

1. Pendahuluan

1.1 Latar Belakang

Text clustering sebagaimana yang biasa diketahui disebut juga sebagai *document clustering* merupakan teknik untuk mengorganisasikan suatu dokumen untuk menghasilkan suatu informasi yang ada di dalamnya maupun sebagai penyaringan informasi [3,8]. Pengelompokan antar dokumen dibagi ke dalam *cluster-cluster* yang membedakan antara satu *cluster* dengan *cluster* lainnya. *Text clustering* ini juga disebut sebagai unsupervised learning, yaitu jumlah *cluster-cluster* yang terbentuk tergantung dari jumlah dokumen yang akan di *cluster*-kan [1,3,5,8].

K-medoids merupakan *Partitioning Clustering Method* berbasis *object-based* [3,6,11,12]. Metode k-medoids ini akan membagi data kedalam beberapa *cluster* dan akan berusaha untuk mencari pusat *cluster* tersebut [7,13]. Selain itu metode k-medoids ini lebih tahan terhadap *noise* dan *outliers* daripada k-means karena meminimalkan jumlah *dissimilarities* [2,6,11]. K-means sendiri sangat sensitif terhadap *outliers* dan set objek yang terdekat dengan pusat *centroid* bisa kosong, dalam kasus ini *centroid* tidak *ter-update* [6]. Karena menggunakan objek paling pusat dalam sebuah *cluster*, k-medoids akan lebih tahan terhadap *noise* dan *outliers* dibandingkan k-means [6,11]. Algoritma untuk K-Medoids yang biasa dipakai adalah *Partitioning Around Medoids* (PAM) [3]. *Partition Around Medoids* ini memiliki waktu eksekusi yang besar apabila data yang dimasukkan juga besar. Untuk menghemat waktu dalam komputasi perhitungan pada K-Medoids, maka muncullah suatu algoritma K-Means-Like yang dapat diterapkan pada metode k-medoids [6]. Algoritma ini menghasilkan performansi yang sama bila dibandingkan dengan algoritma k-medoids *clustering* biasa (PAM), akan tetapi waktu eksekusi untuk K-Means-Like ini lebih baik bila dibandingkan dengan PAM [6]. Karena alasan inilah, pada Tugas Akhir ini menggunakan algoritma K-Means-Like pada metode K-Medoids. Algoritma ini dinamakan dengan 'K-Means-Like' karena tahapan-tahapan serta kompleksitas waktu pada algoritma ini mirip dengan K-Means [6] tetapi apa yang dilakukan di tiap tahapan berbeda dan algoritma ini bukan menggabungkan antara K-Means dan K-Medoids [6].

Algoritma K-Means-Like ini akan diujikan pada data *text* yang memiliki dimensi yang besar untuk mendapatkan performansi cluster yang baik berdasarkan parameter-parameter yang mempengaruhi. Representasi *text* yang ada berasal dari kata-kata atau *term* yang terdapat dalam dokumen tersebut sehingga dokumen yang memiliki *term* dengan nilai similaritas yang tinggi antar *term* pada dokumen lain akan dikelompokkan kedalam 1 *cluster*.

1.2 Perumusan Masalah

Perumusan masalah yang di dapatkan dalam Tugas Akhir ini antara lain :

1. Bagaimana melakukan *text clustering* dengan menggunakan algoritma K-Means-Like pada metode K-Medoids?
2. Bagaimana pengaruh nilai k terhadap waktu eksekusi dan jumlah iterasi pada *text clustering* menggunakan algoritma K-Means-Like pada metode K-Medoids?
3. Bagaimana mengetahui kualitas *cluster* yang terbentuk dengan K-Means-Like pada metode K-Medoids menggunakan *F-Measure* untuk *text clustering*?

1.3 Batasan Masalah

Batasan masalah yang terdapat dalam Tugas Akhir ini antara lain :

1. Tugas Akhir ini hanya fokus ke analisa *text clustering* menggunakan algoritma K-Means-Like pada metode K-Medoids, bukan terhadap studi kasusnya
2. Tidak membandingkan hasil dari *clustering* menggunakan Algoritma K-Means-Like dengan algoritma lain.
3. Data dokumen yang digunakan telah mengalami proses *preprocessing* terlebih dahulu.

1.4 Tujuan

Tujuan dari pembuatan Tugas Akhir ini adalah :

Menerapkan algoritma K-Means-Like pada *text clustering* dengan melihat performansi cluster berdasarkan parameter k, waktu eksekusi, jumlah iterasi, dan F-Measure.

Hipotesa yang di perkirakan dalam Tugas Akhir ini adalah Algoritma K-Means-Like mampu melakukan *text clustering* dengan tingkat similaritas tinggi dan waktu eksekusi yang cepat.

1.5 Metodologi Penyelesaian Masalah

1. Studi Literatur
Studi literatur yang dilakukan dapat dilakukan dengan mencari sumber referensi dari buku dan internet, diskusi dengan mahasiswa lain dan dosen.

2. Pengumpulan data
Data yang digunakan berasal dari Tugas Akhir yang berjudul “Filter-Based Feature Selection pada Kategorisasi Artikel Berita Berbahasa Indonesia”.
3. Analisa dan Perancangan Kebutuhan
Data set yang ada telah dilakukan *preprocessing* terlebih dahulu, lalu merancang algoritma K-Means-Like pada metode K-Medoids.
4. Implementasi dan Pengujian
Sistem akan diuji, hasil yang diharapkan adalah dapat mengkluster data dengan baik serta dapat mengetahui kualitas *cluster* yang terbentuk. Berikut gambaran umum tentang simulasi implementasi dan pengujian
5. Kesimpulan dan Laporan
Tahap akhir akan dilakukan kesimpulan terhadap Tugas Akhir, lalu akan diiringi dengan penulisan laporan.