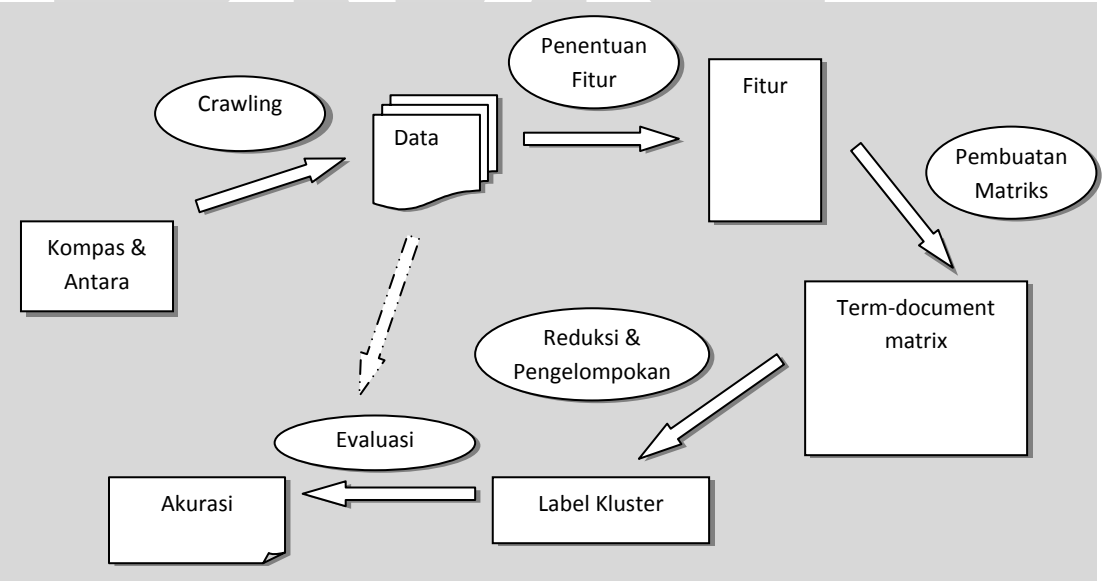
BAB III PERANCANGAN

Pada bab ini akan dijelaskan tahapan yang dilalui dalam melakukan perancangan penelitian yang akan dilakukan dalam tugas akhir ini. Tahapan tersebut meliputi perancangan implementasi *tools* yang akan digunakan dalam percobaan serta perancangan percobaan yang meliputi persiapan data dan penetapan variabel percobaan yang akan digunakan.

3.1. Alur Pengelompokan Dokumen

Rancangan alur pengerjaan penelitian dibuat terlebih dahulu untuk memberikan gambaran secara umum proses-proses yang akan dilakukan dalam penelitian. Gambaran umum alur pengerjaan penelitian ditunjukkan Gambar 3.1. Proses utama pada penelitian ini meliputi pengumpulan data (subbab 3.2), pengolahan data sebelum proses pengelompokan (subbab 3.3 sampai dengan subbab 3.5), dan penerapan teknik pengelompokan data (subbab 3.6) serta evaluasi kinerja teknik pengelompokan data (subbab 3.7).



Gambar 3.1. Alur Pengelompokan Dokumen

Data pada penelitian berupa dokumen dalam bahasa Indonesia yang nantinya akan dikelompokkan ke dalam kluster. Dokumen yang digunakan adalah artikel media massa. Artikel ini diambil dari *website* Kompas dan Antara dimana dokumen-dokumen tersebut sudah dikelompokkan ke dalam kategorinya masing-masing.

Proses pengolahan data dilakukan untuk mempersiapkan data yang berupa dokumen sebelum penerapan teknik pengelompokan. Input utama untuk teknik pengelompokan berupa *term-document matrix* yang berisi informasi mengenai kemunculan kata (*term*) dalam dokumen. Oleh karena itu, perlu adanya suatu tahap pembuatan matriks ini.

Untuk mendapatkan *term-document matrix* diperlukan beberapa tahap pendahuluan seperti mengidentifikasi kata yang muncul dalam koleksi dokumen. Karena banyak kata-kata umum yang frekuensi kemunculannya tinggi tetapi muncul hampir dalam setiap kategori dokumen atau yang biasa disebut dengan *stopword*, maka perlu dilakukan penghapusan kata ini dari koleksi kata yang sudah diidentifikasi. Dengan demikian, kata-kata yang digunakan sebagai koleksi kata untuk membentuk *term-document matrix* hanya kata-kata khusus yang dapat merepresentasikan kategori dokumen yang ada. Setelah beberapa tahap pendahuluan ini selesai dilakukan, proses pembuatan *term-document matrix* baru dilakukan dengan informasi koleksi kata yang sudah dibuat sebelumnya. Pembuatan *term-document matrix* dilakukan untuk setiap variasi percobaan yang dilakukan.

