BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Industri perfilman yang terus berkembang, menjadi suatu nilai tambah bagi perusahaan yang bergerak di bisnis pembuatan film. Dengan terus meningkatnya minat penonton terhadap film akan berbanding lurus dengan kebutuhan perusahaan untuk menghasilkan film yang sesuai dengan permintaan pasar. Apabila film yang diproduksi tidak sesuai dengan minat pasar, dapat menyebabkan film tersebut tidak disukai penonton. Untuk itu perusahaan pembuatan film perlu mengetahui bagaimana segmentasi pasar film saat ini, sehingga dapat membuat film yang disukai oleh banyak penonton.

Salah satu sumber yang dapat dipakai untuk mengetahui pasar film adalah *International Movie Database* (IMDB). IMDB merupakan sebuah *website* yang menyediakan dan menyimpan data mengenai film. Berbagai data film dapat ditemukan disini seperti data genre, tahun tayang, budget, rating dan berbagai data lainnya. Agar data pada IMDB dapat dimanfaatkan untuk mengetahui informasi segmentasi pasar film, dapat dilakukan dengan cara mengelompokan data film tersebut. Akan tetapi masih belum diketahui bagaimana bentuk karakteristik, jumlah kelompok pasar yang terbentuk, dan anggota dari masing-masing kelompok. Apabila pengelompokan salah, maka informasi segmentasi pasar film dapat menjadi tidak tepat. Sebuah teknik yang dapat membantu untuk menyelesaikan masalah pengelompokan data adalah *clustering. Clustering* dapat mengelompokkan data berdasarkan kemiripan antar datanya, sehingga dapat dihasilkan kelompok segmentasi pasar film yang tepat.

Clustering adalah salah satu teknik *data mining* yang digunakan untuk mengelompokkan sejumlah data atau obyek ke dalam k buah *cluster* bedasarkan kemiripan antar datanya [**1**]. Kualitas hasil *clustering* sangat bergantung pada data dan teknik clustering yang digunakan [**2**]. Kesalahan dalam pemilihan teknik *clustering dapat* mengakibatkan kualitas dari hasil *clustering* menjadi buruk dan

berdampak pada pengambilan keputusan yang tidak tepat. Kualitas hasil *clustering* dapat diukur salah satunya dengan menggunakan Davies-Bouldin Index (DBI). Nilai DBI yang semakin mendekati 0 (nol) menandakan kualitas dari hasil *clustering* semakin baik.

Pemanfaatan teknik *Clustering* pada IMDB pernah dilakukan oleh Dharak Shah yang menggunakan algoritma K-Means untuk mengelompokan data film [**3**]. K-Means merupakan algortima dari teknik *clustering* berbasis partisi yang membagi data kedalam K buah *cluster* yang sudah ditentukan sebelumnya. Permasalahan yang sering terjadi pada *clustering* berbasis partisi terletak pada proses pemilihan *centroid* awal yang masih dilakukan secara acak [**4**]. Hal ini menyebabkan kualitas hasil *clustering* pada teknik partisi tidak selalu menghasilkan *cluster* dengan kualitas terbaik, karena hasil dari teknik partisi sangat bergantung pada ketepatan terhadap pemilihan *centroid* awal. Selain teknik partisi, *clustering* mempunyai teknik lainnya yang dapat digunakan untuk menyelesaikan kasus *clustering* yaitu *clustering* berbasis hierarki. *Clustering* berbasis hierarki tidak memerlukan penentuan jumlah *cluster* terlebih dahulu di awal proses. Sebaliknya, algoritma hirarki dapat menduga jumlah *cluster* terbaik dari suatu *dataset* [**5**]. Akan tetapi metode hierarki memiliki kelemahan, yaitu semakin bertambahnya jumlah data dan atribut dapat menyebabkan akurasi hasil *clustering* dari algoritma hierarki menjadi buruk [**5**].

Berbagai penelitian telah dilakukan untuk menutupi kekurangan dari algoritma *clustering*. *KMeans-Particle Swarm Optimization* (K-PSO) merupakan pengembangan dari algoritma partisi *K-Means* yang terinspirasi dari kemampuan *Particle Swarm Optimization* (PSO) dalam pencarian solusi terbaik. Penerapan PSO untuk *clustering* data pertama kali dilakukan oleh Van der Merwe dan Engelbrecht [**4**]. Pada awalnya PSO dibuat untuk menyelesaikan masalah optimasi, namun beberapa tahun terakhir PSO telah banyak digunakan untuk memecahkan berbagai masalah yang berkaitan dengan *clustering.* Hal ini disebabkan karena PSO dapat memperbaiki kelemahan algoritma *clustering* berbasis partisi dalam pemilihan *centroid* awal, sehingga mampu memberikan hasil *clustering* yang lebih stabil. Penelitian yang dilakukan oleh Van der Merwe

[**4**] dan Nanik utami [**6**] membuktikan bahwa implementasi algoritma K-PSO dapat menghasilkan *cluster* dengan akurasi yang lebih baik pada data yang memiliki banyak atribut. Akan tetapi algoritma K-PSO memiliki kelemahan pada kecepatan proses yang cenderung lambat saat jumlah data semakin banyak dan masih memerlukan inputan jumlah *cluster*, sehingga tidak menjamin jumlah *cluster* yang terbentuk adalah jumlah *cluster* yang terbaik dari dataset.

