Recommender System Berbasis Hybrid Filtering di Twitter Menggunakan K-Means Clustering (Studi Kasus: Film di Disney+)

Farid Krida Mukti[1], Erwin Budi Setiawan[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]faridkrida@students.telkomuniversity.ac.id , [2]erwinbudisetiawan@telkomuniversity.ac.id ,

Abstrak

Dengan perkembangan teknologi yang sudah sangat pesat, sehingga menonton film di rumah sudah menjadi sarana hiburan. Disney+ adalah platform menonton film yang menyediakan berbagai berbagai judul film. Namun, karena ada terlalu banyak film terlalu banyak judul film, hal ini menyebabkan kebingungan di antara para pengguna. Untuk menentukan yang mana film untuk ditonton. Solusi untuk masalah ini adalah dengan menyediakan sebuah sistem rekomendasi yang memberikan rekomendasi untuk film yang akan ditonton. Twitter adalah sebuah media sosial yang sosial yang digunakan untuk menulis posting yang disebut *tweet*. Dalam sistem ini, tweet merupakan data yang akan diolah menjadi rating. Penelitian ini dilakukan menggunakan *K-Means Clustering* dan *Hybrid Filtering*. Dengan menggunakan dataset yang diperoleh dari Kaggle dataset berisi IMDb, Disney Rotten Tomatoes, dan Metacritic. Kemudian melakukan preprocessing data dengan *text processing*, *polarity*, dan *labeling*. Dan mendapatkan dataset yang akan digunakan untuk eksperimen. Hasil pengujian dari penelitian ini menunjukkan bahwa *K-Means Clustering* dengan *Hybrid Filtering* mendapatkan hasil yang paling baik dengan mendapatkan nilai MAE sebesar 0.4236, dan nilai RMSE sebesar 0.5246.

Kata kunci : sistem rekomendasi, *hybrid filtering*, *collaborative filtering*, *content-based filtering*, *k-means clustering*.

Abstract

With the rapid development of technology, watching movies at home has become a means of entertainment. Disney+ is a movie-watching platform that provides a wide range of movie titles. However, since there are too many movies too many movie titles, it causes confusion among users. To decide which movie to watch. The solution to this problem is to provide a recommendation system that gives recommendations for movies to watch. Twitter is a social social media that is used to write posts called tweets. In this system, tweets are the data that will be processed into ratings. This research is conducted using K-Means Clustering and Hybrid Filtering. By using a dataset obtained from Kaggle dataset containing IMDb, Disney Rotten Tomatoes, and Metacritic. Then preprocessing the data with text processing, polarity, and labeling. And get a dataset that will be used for experiments. The test results of this study show that K-Means Clustering with Hybrid Filtering gets the best results by getting an MAE value of 0.4236, and an RMSE value of 0.5246.

Keywords: Recommendation System, hybrid filtering, collaborative filtering, content-based filtering, k-means clustering.

1. Pendahuluan

Latar Belakang

Dengan berkembangnya revolusi industri 4.0, jejaring sosial selama beberapa tahun mengalami perkembangan. Media sosial sendiri adalah salah satu alat yang selalu digunakan untuk mengekspresikan berbagai kegiatan dan menyampaikan laporan [1]. Twitter adalah salah satu *platform* media sosial paling populer yang memberikan pengguna untuk membuat cuitan singkat "*Tweet*", mencari informasi popular "*Trending*" dan bisa berkomunikasi secara singkat "*Direct Message*" [2]. Karena fungsinya, banyak pengguna Twitter yang hobi menonton film menggunakan media sosial untuk mencari informasi dan memberikan ulasan terkait film tersebut.

Film merupakan salah satu media hiburan yang menawarkan berbagai macam judul dan *genre* film. Berkat revolusi industri 4.0, kemudian akses terhadap berbagai macam film berkembang pesat. Banyak Perusahaan di industri film telah melakukan pasar *direct-to-customer* untuk menikmati kenyamanan hiburan internet [3]. Salah satu layanan yang sering digunakan adalah Disney+. Disney+ memungkinkan pengguna untuk menonton berbagai judul dan *genre*, bisa menambahkan film ke dalam *watchlist*, dan mengunduh film [2]. Dengan banyaknya judul dan genre yang tersedia di platform Disney+, diperlukan sistem rekomendasi untuk memudahkan pengguna dalam memilih film berdasarkan judul dan genre yang disukai. Di samping itu dapat diterapkan pada platform streaming film, sistem rekomendasi juga dapat diterapkan pada *platform* lain seperti musik, TV, *e-commerce* dan lainnya [4].

