels on classes. Such filter method can not be used on document clustering. This research proposed feature dimension reduction method by using variations of several filter methods in ALOFT algorithm for document clustering. Before dimension reduction process first step that must be done is preprocessing phase then calculate weight of term using tfidf. filter method used in this study are Document Frequency (DF), Term Contribution (TC), Term Variance Quality (TVQ), Term Variance (TV), Mean Absolute Difference (MAD), Mean Median (MM), and Arithmetic Mean geometric Mean (AMGM). Furthermore, the final feature set selected by ALOFT algorithm. The last phase is document clustering using two different clustering methods, k-means and agglomerative hierarchical clustering (HAC). Experiment results showed that the proposed method using k-means algorithm able to improve results of VR methods.

Keywords: *ALOFT, dimension reduction, document clustering, filter method*

I. PENDAHULUAN

PENINGKATAN jumlah informasi digital yang tersedia di internet dapat menyebabkan pengguna mengalami kesulitan untuk menemukan informasi yang relevan sesuai dengan kebutuhan pengguna. Sehingga diperlukan pengelolaan informasi digital. Cara yang dapat digunakan untuk mengelola informasi digital adalah dengan pengelompokan dokumen. Hasil dari pengelompokan dapat dimanfaatkan untuk pencarian informasi [1], sistem pendukung keputusan [2], perbandingan review produk secara otomatis [3], dan peringkasan dokumen [4][5].

Pengelompokan dokumen merupakan salah satu teknik dalam *text mining* dimana salah satu tahap yang dibutuhkan pada *text mining* adalah tahap *preprocessing*, tahap ini terdiri dari proses *tokenizing, stopword removal*, dan *stemming*. Pembentukan kata dasar dengan proses *stemming* pada dokumen berbahasa Indonesia masih

memiliki kendala dimana tidak semua kata dapat terpotong dengan benar. Misalnya kata "penyidikan" setelah di lakukan *stemming* kata yang dihasilkan menjadi "sidi" padahal seharusnya kata dasar dari penyidikan adalah "sidik". Selain itu algoritma *stemming* bahasa Indonesia yang ada saat ini belum bisa mengatasi masalah pada kata bersisipan [6]. Sehingga pada penelitian ini proses pembentukan kata dasar dilakukan dengan menggunakan produk Kateglo (kamus, tesaurus, dan glosarium) bahasa Indonesia.

Setelah dilakukan *preprocesing* setiap dokumen direpresentasikan menjadi vektor menggunakan *Vector Space Model* (VSM) [7]. Pada VSM setiap term yang terdapat didalam dokumen merupakan representasi dari fitur yang berbeda. Sehingga semakin besar dokumen maka akan menghasilkan fitur yang semakin banyak, ratusan bahkan ribuan fitur. Jumlah fitur yang banyak dapat membuat tingginya komputasi, selain itu jika terdapat banyak fitur yang tidak relevan dapat menyebabkan performa yang buruk dari algoritma *clustering*. Salah satu cara untuk mengatasi dimensi fitur yang tinggi adalah dengan reduksi dimensi [8], [9], [10], [11].

Seleksi fitur merupakan salah satu metode reduksi dimensi dimana metode ini bertujuan untuk menghapus fitur – fitur yang tidak relevan berdasarkan kriteria tertentu. Terdapat dua kategori dari seleksi fitur yaitu *wrapper* dan filter. Untuk mendapatkan fitur yang cocok dengan *wrapper* dibutuhkan biaya yang tinggi karena harus berulang kali melakukan pengujian dengan *machine learning.* Sedangkan model filter menggunakan analisis statistik untuk menentukan relevansi fitur tanpa berulang kali menguji dengan *machine learning*, model filter relatif cepat dan lebih efisien.

Penelitian [12] memperkenalkan beberapa metode filter untuk pengelompokan dokumen yaitu *Document Frequency* (DF), *Term Contribution* (TC), *Term Variance Quality* (TVQ), dan *Term Variance* (TV). Penelitian [13] mengusulkan seleksi fitur dengan model filter pada metode *unsupervised* dan *supervised* untuk data dengan dimensi yang besar. Filter-fiter tersebut dapat mengatasi masalah relevansi dengan perhitungan statistik yaitu *Mean Absolute Difference* (MAD), *Mean Median* (MM) dan *Arithmetic Mean Geometric Mean* (AMGM).

Pemilihan fitur pada model filter yang paling umum adalah dengan perangkingan fitur berdasarkan nilai relevansi dari fitur, kemudian memilih sejumlah *n* fitur teratas dari keseluruhan dokumen, cara ini sering disebut dengan *Variable Rangking* (VR) [14]. Karena VR memilih *n* fitur teratas berdasarkan masukan pengguna maka nilai *n* menjadi sangat penting karena jumlah fitur yang berbeda mungkin akan menghasilkan kelompok dokumen yang berbeda. Selain itu fitur yang dipilih mungkin tidak mencakup keseluruhan dokumen karena semua fitur yang ada dalam dokumen tersebut tidak masuk kedalam *n* fitur teratas.

Untuk mengatasi masalah pemilihan fitur secara VR penelitian yang dilakukan oleh [15] mengusulkan sebuah metode seleksi fitur yang diberi nama *At Least One FeaTure* (ALOFT) untuk pengklasifikasian dokumen. Algoritma ALOFT merupakan seleksi fitur untuk klasifikasi dokumen dengan menggunakan metode filter yang dapat menghasilkan sejumlah himpunan fitur secara otomatis tanpa adanya parameter masukan dari pengguna. Karena sebelumnya algoritma ALOFT digunakan pada klasifikasi dokumen, metode filter yang digunakan algoritma ALOFT membutuhkan adanya label pada kelas sehingga metode filter tersebut tidak dapat digunakan untuk pengelompokan dokumen.

