Jurnal UJMC, Volume 6, Nomor 1, Hal. 1 - 8
pISSN : 2460-3333 eISSN : 2579-907X

Implementasi Metode *Improved K-Means* dengan Algoritma *Dbscan* untuk Pengelompokan Film

Annisa Ayunda Permata Sari[1], Muhammad Muhajir[2]

[1]Universitas Islam Indonesia, 16611007@students.uii.ac.id
[2]Universitas Islam Indonesia, mmuhajir@uii.ac.id

Abstract. *The Indonesian film industry continues to experience an increase seen from the number of films that appear in theaters today with a box office increase of 28 percent each year in the past four years. Internet Movie Database (IMDb) is a website that provides information about films around the world, including the people involved in it from actors, directors, writers to makeup artists and soundtracks. In this case the researcher wants to conduct research on the characteristics of the film and the factors that make a film to be included in the IMDb Top 250. The data used in this study uses scraped data from the website. The method used is a non-hierarchical clustering method, namely kmeans and Dbscan. Where the Dbscan algorithm is used to determine the optimum number of clusters then proceed by grouping data based on centroids with k-means algorithm. From the analysis it was found that the factors that could influence a film included in the IMDB Top 250 were duration, number of votes, and films directed by Rajkumar Hirani and the optimal number of clusters using Dbscan algorithm obtained six clusters. With the improved k-means algorithm, the accuracy value for the cluster results is 87.2%.*

***Keywords*:** *Movie, Imdb, Kmeans, Dbscan*

Abstrak. Industri perfilman Indonesia terus mengalami peningkatan dilihat dari banyaknya film-film yang muncul di bioskop saat ini dengan peningkatan *box office* sebesar 28 persen setiap tahun nya dalam kurun waktu empat tahun terakhir. *Internet Movie Database* (IMDb) adalah situs web yang menyediakan informasi mengenai film dari seluruh dunia, termasuk orang-orang yang terlibat di dalamnya mulai dari aktor/aktris, sutradara, penulis sampai penata rias dan soundtrack. IMDb merupakan sumber informasi paling populer dan terpercaya baik untuk film, TV, dan konten selebritas lain. Dalam hal ini peneliti ingin melakukan penelitian mengenai karakteristik film dan faktor yang membuat sebuah film dapat masuk dalam IMDb Top 250. Data yang digunakan pada penelitian ini menggunakan data hasil *scraping* dari website. Metode yang digunakan adalah metode pengelompokan *cluster non-hierarki*, yaitu *kmeans* dan *Dbscan*. Dimana algoritma *Dbscan* digunakan untuk menentukan jumlah *cluster* optimum kemudian dilanjutkan dengan mengelompokkan data berdasarkan *centroid* dengan algoritma *k-means*. Dari hasil analisis diperoleh bahwa faktor yang dapat memengaruhi suatu film masuk dalam IMDB Top 250 adalah durasi, jumlah votes, dan film yang disutradarai oleh Rajkumar Hirani dan untuk jumlah cluster optimal menggunakan algoritma *Dbscan* diperoleh enam *cluster*. Dengan algoritma *improved k-means* didapatkan nilai akurasi untuk hasil *cluster* sebesar 87.2%.

***Kata Kunci*:** *Film, IMDb, K-Means, Dbscan*

1 Latar Belakang

Dunia perfilman memiliki kisah perjalanan yang cukup panjang, mulai dari film bisu dan tidak berwarna hingga saat ini telah menjadi film yang kaya akan efek dan dapat dengan mudah ditemukan di dunia hiburan. Hal ini tidak lepas pada Industri perfilman Indonesia, dimana dalam kurun 4 tahun terakhir mengalami peningkatan 28% setiap tahunnya. Menurut [1] data yang ada mencatat bahwa Indonesia menjadi negara Asia Pasifik dengan perkembangan paling signifikan. Hal ini membuktikan bahwa pasar film Indonesia mempunyai potensi yang layak untuk dapat bersaing dengan film luar lainnya. Sehingga perlu diamati bagaimana perkembangan film-film dari berbagai penjuru dunia yang banyak disukai saat ini. Salah satu media komunikasi antara pembuat dan penikmat film adalah internet. IMDb merupakan sumber informasi paling populer dan terpercaya baik untuk film, TV, dan konten selebritas lain [2]. Diketahui jumlah pengunjung harian IMDb mencapai 40.3 juta pengunjung dan jumlah pengunjung bulanan mencapai 967.5 juta. Dalam hal ini peneliti ingin melakukan penelitian mengenai pengelompokan film dan faktor yang membuat sebuah film dapat masuk dalam IMDb Top 250, sehingga oleh penikmat film dapat digunakan sebagai referensi untuk memilih film sesuai dengan karakteristik yang diinginkan dan oleh para pembuat film dapat menjadi bahan masukan mengenai bagaimana karakter film yang banyak disukai oleh masyarakat dunia. Penelitian ini dilakukan dengan metode cluster non-hierarki yaitu *Dbscan* dan *Kmeans.*