Sistem rekomendasi adalah sistem yang dapat membantu mengatasi banjir informasi dengan memberikan rekomendasi khusus kepada pengguna, dan rekomendasi tersebut diharapkan dapat memenuhi keinginan dan

kebutuhan pengguna [5]. Rekomendasi sistem memiliki beberapa metode yaitu *content-based*, *collaborative filtering* dan *hybrid based* [6]. Salah satu metode sistem rekomendasi yang akan digunakan adalah *collaborative filtering*. *Collaborative filtering* adalah sistem rekomendasi yang menggabungkan semua pengguna untuk memilih produk yang sama, seperti film, berdasarkan rating pengguna [7]. *Collaborative filtering* dibagi menjadi dua bagian, yaitu *user based* dan *item based*. *Collaborative filtering* adalah algoritma sistem rekomendasi yang paling sukses dan popular, tetapi memiliki akurasi yang buruk dan waktu berjalan yang lama, sehingga diperlukan clustering untuk mengatasi masalah ini [8]. Dengan adanya masalah pada *collaborative filtering* yang ditemukan, kami menggunakan salah satu dari cluster untuk menyelesaikannya menggunakan *K-Means clustering*.

Oleh karena itu pada penelitian ini kami akan menguji sistem rekomendasi yang dibangun dengan menggunakan *Hybrid Filtering* dengan metode *K-Means Clustering*. Diharapkan dengan adanya sistem ini rekomendasi film Disney+, penonton akan menemukan film yang sesuai dengan minat mereka, dan akan mendapatkan perangkat performa yang lebih akurat dan tepat.

Topik dan Batasannya

Topik pada penelitian ini yang diteliti oleh penulis adalah mengembangkan metode *hybrid filtering* yang digabungkan dengan *k-means clustering*. Keterbatasan dalam penelitian ini adalah pada Dataset bersumber dari website Kaggle dan menggunakan judul film sebagai crawling data di Twitter. Serta tweet yang di crawling pada Twitter adalah tweet yang menggunakan Bahasa Indonesia.

Tujuan

Tujuan dari penelitian ini adalah untuk menerapkan sistem penggabungan teknik *Hybrid Filtering* dengan metode *K-Means Clustering*. Dengan harapan penerapan klasifikasi menggunakan *K-Means Clustering* setelah diproses menggunakan *Hybrid Filtering* dapat menghasilkan model rekomendasi film yang baik dan dapat memberikan prediksi yang akurat untuk film yang direkomendasikan dan tidak direkomendasikan dari proses penambahan metode klasifikasi.

Organisasi Tulisan

Struktur penelitian pada jurnal ini setelah bagian pendahuluan, bagian yang akan dibahas selanjutnya adalah studi terkait pada bagian 2. Dilanjutkan bagian 3 menyajikan teori dan perancangan sistem yang dibangun. Kemudian bagian 4 menyajikan hasil yang didapatkan dan hasil analisis. Pada bagian akhir yaitu bagian 5 berisikan kesimpulan dari penelitian yang telah dilakukan.

2. Studi Terkait

Bagi Pada penelitian ini penulis menggunakan berbagai jenis referensi penelitian yang pernah dilakukan sebelumnya. Referensi penelitian yang dimaksud agar penulis mendapatkan pengetahuan dari penelitian sebelumnya dan dapat sebagai pembanding dan pedoman pada penulisan Tugas Akhir. Referensi penelitian yang dituju berupa segi penelitian, metode, dan kesimpulan yang diperoleh dari penelitian tersebut.

Dari banyaknya penelitian terkait dengan Sistem Rekomendasi menggunakan *Collaborative Filtering* dan ditambah dengan klasifikasi *K-Means*. Salah satunya penelitian yang ditulis oleh Phorasim, Phongsavanh, and Lasheng Yu. Sharma dengan judul penelitian "Movies Recommendation system using Collaborative Filtering and K-Means" [10]. Pada penelitian tersebut menggunakan *Collaborative Filtering* untuk rekomendasi sistem pada restoran dan membandingkan algoritma *Collaborative* dan *K-Means* pada klasifikasi. Untuk dataset yang digunakan bersumber dari Phorasim dengan data berjumlah 1.093.360 ulasan oleh 162.541 pengguna. Berdasarkan hasil uji algoritma KNN dalam uji MSE memperoleh hasil direntang 0.94 – 1.2 dengan memasukan nilai K dari 2 hingga 50 sementara itu pada algoritma *K-Means* dalam uji MSE memperoleh hasil direntang 0.70 – 0,80 yang menunjukan bahwa *user-based K-Means* dan *item-based K-Means* mengungguli KNN.

