Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods

Gisela Yunanda*, Dade Nurjanah, Selly Meliana

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]giselayunanda@students.telkomuniversity.ac.id, [2]dadenurjanah@telkomuniversity.ac.id,
[3]sellymeliana@telkomuniversity.ac.id
Correspondence Author Email: giselayunanda@student.telkomuniversity.ac.id
Submitted: **11/06/2022**; Accepted: **29/06/2022**; Published**: 30/06/2022**

Abstract−The rapidly growing information causes information overload, so news portals publish information massively. Readers need time to search and read more news, but the time relevance of news wears off quickly. A recommendation system is needed that can recommend news according to the preferences of readers. This study recommends news using the TF-IDF method. TF-IDF gives weight to each word in the news title, and then looks for similarity between stories using cosine similarity. To prove the accuracy of whether the system recommendation results were actually clicked by the reader, the recommendation results were matched with the reader's news history on the online news portal Microsoft News using a hit-rate. The hit-rate result in this study was 80.77%.

Keywords: Recommendation System; News; Microsoft News; TF-IDF; Cosine Similarity

1. INTRODUCTION

The development of information causes information overload, where news portals publish massive amounts of news every day [1]. Information overload causes the relevance of news time to end quickly and then it is replaced with more recent news [1]. This is inversely proportional to readers who need more time to find and read the news they are interested in [1]. Therefore we need a recommendation system that can help readers to find and choose news. A recommendation system is a system designed to assist users by suggesting information that may be useful to achieve user goals, such as suggesting reading articles or choosing certain products so that they are more effective in determining the desired product. The recommendation system filters the data using different algorithms and recommends the most relevant data [2].

An example of an application that implements a recommendation system in its application service is Spotify, in this application a recommendation algorithm is used to personalize songs that each user might like. In addition, there are online commercial websites such as YouTube that recommend viewing to users based on the user's past behavior history. When there are new videos available on YouTube, those videos will be compared to the user's interests based on the user's behavioral history. If the videos is of interest to the user, the system will recommend the videos to the user [2].

The recommendation system acquires user preferences, and uses them to create several types of user models. User preferences can be analyzed through the availability of user data, information about users, and their environment [3]. Feedback provided by users can be divided into two types, namely implicit feedback and explicit feedback. Implicit feedback for example from the user's transaction history or monitoring what items the user has viewed. While explicit feedback, for example by asking users to rate an item, or asking questions at the beginning of accessing the application. Likewise with a news recommendation system that recommends news to readers so that readers can find news quickly and avoid information overload. New news and topics appear and disappear quickly, while old news is no longer interesting. Recommendation systems usually recommend news to readers based on the most popular stories obtained from ratings given by other readers, or based on the similarity of news history between one reader and another. With a recommendation system, it makes it easy for readers to access information and saves time.

The first research is a laptop recommendation system using collaborative filtering and content-based filtering by Wijaya et al. (2018) which can provide laptop recommendations based on interests and needs [4]. Collaborative filtering method in this research is done by utilizing the opinions of other buyers on an item to predict items that may be liked by certain buyers. Meanwhile, content-based filtering utilizes item attributes and TF-IDF to provide recommendations. In this study, the measurement of similarity between one item and another uses cosine similarity based on the rating given by the user. The content-based method has a faster execution time than the collaborative filtering and mixed hybrid methods.

The second research is an exclusive pen product recommendation system using the collaborative filtering method by Putri et al. (2020). In implementing the method, the author uses TF-IDF to weight the content and customer attributes. The recommendation system generates recommendations that have similarities between content attributes and attributes owned by customers. The recommendation system has an accuracy of 96.5% which can be said to be in accordance with customer criteria [5].

Then the third research is an online news recommendation system using TF-IDF and cosine similarity by Indriani et al. (2019). This study aims to help readers who have difficulty choosing news on the Banjarmasin Radar news portal. The TF-IDF method is used to determine the number of words that appear in a document and determine the occurrence of words in the entire document. Meanwhile, cosine similarity is used to calculate the similarity value

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

between two news stories. The news recommendation system in this study resulted in a precision value of 76% of all news [6].

