P-ISSN: 1979-5661
E-ISSN: 2622-321X

Classification of Indonesian Online News Topics Using Text Mining

Gede Putra Aditya Brahmantha[a1], Ema Utami[a2], Ainul Yaqin[a3]

[a]Magister Tenik Informatika, Universitas Amikom
Yogyakarta, Indonesia
[1]putra.aditya@students.amikom.ac.id
[2]ema.u@amikom.ac.id
[3]ainulyaqin@amikom.ac.id

Abstract

Online news is in great demand by many people along with the rapid development of information technology. It needs to develop a system that can automatically classify news according to news categories using text mining method. In this research, the methods used in the classification are K-Nearest Neighbor. News classification is done to classify news into 4 categories, namely automotive, technology, money and food. Each topic contains 300 news data. Before the classification began, data in the form of texts will first be carried out in the preprocessing stage and weighted using TF-IDF. Based on the evaluation using the confusion matrix with the division of the dataset into 960 training data and 240 test data, with the value of k=7 it obtained accuracy of 93.75%, Precision of 94.09%, Recall of 93.56%, and F1-Score of 93.71%.

Keywords: *crawling, online news, text mining, classification*, *k-nearest neighbor*

1. Introduction

The development of Information Technology Affects how to spread information, news that was originally spread using radio, newspaper and television media, currently news is widely spread through web platforms in an up-to-date manner. Based on a reuters institute report, as many as 89% of respondents answered accessing news through online media in 2021, this survey was conducted on 2007 respondents. [1].

News published in the online website are generally categorized based on topics such as automotive, politics, business, food and others. At this time the grouping of news that will be published into the web must be through manually grouping by an editor, an editor must understand the content of the news before grouping the news one by one so it is necessary to do so with accuracy but if the more days the more news that needs to be published into the web, the editor will be overwhelmed by grouping the number of news that will be published. Therefore, on this issue required a system that classifies the news automatically.

Text mining or also known as text data mining or search for information on textual data is a process to search for information that focuses on data in the form of documents or texts that have the purpose of extracting useful information and identifying it. Text mining is similar to data mining. Both have the same goal, which is to obtain knowledge and information from a set of data. The basic difference between text mining and data mining is that in data mining, the data used is structured while the text mining data used is unstructured. [2]

There is related research about classifying news, namely the classification of news using the Support Vector Machine method written by Robbi Nanda, et al. The study used 510 news data with a classification limit of 3 news categories, namely democracy, employment and poverty categories. SVM algorithm get the highest accuracy at 88% for parameter value C =1, linear kernel with the division of test data and training data by 90% and 10%. [3]

Another related research is Indonesian News Classification Using Naive Method Bayesian Classification and Support Vector Machine with the usage of Confix Stripping Stemmer written by Ariadi and Fithriasari. One of the process is confix-stripping stemmer as a way to get the basic

P-ISSN: 1979-5661
E-ISSN: 2622-321X

word of Indonesian news. For the classification method used is Naive Bayes Classifier (NBC) which is frequently used in text data and Support Vector Machine (SVM) which is well renowned for doing exceptionally well on big-dimensional data. The best classification outcomes are determined by contrasting the two approaches. The results showed that linear kernel SVM and RBF kernel produce the same classification accuracy and when compared with NBC, SVM is better, which is obtained for each measurement of accuracy, precision, recall, and F-Measure performance is 88.1%, 89.1%, 88.1%, and 88.3%. [4]

K-Nearest Neighbor (KNN) is one of the classification methods for a set of data based on the majority of categories and the aim is to classify new objects based on attributes and samples from training data. [5], The advantages of the K-Nearest Neighbor method are that it can be applied to large data effectively with accurate results. [6] and K-NN also has the advantage of very fast and simple training so it is easy to learn. [7]

In this study, the classification of news is done to classify indonesian online news into 4 categories, namely automotive, technology, money and food. Each topic contains 1200 news data. The algorithm used in this study is K-Nearest Neighbor and evaluate the classification result with confusion matrix and calculating recall, precision, and f1-score.

