Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 1, Januari 2019, hlm. 196-204 http://j-ptiik.ub.ac.id

Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN)

Nurul Muslimah[1], Indriati[2], Randy Cahya Wihandika[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]nurulmuslimah25@gmail.com, [2]indriati.tif@ub.ac.id, [3]rendicahya@ub.ac.id

Abstrak

Film ialah media komunikasi bersifat audio visual, dimana tersirat pesan yang ingin disampaikan pencipta film. Film memiliki beberapa genre yakni romantis, horor, thriller, komedi, fantasi dan lain sebagainya. Tidak sedikit penikmat film yang masih bingung akan perbedaan dari genre-genre tersebut. Hal tersebut mengakibatkan banyaknya penikmat film yang susah untuk membedakan genre film sehingga pesan pada film tak sepenuhnya dapat tersampaikan kepada penikmat film. Oleh sebab itu melakukan klasifikasi pada film berdasarkan sinopsis film dirasa dapat menjadi salah satu solusi untuk masalah tersebut. Pengklasifikasian pada sinopsis film akan membantu dalam mengelompokan film dengan genre yang sesuai. Proses klasifikasi genre film berdasarkan sinopsis dimulai dengan melakukan preprocessing, kemudian pembobotan term hingga klasifikasi dengan metode Improved K-NN. Berdasarkan implementasi serta pengujian yang dilakukan pada penelitian Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-NN yang mana menggunakan 250 dokumen sebagai data latih dan 50 dokumen sebagai data uji didapatkan hasil terbaik yakni precision = 1, recall = 0,88, f-measure = 0,936170213, serta tingkat akurasi sebesar 88%. Serta dilakukan perbandingan dengan K-NN, terbukti bahwa pengklasifikasian menggunakan metode Improved K-NN lebih baik dibandingkan dengan metode K-NN.

Kata kunci : Text Mining, Klasifikasi, Film, Sinopsis, Improved K-Nearest Neighbor.

Abstract

Movie is audio visual communication media, which imply the message the movie creator wants to convey. Movie has several genres namely romantic, horror, thriller, comedy, fantasy and so on. Not a few movie connoisseurs are still confused about the differences in these genres. This resulted in many movie lovers who were difficult to distinguish the genre of movie so that the message in the movie could not be fully conveyed to the audience of the movie. Therefore, the classification of movies based on the synopsis of the movie can be one of the solutions to the problem. Classification in the movie synopsis will help in grouping movies with the appropriate genre. The genre genre classification process based on the synopsis begins with preprocessing, then weighting the term to classification with the Improved K-NN method. Based on the implementation and testing conducted in the movie classification research based on the synopsis using Improved K-NN which uses 250 documents as training data and 50 documents as the test data the best results are precision = 1, recall = 0.88, f-measure = 0.936170213, and an accuracy rate of 88%. As well as comparison with K-NN, it was proven that classification using the Improved K-NN method was better than the K-NN method.

Keywords : *Text Mining, Classification, Movie, Synopsis, Improved K-Nearest Neighbor*.

1. PENDAHULUAN

Film merupakan salah satu media untuk berkomunikasi yang memiliki sifat audio visual dimana tersirat pesan yang ingin disampaikan oleh pencipta film. Film memiliki beberapa genre film yang beragam yakni romantis, horor, thriller, komedi, fantasi dan lain sebagainya. Dari banyaknya genre yang disediakan oleh film, tidak sedikit penikmat film yang masih bingung membedakan genre-genre film tersebut. Hal tersebut mengakibatkan banyaknya penikmat film susah membedakan genre film sehingga pesan yang ingin disampaikan oleh film tak sepenuhnya dapat

tersampaikan. Oleh karena itu melakukan klasifikasi pada film dirasa penting untuk mempermudah penonton dalam memilih genre yang tepat dan sesuai dengan yang diinginkan.

Klasifikasi teks adalah melakukan pengelompokan pada dokumen atau teks ke dalam kelas-kelas yang memiliki karateristik yang sama atau mirip. Klasifikasi memungkinkan untuk mengelompokan film ke dalam kelas yang sesuai berdasarkan kategorinya. Tidak sedikit penikmat film yang masih bingung membedakan atau menentukan genre film yang sesuai dengan yang diinginkan, serta agar pesan pada film dapat ditujukan dan disampaikan dengan tepat maka melakukan klasifikasi pada sinopsis film dirasa menjadi solusi yang tepat untuk masalah tersebut. Metode Improved K-Nearest Neighbor (K-NN) dapat digunakan untuk melakukan pengklasifikasian. Improved K-NN menggunakan nilai k yang berbeda pada tiap kategori sesuai dengan banyaknya data latih (Puspitasari et all, 2017).

