KLASTERISASI GENRE CERPEN KOMPAS MENGGUNAKAN *AGGLOMERATIVE HIERARCHICAL CLUSTERING- SINGLE LINKAGE*

Zenal Arifin[1] dan Stefanus Santosa[2], M. Arief Soeleman[3]
[123]Progam Magister Teknik Informatika Universitas Dian Nuswantoro

ABSTRAK

Teks merupakan sarana interaksi dalam semua media komunikasi tulisan. Oleh karena peningkatan ukuran dan jenisnya yang sangat cepat, maka proses analisis data teks menjadi sesuatu yang bermakna sangatlah penting. Penggalian teks telah menjadi teknologi yang penting terutama dalam pengolahan dokumen cerpen. Pembaca cerpen saat ini kesulitan untuk memperoleh cerpen yang diinginkan jika cerpen tersebut tidak terkelompok dengan baik. Jika pengelompokan dilakukan secara manual membutuhkan waktu yang sangat lama. Oleh sebab itu, clustering menjadi solusi untuk mengatasi masalah tersebut. Clustering cerpen berfungsi untuk mengelompokkan dokumen cerpen berdasarkan tingkat kemiripan dari dokumen cerpen tersebut. Penelitian ini mengusulkan suatu model klasterisasi berbasis metode Hierarchical Clustering, khususnya Single Linkage Clustering. Metode Hierarchical Aggomerative Clustering terbukti memiliki performansi yang lebih baik daripada pendekatan penelitian sebelumnya yang menggunakan k-Means. Dari 127 dataset cerpen yang telah diujicobakan didapatkan nilai akurasi dari metode Agglomerative Hierarchical Clustering Single Linkage 47,2441 %, sedangkan metode k-Means hanya 37,7953 %.

Kata kunci : Klasterisasi, Hierarchical Agglomerative Clustering, Single Linkage, k-Means

1. PENDAHULUAN

1.1. Latar Belakang Masalah

Salah satu cara untuk memperoleh informasi seimbang adalah dengan membaca secara serial dokumen cerpen yang membahas topik yang sama. Namun hal ini menyulitkan pembaca untuk menangkap topik bahasan utama dari dokumen-dokumen cerpen dengan topik yang sama tersebut karena harus mengingat–ingat isi dokumen cerpen yang telah dibaca sebelumnya. Pembaca harus mengintegrasikan dahulu dokumen–dokumen cerpen yang dia baca di dalam pikirannya sebelum dapat merangkum maksud dan topik utama dokumen–dokumen cerpen tersebut secara keseluruhan.

Sistem pengelompokan dokumen cerpen merupakan sistem penggabungan dokumen cerpen berdasarkan tingkat kemiripan dari dokumen cerpen tersebut. Sistem pengelompokan ini memberikan kemudahan dalam pencarian dokumen cerpen yang diinginkan. Dampaknya pencarian dokumen cerpen yang diinginkan akan semakin cepat dilakukan. Dokumen cerpen yang telah dikelompokkan akan tersusun dengan terstruktur dan rapi sesuai dengan kemiripan dokumen tersebut[1].

Penelitian tentang model klasterisasi cerpen sudah pernah dilakukan, khususnya cerpen berbahasa Indonesia. Hasil pengujian akurasi terhadap Model Klasterisasi Genre Cerpen KOMPAS menggunakan *k-Means* menunjukkan *Index Davies Bouldin* berada pada 0.001. Dari proses klasterisasi yang telah dilakukan diperoleh hasil terdapat perbedaan antara pengelompokan yang dilakukan secara manual dengan pengelompokan yang dilakukan oleh *K-Means Clustering.* Perbedaan terbesar ada pada hasil analisis genre Keluarga, yang diikuti semakin kecil pada genre Sejarah, Percintaan, dan Religius [2]

Penerapan k-Mean pada klasterisasi dokumen yang lain juga memiliki akurasi yang rendah. Yang

mirip dengan itu adalah tentang Klasterisasi dokumen e–jurnal menggunakan algoritma k-Means juga. Model ini menghasilkan akurasi 50% pada uji coba pertama terhadap 5 data dokumen[3].