Penelitian yang selanjutnya yang ditulis oleh M. Irawan, dll berjudul "User-Based dan Item-Based Collaborative Filtering pada Recommender System Buku dengan Metode Naïve Bayes Classification" [11]. Pada penelitian tersebut membandingkan *Collaborative Filtering* dengan *Collaborative Filtering* dikombinasikan dengan *Naïve Bayes Classification* kemudian diimplementasikan pada *user-based* dan *item-based*. Penelitian menggunakan dataset goodbooks-10k dari website kaggle dengan jumlah data rating 981.756, data users 53424 dan data buku 10000. Berdasarkan hasil yang diperoleh dalam menghitung nilai precision dan recall. Menghasilkan bahwa *Collaborative Filtering* dikombinasikan dengan *Naïve Bayes Classification* lebih baik dari *Collaborative Filtering* saja berdasarkan nilai performansi pada setiap metode [9].

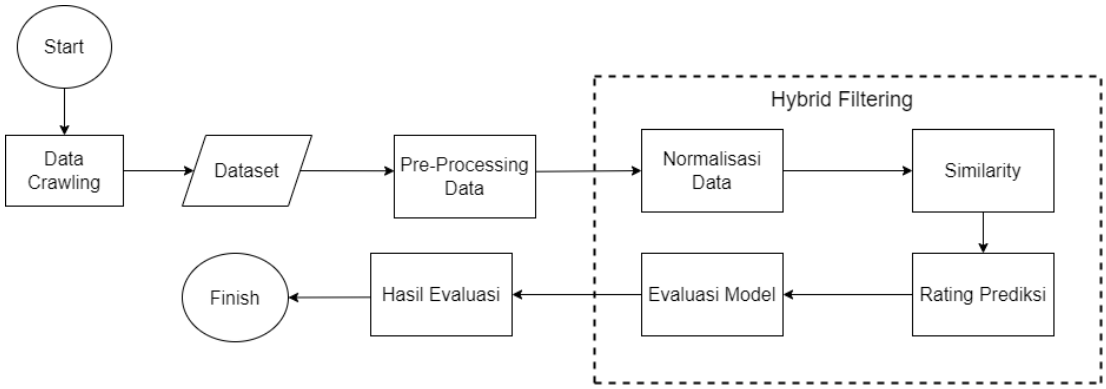
Berdasarkan penelitian Yessica Putri Santoso dll. Dengan topik "Implementasi metode K-Means Clustering pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen". Menampilkan 1 dari 70 data yang diuji terdapat 39 data pembicara yang dapat direkomendasikan sebagian pembicara tetap dan 31 data presenter tidak memenuhi syarat menjadi dosen tetap karena hasil penghitungan akurasi sebesar 55.67% sehingga bisa disimpulkan bahwa algoritma *K-Means* tidak cocok untuk kasus ini [12].

Kemudian berdasarkan penelitian Riyan Alfa Rizkie dan Muhammad Fachrurrozi dengan judul "Sistem Rekomendasi Wisata Kuliner Kota Palembang dengan Metode Collaborative Filtering". Dengan menggunakan metode *collaborative filtering* memiliki nilai relevansi sebesar 80% dan nilai Mae Absolute Error sebesar 0,723948146 menggunakan data makanan sebesar 18 data, pelatihan data sebesar 100 dan pengujian data sebesar 10 [13].

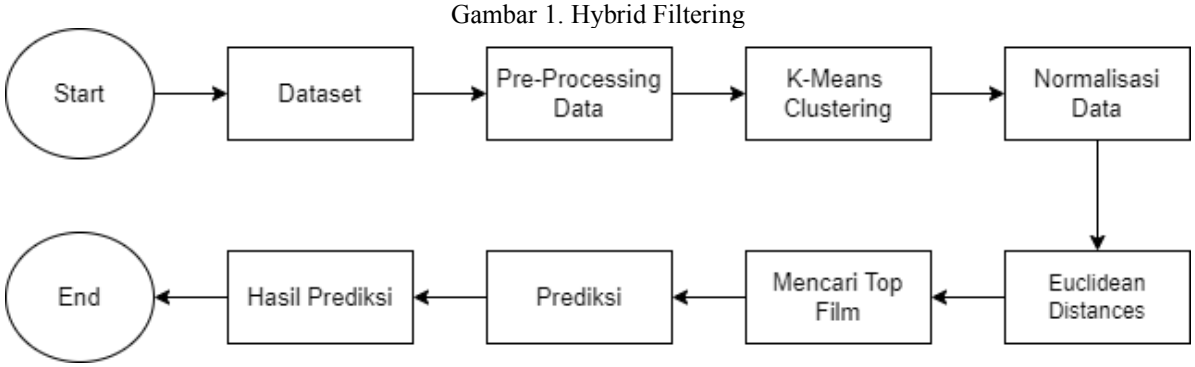
Bersarakan penelitian terkait, penulis akan menggunakan hybrid filtering dan membandingkan dengan tambahan k-means clustering. Karena dinilai lebih efisien dan akurat dari metode klasifikasi lainnya karena menggunakan Mean Absolute Error (MAE) dan Root Mean Square Error (RMSE) untuk memberikan prediksi terkait film [18]. Penulis berharap menggunakan metode tersebut, penelitian yang akan dilakukan mendapatkan nilai performansi yang efisien dan akurat.