Puspitasari dkk. (2017) pada penelitiannya melakukan pembahasan mengenai klasifikasi yang dilakukan pada dokumen tumbuhan obat menggunakan metode Improved K-NN dimana didapatkan perolehan F1-measure sebanyak 70,99%, dan dari pengujian data latih ditemukan bahwa semakin besar jumlah data latih maka nilai akurasi akan semakin tinggi, serta diperoleh F1-measure untuk data latih seimbang sebesar 1,9% lebih baik dari data latih tidak seimbang. Selain itu penelitian lainnya ialah penelitian yang dilakukan oleh Megantara dkk. (2010) yang membahas tentang klasifikasi teks dengan menggunakan Improved K-NN, dari hasil penelitian terbukti bahwa metode Improved K-NN mencapai hasil yang lebih baik dibandikan metode K-NN dalam berbagai kondisi, dan dilihat dari standar deviasi yang didapatkan metode Improved K-NN memiliki kestabilan yang lebih baik dari K-NN. Pada penelitian yang dilakukan oleh Nathania dkk. (2017) yang membahas tentang melakukan klasifikasi spam pada twitter menggunakan metode Improved K-NN, didapatkan hasil rata-rata nilai Precision = 0,8946 dan Recall = 0,9405 serta F-Measure = 0,9155. Hasil akurasi diperoleh ialah 89,57%.

Berdasarkan permasalahan tersebut, penulis memberikan usul untuk melakukan pengklasifikasin pada sinopsis film ke lima genre film yakni aksi, horor, romantis, thriller, dan keluarga dengan menggunakan metode Improved K-NN dengan harapan mendapatkan hasil yang maksimal.

2. KAJIAN PUSTAKA

2.1 *Text Mining*

Proses untuk melakukan penambangan data dari dokumen atau data-data yang tidak terstruktur. Text mining berusaha untuk menghasilkan informasi yang tersirat secara implisit atas informasi yang dengan otomatis diekstrak dari dokumen (Feldman., 2007).

2.2 *Text Preprocessing*

Pada tahap ini data-data yang belum terstruktur dibersihkan agar menjadi lebih terstruktur. Terdapat beberapa proses untuk melakukan *preprocessing* yakni *cleansing*, *case folding*, tokenisasi, *filtering*, dan *stemming*.

1. *Cleansing*
   Membersihkan *text* dari karakter-karakter yang tidak perlu serta menghapus *link* pada dokumen.
2. *Case Folding*
   Mengubah semua huruf pada data atau dokumen sehingga menjadi huruf kecil.
3. Tokenisasi
   Melakukan pemotongan pada string-string yang terdapat dalam dokumen.
4. *Filtering*
   Mengambil kata inti dari dokumen, serta menghilangkan kata yang tidak diperlukan.
5. *Stemming*
   Menghilangkan imbuhan pada suatu kata sehingga hanya tersisa kata akarnya saja.

2.3 Pembobotan

1. *Term Frequency* (TF) dan Pembobotan TF (Wtf)

frekuensi atau tingkat kemunculan *term* pada suatu dokumen, frekuensi untuk setiap *term* bervariasi oleh karena itu frekuensi kemunculan *term* menjadi atribut penting untuk membedakan dokumen satu sama lain (Xia dan Chai, 2011) sedangkan Wtf ialah suatu proses untuk melakukan perhitungan bobot untuk setiap *term* (kata). Berikut ialah persamaan untuk menentukan nilai TF dan Wtf (Manning, Raghavan dan Schutze, 2009).

$$W f_{t,d} = \begin{cases} 1 + log\ tf_{t,d} & tf_{t,d} \geq 1 \\ 0 & otherwise \end{cases} \quad (1)$$

Keterangan :

- $W_{tf_{t,d}}$: Hasil dari pembobotan $tf_{t,d}$

- $tf_{t,d}$: Frekuensi muncul *t* pada dokumen *d*

2. *Document Frequency* (DFt) dan *Inverse Document Frequency* (IDFt)

Document Frequency (DFt) ialah banyaknya dokumen yang terdapat *term* t, dan *Inverse Document Frequency* ialah jumlah dokumen yang terdapat *term* yang dicari pada seluruh dokumen yang ada. Berikut persamaan untuk menghitung IDF yang diusulkan oleh Jones (1972) (Manning, Raghavan dan Schutze, 2009).

$$idf_t = log \frac{N}{df_t} \quad (2)$$

Keterangan :

- idf_t : Hasil invers df_t
- df_t : Banyaknya dokumen yang terdapat *t*
- N : Jumlah seluruh dokumen

3. Pembobotan TF-IDF (Wt,d)

Proses untuk melakukan penggabungan bobot pada tiap term dalam setiap dokumen. Untuk menghitung pembobotan. Proses ini dapat dilakukan dengan mengkalikan TF dan IDFt, untuk menghitung nilai Wt,d dapat menggunakan persamaan berikut (Manning, Raghavan dan Schutze, 2009) :

$$W_{t,d} = W_{tf_{t,d}} * idf_t \quad (3)$$

4. Normalisasi

Normalisasi dilakukan agar dapat lebih mudah melakukan perhitungan nilai *cosine similarity*. Persamaan berikut dapat digunakan untuk melakukan normalisasi (Nathania dkk, 2017) :

$$w_{t,d} = \frac{w_{t,d}}{\sqrt{\sum_{t=1}^{n} {w_{t,d}}^2}} \quad (4)$$

5. *Cosine Similarity*

Metode untuk melakukan perhitungan tingkat kemiripan antar objek (Bagaskoro dkk, 2017). Berikut merupakan persamaan *cosine similarity* setelah melewati tahap normalisasi (Nathania dkk, 2017) :

$$CosSim\left(d_j, q\right) = \overrightarrow{d_j}.\vec{q} = \sum_{i=0}^{t} \left(w_{ij}.w_{iq}\right) \quad (5)$$