Berbagai usaha telah dilakukan untuk memperbaiki model *cluster* dan menghitung jumlah *cluster* yang optimal sehingga dapat dihasilkan *cluster* yang paling baik. Ada dua metode *cluster*ing yang dikenal, yaitu hierarchical *cluster*ing dan partitioning. k-Means *term*asuk dalam metode partitioning.

Metode k-Means merupakan metode *cluster*ing yang paling sederhana dan umum [4]. Hal ini dikarenakan k-Means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien [5]. Namun, k-Means mempunyai mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode k-Means ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan [4]. Hal ini menyebabkan hasil *cluster*-nya berupa solusi yang sifatnya local optimal.

Untuk itu penelitian ini mengusulkan suatu pendekatan berbasis metode hierarchical *cluster*ing, khususnya *Single Linkage cluster*ing. Menurut Handoyo *Single Linkage* memiliki kemampuan lebih baik dibanding k-Means dalam pengelompokan data [6]. Pendekatan ini dilakukan untuk memperbaiki Model klasterisasi *K-Means* dengan menggunakan metode hierarki yang memiliki kemampuan lebih baik dalam penentuan pusat awal *cluster*.

1.2. Rumusan Masalah

Beberapa masalah yang dijadikan bahan penelitian dirumuskan sebagai berikut.

a. Pembaca kesulitan dalam mengintegrasikan dokumen–dokumen cerpen yang dia baca, merangkum maksud dan topik utama dokumen–dokumen cerpen tersebut secara keseluruhan karena sulit mencari kembali cerpen yang diinginkan.
b. Model klasterisasi dokumen cerpen yang penting untuk mencari kembali cerpen yang diinginkan saat ini akurasinya masih rendah.

1.3. Tujuan Penelitian

a. Melakukan pengkategorian dokumen cerpen sehingga memudahkan pencarian kembali dokumen cerpen yang diinginkan.
b. Terciptanya Model Klasterisasi Dokumen Cerpen berbasis metode *Agglomerative Hierarchical Clustering- Single Linkage* (AHC-SL) dengan akurasi yang lebih tinggi melalui pencapaian nilai *Purity* yang baik.

.

1.4. Manfaat Penelitian

Manfaat yang didapat dari penelitian ini adalah :

a. Hasil dari pengelompokan dokumen cerpen dengan metode *Agglomerative Hierarchical Clustering-Single Linkage* (AHC-SL) diharapkan akan membantu pengguna/pembaca menemukan informasi yang relevan, lebih cepat, dan akan memungkinkan untuk pencarian dokumen cerpen yang diinginkan menjadi lebih mudah.
b. Bagi ilmu pengetahuan hasil penelitian merupakan sumbangan model baru, yakni Model Klasterisasi Dokumen Cerpen dengan metode berbasis *Agglomerative Hierarchical Clustering*, khususnya *Single Linkage Clustering* (AHC-SL).

2. MODEL KLASTERISASI GENRE CERPEN KOMPAS MENGGUNAKAN AHC-SL

Cerpen merupakan tulisan tentang kisah pendek yang memberikan kesan tunggal yang dominan dan memusat pada suatu tokoh pada suatu situasi tertentu. Jumlah dokumen cerpen sangat banyak sehingga menyulitkan pembaca dalam mencari dokumen cerpen yang diinginkan. Oleh sebab itu dibutuhkan sebuah metode khusus untuk dapat mengelompokkan cerpen – cerpen tersebut sehingga dapat memudahkan pembaca dalam mencari dokumen cerpen yang diinginkan.

Beberapa jenis cerpen yang diterbitkan oleh penerbit Kompas antara lain cerpen romantis, jenis ini

biasanya menyuguhkan kisah percintaan mulai awal sampai akhir cerita. Tokoh yang ada dalam jenis ini biasanya berkisar antara usia remaja hingga dewasa. Pertemuan antara dua karakter utama ditulis semenarik mungkin, dilanjutkan dengan konflik-konflik percintaan hingga mencapai sebuah titik klimaks, lalu diakhiri dengan sebuah ending yang kebanyakan bercabang jadi tiga : happy ending (dua tokoh utama bersatu), sad ending (dua tokoh utama tidak bersatu), dan ending menggantung (pembaca dibiarkan menyelesaikan sendiri kisah itu).