2. Reseach Methods

2.1. Dataset Collection

In this study, the data obtained by doing web crawling to get news articles on the topic 'Automotive' 'Technology' 'Money' and 'Food' from an Indonesian-language news portal site as many as 300 articles for each topic so that the total data obtained is 1200 articles. The Crawling process is carried out using Selenium program. The Data is labeled according to the topic of the news. From the dataset used, the data is divided into a proportion of 80% for training data and 20% for testing data, the data division process is done randomly.

2.2. Preprocessing

For the first step of text classification, Data that is still regarded as inconsistent and unstructured will undergo preprocessing stage. Preprocessing is done to process raw data so that the data can be fixed and classified by the classification algorithm properly. The advantage of preprocessing is to make the system work more efficiently due to the reduction of words that are not needed in the classification process. Preprocessing stage in this study consists of case folding, cleansing, stopword removal, stemming, and tokenization as seen on Figure 1.

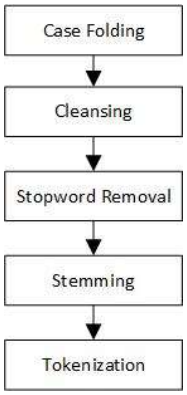

Figure 1. Preprocessing Flowchart

2.2.1. Case Folding

Case folding is the first step in the preprocessing stage that has the objective that each letter turns into a lowercase letter.

Table 1. Case Folding Process

P-ISSN: 1979-5661
E-ISSN: 2622-321X

Before Case Folding	After Case Folding
JAKARTA, KOMPAS.com Elektrifikasi kendaraan menjadi sebuah keniscayaan di tengah tren global dan arah kebijakan	jakarta, kompas.com elektrifikasi kendaraan menjadi sebuah keniscayaan di tengah tren global dan arah kebijakan

2.2.2. Data Cleansing

Data cleansing is a stage in the text cleaning process that eliminates unnecessary characters like usernames, punctuation, and URLs. Additionally, numbers will be eliminated at this time.

Table 2. Data Cleansing Process

Before Data Cleansing	After Data Cleansing
jakarta, kompas.com elektrifikasi kendaraan menjadi sebuah keniscayaan di tengah tren global dan arah kebijakan	jakarta elektrifikasi kendaraan menjadi sebuah keniscayaan di tengah tren global dan arah kebijakan

2.2.3. Stopword Removal

Stopword are list of common words that have little significance and are not used in classification. In this process Stopword removal is done by removing words that are quite common and often appear but do not have a significant effect on the meaning of a text or sentence. [8]

Table 3. Stopword Removal Process

Before Stopword Removal	After Stopword Removal
jakarta elektrifikasi kendaraan menjadi sebuah keniscayaan di tengah tren global dan arah kebijakan	jakarta elektrifikasi kendaraan keniscayaan tengah tren global arah kebijakan

2.2.4. Stemming

Stemming is a process to get the basic words of words that already had affixes, this process is done by removing the affixes in front of the word or behind the word so that it can get the basic words needed for the classification. The stemming proccess is carried out using Sastrawi library.

Table 4. Stemming Process

Before Stemming	After Stemming
jakarta elektrifikasi kendaraan keniscayaan tengah tren global arah kebijakan	jakarta elektrifikasi kendara niscaya tengah tren global arah bijak

2.2.5 Tokenization

Tokenization works by splitting words from text or sentences into multiple tokens. This process will not include spaces.

Table 5. Tokenization Process

Before Tokenization	After Tokenization
jakarta elektrifikasi kendara niscaya tengah tren global arah bijak	[‘jakarta’, 'elektrifikasi', 'kendara', 'tengah', 'tren', 'global', ‘arah’, ‘bijak’]

2.3. Term Frequency Invers Document Frequency (TF-IDF)

The approach used to determine the weight of each extracted word is called Term Frequency-Inverse Document Frequency, or TF-IDF for short. In information retrieval, this method is typically used to count frequent words. TF-IDF weighting model is a technique that integrates model term frequency (tf) and inverse document frequency (idf). Term frequency (tf) is a procedure to count the number of occurrences of a term in a document and inverse document frequency (idf) is used to count the terms that appear in various documents that are considered common terms, and are considered unimportant [4].