The fourth research is the sentence similarity test in the title of the Deli Husada Health Institute student's final project using the cosine similarity method and TF-IDF weighting by Mawanta et al. (2021) [7]. The results of the validation in this study found that 43% of the final project titles were said to not meet and 53% were said to be fulfilled. The fifth study is an Indian food recipe recommendation system by Chippa et al. (2022) using TF-IDF which is proposed on a mobile application to make it easier for users to find recipes using available ingredients [8]. This recommendation system is content based as it recommends recipes to people based on ingredients. TF-IDF is used to extract features and cosine similarity is used to measure the similarity between the ingredients owned by the user and the ingredients in the recipe that the system has. Based on the studies previously mentioned, we utilized TF-IDF and cosine similarity algorithm to help readers choose news stories they might be interested in. Therefore, this research is expected to be able to overcome information overload and minimize the time used by readers to select news. Our contribution in this research is that we first conduct a review of the research literature to design a recommendation system, analyze and design the system, search for datasets, then start implementing the recommendation system. after implementation, the system is tested and the results of the implementation are evaluated based on the availability of data in the dataset.

2. RESEARCH METHODOLOGY

2.1 System Design

In this study, a news recommendation system was built. In building a news recommendation system, there are several steps taken. First, the researcher prepared a dataset in the form of implicit user feedback in the form of a history of news clicks, as well as news titles. then the dataset is analyzed and processed at the preprocessing stage so that the data is structured and uniform. Furthermore, each reader's click history is divided into two with a composition of 70% as history used in system development, and 30% as history that is evaluated with the results of system recommendations. In the news title dataset, each title is assigned a TF-IDF weight and paired with 70% of the news click history, so that each news item in 70% of the news history has a TF-IDF weight. TF-IDF calculates the frequency of occurrence of each word in each word and the frequency of occurrence of words in all documents. After getting the TF-IDF weights, the news click history is divided into two, namely test data and train data, then the similarity value is calculated using cosine similarity. In this news recommendation system, cosine similarity is used to calculate the similarity of two news stories consisting of combined TF-IDF weights for news titles. The recommendation system process using TF-IDF and cosine similarity can be seen in the block diagram of Figure 1.

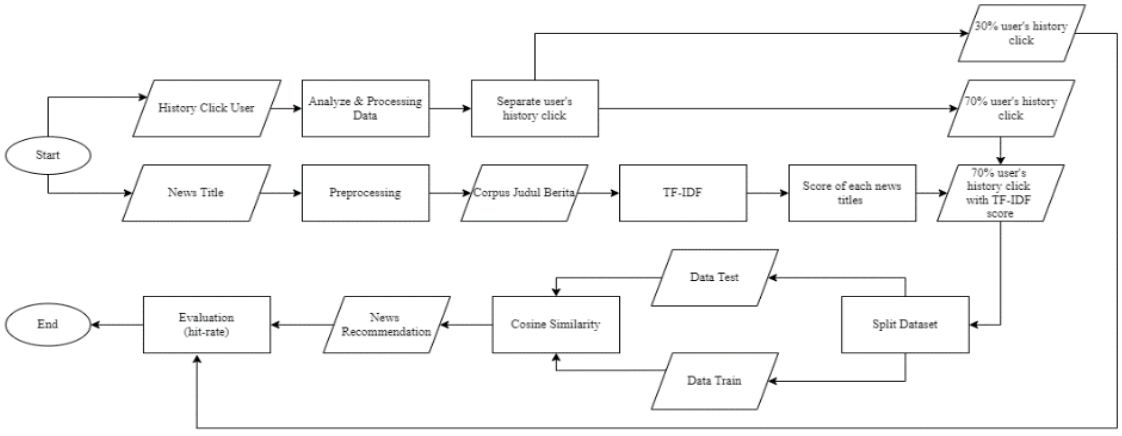

Figure 1. System Development Flow using the TF-IDF method and cosine similarity.

2.2 Dataset

In this research, we use Microsoft News Dataset (MIND) from the Kaggle website. We only uses 5,000 English news titles and 5,286 click history readers who have unique reader IDs and have at least clicked on more than 40 news headlines. The number of datasets used is limited due to limited storage space, memory speed, and the TF-IDF method which can only work on a small corpus.