Cerpen religi, jenis cerpen ini biasanya merupakan kisah romantis atau inspiratif yang ditulis lewat sudut pandang religi. Cerpen horor, cerpen jenis ini biasanya bercerita seputar hantu. Sisi yang menarik dari cerpen ini adalah latar tempatnya, yang kebanyakan sebagai sumber hantu itu berasal. Cerita juga biasa disajikan dalam bentuk perjalanan sekelompok orang ke tempat angker. Cerpen keluarga, cerpen jenis ini bertemakan keluarga, identitas keluarga, masalah yang dihadapi dan kondisi masyarakat yang bersangkutan.

Sebelum dilakukan klasterisasi dokumen teks diperlukan pemrosesan awal yang disesuaikan dengan koleksi, yaitu cerpen religi, horor, romantis, dan keluarga. Pemrosesan awal (preprocessing) dokumen sangat penting peranannya dalam memilih kata yang akan dimasukkan pada dokumen vektor dan untuk menentukan jumlah kata yang terjadi. Pemrosesan awal sangat menentukan kualitas dan performansi dari data yang mewakili dokumen tersebut. Pemrosesan awal ini perlu dilakukan dengan menjalankan beberapa prosedur seleksi dokumen seperti *tokenization*, *stopword removal*, dan *stemming*. Dari hasil pemrosesan awal ini diharapkan proses klasterisasi dapat berlangsung dengan efisien.

Klasterisasi dokumen teks adalah salah satu operasi pada *text mining* untuk mengelompokkan dokumen yang memiliki kesamaan isi. Klasterisasi dapat diaplikasikan untuk menemukan keterkaitan antarcerpen. Klasterisasi dapat digunakan untuk membantu menganalisis cerpen dengan mengelompokkan secara otomatis cerpen yang memiliki kesamaan.

Penentuan Klasterisasi dengan menggunakan algoritma yang keliru dapat dipastikan tidak memenuhi harapan lantaran berlangsung kurang sempurna. Hingga saat ini belum ada pemanfaatan *Agglomerative Hierarchical Clustering* dalam pengelompokan cerpen.

Pengelompokan *Agglomerative Hierarchical Clustering* merupakan metode pengelompokan hierarki dengan pendekatan bawah-atas (bottom up). Proses pengelompokan dimulai dari masing – masing data sebagai satu buah kelompok, kemudian secara rekursif mencari kelompok potensial berdasarkan jarak sebagai pasangan untuk bergabung sebagai satu kelompok yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak ke atas (Agglomerative) membentuk jenjang (hierarki).

Kunci operasi metode *Agglomerative Hierarchical Clustering* adalah penggunaan ukuran kedekatan (proximity) diantara dua kelompok, atau parameter 'kedekatan' kelompok (*cluster* proximity).

Kedekatan dapat didefinisikan sebagai ukuran yang membedakan kelompok - kelompok. Salah satu teknik kedekatan diantaranya adalah *Single Linkage* (jarak terdekat) atau tautan tunggal[7].

Metode *Single Linkage* mengelompokkan dokumen berdasarkan jarak terdekat antardokumen. (nearest neighbor methods) Metode ini menggunakan prinsip jarak minimum yang diawali dengan mencari dua objek terdekat dan keduanya membentuk *cluster* yang pertama. Pada langkah selanjutnya terdapat dua kemungkinan, yaitu 1) objek ketiga akan bergabung dengan *cluster* yang telah terbentuk, atau 2) dua obyek lainnya akan membentuk *cluster* baru. Proses ini akan berlanjut sampai akhirnya terbentuk *cluster* tunggal. Pada metode ini jarak antar*cluster* didefinisikan sebagai jarak terdekat antaranggotanya. *Single lingkage* memberikan hasil bila kelompok-kelompok digabungkan menurut jarak antara anggota-anggota yang paling dekat.