Keterangan :

- d_j = dokumen latih
- q = dokumen uji
- w_{ij} = hasil $W_{t,d}$ dokumen latih
- w_{iq} = hasil $W_{t,d}$ dokumen uji

2.4 Improved K-NN

Metode pengembangan dari metode K-NN, dimana perbedaannya terdapat pada penentuan nilai *k*. Pada Improved K-NN dilakukan modifikasi dalam menentukan nilai *k*, yang mana setiap kategorinya memiliki nilai *k* yang berbeda sesuai dengan banyak atau sedikinya data latih pada setiap kelas atau kategori, sehingga saat nilai *k* semakin tinggi tidak akan mempengaruhi kategori dengan jumlah data latih yang besar (Herdiawan, 2015).

Setelah menghitung nilai *cosine similarity* maka hasil perhitungannya akan diurutkan secara menurun untuk setiap kategori. Setelah itu dilakukan penentuan nilai *k*, selanjutnya akan dilakukan perhitungan untuk mendapatkan nilai *k* baru (*n*), menentukan nilai *k* baru (*n*) dapat dihitung menggunakan persaman berikut (Herdiawan, 2015) :

$$n = \frac{k * N(c_m)}{Maks[N(c_m)|j=1..N_c]} \quad (6)$$

Keterangan :

- n = nilai *k* baru
- k = nilai *k* awal
- $N(c_m)$ = banyak data latih pada kategori m
- $Maks[N(c_m)|j = 1..N_c]$ = banyak data latih terbanyak pada seluruh kategori

Selanjutnya menghitung peluang data uji *X* termasuk dengan dokumen latih d_j sebanyak nilai *n* tetangga untuk tiap kategori dokumen *X* pada dokumen latih d_j sebanyak nilai *n* tetangga untuk *training set*. Persamaan berikut digunakan untuk menghitung peluang dokumen uji *X* pada kategori *m* (Baoli, Shiwen dan Qin, 2003) :

$$p(x, c_m) = argMaks_m = \frac{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j) y(d_j, c_m)}{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j)} \quad (7)$$

Keterangan :

- $p(x, c_m)$: probabilitas data X anggota c_m
- $sim(x, d_j)$: kemiripan antara data X dengan data latih d_j
- $top\ n\ k\ NN$: nilai n terbaik tetangga
- $y(d_j, c_m)$: fungsi atribut yang memenuhi dari salah satu kategori, apabila dokumen latih $d_j$masuk dalam kategori c_m maka akan bernilai 1, dan sebaliknya jika tidak maka akan bernilai 0.

Setelah menghitung peluang pada dokumen uji *X* pada kategori *m* maka akan

dilakukan perbandingan dari hasil peluang pada setiap kategori, nilai peluang terbesar akan menjadi acuan untuk hasil kategori data uji.

2.5 *Confusion Matrix*

Salah satu alat yang bisa membantu dalam melakukan pengujian atau menganalisis hasil klasifikasi. Evaluasi dilakukan dengan menggunakan tabel *confusion matrix* untuk melakukan perbandingan antara kategori aktual dan kategori prediksi (Manning, Raghava dan Schutze, 2009). Tabel 1 menunjukan tabel *confusion matrix* (Ting K.M, 2017).

Tabel 1 *Confusion Matrix*

Actual Class	*Assigned Class*	
	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

Dari Tabel 1 dapat dilihat bahwa terdapat kategori dan prediksi yaitu TP, FP, FN dan TN. Dimana TP, FP, FN, dan TN memiliki arti sebagai berikut (Puspitasari, 2017) :

- TP (*True Positive*) menunjukan banyak dokumen uji masuk kategori x, dan dokumen benar kategori x.
- FP (*False Positive*) menunjukan banyak dokumen uji yang bukan kategori x, dan dokumen tersebut seharusnya masuk kategori x.
- FN (*False Negative*) menunjukan banyak dokumen uji masuk kategori x, dan seharusnya bukan kategori x.
- TN (*True Negative*) menunjukan banyak dokumen uji tidak masuk kategori x, dan memang bukan kategori x.

2.6 *Presicion, Recall,* dan *F-Measure*

1. *Precision*

Nilai keakuratan hasil klasifikasi seluruh data oleh sistem, nilai *precision* dapat dihitung dengan persamaan berikut (Puspitasari, 2017) :

$$Precision = TP/(TP + FP) \quad (8)$$

2. *Recall*

Tingkat kesuksesan sistem mengenali suatu kategori, *recall* dapat dihitung dengan persamaan berikut (Puspitasari, 2017) :

$$Recall = TP/(TP + FN) \quad (9)$$

3. *F-Measure*

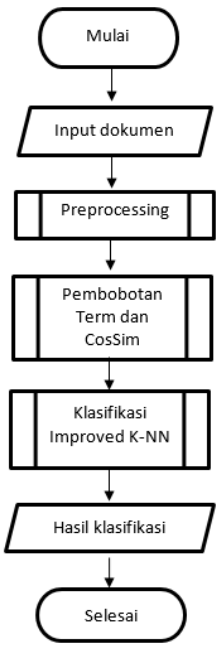
Gambaran dalam pengaruh *relative* antara *precision* dan *recall*. Dalam menentukan F1-*Measure* dapat menggunakan persamaan berikut (Puspitasari, 2017) :

$$F - Measure = (2 * P) * R/(P + R) \quad (10)$$

3. METODELOGI PENELITIAN

Terdapat beberapa tahapan yang dilalui dalam melakukan penelitian ini, yakni diawali dengan mencari literatur yang sesuai, mengumpulkan data, melakukan analisis kebutuhan, merancang sistem, implementasi sistem, melakukan pengujian, menganalisis serta penarikan kesimpulan.