3. Sistem yang Dibangun

Pada penelitian ini, metode yang digunakan adalah metode *hybrid filtering* dengan pendekatan cascade. Pada pendekatan hybrid ini, sistem rekomendasi yang dibuat diminta untuk menghasilkan list item rekomendasi menggunakan metode item-based *hybrid filtering*, yang kemudian dilakukan pengurutan ulang menggunakan metode *k-means clustering* untuk diambil item teratas sebagai hasil rekomendasi akhir, sesuai dengan alur pada Gambar 1 & 2.



Gambar 1. Hybrid Filtering



Gambar 2. K-Means Clustering

3.1 Crawling Data

Dataset yang digunakan dalam penelitian ini berasal dari situs Kaggle, yang berupa file Excel di situs web Kaggle. Dataset berisi nilai genre, movie_title, dan review_content. Dengan review_content akan diolah yang sebelumnya opini tentang film yang mereka tonton akan menjadi rating.

Table 1. Crawling Data

User	Judul Film	Ulasan
36	17713	1048575

3.2 Data Preprocessing

Pre-processing merupakan proses tahap pertama untuk memilih kata dari tweet, sehingga menghasilkan kata yang lebih ringkas, dengan cara memilih dan menghilangkan kata-kata yang tidak diperlukan. Pada penelitian ini dilakukan preprocessing terhadap data awal berupa review Twitter yang kemudian diubah menjadi rating 1-5 sehingga dapat digunakan sebagai sistem rekomendasi. Pada proses, pengubahan tweet menjadi rating beberapa tahap yaitu *Text Processing*, *Polarity*, dan *Pelabelan*.

Text processing merupakan langkah untuk memperoleh data yang lebih terstruktur pada saat pemilihan data teks. Pada tahap ini, teks dibersihkan dan masih mengadung unsur tanda baca, URL, dan angka.

Polarity adalah proses mengidentifikasi teks dengan mengetahui seberapa positif dan negative teks tersebut. Pengguna polarity dapat berguna dalam memprediksi kalimat yang mengandung ekspresi positif atau negative, misalnya "film mana yang terbaik", diikuti kata "terbaik" dengan konteks terbaik [21]. Untuk penelitian yang dilakukan menggunakan polarity dengan library TextBlob. Di library ini, dapat membantu mengolah data teks untuk Menentukan arti kata secara efisien. Kemudian data teks diubah menjadi -1 dan 1. Data teks mempunyai nilai polarity yang mendekati -1 artinya akan dilakukan scoring antara 0-2.4, sedangkan untuk data teks mempunyai nilai polarity mendekati 1, maka data tersebut akan diatur antara 2.6-5 dan data teks menghasilkan polarity 0, maka nilainya akan menjadi 2.5.

Pelabelan untuk mengidentifikasi data hasil polarirty untuk mengecek kembali apakah sesuai dengan konteks symbol yang ada, yaitu hasil data teks berupa symbol yang terdapat nilai 0 sampai 5.

3.3 Hybrid Filtering

Hybrid Filtering adalah pendekatan yang menggabungkan beberapa metode filtering berbeda dalam sistem rekomendasi untuk meningkatkan kualitas rekomendasi yang diberikan kepada pengguna. Dengan menggabungkan beberapa metode, sistem rekomendasi hybrid dapat mencapai keseimbangan antara akurasi dan kecepatan serta mengatasi kelemahan dari masing-masing metode individual. Terdapat beberapa metode filtering yang umum digunakan dalam pendekatan hybrid, yaitu Content-Based Filtering, Collaborative Filtering, dan Hybrid Filtering.

3.3.1 Collaborative Filtering

Collaborative Filtering dapat dibagi menjadi dua jenis metode yaitu, *memory-based* dan *model-based*. *Memory-based collaborative filtering* adalah sistem pemfilteran yang merekomendasikan item yang menarik bagi pengguna dengan menghitung kemiripan ukuran ketertarikan pengguna dan menggunakan perilaku informasi pengguna untuk membuat rekomendasi [14]. *Model-based collaborative filtering* adalah sistem penyaringan berdasarkan preferensi pengguna, melatih model rekomendasi, menghitung dan menempatkan hasil rekomendasi berdasarkan preferensi pengguna secara real [14]. Pada *collaborative filtering*, untuk melakukan menggunakan *item-based similarity* dan *user-based similarity*. Rekomendasi berbasis pengguna mencari kesamaan dari sudut pandang pengguna. Mencari target pengguna terdekat untuk menemukan kesamaan di antara pengguna, kemudian menyarankan tujuan yang serupa dari sudut pandang pengguna. Kemudian, N dengan kemiripan yang sama dapat dipilih untuk membentuk himpunan N tetangga terdekat [15]. Ada banyak cara untuk menemukan kemiripan dan korelasi Pearson, termasuk metode ini. Kemiripan korelasi Pearson antara 1 dan 2 dapat dihitung sebagai berikut [16]:

$$\text{Sim}\ (\boldsymbol{u_1}, \boldsymbol{u_2}) = \frac{\sum \boldsymbol{i_{\epsilon l} u_1 u_2} \left(x_{u1,i} - \bar{x}_{u1}\right)\left(x_{u2,i} - \bar{x}_{u2}\right)}{\sqrt{\sum \boldsymbol{i_{\epsilon l} u_1 u_2}} \left(x_{u1,i} - \bar{x}_{u1}\right)^2 \left(x_{u2,i} - \bar{x}_{u2}\right)^2} \quad (1)$$

Untuk $\boldsymbol{u_1, u_2}$ digunakan untuk menunjuk satu set item yang dilapisi oleh $\boldsymbol{u_1}$ $\boldsymbol{dan}$ $\boldsymbol{u_2}$. $\overline{\boldsymbol{x}}_{\boldsymbol{u}}$ menunjukan peringkat rata rata pengguna $\boldsymbol{u_1}$. Ada cara yang lain dimana peringkat pengguna i untuk item j dapat diselesaikan [15].

3.3.2 Content-Based Filtering

Content-based filtering adalah metode menggunakan informasi konten item yang disarankan untuk Menentukan seberapa cocok item tersebut dengan minat pengguna. Dalam hal ini, sistem menggunakan genre film untuk memberikan rekomendasi kepada pengguna. Berikut rumus yang digunakan [17]:

$$\text{Sim } (\boldsymbol{u_1}, \boldsymbol{u_2}) = \frac{\sum \boldsymbol{u}_{\in \boldsymbol{u}} \left(x_{u1,i1} - \bar{x}_{u1}\right)\left(x_{u,i2} - \bar{x}_{i2}\right)}{\sqrt{\sum \boldsymbol{u}_{\in \boldsymbol{u}}} \left(x_{u,i1} - \bar{x}_{i1}\right)^2 \sum \boldsymbol{u}_{\in \boldsymbol{u}} \left(x_{u,i2} - \bar{x}_{i2}\right)^2} \quad (2)$$

Untuk $\mathbf{u_1}, \mathbf{u_2}$ digunakan untuk menunjuk satu set item yang dilapisi oleh $\mathbf{u_1}$ **dan** $\mathbf{u_2}$. $\bar{\mathbf{x}}_{\mathbf{u}}$ menunjukan peringkat rata rata pengguna $\mathbf{u_1}$. Ada cara yang lain dimana peringkat pengguna i untuk item j dapat diselesaikan [17].

3.4 K-Means Clustering

K-Means Clustering merupakan algoritma pembelajaran tanpa pengawasan, umumnya dipakai pada penambangan data. K-Means clustering mempunyai tujuan membagi nilai pada N data sebagai sejumlah K cluster menggunakan jarak rata-rata terdekat [18]. Data yang mempunyai ciri yang sama akan digabungkan sebagai satu cluster, sedangkan data yang mempunyai ciri tidak sama akan digabungkan menggunakan cluster lainnya [19]. Langkah-langkah yang dilakukan pada K-Means clustering merupakan menggunakan memilih jumlah cluster, menampilkan jumlah maksimum pengguna & maksimum file per cluster, dan menghitung cluster terbaik menggunakan jeda Euclidean. Euclidean distance merupakan rumus untuk menghitung jeda ke setiap titik cluster, yang bisa dihitung menggunakan rumus [19].

$$\boldsymbol{D(i,j)} = \sqrt{\left(\boldsymbol{X_{1i,}Y_{1y}}\right)^2 + \left(\boldsymbol{X_{2i,}Y_{2y}}\right)^2 + \ldots + \left(\boldsymbol{X_{ki,}Y_{ky}}\right)^2} \quad (3)$$

Jarak antara data ke 1 dengan centroid setiap cluster. D$(\boldsymbol{i,j})$ adalah jumlah atribut. $\boldsymbol{X_{ki}}$adalah data ke 1.$\boldsymbol{Y_{ky}}$ adalah data pusat setiap cluster.

3.5 Performance Evaluation

Pengukuran performansi rekomendasi sistem dibagi menjadi beberapa bagian, salah satunya dapat diukur dalam classification accuracy metrics menggunakan Confusion Matrix [20]. Menghitung matrix tersebut membutuhkan menghitung precision, recall.