Selain itu, ada juga penelitian lain yang menggabungkan algoritma *clustering* berbasis hierarki dengan clustering berbasis partisi dengan tujuan untuk meningkatkan kualitas hasil *clustering*. Penelitian yang dilakukan oleh oleh Jung, dkk mencoba menggabungkan algoritma *Hierarchical Agglomerative Clustering* (HAC) dan *K-Means* [**7**]. Algoritma HAC memiliki keunggulan yaitu tidak perlu menduga jumlah *cluster* yang diinginkan. Sebaliknya, algoritma ini dapat mendeteksi jumlah *cluster* yang dapat dibentuk dari suatu *dataset*. Kemampuan dari HAC ini digunakan oleh Jung dalam penelitiannya untuk menentukan jumlah *cluster* terbaik dari dataset dan digunakan sebagai inputan jumlah *cluster* pada proses K-Means. Algoritma dari hasil penelitian Jung mampu menghasilkan *cluster* dengan kualitas yang lebih baik dibandingkan K-Means. Akan tetapi algoritma ini masih memiliki kekurangan, yaitu pemilihan *centroid* awal yang masih dipilih secara acak mengakibatkan kualitas hasil *clustering*nya tidak optimal [**7**].

Dalam menyelesaikan kasus segmentasi pasar film, diperlukan algoritma yang sesuai dengan karakteristik *dataset* IMDB sehingga dapat menghasilkan kelompok segmentasi pasar film yang tepat. Berdasarkan hal yang telah diuraikan sebelumnya, dapat diketahui bahwa algoritma K-PSO dapat menyelesaikan permasalahan algoritma partisi dalam pemilihan *centroid* awal, akan tetapi algoritma ini memiliki kekurangan yaitu tidak mampu menduga jumlah cluster terbaik dari *dataset*. Sedangkan algoritma HAC diketahui dapat menduga jumlah *cluster* terbaik dari suatu *dataset*. Berdasarkan hal tersebut, penggabungan algoritma HAC dan K-PSO (HA-KPSO) dinilai dapat menghasilkan kualitas *clustering* yang lebih baik pada kasus segmentasi pasar film. Hal ini disebabkan karena algoritma HA-KPSO dapat mengatasi masalah penentuan jumlah *cluster*

dan pemilihan *centroid* awal terbaik. HA-KPSO dapat membentuk *cluster* dengan kualitas terbaik saat dievaluasi menggunakan DBI dan antar anggota *cluster*-nya memiliki tingkat kemiripan yang tinggi pada hasil *clustering* dataset IMDB untuk segmentasi pasar film.

1.2 Identifikasi Masalah

Permasalahan yang dikaji pada penelitian ini adalah bagaimana cara menggabungkan algoritma *Hierarchical Agglomerative Clustering* (HAC) dan *KMeans-Particle Swarm Optimization* (K-PSO) untuk menghasilkan *cluster* yang memiliki nilai DBI terbaik pada data IMDB untuk kasus segmentasi pasar film. Algoritma HAC memiliki keunggulan yaitu dapat menduga jumlah *cluster* dari suatu *dataset*. Sementara itu, algoritma K-PSO dapat mencari *centroid* awal terbaik pada dataset dan memiliki akurasi yang baik pada pengelompokan data yang memiliki banyak atribut. Dengan menggabungkan kelebihan dari masing-masing algoritma, diharapkan gabungan HAC dan K-PSO (HA-KPSO) mampu menghasilkan *cluster* dengan nilai DBI terbaik pada *dataset* IMDB untuk menghasilkan kelompok segmentasi pasar film yang tepat.

Pada penelitian sebelumnya, yaitu pada penggabungan algoritma HAC dan K-Means oleh Jung [**7**] maupun penggabungan algoritma K-Means dan PSO oleh van der merwe [**4**], tidak disebutkan dengan jelas bagaimana karakteristik masing-masing algoritma dalam menangani variasi data. Berdasarkan permasalahan tersebut, maka hal berikutnya yang dilakukan dalam penelitian ini adalah meneliti karakteristik algoritma HA-KPSO dalam menangani data dengan variasi jumlah data, jumlah atribut dan jumlah *cluster* menggunakan data model. Sehingga dapat diketahui bentuk karakteristik jumlah atribut dan jumlah data yang menjadi kondisi terbaik dan terburuk untuk algoritma HA-KPSO. Setelah diketahui bentuk karakteristik data dari algoritma HA-KPSO, hal yang dilakukan selanjutnya adalah melakukan implementasi algoritma pada *dataset* IMDB untuk menyelesaikan kasus segmentasi pasar film. Hasil *clustering* dari algoritma HA-KPSO akan dibandingkan dengan beberapa algoritma lainnya menggunakan DBI,

Sehingga dapat diketahui algoritma yang paling tepat untuk menyelesaikan *clustering* pada *dataset* IMDB.