Oleh karena itu, pada penelitian ini diusulkan metode reduksi dimensi fitur dengan menggunakan variasi metode filter pada algoritma ALOFT untuk pengelompokan dokumen. Metode filter yang digunakan adalah metode filter yang tidak membutuhkan label kelas dan berfungsi untuk menghitung / menentukan skor relevansi dari sebuah fitur. Kemudian algoritma ALOFT digunakan untuk memilih sejumlah himpunan fitur secara otomatis. Dengan metode yang diusulkan dapat meningkatkan performa dari algoritma *clustering*. Selain itu dengan penggunaan produk Kateglo untuk pembentukan kata dasar akan terbentuk kata dasar yang sesuai, sehingga hasil *cluster* yang diperoleh akan lebih berkualitas.

II. DATA DAN METODE PENELITIAN

Pada tahap ini pertama akan dibahas mengenai data yang digunakan dan selanjutnya dibahas mengenai detail dari metode yang diusulkan.

A. Data

Data yang digunakan pada penelitian ini diambil dari situs berita online Kompas[1] pada tanggal 21-04-2013 sampai 23-09-2015. Jumlah data yang digunakan sebanyak 1000 dokumen berita dengan rincian 350 berita pada kategori ekonomi, 350 berita pada kategori politik, dan 300 berita pada kategori olahraga. Data yang didapat berupa *xml* dimana di dalam satu dokumen berita kompas masih berisi banyak sekali *tag html* seperti judul, penulis, editor, tanggal terbit, isi berita, dan *link* berita-berita terkait. Bagian yang digunakan pada penelitian ini hanya judul dan isi berita sehingga *tag html* yang tidak digunakan akan dihapus. Data bersih berupa judul dan isi berita selanjutnya disimpan dalam database untuk mempermudah proses selanjutnya.

[1] Berita online Kompas: www.kompas.com

B. Metode

Pada penelitian ini diusulkan metode reduksi dimensi fitur dengan menggunakan variasi metode filter pada algoritma ALOFT untuk pengelompokan dokumen. Tahapan penelitian ini terdiri dari tahap *preprocessing*, perhitungan bobot fitur dengan *tfiidf*, perhitungan relevansi fitur dengan menggunakan metode filter, pemilihan himpunan fitur akhir menggunakan algoritma ALOFT, dan dilanjutkan tahap terakhir yaitu pengelompokan dokumen menggunakan algoritma *k-means* dan Algortima *Hierarchical agglomerative clustering* (HAC). Kemiripan antar dokumen pada tahap pengelompokan dihitung menggunakan *cosine similarity* [16]. Skema dari metode yang diusulkan dapat dilihat pada Gambar 1.

Pada fase *preprocessing* dilakukan proses pembersihan data untuk menghilangkan *tag* html, gambar, tanggal terbit, nama penulis, dan editor sehingga didapatkan data bersih berupa judul berita dan isi berita. Kemudian proses *tokenizing* dilakukan untuk memotong string dokumen input berdasarkan pemisah kata yaitu spasi. Setelah kata (*term*) yang berupa token-token didapat, proses selanjutnya adalah proses *stopword removal*. Pada proses *stopword removal* ini dilakukan penghapusan kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna misalnya kata yang, di, ke, dengan, dan sebagainya. Kamus daftar kata *stopword* yang digunakan pada penelitian ini merupakan kumpulan *stopword* bahasa Indonesia yang didapatkan pada appendix sebuah penelitian [17]. Tahap akhir dari fase *preprocessing* adalah proses pencarian kata dasar dengan menggunakan Kateglo, yaitu dengan cara mencari *root phrase* dari kata turunan.

Setiap term yang dihasilkan dari proses *preprocessing* merupakan representasi dari fitur yang berbeda akan dilakukan perhitungan bobot dengan menggunakan *tfidf*. *Tfidf* merupakan pembobotan sebuah fitur / term (t_k) berdasarkan jumlah frekuensi term (*tf*) yang terdapat pada dokumen dari seluruh koleksi dokumen. Formula dari *tfidf* yang digunakan pada penelitian ini dapat dilihat pada persamaan (1).

$$tfidf(t_k) = tf * \log \frac{N}{df(t_k)} \quad (1)$$

1) Perhitungan Relevansi Fitur dengan Metode Filter

Perhitungan relevansi fitur merupakan fase awal sebelum dilakukan pemilihan fitur untuk mengurangi dimensi dari vektor dokumen. Proses ini bertujuan untuk menentukan skor relevansi dari sebuah fitur dimana semakin tinggi nilai filter dari sebuah fitur maka semakin relevan fitur tersebut, begitu juga sebaliknya. Pada penelitian ini digunakan tujuh penilaian fitur yaitu *Document Frequency* (DF), *Term Contribution* (TC), *Term Variance Quality* (TVQ), *Term Variance* (TV), *Mean Absolute Difference* (MAD), *Mean Median* (MM), dan *Arithmetic Mean Geometric Mean* (AMGM).

- *Document Frequency* (DF)

Perhitungan DF dilakukan dengan cara menghitung jumlah dokumen dimana sebuah fitur muncul. Dalam seleksi fitur metode DF adalah kriteria yang paling sederhana dan mudah untuk dataset yang besar.