Table 2. Confusion Matrix

		True Class	
		Positive	Negative
Kelas Prediksi	Positive	True positives count (TP)	False negative count (FP)
	Negative	False positive count (FN)	True negative count (TN)

Precision merupakan akurasi antara data yang diminta dengan hasil prediksi yang ada pada model. Recall menggambarkan keberhasilan pada model dalam sebuah informasi.

$$Precision = \frac{TP}{TP + TF} \quad (4)$$

$$Recall = \frac{TP}{TF + FN} \quad (5)$$

Pengukuran performansi selanjutnya diukur menggunakan *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE) untuk menghitung perbedaan yang besar untuk kesalah dalam ranting prediksi [18]. Semakin dekat nilai MAE dan RMSE mendekati 0, maka semakin baik. Berikut rumus MAE dan RMSE [18].

$$\text{MAE} = \frac{\sum_{u=i}^{N} |\boldsymbol{P_{u,i}} - \boldsymbol{R_{u,i}}|}{\boldsymbol{N}} \quad (6)$$

n adalah jumlah observasi atau sampel dalam dataset. y adalah nilai aktual (ground truth). ŷ adalah nilai prediksi yang diberikan oleh model. Σ adalah simbol sigma yang menunjukkan penjumlahan dari seluruh perbedaan absolut antara nilai aktual dan nilai prediksi di seluruh sampel.

$$\text{RMSE} = \sqrt{[\frac{1}{n} * \sum(y - \hat{y})^2]} \quad \textbf{(7)}$$

n adalah jumlah observasi atau sampel dalam dataset. y adalah nilai aktual (ground truth). ŷ adalah nilai prediksi yang diberikan oleh model. Σ adalah simbol sigma yang menunjukkan penjumlahan dari kuadrat dari selisih antara nilai aktual dan nilai prediksi di seluruh sampel.

4. Evaluasi

4.1 Hasil Pengujian

Pada penelitian ini langkah yang pertama menghitung nilai rating prediksi menggunakan metode Hybrid filtering kemudian dievaluasi menggunakan nilai MAE dan RMSE untuk memilih top N terbaiknya. Langkah kedua adalah proses klasifikasi menggunakan algoritma *K-Means Clustering* kemudian di optimalkan menggunakan Performance Evaluation dengan tujuan mendapatkan hasil film direkomendasikan atau tidak direkomendasikan kemudian dievaluasi menggunakan nilai precision dan recall.

4.1.1 Data

Pada Data menghasilkan dua jenis simbol (benar dan salah). Rating benar, jika pengguna memberikan rating untuk judul film, dll. Kemudian hasil pengkodean data dari simbol tersebut akan bernilai 1 jika simbol tersebut benar, namun akan bernilai 0 jika catatan tersebut salah.

Table 3. Hasil Data

UserId	Title			
	Twilight	Train to Busan	....	The Disaster Artist
1	1	0	0	0
2	1	0	1	0
3	0	1	0	0

4.1.2 Hybrid Filtering

Setelah menggabungkan content-based filtering dan collaborative filtering, maka didapatkan tabel sebagai berikut. akan dilakukan normalisasi terhadap dataset seperti menghapus kata yang duplikat. dan mengisi nilai 0 pada kolom yang memiliki nilai, Nan. Maka hasil normalisasi akan menjadi seperti pada tabel 7.

Table 4. Normalisasi Data

idUser Film	Twilight	….	The Spider wick Chronicles
Twilight	-0.0915	….	0
….	….	….	….
The Spider wick Chronicles	0	….	0.1344

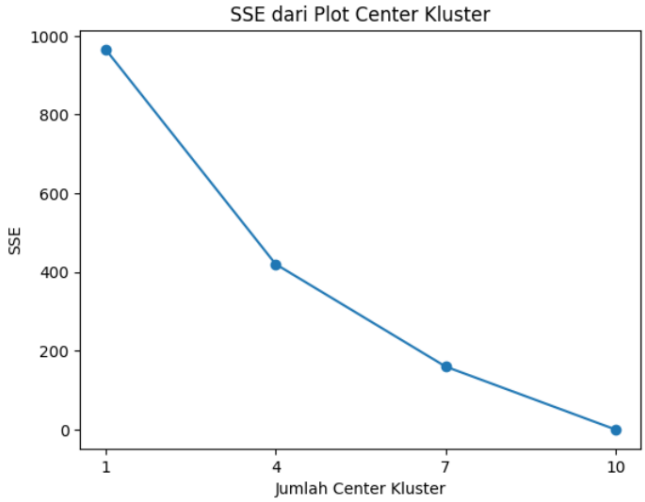
Setelah melakukan normalisasi dataset, dan didapatkan hasil seperti table diatas maka akan mencari nilai kemiripan dengan menggunakan cosine similarity. Kemudian hasil dari cosine similarity untuk pengguna ada pada Tabel 5.