1.3 Premis

Bedasarkan pada permasalahan yang diuraikkan pada latar belakang dan identifikasi masalah, dapat diformulasikan premis-premis sebagai berikut:

1. Algoritma *Hierarchical Agglomerative Clustering* dapat menduga jumlah *cluster* terbaik yang dapat dibentuk dari suatu *dataset* Akan tetapi, penambahan jumlah data dan atribut data dapat menyebabkan akurasi hasil *clustering* menjadi buruk.
2. Algoritma Kmeans-*Particle Swarm Optimization* dapat mencari *centroid* awal terbaik dari suatu *dataset* dan memiliki akurasi yang baik dalam mengelompokan data. Akan tetapi algoritma ini tidak dapat menduga jumlah cluster terbaik yang dapat dibentuk dari suatu *dataset*.

1.4 Hipotesis

Penggabungan algoritma *Hierarchical Agglomerative Clustering* dan *KMeans-Particle Swarm Optimization* (HA-KPSO) dapat menghasilkan *cluster* dengan nilai DBI paling baik pada *dataset* IMDB untuk kasus segmentasi pasar film.

1.5 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menggabungkan algoritma *hierarchical agglomerative Clustering* dan algoritma KMeans - *Particle swarm optimization* untuk menghasilkan *cluster* dengan nilai DBI terbaik.
2. Mengetahui karakteristik dari gabungan algoritma *hierarchical agglomerative Clustering* dan algoritma KMeans - *Particle swarm optimization* pada data.
3. Mengetahui algoritma yang paling tepat untuk digunakan dalam menyelesaikan kasus segmentasi pasar film pada *dataset* IMDB.

1.6 Batasan Masalah

Adapun batasan-batasan yang terdapat pada penelitian ini adalah sebagai berikut :

1. Implementasi kasus *Clustering* dilakukan pada data film yang diperoleh dari IMDB.
2. Algoritma yang diuji adalah gabungan algoritma *Hierarchical Agglomerative Clustering* dan KMeans-*Particle Swarm Optimization* (HA-KPSO).
3. Pengujian dilakukan dengan menggunakan data model dan data kasus yang selengkapnya dijelaskan pada Bab 3.

1.7 Manfaat Penelitian

Adapun manfaat yang diharapkan bisa diperoleh dari penelitian ini adalah sebagai berikut :

1. Bagi sisi keilmuan, penggabungan algoritma HAC dan K-PSO dapat meningkatkan kualitas hasil *clustering* dengan menambahkan proses penentuan jumlah *cluster* dan pencarian *centroid* awal terbaik dari suatu *dataset*.
2. Bagi penelitian di bidang *data mining*, penelitian ini memberikan pengetahuan mengenai karakteristik algoritma dalam melakukan proses *clustering* pada data yang memiliki jumlah data dan jumlah atribut tertentu, sehingga dapat diimplementasikan pada kasus dunia nyata yang tepat untuk menghasilkan hasil *clustering* terbaik dibandingkan algoritma yang lain untuk data tersebut.

1.8 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan dari penelitian ini adalah sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang masalah, identifikasi masalah, , premis, hipotesis, tujuan penelitian, batasan penelitian, manfaat penelitian dan sistematika penelitian.

BAB 2 TINJAUAN PUSTAKA

Bab ini menjelaskan mengenai dasar-dasar teori yang digunakan dalam penelitian ini meliputi teori *data mining*, teknik-teknik dalam *data mining*, Clustering, Algoritma dalam *clustering* dan metode validasi *cluster*.

BAB 3 METODE PENELITIAN

Bab ini menjelaskan mengenai teknik penggabungan algoritma *Hierarchical Agglomerative Clustering* dan K-Means – *Particle Swarm Optimization* serta langkah-langkah penelitian yang terdiri dari pengujian menggunakan data model dan data kasus untuk mengetahui algoritma yang paling baik dalam menyelesaikan kasus *clustering* pada *dataset* IMDB untuk kasus segmentasi pasar film.

BAB 4 HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menjelaskan mengenai hasil dari penelitian pada data model dan data kasus yang disajikan dalam bentuk tabel dan grafik serta analisis terhadap hasil pengujian.

BAB 5 KESIMPULAN DAN SARAN

Bab ini merupakan bab penutup yang di dalamnya menjelaskan mengenai kesimpulan dan saran dari penelitian ini.