TEXT MINING TO CLASSIFY ONLINE NEWS TITLES CASE STUDY OF BANJARMASIN RADAR USING TF-IDF AND K-NN METHODS

Salsabila Anjani[1], Andi Farmadi[2], Dwi Kartini[3], Irwan Budiman[4], M. Reza Faisal[5]
1,2,3,4, Computer Science Study Program FMIPA ULM
Jl. A. Yani Km. 36 Banjarbaru, South of Kalimantan
Email: salsabila.anjani.sa@gmail.com

ABSTRACT

The news media that used to be commonly used were newspapers. However, with the development of the times, the news media is now entering the digital era. Many online news media spread on the internet. The sophistication of the internet makes it easier for readers to choose which news they want to read. Unlike newspapers, online news media have categories where readers can choose. In general, the categorization of a news in online media is determined by the editor. Given the number of news published in a day, of course, makes the editor's job difficult. A category in the news is usually not appropriate because usually the headline is made as attractive as possible to attract the interest of the reader. So there are times when the news title does not match the category that has been entered by the editor. The use of the K-Nearest Neighbor (K-NN) method can be used in determining the categorization of a news. By using a case study of the online media Radar Banjarmasin, a research was conducted to find out how well the Canberra and Euclidean classification methods were using news headline data for categorization. The results obtained in this study are the better classification method is Euclidean and with an accuracy value of 65.00%. Improvements that should be made for further research is to use other methods for comparison.

Keywords **:** *K-Nearest Neighbor, Classification, Online News, TF-IDF, Canberra, Euclidean*

1. INTRODUCTION

Along with the development of the times, there are getting a lot of people who use the facility online on social media, entertainment, and news of the latest from the local to the international. Development that continues to occur in the field of web, libraries digital, news online, technical documentation and other media providers of information have facilitate accessing the data document textual in a number of the very large. If the technologies in these fields are combined, it will result in the development of very useful data sources. It makes text mining or the discovery of knowledge from databases textual become a challenge in itself to have a standard depth of language naturally that is used by most large documents are available. In general, text mining and data mining is considered similar to one each other, with the perception that the technique is to be used in the concept of mining the text [6]. According to Faisal et al. (2019) the difference between data mining and text mining is the data source used, in text mining the data source is text. Another difference is there a process of extraction features to transform the data text into a

data structure [5]. One of the official websites that provides online news is Pro South Kalimantan, which is sheltered by Radar Banjarmasin. Based on website traffic analysis on similarweb (2020), the average website visitor in the last six months was 94.97 thousand visitors, all of whom were Indonesian citizens. News that was published less over as many as 10 news in a day. The published news is divided into several categories, namely 'Banua Bisnis', 'Feature', 'Female', 'Hukum dan Peristiwa', 'Prokaltorial', 'Radar Muda', 'Ragam Info' and 'Sport'. In input process, operator website choose a category in the manual. If the data of news reached hundreds, then the operator requires more time and effort to do inputting the news.

Algorithm K-Nearest Neighbor is one of an algorithm that can be used for the system of classification to a data based on the training data which has a range of the most close to an object. The algorithm K-Nearest Neighbor (K-NN) is doing classify by storing the results of the query instance to the majority of the class that many emerging [1]. The calculation of distance is a method to find the distance between the point of data recently and the collection of data training that exist. Some algorithms calculating the distance of which is like the Manhattan Distance, Canberra Distance, Euclidean Distance and so forth [8]. And the study of this, the calculation of the distance that is used is a Step- step for calculating algorithm K-Nearest Neighbor among others:

a. Giving value to the parameter k.
b. Find the value of the distance between the test data and all training data.
c. If you already get the distance of the data testing, it will be sorted by distance are formed.
d. Determining the value range of the most close in any order parameter k.
e. Results of the numbers (4) to be incorporated into the classroom are appropriate.
f. Determining the number of class neighbors the most close.
g. Class that these will be used as a class of data to be tested [1].

The performance of the K-NN method can be affected by many problems, such as choosing the value of k and choosing the distance measure. To overcome these problems, a large number of machine learning techniques are continuously developed [8].

Comparison of calculations between the distances of two points that have a numerical attribute value of Euclidean distance is usually expressed in equation (1) to determine the difference between the squares of each attribute between the data [7].

$$D\ (d_x, d_y\ ) = \sqrt{(f_{x,1} - f_{y,1})^2 + \cdots + (f_{x,f}\ - f_{y,f}\ )^2} \qquad \ldots (1)$$

The Canberra Distance method is very sensitive to changes that occur between the two coordinates which are close to zero [2] . Below this is the equation of Canberra Distance.