Tujuan sistem yang akan dikembangkan ialah agar dapat mengklasifikasikan film berdasarkan sinopsinya ke dalam beberapa kategori. Proses untuk mengklasifikasikan tersebut dimulai dengan memberikan input berupa dokumen, kemudian tahap selanjutnya ialah preprocessing, yakni proses yang berguna untuk memperbaiki atau menyiapkan dokumen yang di inputkan. Lalu dilakukan pembobotan dan perhitungan cosine similarity, setelah itu dilakukan proses klasifikasi menggunakan metode Improved K-NN. Setelah semua tahap dilakukan maka akan didapatkan hasil kategori atas data uji, dimana nilai yang terbesar dari hasil perhitungan probabilitas merupakan kategori untuk data uji. Gambar 1 menunjukan diagram alir dari sistem.

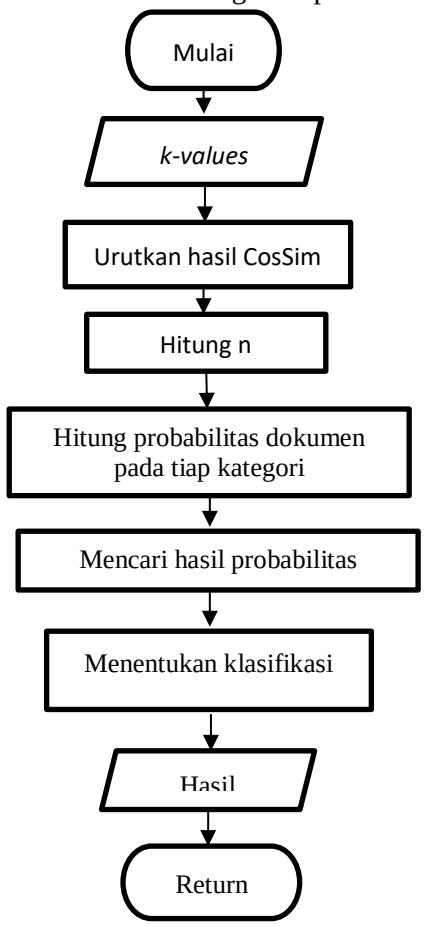


Gambar 1. Diagram Alir Sistem

4. PERANCANGAN DAN IMPLEMENTASI

Setelah melewati preprocessing, pembobotan, perhitungan cosine similarity dan normalisasi, maka selanjutnya akan dilakukan pengklasifikasian menggunakan metode Improved K-NN.

Proses klasifikas menggunakan metode Improved K-NN dimulai dengan menentukan nilai k awal, setelah itu hasil dai perhitungan cosine similarity yang telah di normalisasi di urutkan secara menurun. Selanjutnya dilakukan perhitungan untuk mendapatkan nilai k baru (n) kemudian setelah mendapatkan nilai k baru (n) maka dilakukan perhitungan probabilitas dokumen pada setiap kategori, kategori dengan probabilitas tertinggi menjadi kategori untuk data uji. Setelah semua tahap dilakukan maka akan didapatkan hasil berupa kategori untuk data uji. Gambar 2 menunjukkan diagram alir dari proses klasifikasi dengan Improved K-NN.



Gambar 2. Diagram Alir Klasifikasi dengan Improved K-NN

5. PENGUJIAN DAN ANALISIS

5.1 *Precision, Recall, F-Measure* dan Akurasi

Untuk mengetahui pengaruh dan banyak dokumen latih serta *k-values* pada keberhasilan sistem, maka dilakukan tahap pengujian menggunakan skenario-skenario pengujian. Jumlah data latih serta *k-values* awal pada setiap skenario berbeda-beda yang mana data uji untuk semua skenario sebanyak 50 dokumen yang menjadi data uji. Tabel 2 menunjukan skenario pengujian yang dibangun.

Tabel 2. Skenario Pengujian

Skenario	Data Latih						Data Uji					
	K1	K2	K3	K4	K5	Jumlah	K1	K2	K3	K4	K5	Jumlah
1	20	25	15	10	30	100	10	10	10	10	10	50
2	25	30	35	40	20	150	10	10	10	10	10	50
3	35	20	50	40	30	175	10	10	10	10	10	50
4	40	35	45	30	50	200	10	10	10	10	10	50
5	50	25	40	40	45	200	10	10	10	10	10	50