Langkah-langkah dalam algoritma *Agglomerative Hierarchical Clustering* untuk mengelompokan *N* objek (item/variabel) :

a. AHC dimulai dengan *N* klaster, setiap klaster mengandung entiti tunggal dan sebuah matriks dari jarak (similaritas) $D = \{d_{ik}\}$ dengan tipe *NxN*.
b. Pencarian matriks jarak untuk pasangan klaster yang terdekat (paling mirip). Misalnya jarak antara klaster *U* dan *V* yang paling mirip adalah d_{uv}.

c. Penggabungan klaster *U* dan *V*. Pelabelan klaster yang baru dibentuk dengan (*UV*). Peng-update-an entries pada matrik jarak dengan cara :
   1) Menghapus baris dan kolom yang bersesuaian dengan klaster *U* dan *V*.
   2) Penambahan baris dan kolom yang memberikan jarak-jarak antara klaster (*UV*) dan klaster-klaster yang tersisa.

d. Perulangan langkah 2 dan 3 sebanyak (*N-1*) kali. Semua objek akan berada dalam klaster tunggal setelah algoritma berakhir. Pencatatan identitas dari klaster yang digabungkan dan tingkat-tingkat (jarak atau similaritas) penggabungan terjadi.

Input untuk algoritma *Single Linkage* bisa berwujud jarak atau similaritas antara pasangan-pasangan dari objek-objek. Kelompok-kelompok dibentuk dari entitas individu dengan menggabungkan jarak paling pendek atau similaritas (kemiripan) yang paling besar. Pada awalnya, harus ditemukan jarak terpendek dalam $D = \{d_{ik}\}$ dan menggabungkan objek – objek yang bersesuaian misalnya, *U* dan *V*, untuk mendapatkan klaster (*UV*). Untuk langkah (3) dari algoritma di atas jarak – jarak antara (*UV*) dan klaster *W* yang lain dihitung dengan cara :

$$d_{(uv)w} = min\ \{d_{uw}, d_{vw}\}$$

Di sini besaran – besaran d_{uw} dan d_{vw} berturut – turut adalah jarak terpendek antara klaster – klaster *U* dan *W* dan juga klaster – klaster *V* dan *W*.

Untuk mengukur hasil klasterisasi dapat digunakan nilai purity dari suatu klaster. *Purity* (kemurnian) suatu klaster dipresentasikan sebagai anggota klaster yang paling banyak sesuai (cocok) di suatu kelas[8]. Purity dapat dihitung dengan rumus berikut.

$$Purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_{j} \left| \omega_k \cap c_j \right|$$

Suatu klaster dinilai baik apabila nilai puritynya mendekati nol dan buruk bila puritynya mendekati satu.

3. METODE PENELITIAN

3.1. Pengumpulan Data

Penelitian ini menggunakan data yang diambil dari dokumen cerpen Kompas pada situs https://cerpenKompas.wordpress.com/. Terdapat banyak dokumen cerpen yang terbit setiap minggunya, sehingga perlu dikelompokkan (dokumen *cluster*ing) agar memudahkan dalam proses pencarian dokumen cerpen sesuai dengan topik yang diinginkan.

3.2. Pemrosesan Awal (Preprocessing)

Dalam sistem yang dikerjakan, masukan sistem adalah data cerpen Kompas yang didapat dari website cerpen Kompas. Awalnya cerpen tersebut dimasukkan ke dalam file dokumen cerpen berextensi .doc kemudian tiap file cerpen dimasukkan ke dalam sel file exel untuk dikonversi menjadi file yang berextensi .arf. Setelah dataset cerpen dikonversi semua ke dalam file .arf, maka dataset tersebut diproses dengan tools Weka untuk dilakukan proses selanjutnya.

Pemrosesan awal ini dilakukan dengan menjalankan beberapa prosedur seleksi dokumen seperti *tokenization, stopword removal*, dan *stemming*. Tahapan *preprocessing* menghasilkan kumpulan *term* atau kata yang nantinya akan diberi bobot atau nilai. Bobot tersebut mengindikasikan pentingnya sebuah *term* terhadap dokumen. Semakin banyak *term* tersebut muncul pada koleksi dokumen, semakin tinggi nilai atau bobot *term* tersebut. Pemberian bobot tersebut dinamakan *term weighting*. Setelah tahapan pemberian bobot selesai, maka akan dihasilkan sebuah matrik dokumen dengan dimensi mxn; m adalah jumlah *term* dan n adalah jumlah dokumen.