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|X_{ik}| - |X_{jk}|} \qquad \ldots (2)$$

Firdaus et al. (2019) perform classification on documents news is automated with the help of a computer. The method of classification that is used is the K-Nearest Neighbor (K-NN) combined with algorithms Cosine Similarity and Euclidean Distance. From this research, resulting level of accuracy of the average method K-Nearest Neighbor (K-NN) and Cosine Similarity amounted to 98.12%, and for the method of K-Nearest Neighbor with Euclidean Distance resulting level of accuracy of 56.51% [4].

Wahyono et al. (2019) also did a research using the algorithm K-Nearest Neighbor (K-NN) to perform the classification of data text for the management of documents is automatic. The research carried out by way of changing the representation of the word to the shape of the vector. At the processing stage, the use of calculating the distance value of the K-NN algorithm is essential to determine the distance between data elements. Based on the statement, Wahyono et al. (2019) tried to compare the four pieces of the calculation of the distance that is much used in the K-Nearest Neighbor, namely Euclidean Distance, Chebyshev Distance, Manhattan Distance and Minkowski Distance. The dataset used is Eminem's Youtube comment data as much as 448 data. From the study it was found that the Euclidean Distance and Minkowski Distance in the algorithm K-NN in the data as a representation of the vector of sentence majority produce the level of accuracy that is better than the Chebyshev Distance and Manhattan Distance. From the test, it was also found that the best value for K-Nearest Neighbor was obtained when k = 3 [7].

In this study, the authors performed calculations weights by using a weighting Term Frequency-Inverse Document Frequency (TF-IDF). Weighting TF-IDF in general, ordinary used in the mining of text for retrieval of information. In here the weights are identified to calculate how important word for a document in a single set. The weight increases according to the number of times a word appears in the document [3].

2. RESEARCH METHOD

a. Problem Formulation

The first step can be performed to start the research is determining matter what that wants to be solved. The formulation of the problem serves to assist the author in determining the direction and purpose of the research to be carried out. With the hope of the study can be completed with a more efficient and structured.

b. Literature Study

After doing the formulation of the problem, the stage further that needs to be done is the study of literature. Literature studies are needed to add insight and understanding of the research to be carried out. In terms of this, the authors did the study of research earlier and some literature that relates to research that will be done.

c. Data Collection

As that has been mentioned in the section before, data that is used is the data the title of the page news online Radar Banjarmasin. After all the data collected, the data will be processed and carried out dry with a goal to transform the data of crude into the data ready made.

d. Text Preprocessing

Processing of data that has become only the title and category is entered in the excel file. In line major process preprocessing text that is done are the process of Case Folding, Tokenizing, Stemming, and Stopword. After the data is processed through Case Folding, Tokenizing, Stopword, and Stemming, the next data is ready to be processed because the data has become structured data.

e. TF-IDF Weighting

If the data has been converted into structured data, then the next step can be carried out namely the weighting stage. In the study of this method of weighting that is used is the Term Frequency-Inverse Document Frequency or can be shortened by TF-IDF.

f. K-NN Classification

After the weighting process is done, all data will be classified using the K-NN algorithm. The classification process in this study is divided into two, by calculating the distance using the Canberra and Euclidean distance calculations.

g. Accuracy Results

From the classification of the two calculation distance, then made the comparison results of the level of accuracy between Canberra Distance and Euclidean Distance.

3. RESULTS AND ANALYSIS

3.1 Data Collection

Capturing Data crude done with it manually take the news portal news Radar Banjarmasin. Categories were selected in the study of this there are four pieces starting from the initial selection of categories that exist in the website is 'Banua', 'Bisnis', 'Hukum dan Kriminal' dan 'Sport'. The data collected here is data that is still in the form of a row of 400 URL links that are entered into the excel file.

The first process after collecting links news that has been done by way of the manual, the data are collected in an Excel file for later processed to take the title of the story with the help of the programming language R. The function of the program it is to take the title of the news of each row of links in excel collected into headline news.

Table 1 Comparison after and before document categorization

Information	Class (No)	Class (Yes)
Raw data	400	0
The data has been processed	0	400

The process of converting raw data in the form of a url link in excel then processed by affixing the title data division by category. The distribution of this data is done by comparing one category to represent one hundred news titles. The categorization is done by means of manual and sequential.