Table 1. News Title Dataset

NewsID	Title
N61837	The Cost of Trump’s Aid Freeze in the Trenches of Ukraine’s War
N53526	I Was An NBA Wife. Here’s How It Affected My Mental Health
N38324	How To Get Rid of Skin Tags, According to a Dermatologist

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

Table 2. History Click User

UserID	History
U91836	['N13008', 'N13231', 'N13912', 'N14029', 'N14340', 'N14761', 'N14761', 'N1569', 'N16290', 'N16617', 'N17059', 'N19709', 'N21164', 'N21623', 'N22816', 'N22976', 'N23979', 'N2511', 'N25785', 'N2663', 'N29802', 'N29948', 'N30698', 'N30867', 'N31402', 'N3142', 'N31699', 'N31739', 'N32852', 'N33212', 'N33885', 'N3433', 'N34520', 'N35458', 'N37159', 'N37509', 'N41011', 'N43353', 'N44383', 'N4486', 'N46082', 'N47020', 'N48031', 'N48742', 'N51163', 'N54225', 'N54489', 'N54889', 'N56514', 'N58224', 'N58715', 'N59026', 'N59049', 'N59359', 'N6072', 'N6087', 'N61355', 'N61388', 'N62285', 'N6233', 'N64467','N8129', 'N8275', 'N9798']
U89744	['N10417', 'N11405', 'N12215', 'N12349', 'N13057', 'N13286', 'N13480', 'N14167', 'N16710', 'N18094', 'N18939', 'N22260', 'N22561', 'N23554', 'N24578', 'N2533', 'N25885', 'N26040', 'N2735', 'N28088', 'N29197', 'N306', 'N32182','N33964','N34876','N38105', 'N38779', 'N41777', 'N42620', 'N42634', 'N43608', 'N45794', 'N4693', 'N50748', 'N52622', 'N52631', 'N53428', 'N53494', 'N55051', 'N55556', 'N55911', 'N56638', 'N57528', 'N57967', 'N58090', 'N58226', 'N58264', 'N59289', 'N60550', 'N61342', 'N62058', 'N63411', 'N63658', 'N7132', 'N8331', 'N8448', 'N8519', 'N8629']
U92486	['N10671', 'N11673', 'N12349', 'N1398', 'N14340', 'N14349', 'N1455', 'N14662', 'N14972', 'N15034', 'N18445', 'N20121', 'N21628', 'N22032', 'N23614', 'N24591', 'N24724', 'N27106', 'N2735', 'N27644', 'N2805', 'N28058', 'N28088', 'N28467', 'N28711', 'N30160', 'N31801', 'N31820', 'N32089', 'N33622', 'N35554', 'N36229', 'N38821', 'N4020' 'N40545', 'N41218', 'N42583', 'N42620', 'N43955', 'N44007', 'N4524', 'N45388', 'N46795', 'N50049', 'N51147', 'N51464', 'N51616', 'N52122', 'N52692', 'N53052']

2.2 Preprocessing

Preprocessing data is the process of processing raw data into more structured data. Preprocessing in this study uses Natural Language Processing (NLP) to provide understanding and convenience to computers regarding human language. In this process, the writer uses the Natural Language ToolKit (NLTK) library which makes it easier for the writer to process the text. The steps involved in preprocessing the data can be seen in Figure 2.2

Figure 2. Preprocessing stage

2.3 Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a method of weighting each word by calculating the frequency of occurrence of words in each document and the frequency of occurrence of words in all documents [9]. TF-IDF consists of Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency is a value that shows the frequency of terms that often appear in a document. The greater the number of occurrences of a term in the document, the greater its weight [10]. Inverse Document Frequency (IDF) aims to reduce the weight of the term if it exists in all documents. In contrast to TF, the less the frequency with which words appear in the document, the greater the value [10]. TF-IDF is done to change news titles in textual form to numeric ones so that they can be understood and processed by computers. The author first calculates the term frequency (TF), which is to calculate the frequency of occurrence of words in all news titles in the dataset. Previously, the writer calculated the frequency of occurrence of a word that had been dilemmad into a basic form.

