Movilizer Application with Genre and Rating Classification Using NW-KNN Method

Viny Christanti M. *
Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia

Cindy Winata
Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia

Janson Hendryli
Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia

ABSTRACT

Information about movies can be easily seen in cyberspace. However, not all film sites present relevant and accurate information as examples of high rating films but have bad comments. In addition, there is a review that has not been accompanied by ratings and the genre is unknown. The classification of input data in the form of text will be processed and classified into the same or similar class using the Neighbor-Weighted K-Nearest Neighbor (NW-KNN) method. The NW-KNN method is able to classify well for data that is not evenly distributed by giving weights to each class in the system. The description text of the film will be classified into 10 classes with the number of training data as many as 1028, while the movie review text will be classified into 5 classes with the number of training data as many as 10032. The results of system testing indicate that the NW-KNN method produces an accuracy of 96.6% film genre and 86.85% to classify film reviews into movie ratings.

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Retrieval tasks and goals**; • **Sentiment analysis**;

KEYWORDS

Film Descriptions, Film Reviews, Genre, NW-KNN, Rating

ACM Reference Format:
Viny Christanti M., Cindy Winata, and Janson Hendryli. 2021. Movilizer Application with Genre and Rating Classification Using NW-KNN Method. In *2021 3rd Asia Pacific Information Technology Conference (APIT 2021), January 15–17, 2021, Bangkok, Thailand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3449365.3449371

1 INTRODUCTION

Nowdays, the existence of film is one of the entertainment media, especially teenager. Not only teenagers who like movies, both children and even the elderly like movies. Usually people depend review and rating to buy ticket or online streaming. IMDB is one of website that presents lists movie, TV program, and even biography of famous people in the film industry.

With review and rating in IMDB, people can use it as analysis for of estimated earnings, appearance of sequel movies, determining the number of movie screening schedules, etc. Nowdays IMDB is used as a reference for fans of movies and TV Series because of its accurate rating compared to other sites. People often visit to see their favorite movies from rating or even review. Sometimes if the rating of movie's in IMDB is poor, people don't want to spend their money to buy ticket at the cinema or even spend their time to streaming online.

However, sometimes between reviews and ratings are not balanced. This is because the user must manually press the rating button on the website. The imbalances between reviews and movie's rating will affect the image of the film itself. This will result in a bad rating for that film even the story is good. Because of that problem, a classification process is needed to produce a system that can automatically classify ratings based on reviews.

Movilizer is the answer to the problem above. Movilizer can classify rating automatically based on reviews. Not only that, Movilizer can do genre classifications automatically based on film descriptions. This can balance the input from the user and the output. A method is needed for classification, both for rating and genre classification. NW-KNN is a method for classifying rating from movie review and genre from movie descriptions. NW-KNN can classify unbalanced data by giving weights to each class [10]. The author proposes NW-KNN method to classify movie reviews and description as the promising result of [7].

2 RELATED WORKS

Movilizer website is largely driven by the rapidity of the movie industry in presenting films of various genres. Movie reviews and ratings have been shown to influence people's interest in watching movies. The research in review classifier has been driven by the advances in film industry especially in online movie review, such as the promising results of [3] who comparing several sentiment classification methods, [6] who used Naïve Bayes as the method of classifying review and [8] that used J.48 Algorithm. In [5], the author proposes Improved KNN method for classifying movie description. The author proposes NW-KNN method to classify movie reviews and description as the promising result of [7]. The work resembles the aim of this paper, which is building a website that can automatically classifying review to rating and description to genre, also promote new movie and knowing trending movies.

*viny@untar.ac.id

APIT 2021, January 15–17, 2021, Bangkok, Thailand

ACM ISBN 978-1-4503-8810-8/21/01…$15.00
https://doi.org/10.1145/3449365.3449371

2.1 Pre-Processing

The pre-processing stage is the process of preparing raw data before another process is carried out. In general, data pre-processing is done by eliminating inappropriate data or converting data into forms that are easier to process by the system. Pre-processing is very important in classifying sentiments, especially for social media which mostly contain informal or unstructured words or sentences and have large noise [4].

The pre-processing stage begins by changing uppercase letters to lowercase letters (Case Folding), negation word conversion (Convert Negation), doing word breakdown into a single word (Tokenizing), doing abbreviation checking (Abbreviation Checking), discarding non-essential words (Filtering), and search for basic words (Stemming).