Table 5. Similarity

idUser Film	Twilight	Train to Busan	….	The Spider wick Chronicles
Twilight	1	-0.0984	….	0.0152
Train to Busan	-0.0424	1	….	0.1344
….	….	….	….	….
The Spider wick Chronicles	-0.0847	0.2714	….	1

4.1.3 K-Means Clustering

Untuk mendapatkan cluster yang optimal, maka perlu dilakukan penentuan dengan metode elbow, dimana akan dilakukan perhitungan untuk mendapatkan perbandingan *Sum of Square Error* (SSE) dari masing-masing cluster. Oleh karena itu, jumlah cluster yang akan diambil berdasarkan posisi siku. Berdasarkan Gambar 3 dapat disimpulkan bahwa jumlah cluster yang diambil adalah 4.



Gambar 3. Metode Elbow

Setelah mendapatkan jumlah cluster yang optimal, dilakukan clustering pada dataset. Kemudian dilihat cluster mana yang terbaik dan cluster terbaik ada pada cluster 1. Setelah mendapatkan cluster terbaik maka akan dilakukan tahap berikutnya yaitu normalisasi dengan mengisi 0 pada kolom nilai Nan. Hasil normalisasi dapat dilihat pada Tabel 5.

Table 6. Hasil Normalisasi

idUser Film	Twilight	….	The Spider wick Chronicles
Twilight	0.2861	….	0
….	….	….	….
The Spider wick Chronicles	0.1527	….	0

Setelah dilakukan normalisasi, akan masuk ke tahap berikutnya yaitu mencari nilai kemiripan dengan menggunakan jarak Euclidean. Maka hasil jarak Euclidean per pengguna terdapat pada Tabel 5.

Table 7. Hasil Euclidean Distances

IdUser Film	Twilight	The Spider wick
Twilight	0	1.450349
The Spider wick	1.450349	0

4.2 Analisis Hasil Pengujian

Dari hasil penelitian dapat disimpulkan bahwa hasil pengujian penggabungan hybrid filtering dengan k-means clustering mencapai hasil prediksi sebesar 3.5019. Untuk hybrid filtering hanya mendapat prediksi ranking rendah yaitu 2.8831. Kemudian nilai MAE dan RMSE yang dihasilkan dengan metode hybrid filtering yang dikombinasikan dengan k-means clustering akan mendapatkan nilai MAE sekitar 0.4236 dan untuk RMSE yang diperoleh sekitar 0.5246. Sedangkan dengan hanya menggunakan metode hybrid filtering akan mendapatkan nilai MAE sekitar 4.0257 dan dengan RMSE sekitar 5.1146 berarti gabungan hybrid filtering dan k-means clustering akan lebih baik dibandingkan menggunakan metode hybrid filtering saja karena akurasi dapat dilihat dari nilai rata-rata kesalah MAE dan RMSE. Jika nilainya semakin mendekati 0, maka semakin baik akurasi yang dihasilkan.

5. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dengan menggabungkan hybrid filtering dengan k-means clustering yang digunakan untuk sistem rekomendasi. Dengan menggunakan dataset yang didapat dari Kaggle dengan media Twitter. Yang menghasilkan dataset dengan 36 pengguna, 17713 judul film, dan 1048575 rating ulasan. Kemudian dilakukan preprocessing data dengan text processing, polarity, dan labeling. Dan mendapatkan dataset yang akan digunakan untuk percobaan ini. Setelah itu dilakukan pengujian terhadap dataset tersebut dengan menggabungkan hybrid filtering dengan k-means clustering. Didapatkan bahwa prediksi rating yang dihasilkan dari hybrid filtering memiliki hasil yang lebih rendah dibandingkan dengan k-means clustering saja, yaitu sebesar 2.8831. Kemudian nilai MAE dan RMSE yang dihasilkan oleh hybrid filtering dengan k-means clustering lebih bagus dibandingkan dengan nilai MAE yang dihasilkan oleh hybrid filtering saja. lebih bagus dengan nilai MAE yang dihasilkan sebesar 0.4236 dan untuk RMSE yang dihasilkan sebesar 0.5246 yang dapat diartikan bahwa penggabungan hybrid filtering dan k-means clustering lebih baik dibandingkan dengan hybrid filtering saja, karena akurasi/kinerja dapat dilihat dari nilai rata-rata MAE dan RMSE. Oleh karena itu, penelitian ini diharapkan dapat lebih meningkatkan kinerja sistem rekomendasi dengan Kumpulan data yang lebih besar. Lebih dari itu, dapat dikombinasikan dengan metode lain seperti content base atau algoritma lain.

Daftar Pustaka

[1] D. Das, H. T. Chidananda, dan L. Sahoo, “Personalized movie recommendation system using twitter data,”Advances in Intelligent Systems and Computing, vol. 710, hlm. 339–347, 2018, doi: 10.1007/978-981-10-7871-2_33/COVER.

[2] “Recommender System Berbasis Matrix Factorization di Twitter Menggunakan Random Forest (Studi Kasus: Film di Netflix).”