2.4 Cosine Similarity

The cosine similarity score calculates the cosine of the angle between two objects or vectors in an n-dimensional space [11]. In this study, the cosine similarity score calculates the angle between the reader's news history vector on the test data and the reader on the training data. If the cosine score is 1 (or the angle is 0), then the two vectors are exactly the same. On the other hand, if the cosine score shows -1 (or 180 degree angle) then the two vectors are very different from each other [12].

3. RESULT AND DISCUSSION

3.1 Preprocessing Result

a. Case Folding

Case folding is a stage in preprocessing to change all the letters in the news title into small letters, so that there are no capital letters in the news title. The purpose of this process is to uniform the characters in the news headlines. An example of the case folding process can be seen in the table below.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

Table 3. Case Folding Process on News Titles

Before	After
The Cost of Trump's Aid Freeze in the Trenches of Ukraine's War	the cost of trump's aid freeze in the trenches of ukraine's war
I Was An NBA Wife. Here's How It Affected My Mental Health	i was an nba wife. here's how it affected my mental health
How To Get Rid of Skin Tags, According to a Dermatologist	how to get rid of skin tags, according to a dermatologist

In addition to changing capital letters into small letters, the case folding process also removes punctuation marks or characters other than letters. Punctuation marks and characters other than letters in news headlines are considered as delimiters. The purpose of this process is also to uniform the characters in the news headlines. An example of the process of removing the delimiter is shown in the table below.

Table 4. The Process of Removing Delimiters in News Titles

Before	After
the cost of trump's aid freeze in the trenches of ukraine's war	the cost of trumps aid freeze in the trenches of ukraines war
i was an nba wife. here's how it affected my mental health	i was an nba wife heres how it affected my mental health
how to get rid of skin tags, according to a dermatologist	how to get rid of skin tags according to a dermatologist

b. Filtering

This process is done to eliminate stop words and only retrieve important words. Stop words is a list of words that often appear in news headlines but are ignored because they have no significant meaning. For example, the words "the", "of", "in", "i", "was", "an", "it", "my", and "a" which can be seen in table 3. In addition to reducing the volume In other words, filtering is done as an effort to make the process more precise [13].

Table 5. The Process of Removing Stopwords in News Titles

Sebelum	Sesudah
the cost of trumps aid freeze in the trenches of ukraines war	cost trumps aid freeze trenches ukraines war
i was an nba wife heres how it affected my mental health	nba wife heres affected mental health
how to get rid of skin tags according to a dermatologist	get rid skin tags according dermatologist

c. Stemming

Every word in the news title that has been tokenized or segmented based on spaces then goes to the stemming process. The author uses a stemming process to change each word into its root form. The stemming process removes word affixes at the beginning (prefix), the end (suffix) and insertion (infix) in words. In stemming, the word given in its plural form is changed to its singular form [14]. An example of the application of the stemming process can be seen in table 4.

Table 6. Stemming Process on News Titles

Before	After
cost trumps aid freeze trenches ukraines war	cost trump aid freez trench ukrain war
nba wife heres affected mental health	nba wife here affect mental health
get rid skin tags according dermatologist	get rid skin tag accord dermatologist

3.2 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a method of weighting each word by calculating the frequency of occurrence of words in each document and the frequency of occurrence of words in all documents [9]. TF-IDF is done to change news titles in textual form to numeric ones so that they can be understood and processed by computers. The author first calculates the term frequency (TF), which is to calculate the frequency of occurrence of words in all news titles in the dataset. Previously, the writer calculated the frequency of occurrence of a word that had been dilemmad into a basic form. The following is a frequency table for each word.

Table 7. Term Frequency across Dataset

Term	Frequency
cost	15
trump	313
aid	4

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

…	…
insult	1

To get the TF-IDF weight of a word, the writer uses the TfidfVectorizer module. The TF formula used is as below.

$$tf_{(t)} = \frac{the\ frequency\ of\ occurrence\ of\ the\ term\ t\ in\ the\ news\ title\ j}{number\ of\ terms\ in\ the\ headline\ j} \quad (1)$$

An example of the application of the TF formula on the news headline "cost trump aid freeze trench ukraine war" is as follows.