2.2 Indexing

Indexing is done to speed up the search process so that the location of the searched word can be found quickly. The indexing process is one of the classification documents. Term Frequency - Inverse Document Frequency (TF-IDF) is one way to sort words by frequency in a document. In TF-IDF, TF is a way to count the number of times a word appears in a document. Whereas IDF is used to see the position of word spread on all documents. TF-IDF is obtained by multiplying TF with IDF. Word weight is greater if it often appears in documents and gets smaller if it appears in many documents. TF-IDF formula can be seen in equation number (1) [1].

$$w_{dt} = tf_{td} * idf_t \tag{1}$$

where w_{dt} is TF-IDF weight value for a term in a data, tf_{td} is frequency of occurrence of term t in data and idf_t is IDF value for term.

2.3 Classification

The NW-KNN method is a modification of the K-Nearest Neighbor (K-NN) method to overcome data that is not spread evenly. The NW-KNN method corrects K-NN weaknesses by giving weights for each class. The majority class will be given a small weight and the minority weld will be given greater weight. The class weighting can be seen in equation number (2) [4].

$$Weight_i = \frac{1}{\left(\frac{Num\left(C_i^d\right)}{Min\{Num\left(C_j^d\right)|j=1,\ \ldots,\ k*\}}\right)^{1/exponent}} \tag{2}$$

In second equation, it can be seen that $Weight_i$ is the weight of each class type, $Num(C_i^d)$ is the amount of training data d in class i, $Num(C_j^d)$ is the amount of training data d in class j and *exponent* is a number more than 1.

The weight of each class is used to calculate the similarity of the test data for each class. The results of the calculation of each similarity will be used as a reference to determine the class of the test data. The similarity calculation for the NW-KNN can be seen in equation number (3) [4].

$$score\left(q, c_i\right) = Weight_i x \sum_{d_j \in KNN(q)} sim\left(q, d_j\right) .\delta\left(d_j, c_i\right) \tag{3}$$

Table 1: Number of Word Lists

Word Lists	Number of Words
Word Abbreviation	1009
Stopword	774

Where $score(q, c_i)$ is similarity value, $Weight_i$ is weight of each class and $d_j \in KNN(q)$ is training data d_j that is located in the nearest collection of neighbors from the q test document. We know that $sim(q, d_j)$ is similarities between q test documents and training documents d_j. We can get value of δ classification of d_j data with reference from rules (4):

$$\delta\left(d_j, c_i\right) = \begin{cases} d_j \in c_i = 1 \\ d_j \notin c_i = 0 \end{cases} \tag{4}$$

2.4 Accuracy

System accuracy is needed to determine the level of performance of the system. Confusion Matrix is a tool for testing or analyzing classification results. Accuracy formula can be seen in the equation number (5) [9].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \tag{5}$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

3 EXPERIMENTAL SETTINGS

In this section, the experimental setting is described, including data and the implementation detail.

3.1 Data

Movie description data set consists of 1028 with 10 genres, i.e. action, adventure, comedy, crime, drama, historical, horror, musical, science fiction and war. All data is taken manually from https://filmindonesia.org and crawling from https://themoviedb.org. Data from TheMovieDb will be translate to Indonesia language then save to csv file. Same as movie description, movie review is taken by crawling data from https://youtube.com then save to csv file. Movie review data set consists of 10032 with 5 ratings, i.e. rating 1 until rating 5 with rating 5 is the highest rating and rating 1 is the lowest rating.

3.2 Testing Movie Description

Testing the description of the film is done by determining the value of K. The testing process is done by dividing the training data by 1028 by a ratio of 80:20, which is 80% of 1028 data (822 description data) will remain training data and 20% of 1028 (206 description data) will be validation data. The range of the selected K values is 2-20 and will be done 10 times for each K value. For each iteration, the training and testing data always the same to make it valid. K values that have the highest average accuracy will then be used as K (number of neighbors) in the classification system of film descriptions into the film genre. After cross-validation results, the most optimal K is 20.

Table 2: Amount of Test Data

Genre	Total
Action	21
Adventure	20
Comedy	21
Crime	20
Drama	20
History	20
Horror	21
Musical	21
Science Fiction	20
War	22
Total	206

From 1, there are 1009-word abbreviation to do abbreviation checking of all word in the document. Then there are 774 stopword words to eliminate non-essential words. All abbreviated words will be converted into standard words according to the abbreviation checking dictionary. Likewise, with Stopwords, all non-essential words in the Stopwords dictionary will be deleted. This will make input data more formal and structured.