1. Skenario 1

Tabel 3. Skenario Pengujian 1

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	0	0	0	0	0	1	0,5	0,66666667	50%
4	1	2	1	1	2	1	0,56	0,71794872	56%
6	3	3	2	1	4	1	0,58	0,73417722	58%
8	4	5	3	2	6	1	0,58	0,73417722	58%
10	5	7	4	3	8	1	0,54	0,7012987	54%
15	7	8	5	3	10	1	0,54	0,7012987	54%
20	10	13	8	5	15	1	0,6	0,75	60%
25	13	17	10	7	20	1	0,62	0,7654321	62%
30	17	21	13	8	25	1	0,6	0,75	60%
35	20	25	15	10	30	1	0,58	0,73417722	58%
40	23	29	18	12	35	1	0,54	0,7012987	54%
45	27	33	20	13	40	1	0,52	0,68421053	52%
50	30	38	23	15	45	1	0,5	0,66666667	50%
75	33	42	25	17	50	1	0,52	0,68421053	52%
100	50	63	38	25	75	1	0,5	0,66666667	50%

Pada Tabel 3 menunjukan hasil dari setiap nilai *precision*, *recall*, *f-measure* dan akurasi pada pengujian dengan skenario 1. Dilakukan perhitungan kembali dari *k-values* awal sehingga menjadi nilai *n*, untuk tiap kategori yang dihitung dengan persamaan (6). Skenario 1 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal = 20 dan 25 yakni 0,7654321 dan terendah ada pada *k-values* awal = 2, 50 dan 100 yakni 0,666666667.

2. Skenario 2

Tabel 4. Skenario Pengujian 2

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	1	2	2	2	1	1	0,6	0,75	60%
4	3	3	4	4	2	1	0,72	0,8372093	72%
6	4	5	5	6	3	1	0,72	0,8372093	72%
8	5	6	7	8	4	1	0,7	0,82352941	70%
10	6	8	9	10	5	1	0,7	0,82352941	70%
15	9	11	13	15	8	1	0,76	0,86363636	76%
20	13	15	18	20	10	1	0,78	0,87640449	78%
25	16	19	22	25	13	1	0,72	0,8372093	72%
30	19	23	26	30	15	1	0,6	0,75	60%
35	22	26	31	35	18	1	0,6	0,75	60%
40	25	30	35	40	20	1	0,56	0,71794872	56%
45	28	34	39	45	23	1	0,64	0,7804878	64%
50	31	38	44	50	25	1	0,58	0,73417722	58%
75	47	56	66	75	38	1	0,58	0,73417722	58%
100	63	75	88	100	50	1	0,58	0,73417722	58%

Tabel 4 menunjukan hasil dari nilai precision, recall, f-measure dan akurasi pada pengujian menggunakan skenario 2. Dilakukan perhitungan kembali dari k-values awal untuk mendapatkan nilai n (k-values baru) pada tiap kategori yang dihitung dengan persamaan (6). Skenario 2 menunjukan, nilai f-measure tertinggi ada pada k-values awal = 20 yakni 0,876404494 dan terendah ada pada k-values awal = 40 yakni 0,717948718.

3. Skenario 3

Tabel 5. Skenario Pengujian 3

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	1	1	2	2	1	1	0,58	0,73417722	58%
4	3	2	4	3	2	1	0,66	0,79518072	66%
6	4	2	6	5	4	1	0,7	0,82352941	70%
8	6	3	8	6	5	1	0,72	0,8372093	72%
10	7	4	10	8	6	1	0,66	0,79518072	66%
15	11	6	15	12	9	1	0,7	0,82352941	70%
20	14	8	20	16	12	1	0,68	0,80952381	68%
25	18	10	25	20	15	1	0,7	0,82352941	70%
30	21	12	30	24	18	1	0,66	0,79518072	66%
35	25	14	35	28	21	1	0,72	0,8372093	72%
40	28	16	40	32	24	1	0,7	0,82352941	70%
45	32	18	45	36	27	1	0,68	0,80952381	68%
50	35	20	50	40	30	1	0,72	0,8372093	72%
75	53	30	75	60	45	1	0,74	0,85057471	74%
100	70	40	100	80	60	1	0,7	0,82352941	70%

Tabel 5 menunjukan hasil setiap nilai *precision*, *recall*, *f-measure* dan akurasi dari pengujian dengan skenario 3. Dilakukan perhitungan kembali dari *k-values* awal sehingga mendapatkan nilai *n* (*k-values* baru) pada setiap kategori yang dihitung dengan persamaan (6). Skenario 3 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal = 75 yakni 0,850574713 dan terendah ada pada *k-values* awal = 2 yakni 0,734177215.

4. Skenario 4

Tabel 6. Skenario Pengujian 4

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	2	1	2	2	2	1	0,56	0,71794872	56%
4	4	2	3	3	4	1	0,66	0,79518072	66%
6	6	3	5	5	5	1	0,6	0,75	60%
8	8	4	6	6	7	1	0,68	0,80952381	68%
10	10	5	8	8	9	1	0,72	0,8372093	72%
15	15	8	12	12	14	1	0,74	0,85057471	74%
20	20	10	16	16	18	1	0,74	0,85057471	74%
25	25	13	20	20	23	1	0,74	0,85057471	74%
30	30	15	24	24	27	1	0,76	0,86363636	76%
35	35	18	28	28	32	1	0,76	0,86363636	76%
40	40	20	32	32	36	1	0,76	0,86363636	76%
45	45	23	36	36	41	1	0,74	0,85057471	74%
50	50	25	40	40	45	1	0,76	0,86363636	76%
75	75	38	60	60	68	1	0,76	0,86363636	76%
100	100	50	80	80	90	1	0,76	0,86363636	76%

Tabel 6 menunjukan hasil dari nilai *precision*, *recall*, *f-measure* dan akurasi dari pengujian dengan skenario 4. Dilakukan perhitungan kembali dari *k-values* awal untuk mendapatkan nilai *n* (*k-values* baru) pada tiap kategori yang dihitung dengan persamaan (6). Skenario 4 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal = 30, 35, 40, 50, 75 dan 100 yakni 0,863636364 dan terendah ada pada *k-values* awal = 2 yakni 0,717948718.