- *Term Contribution* (TC)

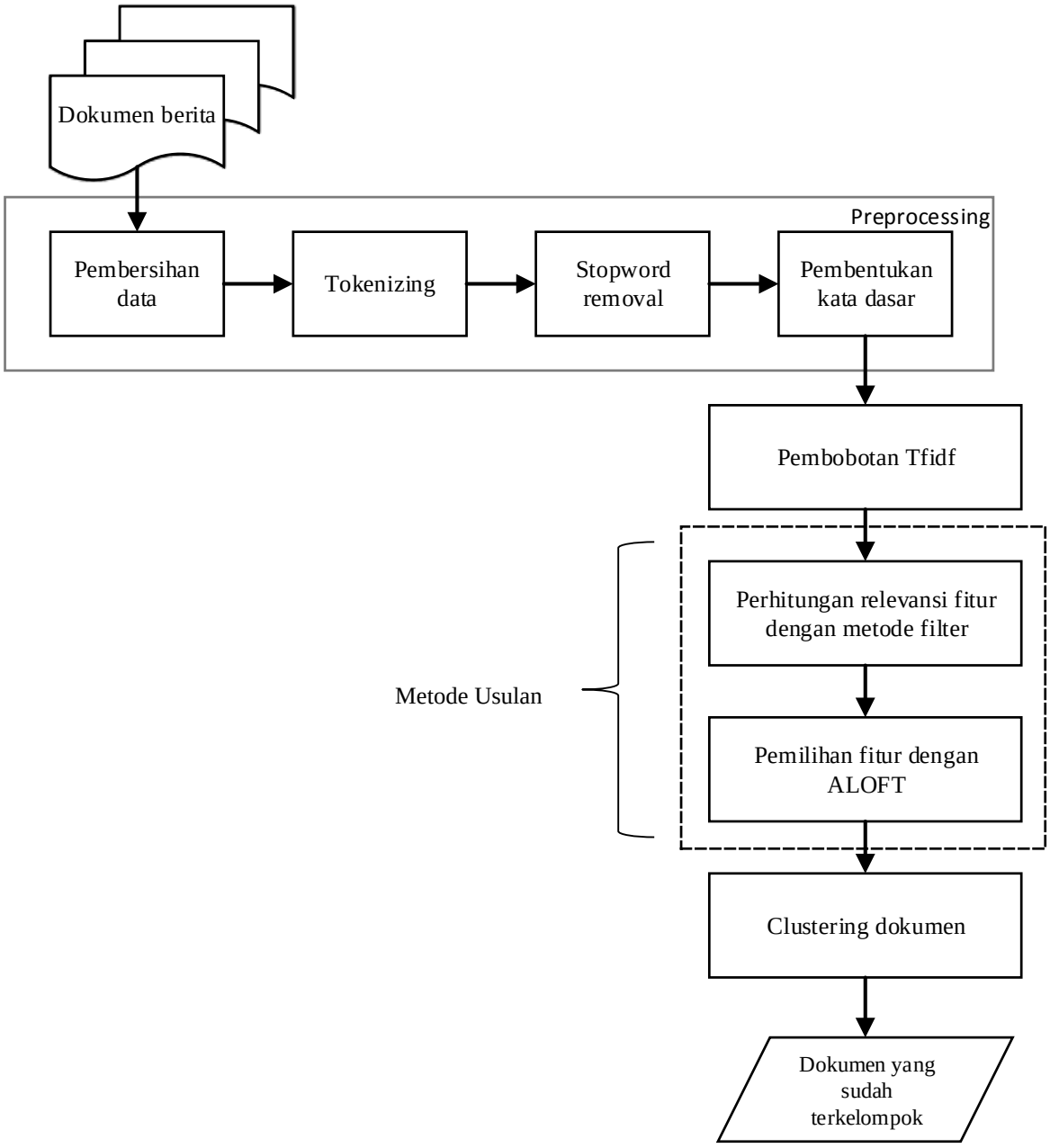
Kontribusi dari term/fitur dapat didefinisikan sebagai kontribusi secara keseluruhan untuk kesamaan dokumen. Jika kemiripan dokumen biasanya dihitung dengan perkalian *dot product* antar dua dokumen maka TC dapat dihitung dengan cara perkalian *dot product* antara dua dokumen (D_i dan D_j) dari kesuluruhan dokumen yang memiliki fitur tersebut dengan syarat bahwa *i* tidak sama dengan *j*. Pada persamaan (2) merupakan formula untuk menghitung filter TC dimana pada filter ini dibutuhkan $tfidf(t_k, D_i)$ yang merupakan *Tfidf* dari fitur t_k pada dokumen D_i dan $tfidf(t_k, D_j)$ yang merupakan *Tfidf* dari term t_k pada dokumen D_j.

$$TC\ (t_k) = \sum_{i,j \cap i \neq j} tfidf(t_k, D_i) * tfidf(t_k, D_j) \quad (2)$$

- *Term Variance Quality* (TVQ)

Term Variance Quality (TVQ) diperkenalkan oleh [18]. Pada persamaan (3) relevansi dari fitur t_k dengan menggunakan filter TVQ membutuhkan nilai *term frekuensi* (f) dari fitur pada dokumen D_i dan juga jumlah dokumen (n) dimana term t_k muncul minimal satu kali.

$$q(t_k) = \sum_{j=1}^{n} f_{kj}^2 - \frac{1}{n} \left(\sum_{j=1}^{n} f_{kj}\right) \quad (3)$$