Proses penyusunan matrik kata dokumen (sering disebut tahap pre-processing) adalah sebagai berikut: tahap awal adalah pengubahan ekspresi kata ke *lower-case* dan penghilangan *stop-word*, seperti cerpen atau preposisi misalnya 'ini','itu','yang','yaitu' dan lain-lain. Penghilangan *stop-word* ini dapat mengurangi frekuensi *feature* 30 sampai 40 persen. Proses leksikal yang lain terhadap feature kata adalah proses *stemming*, yang akan mereduksi semua kata ke dalam akar katanya.

Stemming terhadap *feature* kata dilakukan untuk meningkatkan kinerja klasterisasi dan menurunkan jumlah *feature* kata terindeks. Proses *stemming* dilakukan dengan algoritma Porter. Penyusunan matrik kata dokumen dengan pembobotan ternormalisasi. Program dirancang dengan tools Weka. Selanjutnya metode – metode Klasterisasi yang akan diujicobakan, yaitu : metode *Hierarchical Agglomeratif Clustering- Single Linkage* dan metode partitional (*k-Means*).

3.3. Eksperimen

Pada awalnya eksperimen dilakukan dengan menghitung jarak antara semua pasangan dua data/ kelompok data. Hasil perhitungan ini kemudian disimpan dalam matrik jarak yang menginformasikan jarak antar-semua pasangan dua data. Eksperimen perhitungan jarak dilakukan yang pertama menggunakan pendekatan *Agglomerative Hierarchical Clustering* berikutnya menggunakan algoritma k_Means. Demikian pula untuk penentuan klasternya eksperimen dilakukan melalui kedua pendekatan ini.

Berikutnya adalah pembentukan klaster- klaster. Pembentukan yang pertama menggunakan *Agglomerative Hierarchical Clustering - Single Lingkage*, yakni dengan mengukur jarak terkecil diantara dua data atau dua kelompok data hingga mencakup keseluruhan data. Hasil akhir berupa klaster- klaster yang dapat pula digambarkan dengan dendogram. Yang kedua menggunakan k-Means dengan terlebih dahulu menetapkan pusat klaster, mengukur jarak terhadap pusat klaster hingga terbentuk klaster- klaster.

Dari kedua eksperimen tersebut diperoleh hasil klaster- klaster berdasarkan *Agglomerative Hierarchical Clustering - Single Lingkage* dan berdasarkan *k-Means* yang kemungkinan memiliki hasil yang berbeda.

3.4. Evaluasi

Untuk menguji hasil eksperimen digunakan parameter akurasi. Akurasi masing-masing metode dihitung menggunakan Rand Index, yakni dengan menghitung prosentase terhadap keputusan benar dalam setiap klaster.

Dua dokumen yang sama jika dan hanya jika mereka serupa.

a. TP: dua dokumen serupa dalam satu klaster yang sama
b. TN: dua dokumen tidak serupa dalam satu klaster yang berbeda
c. FP: dua dokumen tidak serupa dalam satu klaster yang sama
d. FN: dua dokumen serupa dalam satu klaster yang berbeda

4. PEMBAHASAN

Hasil *pre-processing* dokumen dengan ekstrak kata dengan menggunakan *treshold* minimal disesuaikan dengan koleksi, yaitu cerpen religi, horor, romantis, dan keluarga. Perlakuan *stemming* terhadap *feature* kata dilakukan untuk meningkatkan kinerja klasterisasi.

Pemrosesan awal (*preprocessing*) dokumen sangat penting peranannya dalam memilih kata yang akan dimasukkan pada dokumen vektor dan menentukan jumlah kata yang terjadi. Pemrosesan awal sangat menentukan kualitas dan performansi dari data yang mewakili dokumen tersebut. Pemrosesan awal ini dilakukan dengan menjalankan beberapa prosedur seleksi dokumen seperti *tokenization*, *stopword removal*, dan *stemming*.