3.2 Preprocessing Text

Before the data entered into the program RapidMiner Previous data have been processed previously must be converted using a nominal To Text. The process is intended so that the data can be processed in RapidMiner as the essentially computers only recognize and be able to process only the data in the form of numbers.

At the stage of preprocessing text, the title file data that has been given the category further carried out four pieces of the process before it entered into the model that has been designed. Starting from the cleaning process, case folding, tokenizing, stopword and stemming. All text preprocessing processes are carried out in the RapidMiner application. There are four processes in processing documents into data in this design, namely transform cases, tokenize, stopword filters , stem (dictionary) and filter tokens. The stages that occur in the transform cases is the change of data heading into letters small that used the research of this.

Data results from the process transform the case is in the form of an excel format and are vertically arranged each title. This file also contains the numbers that will be used for phase weighting so that the numbers are listed now just a figure of zero for each headline news. While the category is in the horizontal right corner.

The next step is to use the tokenize feature. Tokenizing is a process to break down the sentence into terms. Research is using features tokenizing to break the line of the title phrase that previously had conducted the case of folding into a case smaller.

Data results output from tokenizing not much different from the results transform case. The difference in this processed is that the words are broken down into per cell in the excel file. The total number of lengths in excel reaches 1499 rows.

Process stopword done with by the dictionary that previously had in downloaded previously. The stopword dictionary is then entered into the parameter. Its function is to eliminate words that are deemed not necessary.

The data generated by the stopword is data that also has columns and rows that much. But the results of the data it happens reduction because the word is already affected by the filter will be removed. So the amount of data left for the word is 1340.

Stemming is a process to change the words that have been broken down and filtered on a process previously be said basically. Eliminate the word affix initial or suffix the word. This process also requires a dictionary to convert the word into a root word.

The results of the data for the stemming output process are in the form of word recovery into basic words. So the data does not change too much from the previous process.

Limitation on the number of letters in a word here is also determined with the help of using the RapidMiner application. Restrictions are used in research this is the minimum for the number of letters there are three, and limits the maximum there at 99.

The last process in preprocessing the data is to set the number of letters to be processed. In the process of this data, the data shrink again to 1321 data. Furthermore, this data is ready to be processed by weighting.

The whole process lasted start of transform case, tokenizing, stopword, stemming and filtering. The next process is to provide an assessment of the data that has been processed using TF-IDF weighting in accordance with this research design.

Data results are issued by TF-IDF is the data in the form of weighting that in the excel file. Matrix results from TF-IDF weighting Reach 400x400 Data generated. This weighting is calculated using the Rapid Miner program. The process that has been going on from preprocessing text to TF-IDF, at this stage the data collected is ready to be continued in the next process, which is entered into the model and compared. So that the accuracy results can be compared.

3.3 Modeling

This research will using the distance calculation method that exist in the K-Nearest Neighbor namely Canberra and Euclidean. RapidMiner can process the two pieces of the algorithm in one go. This research, using the ratio of 80% of the data training (320 data) and 20% of data testing (80 data). Word weighting is carried out using the TF-IDF method so that 1,265 attributes are obtained. To classify the headline, the writer used the K-NN with distance measurement method Euclidean and Canberra. For the value of *k* on the K-NN, the researchers used the values of 3, 5, 7, 9, 11,13 and 15.

The use of values in the K-Nearest Neighbor method refers to previous research. Starting from the stage of preprocessing text Data crude up by comparing two pieces of a method of classification that exist in the K-Nearest Neighbor is Canberra and Euclidean.

3.4 Evaluation

After the text preprocessing and modeling process is carried out, the next step is to run the model that has been designed in the RapidMiner program. Experiments carried out by running the model beginning were submitted and evaluate models to achieve the result that desired. The results of the evaluation of the model that has been designed to run with a well and getting results are desired.

3.5 Results

The results of testing the model by using the model proposed tools application RapidMiner produce output in the form of accuracy, recall, and precision. Tests were conducted using a K-Nearest Neighbor with the method of classification Canberra issued the results of an accuracy of 25.00%.