From 2, there are 206 testing data that will be used for testing experiment in 3. From 3, the first test up to the fourth test uses training data of 1028 and the test data is 206. The first test uses all stages of pre-processing. But in testing two to four, one of the processes will be eliminated as in the second test the Stemming process is removed, the third test is not done the Stopword process and the fourth test does not do the Stemming and Stopword process. The fifth test uses training data of 822 and the test data is 206 but the test data is not in the training data.

3.3 Testing Movie Review

Testing the review of the film is done by determining the value of K first. The testing process is done by dividing the training data by 10032 by a ratio of 95:5, which is 95% of 10032 data will remain training data and 5% of 10032 will be validation data. The range of the selected K values is 2-20 and will be done 10 times for each K value. For each iteration, the training and testing data always the same to make it valid. After cross-validation results, the most optimal K is 2. The list of words used for pre-processing can be seen in 4.

From 4, there are 502 testing data that will be used for testing experiment in 5

Table 4: Amount of Test Data

Rating	Total
Rating 1	100
Rating 2	100
Rating 3	101
Rating 4	100
Rating 5	101
Total	502

Based on 5, the first test up to the fourth test uses training data of 10032 and test data of 502. The first test uses all stages of pre-processing. But in testing two to four, one of the processes will be eliminated as in testing both the Stemming process is omitted, the third test is not done the Stopword process and the fourth test does not do the process of checking the word abbreviation. The fifth test uses training data of 9530 and test data is 502 but the test data is not in the training data.

4 RESULTS

4.1 Movilizer User Interface

1 1 shows the user interface of Movilizer website. There are 4 main modules, i.e. Home, Add Movie, Edit Movie and Add Review. Module Home is the first module that users see when accessing the website. Add Movie is using to add new movie and Edit Movie is using to edit movie information. Then, Add Review is used for users adding a review of a movie.

From Figure. 1, we can see User Interface from Home Modul. It contains of some movie information, like best movie, new movie and top commented movie.

From Figure. 2, we can see User Interface from Movie Modul. It contains of list movie, like movie poster, movie title, movie genre and movie rating.

From Figure. 3, we can see User Interface from Movie Information Modul. It contains of movie information, like movie poster, movie trailer, movie title, movie genre, movie description, movie rating and movie reviews. In this module we can add new review if you login as user.

4.2 Movie Description Classification

Based on 3, the most effective test if found in the fourth test with accuracy value 96.6%. Summary of each test can be seen in the 6.

Table 3: Testing Movie Description

Testing Number	Training Data	Testing Data	Test Data in Training Data	Stopword	Stemming
1	1028	206			
2	1028	206			-
3	1028	206		-	
4	1028	206		-	-
5	822	206	-		

Table 5: Testing Movie Review

Testing Number	Training Data	Testing Data	Test Data in Training Data	Abbrevation Checking	Stopword	Stemming
1	10032	502				
2	10032	502				-
3	10032	502			-	
4	10032	502		-		
5	9530	502	-			

Table 6: Summary Movie Description Testing

Testing Number	Amount of Correct Classification	Accuracy (%)	Classification Time (minutes)
1	205	99.51	25
2	200	97.09	14
3	202	98.06	16
4	199	96.60	10
5	130	63.11	12

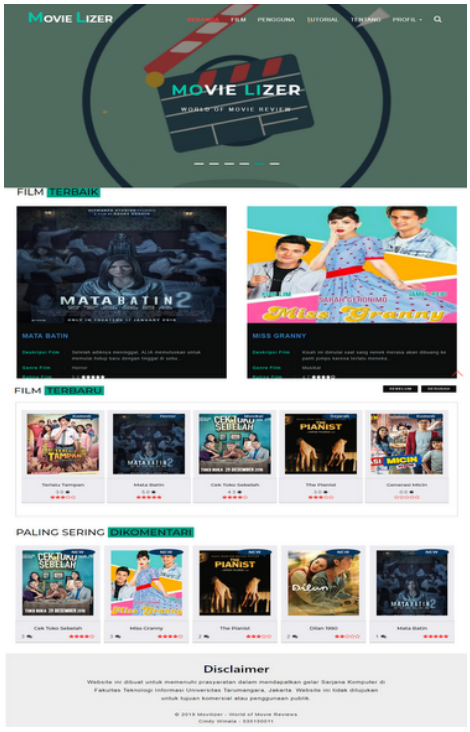

Figure 1: User Interface Home Modul

Based on the table above, the most effective test is testing fourth with an accuracy of 96.60%. In 7 can be seen the confusion matrix of fourth testing.