Tahapan *preprocessing* menghasilkan kumpulan *term* atau kata yang nantinya akan diberikan bobot atau nilai. Bobot tersebut mengindikasikan pentingnya sebuah *term* terhadap dokumen. Semakin banyak *term* tersebut muncul pada koleksi dokumen, semakin tinggi nilai atau bobot *term* tersebut.

Untuk meningkatkan kemampuan *term* sebagai pembeda dokumen pembobotan atas *term* perlu

dilakukan. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen karena dipercaya bahwa frekuensi kemunculan *term* merupakan petunjuk sejauhmana *term* tersebut mewakili isi dokumen. Kekuatan pembeda terkait dengan frekuensi *term* (*Term*-Frequency, TF), *Term* yang memiliki kekuatan diskriminasi adalah *term* dengan frekuensi sedang. Untuk itu pemotongan *frekuensi* bawah ditempuh dengan memberikan *treshold* tertentu untuk minimal jumlah dokumen yang memuat *term* tersebut. Pemotongan *term* dengan frekuensi tinggi dilakukan dengan membuang *stop-word*.

Penggunaan hanya frekuensi *term* dalam dokumen sebagai bobot *term* tersebut dalam representasi dokumen tidaklah memadai. Hal ini karena dapat muncul dari faktor lain, misalnya banyaknya dokumen yang memuat *term* tersebut, atau faktor panjang dokumen tempat *term* tersebut muncul. Faktor panjang dokumen dalam koleksi berakibat seolah-olah *term* yang sering muncul pada dokumen panjang lebih penting daripada *term* yang kurang sering muncul pada dokumen pendek.

Hasil pengujian dari 127 dataset cerpen didapatkan akurasi 47,2441 % dengan 2155 *term*. Hasil uji coba dengan 5 dataset cerpen mendapatkan hasil akurasi 80% dengan 385 *term*. Ini berarti semakin sedikit dataset yang diolah berpengaruh juga pada hasil akurasinya. *Term* 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. Semakin banyak dataset yang diolah maka nilai akurasinya semakin rendah.

Setelah dilakukan *preprocessing* maka langkah selanjutnya menentukan nilai *Purity* dengan *Agglomerative Hierarchical Clustering* dan algoritma *k-Means.*

$$Purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_{j} \left| \omega_k \cap c_j \right|$$

Purity : rasio antara *class* dominan dalam klaster c_j dan ukuran klaster ω_k

Keterangan :

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ adalah himpunan klaster

ω_k adalah himpunan dokumen dalam ω_k

$C = \{c_1, c_2, \ldots, c_j\}$ *adalah himpunan class*

c_j *adalah dokumen dalam* c_j

Hasil keseluruhan dari algoritma *Agglomerative Hierarchical Clustering- Single Lingkage* secara grafik dapat digambarkan sebagai *tree*, yang disebut dengan dendogram. *Tree* ini secara grafik menggambarkan proses penggabungan dari klaster–klaster yang ada, sehingga menghasilkan klaster dengan level yang lebih tinggi. Cabang–cabang dalam pohon menyajikan klaster. Kemudian cabang–cabang bergabung pada node yang posisinya sepanjang sumbu jarak (*similaritas*) menyatakan tingkat penggabungan terjadi. Kemiripan antardokumen ditentukan dengan mengukur jarak antardokumen. Dua dokumen yang mempunyai jarak paling kecil dikatakan mempunyai kemiripan paling tinggi, dan dikelompokkan ke dalam satu klaster yang sama. Sebaliknya dua dokumen yang mempunyai jarak paling besar dikatakan mempunyai kemiripan paling rendah, dan dimasukkan ke dalam klaster yang berbeda.

Hasil implementasi dari metode *Agglomerative Hierarchical Clustering - Single Lingkage* terhadap 127 dataset cerpen dapat dijabarkan sebagai berikut:

Cluster*ed *instance :

0	124
1	1
2	1
3	1

Keterangan:

Cluster 0 ada 124 dokumen cerpen yang memiliki kemiripan.

Cluster 1 ada 1 dokumen cerpen yang memiliki kemiripan.
Cluster 2 ada 1 dokumen cerpen yang memiliki kemiripan.
Cluster 3 ada 1 dokumen cerpen yang memiliki kemiripan.