Table 8. TF Weight of Each Term on the Entire Dataset

TF weight
$tf_{(cost)} = \frac{1}{7}$ = 0,14285714
$tf_{(trump)} = \frac{1}{7}$ = 0,14285714
$tf_{(aid)} = \frac{1}{7}$ = 0,14285714
$tf_{(freeze)} = \frac{1}{7}$ = 0,14285714
$tf_{(trench)} = \frac{1}{7}$ = 0,14285714
$tf_{(ukraine)} = \frac{1}{7}$ = 0,14285714
$tf_{(war)} = \frac{1}{7}$ = 0,14285714
$tf_{(ukraine)} = \frac{1}{7}$ = 0,14285714
$tf_{(war)} = \frac{1}{7}$ = 0,14285714

After getting the TF weight for each word, the inverse document frequency (IDF) is calculated with the following formula.

$$idf_{(t)} = \frac{\log\ (number\ of\ news\ titles\ in\ the\ dataset)}{the\ number\ of\ news\ titles\ containing\ the\ word\ t} \quad (2)$$

The following is the IDF weight for each word in the news title "cost trump aid freeze trench ukraine war"

Table 9. IDF Weight of Each Word in the Entire Dataset

IDF weight
$idf_{(cost)} = log\frac{5000}{15}$ = 1,176091259055681
$idf_{(trump)} = log\frac{5000}{310}$ = 2,491361693834273
$idf_{(aid)} = log\frac{5000}{4}$ = 0,602059991327962
$idf_{(freeze)} = log\frac{5000}{5}$ = 0,698970004336019
$idf_{(trench)} = log\frac{5000}{1}$ = 0
$idf_{(ukraine)} = log\frac{5000}{42}$ = 1,6232492903979
$idf_{(war)} = log\frac{5000}{22}$ = 1,342422680822206

The high IDF weight for each word indicates that the word rarely appears in the news headlines, while the low IDF weight indicates that the word often appears in the news headlines.
After getting the TF and IDF weights, the TF-IDF weights were calculated for each word using the formula below.

$$TF - IDF = TF * IDF \quad (3)$$

Table 10. TF-IDF Weight of Each Term in the Entire Dataset

TF-IDF weight
$tf - idf_{(cost)}$ = 0.34735415982506532
$tf - idf_{(trump)}$ = 0.31540157342714026
$tf - idf_{(aid)}$ = 0 .4835107236469051
$tf - idf_{(freeze)}$ = 0.4233139857894654
$tf - idf_{(trench)}$ = 0.4333040096343576
$tf - idf_{(ukraine)}$ = 0.20628743826957918
$tf - idf_{(war)}$ = 0.36957098470030486

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

The TF-IDF weight of each word that composes the news title is then calculated on average to get the news title score. News title score is used to measure the similarity between titles.

Table 11. TF-IDF Weight News Title

Title	TF-IDF Weight
cost trump aid freeze trench ukraine war	29.250713195245208
i was an nba wife heres how it affected my mental health	27.340397135140552
how to get rid of skin tags according to a dermatologist	27.295078734106244

3.3 Cosine Similarity

Cosine similarity is used to calculate the similarity between two news headlines. The entire news history list is divided into two, namely 70% as training data and 30% as test data. The value of cosine similarity ranges from -1 to 1. The similarity value of cosine similarity is equal to the value of cos θ where here is the angle between reader's history click [15]. Experimentally, the author limits the similarity value of 0.95 so that the recommended news has high similarity. The formula used in calculating cosine similarity is as follows.

$$similarity(A,B) = \cos(\theta) = \frac{\mathbf{A}.\mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} \mathbf{A}i\mathbf{B}i}{\sqrt{\sum_{i=1}^{n} \mathbf{A}i^2}\sqrt{\sum_{i=1}^{n} \mathbf{B}i^2}} \quad (4)$$

The following is a table containing the results of the similarity between the news history vectors from one reader to another.