2 Tinjauan Pustaka

2.1 Film

Menurut Ibrahim [3], pada hakikatnya semua film adalah dokumen sosial dan budaya yang membantu mengkomunikasikan zaman ketika film itu dibuat bahkan sekalipun ia tak pernah dimaksudkan untuk itu.

2.2 *Text Mining*

Text mining adalah suatu proses analisis data berupa teks dimana sumber datanya didapatkan dari dokumen. Tahapan dalam melakukan analisis pada *text mining* adalah mengumpulkan data kemudian melakukan ekstraksi terhadap fitur yang digunakan. Ekstraksi fitur dilakukan dengan melakukan pembersihan data mulai dari *tokenizing*, *stop words removal*, dan *stemming* [4].

2.3 Algoritma *Term Frequency Inverse Document Frequency (TF-IDF)*

Metode TF-IDF adalah sebuah cara pemberian bobot hubungan suatu kata *(term)* terhadap dokumen. Dimana TF-IDF ini merupakan ukuran statistik yang digunakan untuk mengevaluasi seberapa penting peran kata tersebut dalam sebuah dokumen atau kelompok kata [5].

$$W_{dt} = tfdt * log\left(\frac{N}{df}\right) \quad (1)$$

Dimana:
W_{dt} = bobot dokumen ke-d terhadap kata ke-t

$tfdt$ = banyaknya kata yang dicari pada sebuah dokumen
N = total dokumen
df = banyak dokumen yang mengandung kata yang dicari [6].

2.5 Analisis Faktor

Analisis faktor merupakan suatu metode yang digunakan untuk mereduksi atau meringkas data, dari variabel yang banyak menjadi variabel yang lebih sedikit tanpa menghilangkan informasi yang terkandung dalam variabel asli [7].

2.6 *Principal Component Analysis (PCA)*

Menurut Jolluffe [8], prosedur PCA bertujuan untuk menyederhanakan dan menghilangkan faktor atau indikator skrining yang kurang dominan dan kurang relevan tanpa mengurangi arti dan tujuan dari data asli dari variabel random x (matriks berukuran n x n).

2.7 Analisis Clustering

Menurut Supranto [7] analisis *clustering* merupakan suatu kelas teknik yang dipergunakan untuk mengklasifikasikan objek kedalam kelompok yang relatif homogen. Analisis *cluster* dibagi menjadi dua, yaitu metode hierarki dan non-hierarki. Metode non-hierarki menghasilkan partisi dari data sehingga objek dalam satu *cluster* lebih mirip satu sama lain dibandingkan dengan objek dalam *cluster* lain [9].

2.7.1 *Dbscan Clustering*

Dbscan adalah metode pengelompokan berdasarkan tingkat kepadatan data (*density-based*). Tidak seperti *k-means clustering*, *Dbscan* tidak perlu menentukan jumlah kelompok secara manual. Namun, diperlukan jumlah minimum tetangga untuk dipertimbangkan dalam kelompok dan jarak maksimum yang diperbolehkan antara titik manapun untuk menjadi bagian dari kelompok yang sama [10]. Salah satu keuntungan *Dbscan* dibanding *k-means* adalah *Dbscan* tidak terbatas pada jumlah *cluster* yang ditetapkan saat inisialisasi. Algoritmanya akan menentukan jumlah *cluster* berdasarkan kepadatan suatu daerah, sehingga dapat digunakan untuk optimalisasi cluster [11].