[3] C. T. Havard, “Findings in Sport, Hospitality, Entertainment, and Event Management Analysis-Entertainment Disney, Netflix, and Amazon Oh My! An Analysis of Streaming Brand Competition and the Impact on the Future of Consumer Entertainment.”

[4] F. Ortega, A. Hernando, J. Bobadilla, dan J. H. Kang, “Recommending items to group of users using Matrix Factorization based Collaborative Filtering,” Inf Sci (N Y), vol. 345, hlm. 313–324, Jun 2016, doi: 10.1016/J.INS.2016.01.083.

[5] G. Kumar, S. Rathod, and A. Laha, “Sentiment Analysis on Micro-Blogs,” SSRN Electron. J., vol. 4, no. 11, pp. 121–126, 2021, doi: 10.2139/ssrn.3867142.

[6] L. R. Dharmawan, I. Arwani, and D. E. Ratnawati, “Analisis Sentimen pada Sosial Media Twitter Terhadap Layanan Sistem Informasi AkademikMahasiswa Universitas Brawijaya dengan Metode K-Nearest Neighbor,”J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 3, pp. 959–965, 2020, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j- ptiik/article/view/7099.

[7] V. L. Jaja, B. Susanto, and L. R. Sasongko, “Penerapan Metode Item-Based Collaborative Filtering Untuk Sistem Rekomendasi Data MovieLens,” d’CARTESIAN, vol. 9, no. 2, p. 78, 2020, doi: 10.35799/dc.9.2.2020.28274.

[8] P. G. Padti, K. Hegde, and P. Kumar, “Hybrid Movie RecommenderSystem,” vol. 4, no. 7, pp. 311–314, 2021.

[9] A. Halim, H. Gohzali, D. M. Panjaitan, and I. Maulana, “Sistem Rekomendasi Film menggunakan Bisecting K-Means dan CollaborativeFiltering,” Citisee, vol. 1, no. 3, pp. 37–41, 2017.

[10] P. Phorasim and L. Yu, “Movies recommendation system using collaborative filtering and k-means,” Int. J. Adv. Comput. Res., vol. 7, no.29, pp. 52–59, 2017, doi: 10.19101/IJACR.2017.729004.

[11] Phorasim, Phongsavanh & Yu, Lasheng. (2017). Movies recommendation system using collaborative filtering and k-means. International Journal of Advanced Computer Research. 7. 52-59. 10.19101/IJACR.2017.729004.

[12] Y. P. Santoso, M. Marlina, and H. Agung, “Implementasi Metode K-Means Clustering pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen,” J. Inform. Univ. Pamulang, vol. 3, no. 4, p. 228, 2018, doi: 10.32493/informatika.v3i4.2133.

[13] R. A. Rizkie and M. Fachrurrozi, “Sistem Rekomendasi Wisata Kuliner Kota Palembang Menggunakan Metode Collaborative Filtering,” Generic, vol. 12, no. 1, pp. 1–3, 2020, [Online]. Available: http://generic.ilkom.unsri.ac.id/index.php/generic/article/view/101.

[14] X. Wang, Z. Dai, H. Li, and J. Yang, “A New Collaborative Filtering Recommendation Method Based on Transductive SVM and Active Learning,” Discret. Dyn. Nat. Soc., vol. 2020, no. 1, 2020,

doi: 10.1155/2020/6480273.
[15] P. Thakkar, K. Varma, V. Ukani, S. Mankad, and S. Tanwar, Combining User-Based and Item-Based Collaborative Filtering Using Machine Learning. Springer Singapore, 2019.
[16] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," vol. 22, no. 1, pp. 5–53, 2004.
[17] V. Subramaniyaswamy, R. Logesh, M. Chandrashekhar, A. Challa, and V. Vijayakumar, "A personalised movie recommendation system based on collaborative filtering," Int. J. High Perform. Comput. Netw., vol. 10, no. 1–2, pp. 54–63, 2017, doi: 10.1504/IJHPCN.2017.083199.
[18] M. Garanayak, S. N. Mohanty, A. K. Jagadev, and S. Sahoo, "Recommender system using item based collaborative filtering (CF) and K-means," Int. J. Knowledge-Based Intell. Eng. Syst., vol. 23, no. 2, pp. 93–101, 2019, doi: 10.3233/KES-190402.
[19] Y. P. Santoso, M. Marlina, and H. Agung, "Implementasi Metode K- Means Clustering pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen," J. Inform. Univ. Pamulang, vol. 3, no. 4, p. 228, 2018, doi: 10.32493/informatika.v3i4.2133.
[20] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," vol. 22, no. 1, pp. 5–53, 2004.
[21] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," Int. J. Inf. Technol., 2020, doi: 10.1007/s41870-019-00409-4.

Lampiran