Table 12. Similarities Between Two News History Readers

User U39696's History Weight (from Data Test)	User U91836's History Weight (from Data Train)	Similarity
[33.76; 27.45; 29.80; 29.59; 29.46; 32.05; 27.54; 27.68; 33.95; 25.17; 35.37; 29.96; 35.75; 33.80; 33.78; 35.01; 31.65; 33.72; 31.32; 27.47; 25.31; 27.78; 29.52; 37.01; 33.20; 38.96; 35.33; 40.52; 19.38; 31.78; 33.70; 29.77; 27.67; 35.45]	[29.60; 35.47; 29.93; 35.64; 33.58; 29.30; 29.30; 29.45; 33.21; 29.72; 29.93; 37.01; 33.95; 31.40; 27.63; 43.13; 29.32; 35.67; 35.65; 27.46; 25.30; 25.20; 29.96; 33.85; 21.91; 29.64; 29.28; 29.05; 29.83; 31.96; 31.60; 29.64; 31.70; 31.92; 33.70]	0.98595959

3.4 Results

Based on the tests carried out on each news history reader, test data on all news history readers on training data, the highest similarity value between readers with ID U15033 and readers U73303 is 0.9963, while the lowest similarity value is between U18147 readers and U80573 readers, which is 0.9082. If the similarity value between the two news histories of readers is more than 0.95, then the news history of the readers of the training data is stored to later become a news candidate who will be recommended to the readers of the test data. In news candidates, the frequency of occurrence of each news is calculated and then the top-10 is selected which will be recommended to readers on the test data. Top-10 news recommendations are a collection of news stories that readers have never clicked on before.

Table 13. Click Frequency of Each News ID

News ID	Click Frequency
N871	978
N55189	780
N59704	707
N54827	647
N619	630
N55743	624
N5978	593
N61388	558
N55846	527

From the 1,586 readers in the test data, the recommendation result was clicked by 1,281 readers.

Table 14. Results of Recommendations

User ID	User Click History	Recommendations	Watched
U39696	['N1085', 'N11249', 'N12349', 'N13008', 'N13138', 'N13836', 'N14150', 'N14522', 'N14742', 'N15788', 'N16292', 'N16695', 'N16715', 'N18445', 'N18642', 'N19028', 'N19110', 'N20592', 'N20798', 'N2203',	['N871', 'N55189', 'N59704', 'N54827', 'N619', 'N55743', 'N5978', 'N61388', 'N55846', 'N60702']	['N54827','N60702']

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

	'N22079', 'N22397', 'N23571', 'N24999', 'N25378', 'N26179', 'N26364', 'N27235', 'N27501', 'N30110', 'N30353', 'N306', 'N33013', 'N33198', 'N33458', 'N35926', 'N36530', 'N37920', 'N37942', 'N4020', 'N42691', 'N4408', 'N45020', 'N46023', 'N46131', 'N46846', 'N47563', 'N47672', 'N49285', 'N49289', 'N49953', 'N50047', 'N50069', 'N50475']		
U17999	['N10865', 'N11855', 'N12907', 'N13008', 'N14275', 'N15603', 'N1848', 'N19591', 'N1985', 'N21623', 'N22562', 'N24233', 'N26136', 'N2805', 'N28862', 'N28952', 'N28992', 'N30092', 'N30955', 'N33704', 'N33976', 'N34937', 'N35685', 'N36602', 'N38046', 'N39084', 'N4020', 'N40442', 'N40508', 'N41089', 'N42458', 'N43671', 'N43903', 'N44442', 'N44454', 'N44589', 'N45286', 'N45794', 'N45896', 'N47558', 'N49088','N49285', 'N51706', 'N52307']	['N871', 'N55189', 'N59704', 'N54827', 'N619', 'N55743', 'N5978', 'N61388', 'N55846', 'N60702']	[]

To measure the quality of the recommendation system and determine whether the recommendation results were clicked by the reader, the author uses a hit-rate measure to match whether there are news recommendation results in the reader's click history that are intentionally omitted. The calculation of accuracy using hit-rate is measured by counting the number of clicks on top-N news recommendations made by readers. If the reader clicks 1 of the top-N news recommendations, the number of hits is 1. The formula for calculating the hit-rate is as follows:

$$Hit\text{-}Rate = \frac{jumlah\ hit}{jumlah\ pembaca\ yang\ direkomendasikan} = \frac{1281}{1586} = 0{,}8077 = 80{,}77\%$$

In this research, the writer counts the number of hits from all readers and divides it by the total number of readers. Closer to 1 indicates that the algorithm can recommend news to readers, while closer to 0 indicates that the algorithm cannot recommend news to readers.