Based on 7, it can be seen that there is a class placement error with accuracy 96.60% which means that the system successfully classifies 199 out of 206 sentences. Genre Action has a lot of error with 3 of incorrect data predicted from a total of 21 data. So, it can

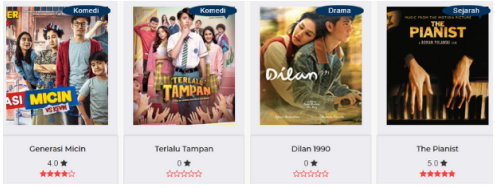

Figure 2: User Interface Movie Modul

be concluded that Stopword and Stemming have an effect on this test.

4.3 Movie Review Classification

Based on 5, the most effective test if found in the fourth test with accuracy value 96.6%. Summary of each test can be seen in the Table 8.

Based on the table above, the most effective test is testing first with an accuracy of 86.85%. In 9 can be seen the confusion matrix of first testing.

Based on 9, we can see that there is a class placement error with accuracy 0.8685 or 86.85% which means the system successfully classifies as many as 436 out of 502 sentences. Rating 5 has a lot of error with 27 of incorrect data predicted from a total of 101 data. So, it can be concluded that without the deletion of the abbreviation, Stopword and Stemming the system can classify the movie review data correctly.

5 DISCUSSIONS

The most effective percentage of film genre classification among the five tests is in fourth testing with an accuracy of 96.60% with training data of 1028 sentences and test data of 206 sentences at K = 20 and time for 10 minutes. Although there were still errors in the class placement of the genre, the decrease in accuracy from testing first until fourth was not significant. In addition, fourth testing time

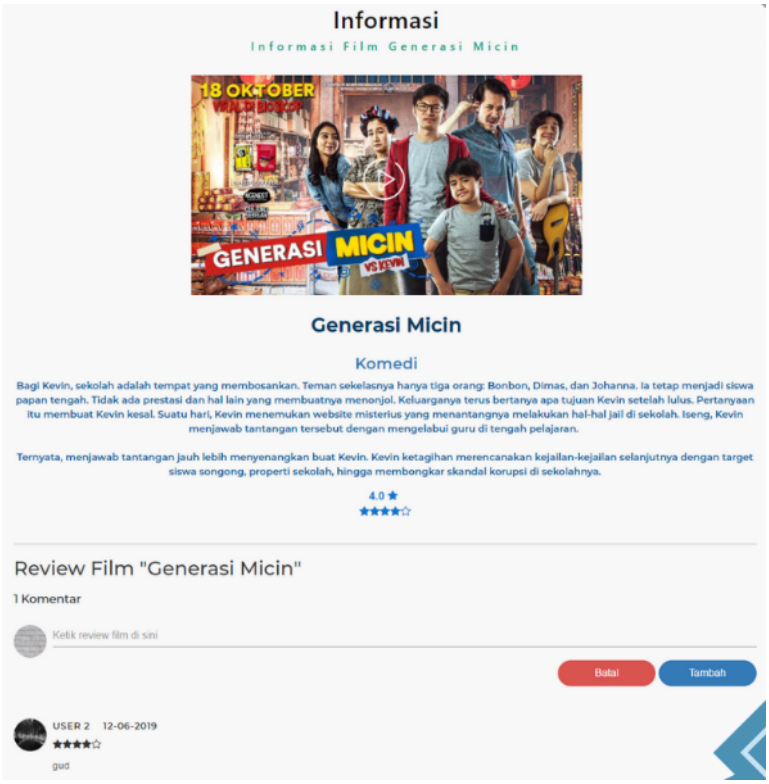

Figure 3: User Interface Movie Information Modul

Table 7: Confusion Matrix Movie Description

Predict Class												
Actually		G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Total
Class	G1	18	0	0	0	1	0	0	1	1	0	21
	G2	0	20	0	0	0	0	0	0	0	0	20
	G3	0	0	21	0	0	0	0	0	0	0	21
	G4	1	0	0	18	0	0	0	0	1	0	20
	G5	0	0	0	0	20	0	0	0	0	0	20
	G6	0	0	0	0	0	20	0	0	0	0	20
	G7	0	0	0	0	0	0	20	0	0	1	21
	G8	0	0	0	0	0	0	0	21	0	0	21
	G9	0	0	0	0	0	0	0	0	20	0	20
	G10	0	0	0	0	0	0	0	0	1	21	22
												206