Gambar 1. Diagram alir dari metode yang diusulkan

- *Term Variance* (TV)

Term variance menghitung varian dari semua fitur yang ada pada dataset. Suatu fitur yang muncul dalam sedikit dokumen atau memiliki distribusi umum di seluruh dokumen akan memiliki nilai $v\,(t_k)$ yang rendah. Dari persamaan 4 dapat dilihat bahwa nilai $v\,(t_k)$ sangat dipengaruhi oleh frekuensi kemunculan (*f)* dari fitur.

$$v\,(t_k) = \sum_{j=1}^{n} (f_{kj} - \bar{f}_k)^2 \quad (4)$$

- *Mean Absolute Difference* (MAD)

Metode ini memberikan nilai relevansi dari setiap fitur dengan menghitung perbedaan sampel dari nilai rata-rata. MAD didefinisikan sebagai menghitung selisih mutlak *tfidf* fitur dari nilai rata-rata *tfidf* fitur.

$$MAD\,(t_k) = \frac{1}{n} \sum_{j=1}^{n} |tfidf_{kj} - \overline{tfidf_k}| \quad (5)$$

Algoritma: ALOFT
Input :
D = { d1, d2, d3 ,d4.... dn }
Output :
himpunan fitur
Metode :

1. for *h* =1 sampai *V* hitung nilai filter untuk setiap term
2. S_h = filter (*Wh*)
3. End for
4. *m*=0
5. set *VF* sebagai vektor kosong
6. for semua *Di* ∈ *D* pilih nilai filter tertinggi untuk setiap dokumen
7. nilai terbaik = 0.0
8. for *h*=1 sampai *V*
9. if $W_{h,i} > 0$ & $S_h >$ nilai terbaik maka
10. nilai terbaik = S_h
11. fitur terbaik = *h*
12. end if
13. end for
14. if fitur terbaik bukan anggota *VF* maka
15. *m*= *m*+1
16. VF_m = fitur terbaik
17. End if
18. End for

Gambar 2. Pseudocode Algoritma ALOFT

- *Mean Median* (MM)

Nilai *mean median* (MM) yang dihasilkan memberikan nilai relevansi untuk setiap fitur berdasarkan perbedaan mutlak antara mean dan median dari *tfidf.*

$$MM(t_k) = |\overline{tfidf_k} - median\ (tfidf_k)| \quad (6)$$

- *Arithmetic Mean Geometric Mean* (AMGM)

AMGM yang diusulkan oleh [13] merupakan filter untuk mengatasi masalah pada *Arithmetic Mean* dan *Geometric Mean* dengan menerapkan fungsi eksponensial untuk setiap fitur berdasarkan nilai *tfidf*-nya.

$$AMGM(t_k) = \frac{\frac{1}{n}\sum_{j=1}^{n} \exp(tfidf_{kj})}{(\prod_{j=1}^{n} \exp(tfidf_{kj}))^{\frac{1}{n}}} \quad (7)$$

2) Pemilihan fitur dengan algoritma ALOFT

Himpunan fitur yang akan digunakan untuk pengelompokan dokumen dipilih dengan menggunakan metode ALOFT [15]. Dengan metode ini dipastikan bahwa setiap dokumen akan berkontribusi untuk pemilihan akhir himpunan fitur. Setidaknya terdapat satu fitur yang mewakili dokumen. Keuntungan lain yang didapat dari algoritma ALOFT adalah dengan menggunakan ALOFT tidak diperlukan masukan pengguna untuk pemilihan fiturnya. Gambar 2 merupakan prosedur dari algoritma ALOFT. Dari prosedur dapat dilihat bahwa proses pada baris 1 - 3 adalah menghitung nilai relevansi dari fitur menggunakan metode filter kemudian disimpan dalam S_h dimana *V* merupakan jumlah fitur dari keseluruhan dokumen. Selanjutnya baris 4-18 adalah proses membuat vektor fitur (*VF*) yang baru. Fitur ke *h* dimasukkan di *VF* jika nilai S_h tertinggi di antara semua fitur. Namun, jika fitur ini sudah terdapat di *VF*, fitur tersebut akan diabaikan dan algoritma berjalan ke dokumen berikutnya. Pada akhir fase ini, *VF* harus menjadi vektor dengan nilai-nilai *m*, dan nilai-nilai ini merupakan indeks dari fitur yang dipilih.

TABEL I
HASIL UJI COBA METODE MENGGUNAKAN K-MEANS

jumlah *k*	*Average Silhouette Width* (ASW)						
	DF+ALOFT	TC+ALOFT	TV+ALOFT	TVQ+ALOFT	MAD+ALOFT	MM+ALOFT	AMGM+ALOFT
3	**0.487**	**0.534**	**0.534**	**0.544**	**0.553**	0.280	0.145
6	0.428	0.511	0.428	0.527	0.542	0.256	0.143
9	0.409	0.335	0.335	0.439	0.374	0.278	0.159
12	0.410	0.344	0.343	0.455	0.390	0.296	0.174
15	0.423	0.356	0.359	0.475	0.400	0.315	0.195
18	0.405	0.355	0.352	0.460	0.372	0.277	0.207
21	0.399	0.327	0.320	0.454	0.369	0.290	0.218
24	0.355	0.335	0.307	0.445	0.355	**0.361**	**0.223**
Jumlah fitur	19	16	15	16	15	168	119