Setiap teknik pengelompokan dokumen menggunakan input berupa *term-document matrix*. Setelah pengelompokan, dokumen akan berada pada kluster-kluster yang paling cocok merepresentasikan dokumen tersebut. Kluster dibuat berdasarkan pola yang ada pada koleksi dokumen dan sesuai dengan jumlah kluster yang diinginkan.

Untuk mengetahui kinerja dari teknik pengelompokan yang digunakan, maka dilakukan evaluasi berupa nilai akurasi pengelompokan tersebut dengan membandingkan hasil pengelompokan dengan informasi kelompok atau kluster yang ada. Nilai akurasi adalah total dokumen yang dikelompokkan dengan benar dibagi dengan total dokumen.

3.2. Data

Data yang digunakan dalam percobaan untuk pengelompokan dokumen adalah artikel media massa yang diambil dari *website* Kompas dan Antara. Pemilihan penggunaan artikel dari Kompas dan Antara karena kemudahan akses dan pemakaian

bahasa yang cukup baku dalam setiap artikelnya. Dengan demikian, diharapkan percobaan dengan menggunakan data ini dapat memberikan hasil yang lebih akurat. Selain itu, artikel-artikel ini sudah dikelompokkan ke dalam kategori oleh pihak *website*. Hal ini dapat mendukung kemudahan dalam proses evaluasi.

Setiap artikel yang digunakan termasuk dalam satu kategori atau kluster yang sudah ditentukan sebelumnya. Informasi tentang kluster ini disimpan untuk keperluan evaluasi setelah penerapan teknik pengelompokan dokumen. Artikel tanpa informasi kluster akan dikumpulkan dalam satu koleksi dokumen yang digunakan sebagai input bagi teknik pengelompokan dokumen. Artikel Kompas yang diambil berjumlah 1971 dan terdiri dari 8 kategori. Sedangkan artikel Antara berjumlah 302 dan terdiri dari 1 kategori.

3.3. Persiapan Data

Data yang digunakan untuk percobaan adalah artikel media massa yang diambil dari *website* Kompas yaitu www.kompas.com dan *website* Antara yaitu www.antara.co.id. Untuk keperluan evaluasi kinerja teknik pengelompokan dokumen, diperlukan informasi mengenai kategori atau kluster dari setiap artikel yang digunakan dalam setiap percobaan. Data artikel yang didapat dari *website* ini sudah terkelompok dalam klusternya masing-masing yang sesuai. Untuk setiap artikel yang telah diambil akan simpan dalam direktori yang sesuai dengan kategori artikel tersebut. Setiap artikel disimpan dalam sebuah berkas teks.

Artikel Kompas yang digunakan diambil dari delapan kategori atau kluster yang ada pada Kompas yaitu bisnis keuangan, kesehatan, olahraga, perempuan, sains, travel, properti, dan politik hukum. Sedangkan, artikel Antara yang digunakan hanya artikel dari kategori olahraga. Informasi lengkap mengenai artikel dijelaskan pada subbab 4.1. Karena artikel diambil secara otomatis oleh sebuah program, maka diperlukan proses pengecekan tersendiri apakah artikel yang didapat layak untuk digunakan sebagai input bagi proses selanjutnya. Misalnya dengan mengecek apakah isi berkas teks kosong atau tidak (penjelasan pada subbab 4.1).

3.4. Penentuan Fitur

Pada (Berry & Shahnaz, 2004) dijelaskan bahwa pada *vector space model* untuk data teks, dokumen direpresentasikan sebagai vektor dalam dimensi m dimana m adalah jumlah *term* yang digunakan. Setiap komponen dalam vektor merefleksikan tingkat dimana *term* yang bersangkutan memberikan arti (*semantic*) pada dokumen tersebut. Maka koleksi dokumen dapat dinyatakan dengan *term-document matrix* yang merupakan kumpulan dari masing-masing vektor dokumen. Dari sini dapat ditarik kesimpulan bahwa kata atau *term* dapat merepresentasikan arti atau *semantic* dari suatu dokumen yang dapat berguna untuk membedakan dokumen tersebut dengan dokumen lain.