2.7.2 *K-means Clustering*

K-means merupakan jenis algoritma yang digunakan untuk mengelompokkan data berdasarkan titik pusat *cluster* (*centroid)* data. Pengelompokkan data dilakukan dengan cara memaksimalkan kesamaan data pada satu *cluster* dan meminimalkan kesamaan data antar *cluster*. Fungsi jarak dalam *cluster* digunakan sebagai ukuran kemiripan. Sehingga proses pemaksimalan kemiripan data didapatkan dari jarak terpendek antara data terhadap titik *centroid* [12].

2.8 *Improved K-means*

Konsep dasar yang digunakan pada *improved k-means* adalah metode *k-means clustering.* Pada metode *k-means,* pencarian titik pusat awal dilakukan dengan cara acak. Namun pada *improved k-means,* dilakukan modifikasi pada tahapan algoritma dengan menambahkan beberapa tahapan dalam pencarian titik pusat awal sehingga titik pusat awal ditemukan tanpa pengacakan [13]. Berikut merupakan langkah-langkah dalam algoritma *improved k-means*:

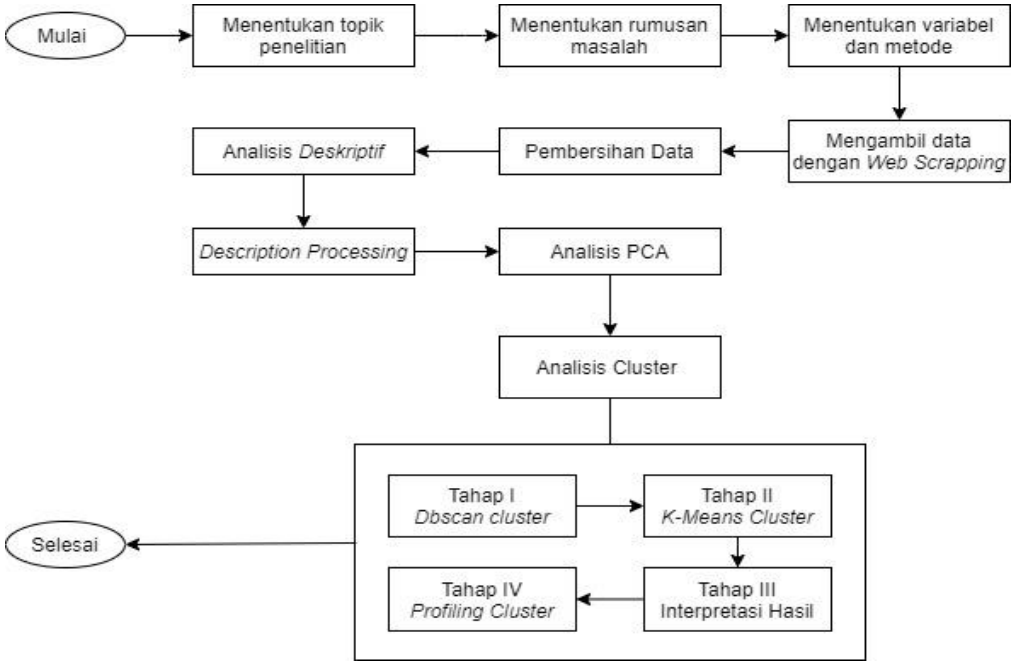
1. Memasukkan jumlah *cluster* (k) yang telah diperoleh dengan algoritma *Dbscan,* dan menetapkan pusat *cluster* sembarang.
2. Menghitung jarak setiap data ke pusat *cluster* menggunakan persamaan.
3. Mengelompokkan data ke dalam *cluster* dengan jarak paling dekat menggunakan persamaan.
4. Menghitung pusat *cluster* yang baru dengan persamaan.
5. Mengulangi langkah 2 sampai 4 sampai sudah tidak ada lagi data yang berpindah ke *cluster* yang lain [14].

3 Metodologi Penelitian

3.1 Metode Analisis Data

Analisis data diawali dengan metode PCA kemudian dilanjutkan dengan *clustering.* Metode *clustering* yang digunakan oleh peneliti adalah metode non-hierarki *Dbscan* dan *K-means. Dbscan* digunakan untuk menentukan jumlah optimal kelompok dan algoritma *k-means* untuk mengelompokkan obyek film.