III. UJI COBA DAN ANALISA

Sebanyak 1000 dokumen berita dari situs berita online Kompas digunakan untuk memastikan efektivitas dari metode yang diusulkan. Selanjutnya untuk mengevaluasi kualitas *cluster* yang dihasilkan dari metode yang diusulkan digunakan *Silhouette Coefficient* [19]. Selain itu s*ilhouette coefficient* juga mengindikasikan derajat kepemilikan setiap objek yang berada di dalam *cluster*. Dokumen D_j yang berada pada *cluster* memiliki rentang nilai *Silhouette* antara -1 sampai 1. Semakin dekat nilai s*ilhouette* ke 1 maka semakin tinggi derajat D_j di dalam *cluster*. Pada persamaan (8) dan (9) merupakan perhitungan nilai *Silhouette (s(i))*. Selanjutnya setiap *cluster* yang telah dihitung nilai *s(i)* akan dihitung nilai rata-rata dari *s(i)*. Perhitungan ini lebih dikenal dengan nama *Average Silhouette Width* (ASW).

$$b(i) = max_{cj \neq a}\, d(i, c_j) \tag{8}$$

$$s(i) = \frac{b(i) - a\,(i)}{max\,\{a(i), b(i)\}} \tag{9}$$

Uji coba pada penelitian ini dilakukan pada PC dengan spesifikasi CPU i5-3230 M 2.60 GHz dan RAM 4GB. Proses uji coba dilakukan dengan mengestimasi parameter *k* yang merupakan jumlah *cluster* yang dihasilkan. Parameter *k* yang digunakan diestimasi dan diubah – ubah untuk mendapatkan nilai yang optimal sehingga memberikan hasil pengujian yang terbaik. Dalam penelitian ini nilai *k* yang digunakan dimulai dari *k* = 3-24. Pada Tabel I dapat dilihat hasil uji coba dari reduksi dimensi menggunakan variasi metode filter dan fitur akhir dipilih menggunakan algoritma ALOFT dan proses pengelompokan dokumen yang digunakan adalah algoritma *k-means.* Dapat dilihat pada Tabel I bahwa pada metode filter DF, TC, TV, TVQ memiliki *Average Silhouette Width* (ASW)

TABEL II
HASIL UJI COBA METODE MENGGUNAKAN HAC

jumlah *k*	*Average Silhouette Width* (ASW)						
	DF+ALOFT	TC+ALOFT	TV+ALOFT	TVQ+ALOFT	MAD+ALOFT	MM+ALOFT	AMGM+ALOFT
3	0.004	-0.065	0.220	-0.042	0.384	-0.016	-0.014
6	-0.069	0.042	0.215	0.136	**0.423**	-0.030	-0.023
9	0.106	**0.260**	0.184	0.323	0.387	-0.029	-0.033
12	0.285	0.252	0.303	0.348	0.270	-0.021	-0.037
15	0.346	0.243	0.303	0.319	0.273	-0.016	-0.025
18	**0.358**	0.237	**0.305**	0.350	0.272	**0.185**	**-0.005**
21	0.357	0.218	0.276	0.341	0.282	0.174	-0.009
24	0.331	0.219	0.272	**0.352**	0.273	0.179	-0.012
Jumlah fitur	19	16	15	16	15	168	119

TABEL III
JUMLAH FITUR MENGGUNAKAN METODE VR

Metode	Jumlah Fitur	Jumlah Dokumen	Jumlah Fitur	Jumlah Dokumen
Document Frequency (DF) + VR	19	954	50	1000
Term Contribution (TC) + VR	16	879	30	1000
Term Variance (TV) + VR	15	850	30	1000
Term Variance Quality (TVQ) + VR	16	931	30	1000
Mean Absolute Difference (MAD) + VR	15	997	30	1000
Mean Median (MM) + VR	168	856	450	1000
Arithmetic Mean Geometric Mean (AMGM) + VR	119	832	1150	1000

tertinggi pada $k = 3$. Dimana hasil ini telah sesuai dengan *ground truth* bahwa terdapat tiga buah kategori dari data seperti yang telah dijelaskan pada bagian sebelumnya. *Average Silhouette Width* (ASW) tertinggi dimiliki oleh metode filter MAD dengan fitur akhir sejumlah 15 fitur. Pada metode filter MM dan AMGM nilai fitur tertinggi didapat pada $k = 24$, jika nilai k diperbanyak maka MM dan AMGM akan memiliki ASW yang lebih tinggi. Hal ini terjadi karena jumlah fitur yang dihasilkan oleh MM dan AMGM terlalu banyak.

Pada Tabel II terlihat bahwa hasil dari algortima *Hierarchical Agglomerative Clustering* (HAC) memiliki nilai terbaik pada k yang berbeda-beda. Sedangkan jika dilihat data dari *ground truth* yang ada bahwa nilai k yang optimal adalah pada $k = 3$. Hal ini karena proses pengelompokan menggunakan algoritma HAC menghasilkan kelompok dokumen yang berkumpul menjadi satu dan hanya beberapa dokumen yang terdapat di *cluster* yang berbeda. Sehingga didapatkan nilai *Average Silhouette Width* (ASW) yang rendah pada $k = 3$.