Pada percobaan ini digunakan fitur berupa *term* atas dasar pemikiran bahwa *term* dapat merepresentasikan *semantic* dari dokumen sehingga pembedaan antara tiap dokumen dapat dilakukan yang akan berujung pada pengelompokan dokumen ke dalam kluster. Dokumen yang memiliki kemiripan akan dikelompokkan ke dalam kluster yang sama.

Hal diatas dapat dijelaskan dengan fenomena bahwa manusia dapat mengelompokkan sekumpulan dokumen dengan membaca dan mengidentifikasi kata-kata khusus yang muncul di dalamnya. Dari kata-kata khusus ini manusia kemudian dapat menyimpulkan dokumen yang bersangkutan adalah dokumen dengan topik apa. Atas dasar pemikiran inilah, maka teknik pengelompokan dokumen menggunakan kata-kata khusus tersebut untuk mengelompokkan dokumen.

Pemilihan fitur dilakukan dengan cara mengumpulkan dan menghitung frekuensi kemunculan setiap kata yang ada pada koleksi dokumen. Kata yang diambil hanyalah kata yang unik (*type*). Jadi, pada saat melakukan *scanning* pada setiap dokumen dalam koleksi, suatu kata yang ditemukan tidak akan dimasukkan ke dalam koleksi *term* jika kata tersebut sudah terlebih dahulu ada di dalam koleksi *term* atau kata. Untuk menghindari pemakaian kata-kata yang tidak penting dalam percobaan, perlu satu parameter tambahan yaitu minimum frekuensi. Dengan parameter ini, kata-kata yang akan dipakai dalam percobaan bisa dibatasi. Jadi, kata tidak akan dipakai jika frekuensi kemunculannya lebih kecil dari parameter minimum frekuensi.

Pada percobaan ini juga akan diuji kinerja dari teknik pengelompokan dokumen jika fitur yang digunakan merupakan kata-kata khusus yang memang dapat merepresentasikan suatu kluster atau topik. Jadi kata-kata umum yang sering muncul dalam dokumen tetapi tidak dapat digunakan untuk membedakan satu dokumen dengan dokumen lain akan dihapus dari koleksi *term*. Kata-kata tersebut seperti kata-kata penghubung dan kata-kata keterangan. Koleksi kata-kata ini disebut dengan *stopwords*.

Fitur yang digunakan pada percobaan ini adalah fitur 1-*gram* atau yang lebih sering disebut dengan *unigram*. Dengan fitur *unigram* maka yang digunakan sebagai fitur adalah 1 kata per fiturnya. Pemakaian fitur ini sesuai dengan pemikiran bahwa setiap kata yang ada pada dokumen berkontribusi dalam merepresentasikan arti dari dokumen tersebut dan dapat digunakan untuk pengelompokan dokumen yang bersangkutan. Selain itu, sesuai dengan percobaan yang telah dilakukan Berry & Shahnaz (2004) pada domain bahasa inggris, fitur ini memberikan hasil akurasi yang cukup baik.

3.5. *Term-Document Matrix*

Dalam teknik pengelompokan dokumen, input yang digunakan adalah *term-document matrix*. Matriks ini memberi informasi kemunculan fitur yang digunakan atau *term* dalam koleksi dokumen yang ada. Setiap baris merepresentasikan *term* yang ada dalam koleksi. Sedangkan setiap kolom merepresentasikan dokumen yang digunakan. Gambaran mengenai *term-document matrix* ditunjukkan oleh Gambar 3.2.

	d_1	d_2	.	.	.	d_n
t_1	f_{11}	f_{12}				f_{1n}
t_2	f_{21}	f_{22}				f_{2n}
.	.	.				.
.	.	.				.
.	.	.				.
t_m	f_{m1}	f_{m2}	.	.	.	f_{mn}

Gambar 3.2. *Term-document Matrix*

Setiap nilai f_{ij} menunjukkan bobot *term* ke *i* pada dokumen *j*. Masing-masing dokumen dapat direpresentasikan sebagai vektor bobot *term-term* yang ada pada dokumen tersebut. Vektor dokumen ke *i* dapat direpresentasikan sebagai

$$d_i = [f_{1i}, f_{2i}, f_{3i}, ..., f_{4i}], i = 1,2,...,m; m = jumlahterm$$

Ada beberapa variasi dari informasi yang direpresentasikan f_{ij} (bobot) yang digunakan pada percobaan ini. Variasi yang pertama adalah *presence*. Dengan informasi *presence*, informasi yang direpresentasikan oleh f_{ij} berupa ada atau tidak *term* ke *i* pada dokumen *j* yang dinyatakan dengan bilangan biner. Nilai 1 digunakan jika *term* ke *i* muncul pada dokumen ke *j*, sedangkan nilai 0 digunakan jika *term* ke *i* tidak muncul pada dokumen ke *j*.