3.2 Langkah-Langkah Penelitian



Gambar 1. Flowchart Penelitian

3.3 Obyek Penelitian

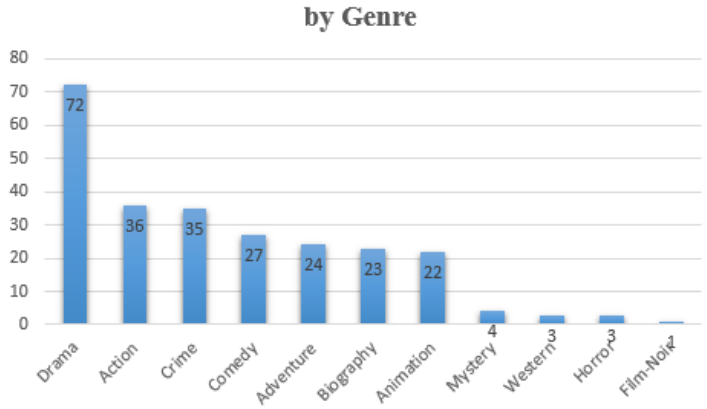
Populasi penelitian ini adalah seluruh informasi film di website IMDb. Sampel data yang digunakan adalah data judul film, tahun *release*, durasi, deskripsi, *votes*, *rating*, *genre*, *actor*, dan *director* film dalam IMDb Top 250 per 5 Desember 2019.

3.4 Jenis dan Sumber Data Penelitian

Data yang digunakan dalam penelitian ini adalah berupa data sekunder yang diperoleh dari website *Internet Movie Database* (IMDb) dengan teknik *web scraping*.

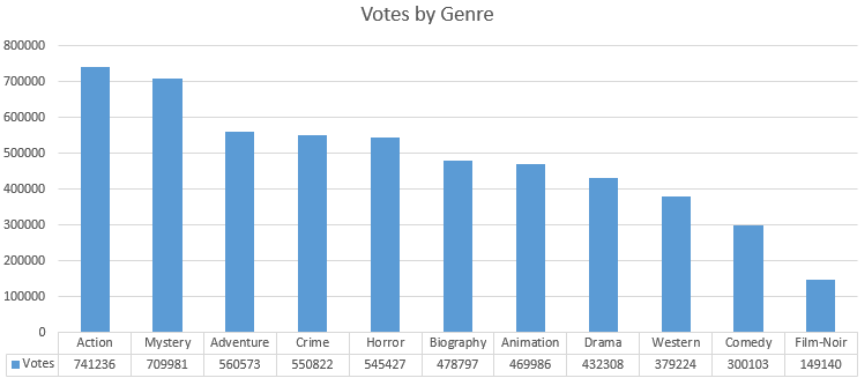
4 Hasil dan Pembahasan

Dalam penelitian ini terdapat sebelas macam genre (*Drama, Action, Crime, Comedy, Adventure, Biography, Animation, Mystery, Western, Horror, Film-Noir)* dengan frekuensi dapat dilihat pada plot histogram dibawah.



Gambar 2. Plot *Genre*

Film dengan *genre Drama* merupakan yang paling banyak muncul dalam *Top 250* dengan lebih dari 60 film, diikuti dengan *genre Action* dan *Crime* yang masing-masing muncul sebanyak 36 dan 35 film. Sedangkan frekuensi *genre* yang muncul paling sedikit dalam *Top 250* adalah film dengan *genre Film-Noir* sebanyak 1 film.



Gambar 3. Rata-rata *Votes* Berdasarkan Genre

Berdasarkan **Gambar 3.** jumlah *votes* setiap genre, film dengan genre *Action* memiliki suara terbanyak dengan 741,236 *votes* dan untuk genre *Film-Noir* memiliki suara terendah dibandingkan genre lainnya karena dalam list ini hanya terdapat satu film dengan genre ini. Posisi kedua dengan jumlah suara terbanyak terdapat genre *Mystery* dengan 709,981 *votes*.