5. Skenario 5

Berikut merupakan hasil pengujian dari skenario 5, dimana menggunakan data latih sebanyak 200 dokumen dan 50 data uji. Setiap kategori memiliki jumlah data latih yang berbeda.

Tabel 7. Skenario Pengujian 5

k-values	n (*k-values* Baru)					Precision	Recall	F-Measure	Akurasi
	K1	K2	K3	K4	K5				
2	2	1	2	1	2	1	0,54	0,701298701	54%
4	3	3	4	2	4	1	0,7	0,823529412	70%
6	5	4	5	4	6	1	0,58	0,734177215	58%
8	6	6	7	5	8	1	0,66	0,795180723	66%
10	8	7	9	6	10	1	0,68	0,80952381	68%
15	12	11	14	9	15	1	0,66	0,795180723	66%
20	16	14	18	12	20	1	0,7	0,823529412	70%
25	20	18	23	15	25	1	0,76	0,863636364	76%
30	24	21	27	18	30	1	0,78	0,876404494	78%
35	28	25	32	21	35	1	0,76	0,863636364	76%
40	32	28	36	24	40	1	0,88	0,936170213	88%
45	36	32	41	27	45	1	0,88	0,936170213	88%
50	40	35	45	30	50	1	0,86	0,924731183	86%
75	60	53	68	45	75	1	0,86	0,924731183	86%
100	80	70	90	60	100	1	0,86	0,924731183	86%

Hasil dari masing-masing nilai *precision*, *recall*, *f-measure* dan akurasi dari pengujian dengan menggunakan skenario 5. Dilakukan perhitungan kembali dari *k-values* awal sehingga menjadi n (*k-values* baru) untuk setiap kategori yang dihitung menggunakan persamaan (6). Skenario 5 menunjukan, nilai *f-measure* tertinggi ada pada *k-values* awal dengan nilai 40 dan 45 yakni sebesar 0,936170213 dan terendah ada pada *k-values* awal dengan nilai 2 yakni sebesar 0,7012987.

Dari pengujian yang dilakukan dengan 5 skenario, hasil terbaik didapatkan pada skenario 5. Pada skenario 5 nilai terbaik berada pada *k-values* 40 dan 45 yang menunjukan nilai *f-measure* sebesar 0,936170213 dan tingkat akurasi sebesar 88%

5.2 PERBANDINGAN DENGAN K-NN

Data latih pada skenario 5 digunakan untuk melakukan perbandingan hasil pengujian dengan metode K-NN pada skenario pengujian dengan metode K-NN. Skenario 5 dipilih karena hasil pengujiannya memiliki hasil yang terbaik dibandingkan dengan skenario lainnya yakni menggunakan 200 dokumen yang mana untuk kategori romantis berjumlah 40, untuk kategori keluarga berjumlah 35, untuk kategori

aksi berjumlah 45, untuk kategori horor berjumlah 30, dan untuk kategori *thriller* berjumlah 50, dengan menggunakan 50 data uji. Tabel 8 ditunjukan hasil nilai dari *precision, recall, f-measure* dan akurasi yang dihasilkan.

Tabel 8. Skenario Pengujian KNN

K-values	*Precision*	*Recall*	*F-Measure*	Akurasi
2	1	0,86	0,924731	86%
4	1	0,62	0,765432	62%
6	1	0,62	0,765432	62%
8	1	0,64	0,780488	64%
10	1	0,6	0,75	60%
15	1	0,66	0,795181	66%
20	1	0,6	0,75	60%
25	1	0,56	0,717949	56%
30	1	0,56	0,717949	56%
35	1	0,54	0,701299	54%
40	1	0,58	0,734177	58%
45	1	0,52	0,684211	52%
50	1	0,5	0,666667	50%
75	1	0,54	0,701299	54%
100	1	0,28	0,4375	28%

Hasil pengujian menunjukan bahwa pada nilai *k*=2 mengalami peningkatan nilai *f-measure* sedangkan pada nilai *k*=100 nilai *f-measure* mengalami penurunan. Hal ini menunjukan bahwa pada nilai *k* yang rendah maka nilai *f-measure* akan semakin tinggi sedangkan semakin besar nilai *k* maka nilai *f-measure* akan semakin rendah, ini membuktikan bahwa hasil yang didapatkan dengan menggunakan metode K-NN memiliki tingkat akurasi yang rendah, serta penentuan nilai *k* merupakan hal yang perlu diperhatikan karena sangat berpengaruh pada tingkat akurasi.