Variasi yang kedua adalah *frequency*. Dengan informasi *frequency*, informasi yang direpresentasikan oleh f_{ij} adalah jumlah kemunculan *term* ke *i* pada dokumen ke *j*. Variasi yang ketiga adalah *frequency normalized term frequency* (TF). Dengan informasi ini, nilai yang direpresentasikan oleh f_{ij} adalah jumlah kemunculan *term i* pada dokumen *j* dibagi dengan jumlah semua fitur yang ada pada dokumen ke *j*. Dengan demikian, panjang dokumen yang berbeda-beda tidak akan membawa efek yang dapat mempengaruhi hasil. Secara matematis, nilai dari f_{ij} dapat ditulis sebagai

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (21)$$

Dimana $n_{i,j}$ adalah jumlah kemunculan *term i* pada dokumen *j* dan pembaginya adalah total kemunculan semua fitur pada dokumen *j*.

Variasi keempat adalah *frequency normalized term frequency-inverse document frequency* (TF-IDF) (Ramos). Informasi ini mengukur seberapa penting (*important*) suatu kata terhadap suatu dokumen pada korpus atau koleksi. Tingkat *importance* akan meningkat secara proporsional terhadap jumlah kemunculan kata pada dokumen tetapi diimbangi dengan jumlah kemunculan kata tersebut pada keseluruhan dokumen. Dengan demikian, kata-kata umum seperti kata penghubung dan lain lain yang sering muncul pada dokumen tidak akan memberikan informasi

yang signifikan dan sebaliknya kata-kata khusus yang memang bisa menggambarkan informasi yang terkandung dalam dokumen akan memberikan informasi yang lebih signifikan. Secara matematis TF-IDF ditulis sebagai

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$|D|$: total dokumen dalam koleksi

$|\{d : t_i \in d\}|$: jumlah dokumen dimana *term i* muncul

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (22)$$

Sebelum *term-document matrix* dibentuk, perlu dilakukan pengumpulan fitur yang digunakan, dalam hal ini fitur berupa *term.* Kumpulan *term* yang telah diidentifikasi akan digunakan sebagai input bagi proses pembentukan *term-document matrix.*

Term-document matrix akan dibentuk untuk setiap variasi percobaan. Jadi satu variasi percobaan akan memiliki *term-document matrix* sebagai input yang berbeda dengan variasi percobaan yang lain. *Term-document matrix* dibentuk dengan menghitung jumlah kemunculan setiap *term* dalam koleksi *term* pada setiap dokumen. Matriks ini akan menjadi input langsung bagi proses pengelompokan dokumen.

3.6. Teknik Pengelompokan Dokumen

Pada tugas akhir ini digunakan dua teknik reduksi dimensi untuk aplikasi pengelompokan dokumen yaitu Non-Negative Matrix Factorization dan Random Projection. Untuk Non-Negative Matrix Factorization, metode yang digunakan sebagai *update rule* adalah metode GD-CLS. Untuk Random Projection sendiri, diperlukan proses pengelompokan tambahan. Teknik yang digunakan untuk pengelompokan adalah K-Means. Pada subbab ini akan dijelaskan perancangan dari penerapan dua teknik diatas yang digunakan untuk mengelompokkan dokumen. Matriks yang telah dihasilkan pada subbab 3.5 digunakan sebagai input untuk teknik pengelompokan dokumen.

3.6.1. Nonnegative Matrix Factorization

Input yang berupa *term-document matrix V* yang berukuran m x n, dimana m adalah jumlah *term* dan n adalah jumlah dokumen, akan difaktorisasi menjadi matriks *W* dan *H* dengan menggunakan *update rule* yang telah dijelaskan pada subbab 2.5. *Update rule* untuk masing-masing matriks *W* dan *H* adalah

Update rule untuk matriks *W*

$$W_{ij} \leftarrow W_{ij} \frac{V_{ij} H_{ij}^T}{W_{ij} H_{ij} H_{ij}^T + \varepsilon}, \varepsilon = 10^{-9}$$

Update rule untuk matriks *H*

$$H_j = \frac{W_{ij}^T V_j}{W_{ij}^T W_{ij} + \lambda I_{ij}}, I = matriksIdentitas_{kxk}$$