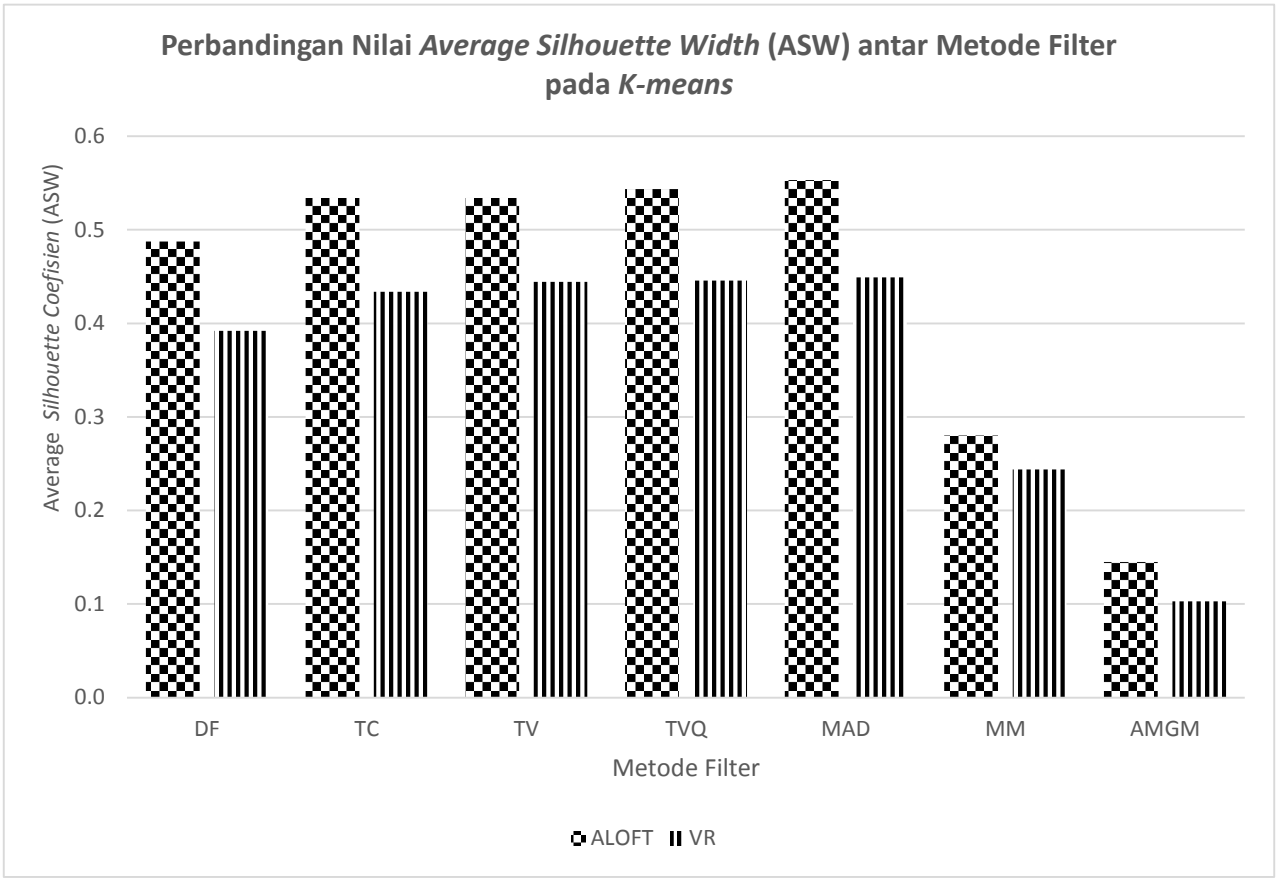
Pemilihan fitur menggunakan algoritma ALOFT memiliki kelebihan dimana dengan menggunakan algoritma ALOFT pengguna tidak perlu memasukkan jumlah fitur yang ingin diproses. Berbeda dengan pemilihan fitur dengan metode *Variable Ranking* (VR) yang memilih fitur dengan cara memasukkan sejumlah n fitur teratas. Nilai n ini menjadi sangat penting karena jumlah fitur yang berbeda mungkin akan menghasilkan kelompok dokumen yang berbeda. Karena harus melakukan estimasi jumlah n yang optimal. metode VR menjadi sangat tidak efisien karena waktu yang akan dibutuhkan untuk memilih fitur yang tepat pasti akan lebih lama. Selain itu sejumlah fitur dari n fitur teratas belum tentu mencakup pada semua dokumen. Pada penelitian ini dilakukan perbandingan dengan metode VR. Komparasi dilakukan pada $k = 3$ dimana pada jumlah cluster ini adalah jumlah yang sesuai dengan *ground trurth* bahwa data yang digunakan terdiri dari tiga kategori.

Jumlah fitur yang dihasilkan dari pemilihan fitur menggunakan metode *Variable Ranking (VR)* dapat dilihat pada Tabel III. Terlihat pada Tabel III bahwa jumlah fitur untuk metode VR lebih banyak jika dibandingkan dengan metode ALOFT hal ini karena pada jumlah fitur yang sama dengan metode ALOFT, metode VR belum bisa mencapai keseluran dokumen. Masih terdapat beberapa dokumen yang tidak terwakili oleh fitur sehingga mengakibatkan kesulitan menentukan kemiripan dari dokumen yang tidak terwakili fitur tersebut. Oleh karena itu, jumlah fitur yang dipilih dengan metode VR adalah jumlah minimal untuk dapat mencakup keseluruhan dokumen.