Agar perbandingan hasil pengujian lebih jelas maka ditunjukan perbandingan antara hasil pengujian dari improved K-NN dengan menggunakan skenario 5 yang merupakan skenario terbaik dengan hasil pengujian dari K-NN. Pada Tabel 9 menunjukan perbandingan dari hasil pengujian antara improved K-NN dengan skenario 5 dan K-NN.

Tabel 9. Skenario Pengujian KNN

k-values	Improved K-NN				K-NN			
	Precision	*Recall*	*F-Measure*	Akurasi	*Precision*	*Recall*	*F-Measure*	Akurasi
2	1	0,54	0,701298701	54%	1	0,86	0,924731	86%
4	1	0,7	0,823529412	70%	1	0,62	0,765432	62%
6	1	0,58	0,734177215	58%	1	0,62	0,765432	62%
8	1	0,66	0,795180723	66%	1	0,64	0,780488	64%
10	1	0,68	0,80952381	68%	1	0,6	0,75	60%
15	1	0,66	0,795180723	66%	1	0,66	0,795181	66%
20	1	0,7	0,823529412	70%	1	0,6	0,75	60%
25	1	0,76	0,863636364	76%	1	0,56	0,717949	56%
30	1	0,78	0,876404494	78%	1	0,56	0,717949	56%
35	1	0,76	0,863636364	76%	1	0,54	0,701299	54%
40	1	0,88	0,936170213	88%	1	0,58	0,734177	58%
45	1	0,88	0,936170213	88%	1	0,52	0,684211	52%
50	1	0,86	0,924731183	86%	1	0,5	0,666667	50%
75	1	0,86	0,924731183	86%	1	0,54	0,701299	54%
100	1	0,86	0,924731183	86%	1	0,28	0,4375	28%

Hasil pengujian dari improved K-NN dengan menggunakan skenario 5 menunjukan, nilai f-measure tertinggi ada pada k-values awal dengan nilai 40 dan 45 yakni sebesar 0,936170213 dan terendah ada pada k-values awal dengan nilai 2 yakni sebesar 0,701298701. Sedangkan pada hasil pengujian dari K-NN dengan menggunakan skenario 5 menunjukan, nilai f-measure tertinggi ada pada k-values awal dengan nilai 2 yakni sebesar 0,924731 dan terendah ada pada k-values awal 100 yakni sebesar 0,4375.

Hal tersebut menunjukan bahwa, pada metode K-NN semakin kecil k-values awal maka semakin tinggi nilai f-measure dan akurasi yang didapatkan. Terbukti pada k-values awal 2 tingkat akurasi pada metode K-NN sebesar 86% dan pada k-values 100 tingkat akurasinya sebesar 28%. Sedangkan metode Improved K-NN pada nilai f-measure dan akurasi untuk k-values awal 2 hasil yang didapatkan merupakan hasil yang terendah, dimana tingkat akurasinya sebesar 54% dan pada k-values 100 tingkat akurasi meningkat menjadi 76%. Selain itu pada pengujian dengan menggunakan metode Improved K-NN tingkat akurasi terbesar berhasil didapatkan pada k-values 40 dan 45 yakni sebesar 86%. Pada metode Improved K-NN tingkat akurasi yang didapatkan untuk masing-masing k-values tidak ada yang dibawah 50% sedangkan pada metode K-NN terdapat akurasi dibawah 50%. Hal ini membuktikan bahwa hasil yang dihasilkan oleh metode Improved K-NN lebih baik dibandingkan hasil dari metode K-NN, serta metode Improved K-NN lebih stabil dibandingkan dengan metode K-NN.

Gambar 7. Grafik Perbandingan Akurasi Improved K-NN dan K-NN

Gambar 7 menunjukan grafik dari perbandingan hasil pengujian menggunakan Improved K-NN dan K-NN, garis berwarna biru mewakili Improved K-NN dan garis berwarna merah mewakili K-NN. Grafik

menunjukan bahwa pada nilai *k*=2 pada Improved K-NN tingkat akurasinya berada pada 54% sedangkan pada K-NN tingkat akurasinya berada pada 86%. Pada Improved K-NN untuk setiap nilai *k* tingkat akurasinya mengalami kenaikan sedangkan pada K-NN semakin tinggi nilai *k* semakin rendah pula akurasi yang didapatkan.

Pada Improved K-NN akurasi tertinggi berada pada nilai 88% pada nilai *k* = 40 dan *k* = 45, sedangkan pada K-NN akurasi tertinggi berada pada nilai 86 persen pada nilai *k* = 2. Selain itu untuk nilai akurasi terendah pada Improved K-NN berada pada nilai 54% pada nilai *k*=2, sedangkan pada K-NN nilai akurasi terendah berada pada nilai 28% pada nilai *k*=100. Hal ini membuktikan bahwa hasil yang didapatkan dengan menggunakan Improved K-NN lebih baik dibanding dengan K-NN, karena tingkat akurasi Improved K-NN tidak berada dibawah 50% sedangkan K-NN memiliki tingkat akurasi yang rendah yaitu berada dibawah 30%.