Matriks *W* berukuran m x k dimana m adalah jumlah *term* dan k adalah jumlah kluster yang ingin dibentuk. Matriks *H* berukuran k x n dimana k adalah jumlah kluster yang ingin dibentuk dan n adalah jumlah dokumen. Sesuai penjelasan pada subbab 2.5, untuk masing-masing kolom matriks *H*, nilai maksimum menunjukkan kluster dimana dokumen yang bersangkutan akan dimasukkan. Label kluster yang didapat untuk setiap dokumen disimpan dalam vektor *L* 1 x n, dimana n adalah jumlah dokumen. Label ini akan dikonversi ke label kluster yang sebenarnya menggunakan informasi yang didapat dari penerapan formula *similarity* (penjelasan pada subbab 3.7).

Proses evaluasi atau penentuan nilai akurasi dilakukan dengan membandingkan setiap nilai pada vektor hasil konversi dengan informasi kluster untuk setiap dokumen yang telah disimpan dalam sebuah vektor *K* terlebih dahulu.

3.6.2. Random Projection

Sesuai dengan penjelasan pada subbab 2.6, perlu dibuat matriks acak berukuran k x m yang nantinya akan dikalikan dengan *term-document matrix* yang berukuran m x n sebagai matriks input. Matriks acak dibentuk dengan aturan sebagai berikut:

$$\text{Tipe 1:} \quad R_{ij} = \sqrt{3}x \begin{cases} +1, prob = 1/6 \\ 0, prob = 2/3 \\ -1, prob = 1/6 \end{cases}$$

atau

$$\text{Tipe 2:} \quad R_{i,j} = \begin{cases} +1, prob = 1/2 \\ -1, prob = 1/2 \end{cases}$$

Setelah matriks acak dibentuk, matriks ini dikalikan dengan *term-document matrix* untuk menghasilkan matrik yang berdimensi lebih kecil k x n. Matriks hasil perkalian ini digunakan sebagai model untuk melakukan pengelompokan. K-Means diterapkan pada matriks hasil perkalian matriks acak dan matriks input. Hasil penerapan K-Means disimpan dalam vektor *C* 1 x n. Label kluster hasil penerapan K-Means perlu untuk dikonversi menjadi label kluster yang sebenarnya sehingga proses evaluasi bisa dilakukan. Proses konversi dilakukan menggunakan informasi yang didapat dari penerapan formula *similarity* seperti yang dijelaskan pada subbab 3.7.

Proses evaluasi atau penentuan nilai akurasi dilakukan dengan membandingkan setiap nilai pada vektor hasil konversi dengan informasi kluster untuk setiap dokumen yang telah disimpan dalam sebuah vektor *K* terlebih dahulu.

3.7. Pemetaaan Kluster yang Dibangun ke Kluster yang Sebenarnya

Untuk setiap nilai *i* yang berbeda pada vektor label *L* hasil pengelompokan, akan ditentukan nilai *similarity*-nya dengan setiap kluster *k* yang ada dengan formula *similarity.*

Formula *similarity* :

similarity(Cluster$_i$,Cluster$_k$) = jumlah dokumen di Cluster i yang muncul di Cluster k

Contoh dari penerapan formula *similarity* diatas adalah sebagai berikut:

Misalkan label kluster yang sebenarnya adalah $T = \{A, B\}$, ada 5 dokumen yang akan dikelompokkan $D = \{d_1, d_2, d_3, d_4, d_5\}$, dan pengelompokan yang benar

adalah $Cluster_A = \{d_2, d_3\}$ dan $Cluster_B = \{d_1, d_4, d_5\}$. Label kluster yang dihasilkan dari penerapan teknik NMF adalah $Cluster_1 = \{d_2, d_3, d_5\}$ dan $Cluster_2 = \{d_1, d_4\}$. Matriks *S* hasil penerapan formula *similarity* dimana S_{ix} = *similarity*($Cluster_i$, $Cluster_x$) = jumlah dokumen di $Cluster_i$ yang muncul di $Cluster_x$, $i = (1,2)$ dan $x = (A,B)$ adalah

	$Cluster_A$	$Cluster_B$
$Cluster_1$	2	1
$Cluster_2$	0	2

Cluster i diberi nama label kluster yang sebenarnya dimana mereka mempunyai nilai *similarity* yang terbesar diantara *k* nilai yang ada. Informasi ini disimpan dalam vektor 1 x k dimana k adalah jumlah kluster yang ingin dibentuk. Dalam contoh diatas vektor ini akan berisi {*A, B*} yang artinya semua label 1 akan dikonversi menjadi *A* dan semua label 2 akan dikonversi menjadi *B*. Informasi ini akan dijadikan dasar dalam penentuan aturan konversi label hasil penerapan teknik pengelompokan ke label kluster yang sebenarnya.