Explanation:

G1	Action	G6	Musical
G2	Drama	G7	War
G3	Horror	G8	Adventure
G4	Crime	G9	Science Fiction
G5	Comedy	G10	Historical

is faster and produces a fairly high accuracy of 96.60%. For example, movie “Soekarno” genre actually is historical but system has been classified into war genre. This because historical and war genre’s word almost the same. The fifth case of movie review description testing’s accuracy below 66% because there are so many words that have no weight, so system cannot determine the classes of movie description.

Table 8: Summary Movie Review Testing

Testing Number	Amount of Correct Classification	Accuracy (%)	Classification Time (minutes)
1	436	86.85	55
2	401	79.88	45
3	383	76.29	40
4	384	76.49	42
5	121	24.10	50

Table 9: Confusion Matrix Movie Review

Predict Class							
Actually Class		1	2	3	4	5	Total
	1	100	0	0	0	0	100
	2	7	92	1	0	0	100
	3	6	3	88	4	0	101
	4	3	1	14	82	0	100
	5	3	6	5	13	74	101
							502

The highest percentage rating of film among the five tests is in first test with an accuracy of 86.85% with training data of 10032 sentences and test data as many as 502 sentences at K = 2 and time for 55 minutes. Although there were still errors in the placement, the decrease in accuracy from testing first until fifth was very far and significant. Although testing time first is longer than other tests, it can produce the highest accuracy, which is 86.85%. For example, sentence "Film ini bagus banget" has been classified as rating 4. This is because word "bagus" in rating 4 more than rating 5. So, system will classify into rating 4. Same as fifth case of movie review testing's accuracy below 25% because system cannot determine the rating classes of new word.

ACKNOWLEDGMENTS

The author would like to acknowledgement the support of Faculty of Information Technology, Universitas Tarumangara for this research.

REFERENCES

[1] Amrizal, Victor. "PENERAPAN METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN COSINE SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI UNTUK MENGETAHUI SYARAH HADITS BERBASIS WEB (STUDI KASUS: HADITS SHAHIH BUKHARI-MUSLIM)." JURNAL TEKNIK INFORMATIKA 11.2 (2018): 149-164.

[2] Indriati, Indriati, and Achmad Ridok. "Sentiment Analysis For Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn)." Journal of Environmental Engineering and Sustainable Technology 3.1 (2016): 23-32.

[3] Martha, M., *et al.* "Perbandingan Pengklasifikasi k-Nearest Neighbor dan Neighbor-Weighted k-Nearest Neighbor Pada Sistem Analisis Sentimen dengan Data Microblog." *FRONTIERS: JURNAL SAINS DAN TEKNOLOGI* 1.1 (2018).

[4] Mujilahwati, Siti. "Pre-Processing Text Mining Pada Data Twitter." Semin. Nas. Teknol. Inf. dan Komun 2016.Sentika (2016): 2089-9815.

[5] Muslimah, Nurul; Indriati; dan Wihandika, Randy Cahya. "Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbour (K-NN). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol.3, No.1, Januari 2019.

[6] Nugroho, Yusuf Sulistyo. "Prediksi Rating Film Menggunakan Metode Naïve Bayes." *Jurnal Teknik Elektro* 8.2 (2016): 60-63.

[7] Ridok, Achmad, and Retnani Latifah. "Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN." *Proceedings Konferensi Nasional Sistem dan Informatika (KNS&I)* (2015).

[8] Rintyarna, Bagus Setya. "Sentiment Analysis pada Movie Review dengan Pendekatan Klasifikasi dalam Algoritma J. 48." *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)* 1.2 (2017).

[9] Solichin, Achmad. Mengukur Kinerja Algoritma Klasifikasi dengan Confusion Matrix. https://achmatim.net/2017/03/19/mengukur-kinerja-algoritma-klasifikasi-dengan-confusion-matrix/, Accessed on 28 July 2019.

[10] Yudha, Bayu Laksana, Lailil Muflikhah, and Randy Cahya Wihandika. "Klasifikasi Risiko Hipertensi Menggunakan Metode Neighbor Weighted K-Nearest Neighbor (NWKNN)." Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2548 (2017): 964X.