Setelah diadakan klastering dokumen cerpen berdasarkan *class* didapatkan hasil sebagai berikut.

Classes* to *cluster

0	1	2	3	
9	0	1	0	Religi
16	0	0	1	Horor
41	0	0	0	Romantis
58	1	0	0	Keluarga

Cluster 0 yang memiliki jumlah data tertingginya adalah 58, *term*asuk *class* keluarga. *Cluster* 1 yang memiliki jumlah data tertingginya adalah 1, *term*asuk no *class*, karena 1 sudah masuk *class* keluarga. *Cluster* 2 yang memiliki jumlah data tertingginya adalah 1, *term*asuk *class* religi. *Cluster* 3 yang memiliki jumlah data tertingginya adalah 1, *term*asuk *class* horor. Tidak dijumpai data yang *term*asuk dalam *class* romantis,

Jadi jumlah data yang benar sesuai *class*nya ada 60 data cerpen (58 berasal dari *class* keluarga, 1 dari *class* religi, dan 1 dari *class* horor) atau 47,2441 %. Data yang tidak benar (*incorrecly cluster*ed *instance*) ada 67 data cerpen atau 52,7559 %.

Hasil implementasi dari metode *k-Means* terhadap 127 dataset cerpen dapat dijabarkan sebagai berikut.

Clustered instance :

0	1
1	22
2	73
3	31

Keterangan:
Cluster 0 terdapat 1 dokumen cerpen yang memiliki kemiripan.
Cluster 1 ada 22 dokumen cerpen yang memiliki kemiripan.
Cluster 2 ada 73 dokumen cerpen yang memiliki kemiripan.
Cluster 3 ada 31 dokumen cerpen yang memiliki kemiripan.

Setelah diadakan *cluster*ing dokumen cerpen berdasarkan *class* didapatkan hasil sebagai berikut.

Classes to cluster

0	1	2	3	
0	1	7	2	Religi
0	4	9	4	Horor
0	8	23	10	Romantis
1	9	34	15	Keluarga

Cluster 0 yang memiliki jumlah data tertingginya adalah 1, *term*asuk no *class*, karena terdapat nilai kembar. *Cluster* 1 yang memiliki jumlah data tertingginya adalah 4, *term*asuk *class* horor. *Cluster* 2 yang memiliki jumlah data tertingginya adalah 34 *term*asuk *class* keluarga. *Cluster* 3 yang memiliki jumlah data tertingginya adalah 10 *term*asuk *class* romantis.

Jadi jumlah data yang benar sesuai *class*nya ada 48 data cerpen atau 37,7953 %. Data yang tidak benar (incorrecly *cluster*ed *instance*) ada 79 data cerpen atau 62,2047 %.

Setelah dilakukan pengujian secara keseluruhan, diperoleh hasil bahwa algoritma *Agglomerative Hierarchical Clustering- Single Lingkage* memiliki performansi yang lebih baik jika dibandingkan dengan algoritma *k-Means*. Nilai *purity* metode *Agglomerative Hierarchical Clustering- Single Lingkage* selalu lebih tinggi jika dibandingkan dengan metode *k-Means*. Dari 127 dataset cerpen yang telah diujicobakan didapatkan nilai akurasi dari metode *Agglomerative Hierarchical Clustering* dan *k-Means* dapat dilihat pada Tabel 1.

Tabel 1. Nilai Akurasi

No	Metode	Akurasi
1.	*Agglomerative Hierarchical Clustering - Single Lingkage*	47,2441 %
2	*k-Means*	37,7953 %

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Hasil penelitian ini menunjukkan bahwa Metode *Agglomerative Hierarchical Clustering - Single Lingkage* memiliki performansi yang lebih baik dibandingkan dengan metode *k-Means*. Dari 127 dataset cerpen yang telah diujicobakan didapatkan nilai akurasi dari metode *Agglomerative Hierarchical Clustering - Single Lingkage* 47,2441 %, sedangkan metode *k-Means* hanya 37,7953 %.