3.5 Discussion

The news recommendation system using the TF-IDF method and cosine similarity measurement is implemented on Microsoft News Dataset data with a comparison of 70% training data and 30% test data resulting in an accuracy measured by a hit-rate of 80.77%. The news recommendations generated to 1,586 readers contain the same news because they are taken from the top-10 news candidates that are often clicked by readers on the training data.

4. CONCLUSIONS

Based on the research that has been done, it can be found that the news recommendation system using TF-IDF and the similarity of the algorithm on the Microsoft News Dataset uses 5,286 reader histories who have at least clicked on 40 news stories and 5,000 news titles can recommend the 10 best news stories and produce the best performance with an accuracy of 80.77. Suggestions that can be given by the author for further research is to use a dataset that does not have feedback in the form of a history of news clicks, but implicit feedback in the form of the time used by readers to read the news, as well as explicit feedback such as ratings given to news readers who have read. , so that the results of recommendations for each reader can be different (personalization).

REFERENCES

[1] F. Wu *et al.*, "MIND: A Large-scale Dataset for News Recommendation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 3597–3606. doi: 10.18653/v1/2020.acl-main.331.

[2] S. N. Mohanty, J. M. Chatterjee, S. Jain, A. A. Elngar, and P. Gupta, Eds., *Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries*, 1st ed. Wiley, 2020. doi: 10.1002/9781119711582.

[3] A. Susanto, "Program Studi Informatika Fakultas Teknologi Informasi dan Elektro Universitas Teknologi Yogyakarta," p. 14.

[4] A. E. Wijaya and D. Alfian, "Sistem Rekomendasi Laptop Menggunakan Collaborative Filtering dan Content-Based Filtering," p. 17.

[5] M. W. Putri, A. Muchayan, and M. Kamisutara, "Sistem Rekomendasi Produk Pena Eksklusif Menggunakan Metode Content-Based Filtering dan TF-IDF," *JOINTECS J. Inf. Technol. Comput. Sci.*, vol. 5, no. 3, p. 229, Sep. 2020, doi: 10.31328/jointecs.v5i3.1563.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 277−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1670

[6] F. Indriani and M. R. Faisal, "Sistem Rekomendasi Berita Online dengan Menggunakan Pembobotan TF-IDF dan Cosine Similarity," vol. 2, p. 10, 2019.
[7] I. Mawanta and T. S. Gunawan, "Uji Kemiripan Kalimat Judul Tugas Akhir dengan Metode Cosine Similarity dan Pembobotan TF-IDF," vol. 5, p. 13, 2021.
[8] S. Chhipa, V. Berwal, T. Hirapure, and S. Banerjee, "Recipe Recommendation System Using TF-IDF," *ITM Web Conf.*, vol. 44, p. 02006, 2022, doi: 10.1051/itmconf/20224402006.
[9] A. Irvandani, K. Auliasari, and R. Primaswara Prasetya, "Sistem Rekomendasi Pemilihan Fotografer dengan Metode Haversine dan TF-IDF di Malang Raya," *JATI J. Mhs. Tek. Inform.*, vol. 4, no. 1, pp. 137–146, Aug. 2020, doi: 10.36040/jati.v4i1.2330.
[10] C. H. Yutika and S. A. Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes," p. 9.
[11] M. I. Huda, N. D. W. Cahyani, and H. Nurrahmi, "Pemanfaatan Metode Cosine Similarity untuk Mengidentifikasi Cyberbullying pada Twitter," p. 9.
[12] R. Banik, *Hands-On Recommendation Systems with Python*. 2018. Accessed: Jun. 26, 2022. [Online]. Available: http://sbiproxy.uqac.ca/login?url=https://international.scholarvox.com/book/88861422
[13] D. Oleh, "Sistem Rekomendasi Buku Menggunakan Metode Content Based Filtering," p. 59.
[14] T. Jo, *Text Mining*, vol. 45. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-319-91815-0.
[15] P. C. Purnama and S. A. Faraby, "Analisis Perbandingan Metode Similarity Pearson dan Cosine pada Sistem Rekomendasi Film dengan Pendekatan User-Based Collaborative Filtering," p. 22.