Dengan informasi aturan konversi, maka konversi dilakukan untuk setiap nilai pada vektor label *L* dan menghasilkan vektor *C* 1 x n dimana n adalah jumlah dokumen yang digunakan.

3.8. Evaluasi Kinerja

Kinerja dari masing-masing teknik pengelompokan dokumen ditentukan dari tingkat atau nilai akurasi masing-masing teknik dalam mengelompokkan dokumen yang diberikan. Akurasi untuk evaluasi kinerja sudah umum digunakan. Untuk menentukan akurasi diperlukan informasi mengenai kategori atau kluster yang sebenarnya dari dokumen yang digunakan sebagai input. Oleh karena itu, perlu dilakukan pengelompokan manual terlebih dahulu. Namun, pada percobaan ini, dokumen yang digunakan, yang diambil dari Kompas dan Antara, telah dikelompokkan ke dalam klusternya masing-masing oleh pihak Kompas dan Antara. Jadi yang perlu dilakukan adalah menyimpan informasi kluster ini untuk nantinya

digunakan sebagai tolak ukur penentuan kinerja teknik pengelompokan dokumen yang diuji.

Hal yang dilakukan untuk mengevaluasi kinerja teknik pengelompokan dokumen adalah mencocokkan hasil penerapan teknik dengan informasi kluster sebenarnya. Pencocokan ini dilakukan dokumen per dokumen. Pencocokan ini dilakukan setelah hasil pengelompokan telah dikonversi ke dalam label yang digunakan. Akurasi pengelompokan dokumen dihitung menggunakan formula sebagai berikut (Berry & Shahnaz, 2004)

$$AC = \sum_{i=1}^{n} \delta(d_i)/n \tag{23}$$

$\delta(d_i)$ diberi nilai 1 jika dokumen d_i mempunyai label yang sama antara hasil konversi dengan label yang sebenarnya dan diberi nilai 0 jika tidak. n adalah jumlah dokumen yang dipakai.

3.9. Perancangan Eksperimen

Eksperimen bertujuan untuk melihat kinerja teknik reduksi dimensi Nonnegative Matrix Factorization dan Random Projection dalam aplikasi pengelompokan dokumen. Pada akhirnya, perbandingan kinerja dari masing-masing teknik akan memperlihatkan kualitas dari masing-masing teknik. Ada banyak parameter yang mempengaruhi masing-masing teknik. Untuk dapat membandingkan teknik tersebut, perlu dilihat terlebih dahulu pengaruh masing-masing parameter pada teknik sehingga perbandingan yang dilakukan adalah perbandingan yang valid.

Secara garis besar ada beberapa parameter yang akan diujicoba untuk melihat pengaruhnya pada akurasi pengelompokan dokumen. Parameter tersebut adalah parameter yang berhubungan dengan fitur, parameter khusus masing-masing teknik (Nonnegative Matrix Factorization dan Random Projection), serta parameter yang berhubungan dengan dokumen. Parameter yang berhubungan dengan fitur adalah jenis informasi fitur, jumlah fitur yang digunakan, dan penggunaan *stopwords*. Parameter khusus teknik Nonnegative Matrix Factorization adalah nilai *lambda* dan jumlah iterasi. Parameter khusus teknik Random Projection adalah jumlah

pengurangan dimensi dan tipe distribusi matriks acak. Paramater yang berhubungan dengan dokumen adalah jumlah kluster, ukuran kluster, dan keseragaman ukuran kluster.

Selain percobaan terhadap masing-masing parameter dan perbandingan kinerja masing-masing teknik, percobaan juga akan dilakukan pada dokumen dengan sifat kluster yang berbeda, serta dari sumber yang berbeda. Pembahasan lebih lanjut tentang percobaan yang dilakukan dijelaskan pada bab 5.

Karena untuk setiap teknik terdapat unsur acak (*random*), maka, untuk menjaga validitas hasil yang didapat, setiap percobaan akan dijalankan 3 kali. Hasil yang didapat kemudian akan dirata-ratakan untuk mendapatkan hasil seperti yang ditampilkan dalam laporan ini.