Selain itu juga dapat disimpulkan dokumen yang membahas topik yang sama cenderung untuk mengelompok menjadi satu klaster, dan klaster dapat membantu mendapatkan dokumen yang relevan. Sedangkan *Purity* (kemurnian) suatu klaster direpresentasikan sebagai anggota klaster yang paling banyak sesuai (cocok) di suatu kelas. Jumlah klaster memberikan pengaruh terhadap nilai *Purity*. Jika jumlah klaster bertambah, maka nilai Purity akan mengecil. Jika jumlah klaster berkurang, maka nilai Purity semakin membesar.

5.2. Saran

Berdasarkan penelitian yang telah dilakukan, maka beberapa saran yang dapat diajukan adalah sebagai berikut :

a. Model Klasterisasi (pengelompokan) ini dapat dikembangkan menjadi sistem pemeriksaan kemungkinan adanya plagiarisme di dalam penulisan karya ilmiah.
b. Penelitian ini selanjutnya dapat dikembangkan untuk membandingkan metode-metoe *hierarchical* dan *partitioned* yang lainnya seperti *average linkage*, *complete linkage* dan lain-lain.

ACKNOWLEDGMENTS

Karya ilmiah dengan judul "Klasterisasi Genre Cerpen Kompas Menggunakan *Agglomerative Hierarchical Clustering - Single Lingkage*" ini dapat penulis selesaikan sesuai rencana karena dukungan dari berbagai pihak yang tidak ternilai besarnya. Oleh karena itu penulis menyampaikan terima kasih kepada para dosen dan staf administrasi UDINUS yang telah memberikan bantuannya. Selain itu juga kepada Kepala Badan Kepegawaian Daerah (BKD) Kota Tegal, Kepala Dinas Kesehatan Kota Tegal, Direktur Akper Pemkot Tegal.

PERNYATAAN ORIGINALITAS

"Saya menyatakan dan bertanggungjawab dengan sebenarnya bahwa artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing – masing telah saya jelaskan sumbernya " [ZENAL ARIFIN – P31.2013.01369]

DAFTAR PUSTAKA

[1] Milatina, Abdul Syukur, Catur Supriyanto, "Pengaruh Teks Preprocessing Pada *Cluster*ing Dokumen Teks Berbahasa Indonesia", *Pascasarjana Teknik Informatika Universitas Dian Nuswantoro*, Volume 8, No. 1, April 2012.

[2] Hario Guritno, Stefanus Santosa, Model Klasterisasi Genre Cerpen Kompas Menggunakan K-Means, Vol 13.1 Januari 2017 –ISSN 1907-3380, *Pascasarjana Teknik Informatika Universitas Dian Nuswantoro*

[3] Ernie Kurniawan, Maria Fransiska, Tinaliah. "Penerapan Algoritma K-Means Untuk *Cluster*ing Dokumen E-Jurnal STMIK GI MDP", *Program Studi Teknik Informatika, STMIK GI MDP*.

[4] B. Santosa, Data Mining. Teknik Pemanfaatan Data untuk Keperluan Bisnis, First Edition ed. Yogyakarta: Graha Ilmu, (2007). [2] K. Arai and A. R. Barakbah, "Hierarchical *k-Means*: an algorithm for centroids initialization for *K- Means*," (2007).

[5] Mohammad Rizal Arief, Daniel O Siahaan, Isyee Arieshanti, "Klasterisasi Teks Menggunakan Metode Max-Max Roughness (MMR) Dengan Pengayaan Similaritas Kata",*Program Pasca Sarjana, Jurusan Teknik Informatika , ITS*, Vol. 5, No. 4, Juli 2010.

[6] Rendy Handoyo, R. Rumani M, Surya Michrandi Nasution, Perbandingan Metode *Cluster*ing Menggunakan Metode *Single Linkage* dan K - Means pada Pengelompokan Dokumen, JSM STMIK Mikroskil, Vol 15, No 2, Oktober 2014 ISSN. 1412-0100

[7] J. Sankari, Dr. R. Manavalan, "Document Retrieval using *Agglomerative Hierarchical Clustering* with Multi-view ," *International Journal of Scientific Engineering and Technology*, vol. II, no. 9, pp. 861-865, 1 Sept 2013.

[8] Entin Hartini, *Metode Clustering Hirarki*, Pusat Pengembangan Teknologi Informasi dan Komputesi, Batan.