5.3 Analisis

Dari hasil pengujian yang telah dilakukan pada setiap skenario pengujian, diketahui bahwa beberapa faktor dapat memberikan pengaruh terhadap keakuratan dari hasil pengklasifikasian yang dilakukan menggunakan metode Improved K-NN. Berdasarkan evaluasi yang dilakukan dengan data uji sebanyak 50 dokumen, dapat diketahui bahwa semakin besar jumlah data latih yang digunakan maka semakin baik pula nilai *f-measure* yang dihasilkan. Skenario 1 hingga skenario 5 menunjukan peningkatan rata-rata nilai *f-measure*.

Rata-rata nilai *f-measure* yang paling rendah terdapat pada skenario 1, yang disebabkan karena data latih yang digunakan pada skenario 1 memiliki jumlah yang paling sedikit, serta perbandingan data latih untuk tiap kategori pada skenario 1 menggunakan data latih yang paling sedikit untuk setiap kategorinya.

Nilai *precision, recall, dan f-measure* yang paling rendah didapatkan apabila *k-values* awal yang digunakan terlalu kecil atau terlalu banyak seperti yang ditunjukan oleh hasil setiap skenario yang mengakibatkan terjadinya kesalahan pada hasil pengklasifikasian. Hal ini membuktikan bahwa diperlukan ketelitian dalam menentukan *k-values* awal yang terbaik sehingga dapat menghasilkan hasil kategori yang tepat.

6. KESIMPULAN

1. Metode Improved K-Nearest Neighbor dapat dimanfaatkan dalam proses pengklasifikasian film dengan masukan berupa sinopsis film. Dokumen berupa sinopsis film akan melewati beberapa proses yakni *preprocessing*, pembobotaan *term*, hingga perhitungan nilai *cosine similarity* pada data latih yang digunakan. Kemudian proses selanjutnya ialah dengan mengurutkan tingkat kemiripan, menentukan *k-values* yang baru hingga mendapatkan hasil klasifikasi berupa kategori terhadap dokumen.
2. Dari hasil pengujian didapatkan hasil terbaik yakni *precision* sebesar 1, *recall* sebesar 0,88, *f-measure* sebesar 0,936170213 dan akurasi sebesar 88%. Yang mana jumlah dokumen, perbandingan data latih serta *k-values* yang digunakan memiliki pengaruh atas baik atau tidak baiknya proses pengklasifikasian pada dokumen yang berupa sinopsis.
3. Dari hasil pengujian yang telah dilakukan maka dapat ditarik kesimpulan yakni metode Improved K-NN dapat menghasilkan hasil yang lebih baik dengan hasil akurasi sebesar 88%.

7. DAFTAR PUSTAKA

Sremanthy, J., & Balamurugan, P.S. (2012). An efficient text classification using knn and naive bayesian. International Journal on Computer Science and Engineering (IJCSE). Coimbatore, India.

Megantara, G., Kurniati, A.P., & Suryani, A.A., (2010). Klasifikasi teks dengan menggunakan improved k-nearest neighbor algorithm. Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung.

Puspitasari, A.A., Santoso, E., & Indriati. (2018). Klasifikasi dokumen tumbuhan obat menggunakan metode improved k-nearest neighbor. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 2, Oktober 2018, hlm. 3948-3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Nathania, D.Z., Indriati., & Bachtiar, F.A. (2018). Klasifikasi spam pada twitter menggunakan metode improved k-nearest neighbor. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 10, Oktober 2018, hlm. 3948-

3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Feldman, R., & Sanger, J. (2007). The text mining handbook advance approaches in analyzing unstructured data . CAMBRIDGE UNIVERSITY PRESS.

Wahyudi, D., Susyanto, T., & Nugroho, D. (2013). Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia. Jurnal Ilmiah SINUS STMIK Sinar Nusantara Surakarta. Program Studi Teknik Informatika, STMIK Nusantara Surakarta, Surakarta.

Xia, T., & Chai, Y. (2011). An improvement to tf-idf: term distribution based

term weight algorithm. Journal of Software, 6(3), pp.413–420.

Manning, C.D., Raghavan, P., & Schutze, H. (2009). An introduction to

information retrieval. Cambridge, England: Cambridge University

Press.

Bagaskoro, G.N., Fauzi, M.A., & Adikara, P.P. (2018). Penerapan klasifikasi tweets pada berita twitter menggunakan metode k-nearest neighbor dan query expansion berbasis distributional semantic. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 10, Oktober 2018, hlm. 3948-3956. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Zheng, W., Wang, H., Ma, L., & Wang, R. (2015). An improved k-nearest

neighbor classification algorithm using shared nearest neighbor

similarity. 26(10), pp.133–137.

Prayoga, F., Pinandito, A., & Perdana, R. (2017). Rancang bangun aplikasi deteksi spam twitter menggunakan metode naive bayes dan knn pada perangkat bergerak android. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 2, p. 554-564, agu. 2017. Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

Herdiawan. (2015). Analisis sentimen terhadap telkom indihome berdasarkan

opini publik menggunakan metode improved k-nearest neighbor.

Baoli, L., Shiwen, Y., & Qin, L. (2003). An improved k-nearest neighbor algorithm for text categorization. Reading, p.678.

Ting K.M. (2017). Confusion matrix. In: sammut c., webb g.i. (eds) encyclopedia of machine learning and data mining. Springer, Boston, MA.