Tantangan utama pada penelitian ini adalah mengekstrak kata-kata relevan dari variabel *description* yang dapat digunakan sebagai kelompok atau variabel *(features)* dalam *cluster*. Dengan algoritma TF-IDF diperoleh 20 kata penting kemudian dilakukan peng-kode-an untuk setiap kata dan dibuat kolom *dummy* masing-masing, jika sebuah film memiliki kata itu diberi nilai 1 dan 0 jika tidak. Hal tersebut juga berlaku untuk variabel *Actor* dan *Director* yang diambil masing-masing 30 dan 20 terbaik berdasarkan jumlahnya. Sehingga didapatkan *data frame* dengan 250 film dan 86 kolom.

Sebelum melakukan pencocokan dengan algoritma *clustering* perlu dilakukan reduksi untuk jumlah *features* dengan *PCA*. Langkah awal *PCA* adalah menstandarisasi *features* atau kolom. Diperoleh 45 komponen yang menjelaskan 81.54% variansi dalam *dataset.* Berikut merupakan *features* yang paling penting dalam *principal component* pertama.

Tabel 1. Tabel *Principal Component*.

Features	***Abs_weights***
Runtime	0.44846
Votes	0.326861
Director:Rajkumar Hirani	0.242888

Pada tabel diatas *Runtime* menempati posisi pertama dengan variansi terbesar, yang artinya durasi sebuah film memiliki peran penting untuk dapat masuk dalam kategori *IMDB's Top 250*. *Votes* muncul dengan nilai variansi terbesar kedua, dimana variabel tersebut merupakan satu-satunya variabel kontinu sehingga cukup mengganggu. Wajar jika *votes* memiliki variansi besar karena kontinu, tetapi fakta bahwa bobot nya secara signifikan jauh lebih tinggi dibanding *features* lainnya juga berarti bahwa *PCA* tidak bekerja dengan baik dengan *dataset* yang didominasi oleh kolom *dummy.*

Langkah selanjutnya adalah memasukkan algoritma *Dbscan cluster* pada data. Untuk menemukan jumlah *cluster* yang optimal, perlu mengacak nilai epsilon dan *min_samples* untuk mendapatkan *shilhouette score* yang baik. *Shilhouette score* digunakan sebagai metriks untuk mengukur kinerja *dbscan.* Pasangan nilai yang memiliki *shilhouette score* terbaik dan dapat dijelaskan lah yang dipilih. Output untuk model terbaik untuk *eps*=11 dan *min_samples*=1. Berdasarkan performa model terbaik algoritma *Dbscan* mengelompokkan data menjadi 6 kelompok (*cluster).*

```
Estimated number of clusters: 6
Silhouette Coefficient: 0.271
```

Gambar 4. Jumlah *Cluster* dengan *Dbscan*

Jumlah *cluster* yang diperoleh dari *Dbscan* akan diterapkan pada *K-means Clustering.* Dalam penelitian ini perlu ditentukan karakteristik umum setiap *cluster* film yang dibentuk dengan mencocokan dan menafsirkan atribut dari film

yang dapat dikelompokkan bersama. Berikut merupakan hasil pengelompokan film dengan algoritma *K-means*:

Tabel 2. Hasil Pengelompokan Film

Cluster	***Interpretasi***
1	Kelompok "***Al Pacino*-C*rime***". Dimana film dalam kelompok ini bergenre *Crime* dan dibintangi oleh aktor *Al Pacino*. Film dalam kelompok ini juga mempunyai durasi >125 menit.
2	Kelompok "Charlie Chaplin" karena semua film dalam kelompok ini dibintangi dan disutradarai oleh Charlie Chaplin. Dengan genre Comedy dan merupakan film-film lama yang di rilis sebelum tahun 1990. Karena di masa sekarang semakin jarang film ber-genre *Comedy* sehingga disarankan bagi pembuat film untuk dapat memproduksi film-film *Comedy* atau me-remake nya.
3	Kelompok "Martin Scorsese". Hampir seluruh film dalam kelompok ini disutradarai oleh Martin Scorsese dan semua film masuk dalam ranking 60 teratas.
4	Kelompok "Most Voted-Christopher Nolan". Karena semua film dalam kelompok ini memiliki jumlah votes lebih dari 1 juta suara dan disutradarai oleh Christopher Nolan. Kelompok ini didominasi oleh film dengan genre Action.
5	Kelompok Drama.
6	Kelompok "Lord of the Rings" dengan genre Adventure. Kelompok ini berisi series film Lord of the Rings yang merupakan film terkini, sehingga kelompok ini memiliki sutradara dan aktor film yang sama yaitu Peter Jackson dan aktor Elijah Wood.