Tests were conducted using a K-Nearest Neighbor with the method of classification Canberra issued the results of precision at 6:25%. Tests were conducted using a K-Nearest Neighbor with the method of classification Canberra issued the results of a recall of 25.00%. Weight that is used in the testing of this are 1,1,1,1. The value of k is obtained by optimizing the Canberra parameter. The optimization results from the Canberra parameters are presented in the table below:

Table 2 Results of Canberra Parameter Optimization

Iteration	*k Canberra*	Accuracy
4	9	0.250
6	13	0.250
2	5	0.250
5	11	0.250
3	7	0.250
1	3	0.250
7	15	0.250

Tests were conducted using a K-Nearest Neighbor with the method of classification euclidean issued the results of an accuracy of 65.00%.

Tests were conducted using a K-Nearest Neighbor with the method of classification euclidean issued the results of precision of 70.52%, the value of a recall of 25.00% and an accuracy that is obtained is at 65.00%. Here's a comparison chart :

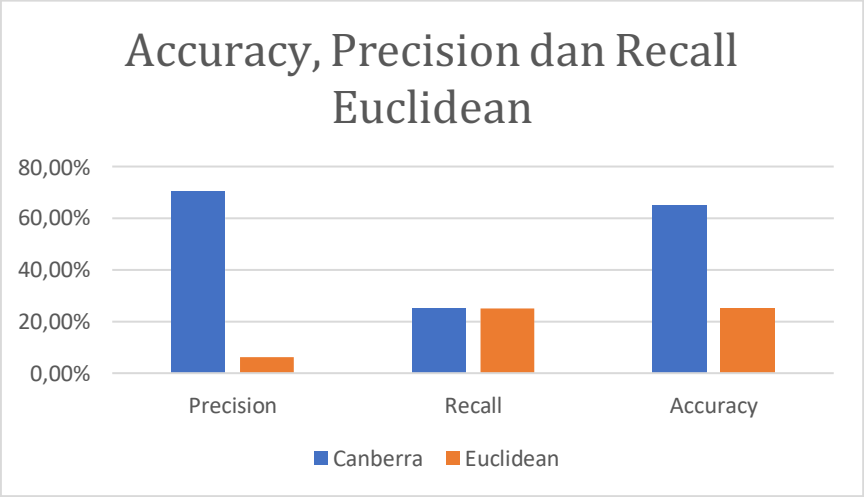

Figure 1 Comparison of Accuracy, Precision, Recall between Canberra and Euclidean

K value obtained by doing optimization on parameters Euclidean. Optimization of the results of the parameters Euclidean presented in the table below is :

Table 3 Euclidean Parameter Optimization Results

Iteration	*k* Euclidean	Accuracy
2	5	0.650
3	7	0.600
5	11	0.575
1	3	0.575
4	9	0.550
7	15	0.525
6	13	0.475

3.6 Analysis

The result of the research that has been going on is to compare the results of the accuracy of the two classification methods. The following table compares the accuracy :

Table 4 Results of comparison of the two pieces of the method of classification

Iteration	*k*	Euclidean's Accuracy	Canberra's Accuracy
1	3	0.575	0.25
2	5	0.65	0.25
3	7	0.6	0.25
4	9	0.55	0.25
5	11	0.575	0.25
6	13	0.475	0.25
7	15	0.525	0.25

Research that has been carried out with the use of RapidMiner it can be seen that the results of the accuracy of the most high- obtained by the K-Nearest neighbor by using Euclidean.

Table 5 Confusion matrix using the K-NN algorithm with Euclidean

	True Banua	True Hukum dan Peristiwa	True Bisnis	True Sport	Class Precision
Pred. banua	15	3	2	2	68.18%
Pred. hukum dan peristiwa	2	15	1	1	78.95%
Pred. bisnis	2	2	15	2	71.43%
Pred. *sport*	1	0	2	15	83.33%
Class recall	75%	75%	75%	75%	

The results of the confusion matrix table above can be taken that the accuracy of the best K-NN algorithm using the Euclidean distance measurement method is 75% using value of k = 3. For a recall value is 75%, and a precision value is 75.47%. The following is a confusion matrix on the distance measurement method using Canberra.

Table 6 Confusion matrix using the K-NN algorithm with Canberra

	True Banua	True Hukum dan Peristiwa	True Bisnis	True Sport	Class Precision
Pred. banua	0	0	0	0	0%
Pred. hukum dan peristiwa	20	20	20	20	25%
Pred. bisnis	0	0	0	0	0%
Pred. sport	0	0	0	0	0%
Class recall	0%	100%	0%	0%	

The best value of K-NN algorithm accuracy using canberra method obtained by 25% with value of k = 5. Value recall obtained using methods canberra by 25% and the precision is 6.25%. There are differences in accuracy, recall and precision are great between the distance measurement method euclidean and canberra . Here

is a graph comparing the results of accuracy between the two classifications Canberra and Euclidean.