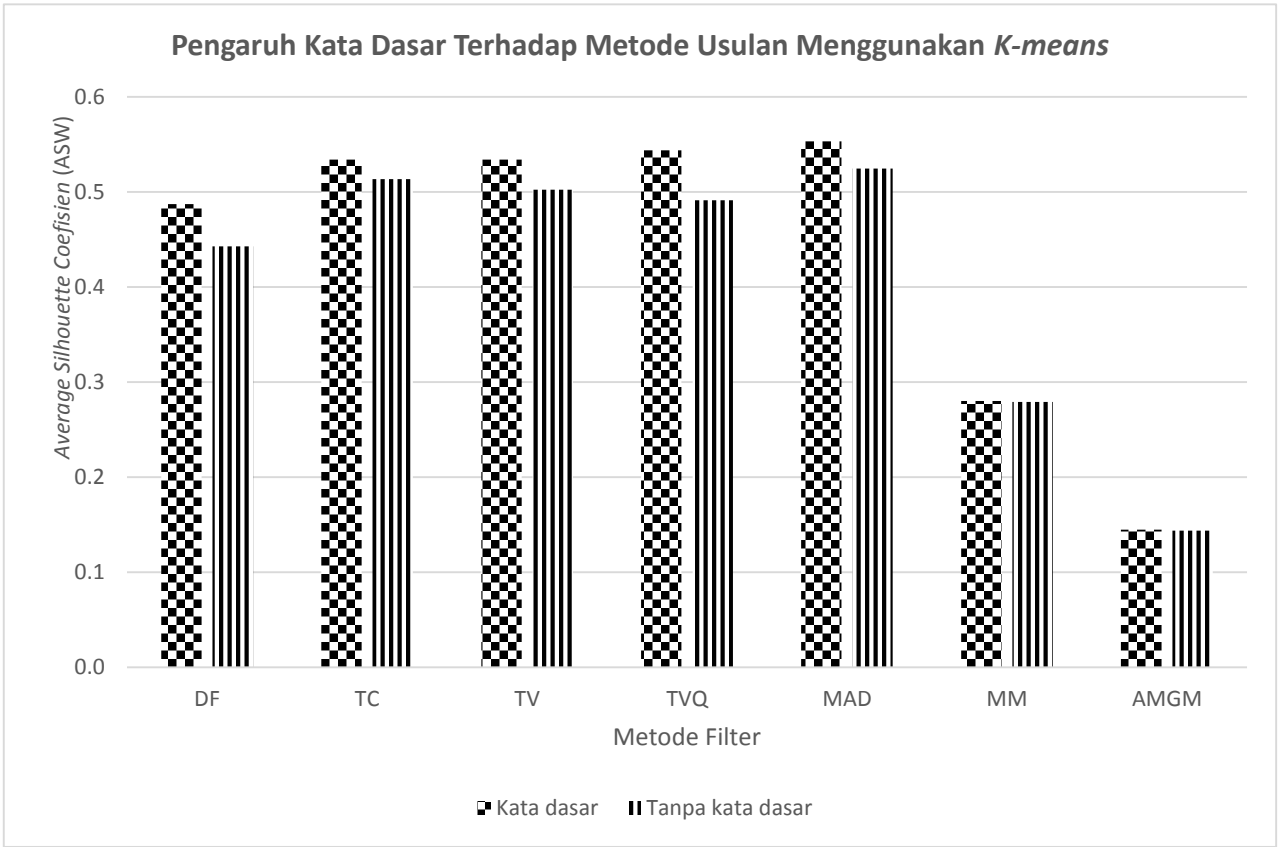
Gambar 3 merupakan grafik perbandingan kualitas *cluster* dengan menggunakan algoritma *k-means*. Dimana pada uji coba menunjukkan bahwa pemilihan fitur menggunakan variasi metode filter pada ALOFT dengan menggunakan algoritma *k-means* memiliki hasil yang lebih baik jika dibandingkan dengan algoritma HAC. Dari gambar dapat terlihat bahwa metode filter MM dan metode filter AMGM memiliki nilai *silhouette* yang lebih rendah diantara keseluruhan metode filter yang lain. Hal ini karena metode filter AMGM dan metode filter MM menghasilkan fitur yang lebih banyak dan terdapat fitur-fitur umum yang memungkinkan muncul di keseluruhan dokumen seperti fitur “sore”, “siang”, “isyarat”, “tiga”, “sisa”, “batas”, dan lain-lain.

Pada penelitian ini digunakan produk Kateglo untuk mencari kata dasar dengan cara memanfaatkan *root phrase* dan kata turunan yang telah disediakan oleh Kateglo. Jumlah fitur yang terbentuk ketika dilakukan pencarian kata dasar dengan menggunakan produk Kateglo adalah sejumlah 12.045 fitur. Sedangkan ketika proses pencarian kata dasar tidak dilakukan maka dihasilkan fitur sebanyak 13.859 dimana terdapat selisih 1.814 fitur yang merupakan kata turunan.

Gambar 4 merupakan grafik perbandingan rata – rata nilai *silhoueete coefficient* antara proses yang dilakukan pencarian kata dasar dengan yang tanpa kata dasar. Perbandingan ini dilakukan untuk mengetahui pengaruh pencarian kata dasar terhadap kualitas *cluster* yang dihasilkan. Pada Gambar 4 terlihat bahwa penggunaan kata dasar dengan menggunakan Kateglo tidak berpengaruh secara signifikan terhadap kualitas *cluster* yang dihasilkan.