Berdasarkan hasil *cluster* dengan algoritma *improved k-means* didapatkan nilai akurasi sebesar 0.872 atau 87.2%, yang artinya hasil *cluster* cukup baik.

5 Kesimpulan

Dari analisis pada bab sebelumnya, diperoleh kesimpulan sebagai berikut:

1. Faktor yang dapat memengaruhi suatu film dapat masuk dalam IMDB Top 250 adalah Durasi dan Jumlah *Votes*. Dimana kedua variabel tersebut merupakan *features* yang paling berpengaruh dibanding *features* lain pada komponen utama ini.
2. Hasil pengelompokan film dalam list IMDB Top 250 adalah sebagai berikut:
   Cluster 1: kelompok "***Al Pacino*-C*rime***"
   Cluster 2: kelompok "***Charlie Chaplin***"
   Cluster 3: kelompok "***Martin Scorsese***"
   Cluster 4: kelompok "***Most Voted-Christopher Nolan***"
   Cluster 5: kelompok "**Drama"**
   Cluster 6: kelompok "***Lord of the Rings***".

Adapun saran yang dapat diberikan penulis berdasarkan hasil penelitian guna menambah kesempurnaan penelitian selanjutnya yaitu dapat lebih mengenali preferensi film kesukaan penonton berdasarkan atribut film nya dan lebih memperhatikan durasi film dilihat dari faktor yang paling berpengaruh yang menjadikan suatu film dapat masuk dalam IMDB Top 250 ini dan berusaha untuk mendapatkan votes dari penonton karena dengan banyaknya votes dari penonton dapat meyakinkan pengunjung IMDB lain untuk menonton film tersebut.

6 Daftar Pustaka

[1] Portal Informasi Indonesia. (2019). *Tren Positif Film Indonesia*. https://indonesia.go.id/ragam/seni/sosial/tren-positif-film-indonesia. Diakses pada 13 Februari 2020.

[2] Hype Stat. (2020). *Imdb.Com – Info*. https://hypestat.com/info/imdb.com#info. Diakses pada 13 Februari 2020.

[3] Ibrahim, I. S. (2011). *Budaya Populer sebagai Komunikasi; Dinamika Popscape dan.* Yogyakarta: Jalasutra.

[4] Feldman, Ronen, Sanger, & dkk. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

[5] Putra, A. A. (2016). *Implementasi Text Summarization Menggunakan Metode Vector Space Model Pada Artikel Berita Berbahasa Indonesia*.

[6] Fitri, Meisya. (2013). *Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia*. Universitas Tanjungpura: Semarang.

[7] Supranto, J. (2004). *Analisis Multivariat: Arti dan Interpretasi.* Jakarta: PT. Rineka Cipta.

[8] Jolluffe, I. T. (2002). *Principal Component Analysis 2nd Edition.* New York: Springer-Verlag

[9] Triyanto, W. A. (2015). Algoritma K-Medoids untuk Penentuan Strategi. *Jurnal SIMETRIS*, Vol. 6 No.1 April 2015 183-188.

[10] Misra, S., Li, H., & He, J. (2019). *Machine Learning for Subsurface Characterization.* Cambridge: Elsevier Inc.

[11] Bari, A., Chaouchi, M., & Jung, T. (2014). *Predictive Analytics For Dummies.* New Jersey: John Wiley & Sons, Inc.

[12] Asroni, & Adrian, R. (2015). *Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang*. Jurnal Ilmiah Semesta Teknika, 78

[13] Guang-ping, C., & Wen-peng, W. (2012). Improved K-means Algorithm with Meliorated Initial Center. *The 7th International Conference on Computer Science & Education*, Volume 12, 150-153.

[14] Narwati. (2010). Pengelompokan Mahasiswa Menggunakan Algoritma K-Means. *Jurnal Dinamika Informatika*, Vol 2 No. 2.