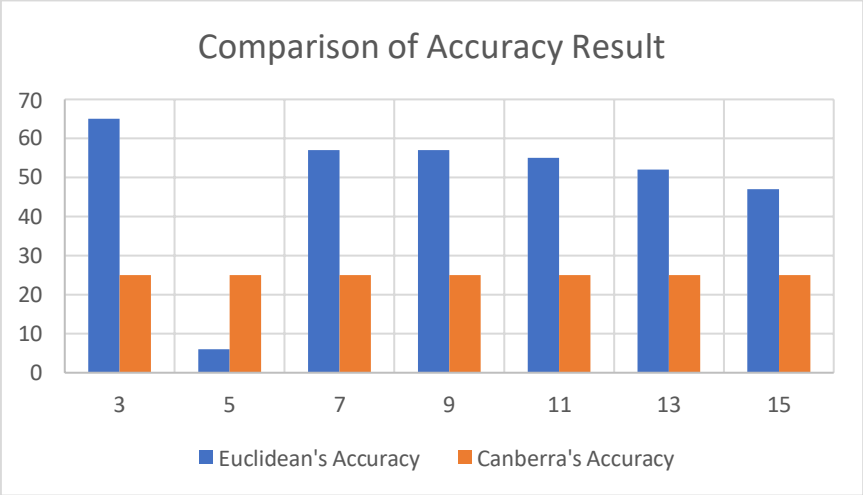

Figure 2 Comparison of results on the classification of the two methods

This difference is caused by the RapidMiner application which does not succeed in dividing the prediction labels evenly, only focusing on predictions on the "hukum dan peristiwa" label. The difference in the values of accuracy, recall, and precision in K-NN is also influenced by the value of *k*, the amount of training and testing data, and the method of calculating the distance.

4. CONCLUSION

The comparison of the results of the method of classification on the method of K-Nearest Neighbor namely Canberra and Euclidean, modeling which was designed in tools RapidMiner successfully run as expected. The results of the accuracy of the most high- obtained by Euclidean method get results accuracy of 65.00%. Results of precision of the most high result is also the Euclidean method at 75.00%. Recall the results of the most high is also contained by the Euclidean method at 65.00%. Data to study the case of Radar Banjarmasin using text mining, Euclidean is the method that is most excellent in the handling of data headline news.

REFERENCES

[1] Banjarsari, M.A., I. Budiman, & A. Farmadi. 2015. Penerapan K-Optimal pada Algoritma KNN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4. *Kumpulan Jurnal Ilmu Komputer (KLIK),* Volume 2, pp. 50-64.

[2] Diaz, R.A.N. 2018. Pengelompokan Artikel Bahasa Bali Menggunakan Algortima K-Means Clustering. *Seminar Nasional Sistem Informasi dan Teknologi Informasi,* Volume 1, pp. 224-228.

[3] Faisal, M. Reza & A.R. Arrahimi. 2020. *Belajar Data Science Klasifikasi dengan Bahasa Pemrograman R*. Banjarbaru: Scripta Cendikia.

[4] Firdaus, Pasnur & Wabdillah. 2019. Implementasi Cosine Similarity Untuk Peningkatan Akurasi Pengukuran Kesamaan Dokumen Pada Klasifikasi Dokumen Berita Dengan K Nearest Neighbor . *Jurnal Teknologi Informasi dan Komunikasi*, Vol. 9, No. 1 : 69-74.

[5] Qaiser, S. & R. Ali. 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents*. International Journal of Computer Applications,* Vol. 8, No. 1 : 25-29.

[6] Salloum, S.A., M. Al-Emran, A.A. Monem & K. Shaalan. 2018. Using Text Mining Techniques for Extracting Information from Research Articles. *Springer International Publishing AG*, Volume 1, pp. 373-397.

[7] Wahyono, I N.P. Trisna, S.L. Sariwening. M. Fajar & D. Wijaya. 2019. Perbandingan Perhitungan Jarak pada K-Nearest Neighbor dalam Klasifikasi Data Tekstual. *Jurnal Teknologi dan Sistem Komputer*, Vol. 8, No. 1 : 54-58.

[8] Zhang, S., X. Li, M. Zong, X. Zhu, & R. Wang. 2018. Efficient kNN Classification with Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 5, pp. 1774-1785.