Gambar 3. Grafik Perbandingan Metode Usulan dengan VR pada *K-means*



Gambar 4. Grafik Pengaruh Kata Dasar Terhadap Metode Usulan Menggunakan *Kmeans*

Hal ini terjadi karena sejumlah 1.814 fitur yang merupakan kata turunan tidak termasuk ke dalam fitur – fitur yang memiliki nilai relevansi tinggi dalam filter DF, TV, TVQ, MAD, MM, dan AMGM sehingga tidak terpilih oleh algoritma ALOFT sebagai himpunan fitur akhir. Selain itu kata turunan yang ada pada Kateglo tidak mencakup keseluruhan kata berimbuhan yang ada pada bahasa Indonesia. Beberapa kata berimbuhan yang tidak terdapat pada Kateglo seperti imbuhan dengan awalan di(-) dan akhiran (–i) misalkan kata "dipakai", "ditutupi", "dijalani", "dipadati", "dipanasi" dan lain-lain. Selanjutnya kata – kata yang memiliki imbuhan lebih dari satu dan juga akhiran lebih dari satu juga tidak terdapat di dalam daftar kata turunan Kateglo seperti kata "mempermainkannya", "memperjuangkannya", dan lain lain. Kata turunan kateglo juga tidak memiliki kata – kata berisisipan seperti "jelajah", "geligi", "selidik", "melaju", dan lain-lain. Kata yang memiliki akhiran (-i) juga tidak terdapat pada kata turunan Kateglo seperti kata "sukai", "tanami", "fasilitasi", dan lain lain.

IV. KESIMPULAN

Pada penelitian ini telah dilakukan reduksi dimensi fitur dengan menggunakan variasi metode filter pada algoritma ALOFT untuk pengelompokan dokumen. Reduksi dimensi dapat mengurangi tingginya dimensi dari data dengan cara menghapus fitur – fitur yang tidak relevan. Pada penelitian ini nilai relevansi fitur dihitung dengan beberapa metode filter yang kemudian dilakukan pemilihan himpunan fitur akhir dengan menggunakan Algoritma ALOFT. Metode filter yang digunakan meliputi *Document Frequency* (DF), *Term Contribution* (TC), *Term Variance Quality* (TVQ), *Term Variance* (TV), *Mean Absolute Difference* (MAD), *Mean Median* (MM), dan *Arithmetic Mean Geometric* Mean (AMGM). Dari keseluruhan metode filter yang digunakan MAD memiliki nilai *Average Silhouette Width* (ASW) yang paling tinggi. Namun metode filter MM dan AMGM pada Algoritma ALOFT dapat menyebabkan berkurangnya kualitas dari *cluster* akibat terlalu banyaknya fitur yang terpilih. Selanjutnya untuk penggunaan produk Kateglo pada proses pembentukan kata dasar dapat meningkat kualitas *cluster* akan tetapi peningkatan yang terjadi tidak terlalu signifikan. Penelitian selanjutnya yang dapat dikembangkan dari metode yang diusulkan adalah dengan menggunakan kombinasi dari beberapa metode filter sehingga nilai relevansi dari sebuah fitur tidak hanya tergantung pada satu metode filter saja.

DAFTAR PUSTAKA

[1] A. Leuski, "Evaluating document clustering for interactive information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management*, ACM, 2001.

[2] D. G. Roussinov and H. Chen, "Document clustering for electronic meetings: an experimental comparison of two techniques," *Decision Support Systems*, vol. 27, no. 1, pp. 67-79, 1999.

[3] F. Rozi, S. H. Wijoyo, S. A. Isanta, Y. Azhar and P. Diana, "Pelabelan Klaster Fitur Secara Otomatis pada Perbandingan Review Produk," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 1, no. 2, pp. 55-61, 2014.

[4] N. Hayatin, C. Fatichah and D. Purwitasari, "Pembobotan Kalimat Berdasarkan Fitur Berita dan Trending Issue untuk Peringkasan Multi Dokumen Berita," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 13, no. 1, pp. 38-44, 2015.

[5] I. Lukmana, D. Swanjaya, A. Kurniawardhani, A. Z. Arifin and D. Purwitasari, "Multi-Document Summarization Based On Sentence Clustering Improved Using Topic Words," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 12, no. 2, pp. 1-8, 2014.

[6] A. Z. Arifin, I. P. A. K. Mahendra and H. T. Ciptaningtyas, "Enhanced confix stripping stemmer and ants algorithm for classifying news document in indonesian language," 2009.

[7] G. Salton, A. Wong and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, p. 613–620, 1975.

[8] S. Tabakhi, P. Moradi and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Engineering Applications of Artificial Intelligence*, vol. 32, pp. 112-123, 2014.

[9] W. Song and S. C. Park, "Genetic algorithm for text clustering based on latent semantic indexing," *Computers & Mathematics with Applications*, vol. 57, no. 11, pp. 1901-1907, 2009.

[10] K. K. Bharti and P. K. Singh, "A three-stage unsupervised dimension reduction method for text clustering," *Journal of Computational Science*, vol. 5, no. 2, pp. 156-169, 2014.

[11] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Systems with Applications*, vol. 42, p. 3105–3114, 2015.

[12] L. Liu, J. Kang, J. Yu and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," 2005.

[13] A. J. Ferreira and M. A. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1794-1804, 2012.

[14] H. Liu, H. Motoda and eds, Computational methods of feature selection, CRC Press, 2007.

[15] R. H. Pinheiro, G. D. Cavalcanti, R. F. Correa and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Systems with Applications*, vol. 39, no. 17, pp. 12851-12857, 2012.

[16] C. C. Aggarwal and C. Zhai, Mining text data, Springer Science & Business Media, 2012.

[17] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," Institute for Logic, Language and Computation Universeit Van Amsterdam, 2003.

[18] I. Dhillon, J. Kogan and C. Nicholas, "Feature selection and document clustering," New York, 2004.

[19] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53-65, 1987.