The steps to perform TF-IDF weighting are as follows :

a. Firstly, Calculating frequency term (tf)

b. Calculating the weighting term frequency (W_{tf})

$$W_{tf} = \begin{cases} 1 + log10tft, d, if\ tft, d > 0 \\ 0, \end{cases} \quad (1)$$

c. Calculating the document frequency (df)

d. Calculating the inverse document frequency (idf)

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

e. TF-IDF weighting is done with following formula :

$$W_{t,d} = Wtf_{t,d}\ x\ idf_t \quad (3)$$

Details :

$tf_{t,d}$ = term frequency

$Wtf_{t,d}$ = weighting from term frequency

df = how many documents contain the term

N = the total number of documents

$W_{t,d}$ = TF-IDF weighting.

2.4 K-Nearest Neighbor

The closest proximity or similarity to the item is used in the K-Nearest Neighbor algorithm. This method may save each feature vector and categorization of learning data since K-Nearest Neighbor is a machine learning methodology. The same features are calculated for the test data during the classification step (whose classification is unknown). The distance from this new vector to the learning data vector is calculated. Next will be taken the nearest K number. The newly classified points are predicted to be included in the most classifications of these points. Each near or Far Point will be calculated using the Euclidean Distance technique. [9].

The steps are carried out as follows :

a. Determine the amount of k values.

b. Determine the object distance for each group of data. The Euclidean distance equation is used to calculate distance.

P-ISSN: 1979-5661
E-ISSN: 2622-321X

$$D(x,y) = \sqrt{\sum_{1=1}^{n}(x_i - y_i)^2} \quad (4)$$

Details :

D = Distance

x = Data Train

y = Data Test

c. Sort all the data from nearest first to farthest

d. Classification results obtained by taking the nearest data label as much as the value of k and the labels with the most occurences will be selected as the result of prediction

2.4 Confusion Matrix

The principle of data mining and text mining accuracy calculations are typically performed using the confusion matrix approach. Recall, accuracy, and f1-score are the three outputs of the calculations carried out by this formula. [10]

1. Recall is the proportion of correctly identified positive cases. Formula of recall:

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

2. Precision is the proportion of cases with the correct positive result. Formula of Precision :

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

3. Accuracy is the ratio of the correctly identified case to the sum of all cases of the formula of accuracy. Formula of Accuracy :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

4. F1-Score is the harmonic mean value of recall and precision. Formula of F1-Score :

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (8)$$

The following table 1 is an example of a confusion matrix of binary consisting of only 2 classes :

Table 6. Confusion Matrix Example

	Prediction 0	Prediction 1
Label 0	TN (True Negative)	FP (False Positive)
Label 1	FN (False Negative)	TP (True Positive)

P-ISSN: 1979-5661
E-ISSN: 2622-321X

3. Result and Discussion

This study used data obtained by doing web crawling to get news articles from indonesian news website, there are 1200 data collected from web crawling using Selenium as a crawler, all of these data are divided into 300 automotive topics, 300 technology topics, 300 money topics, and 300 food topics. From the dataset used, the data is divided into a proportion of 80% for training data and 20% for testing data, the data division process is done randomly.

The Data had to go through the preprocessing stage, specifically change all letters to lowercase, and then delete irrelevant text, remove common words(stopwords), find words to their basic form, and at the last stage separate all sentences into separated words (tokenization). Following preprocessing stage, the TF-IDF weighting is carried out using the following formula (3). Once the weights have been determined, the classification is done using the k-Nearest Neighbors (KNN) algorithm. The k values used in this study is k=7. All of steps above are written as a code in Python programming language.

The evaluation results of the system are obtained by calculating the total size of data that are correctly classified divided by the sum of all test data. Classification results obtained are presented in Table 7 confusion matrix that displays information about the prediction by the system of each class.

Table 7. Confusion Matrix Result

PREDICTED	Automotive	Technology	Money	Food
ACTUAL				
Automotive	49	1	5	1
Technology	0	55	3	0
Money	1	1	57	2
Food	0	0	1	64

Referring to Table 7, The occurrence of errors or failures in the classification due to the similarity between the words composing a class with other classes as well as the wording of each data used as an example in the automotive class has the highest error rate because the wording of the automotive news topics similar to the wording of the money news topics, thus causing errors in the classification.

Evaluation was conducted to test whether the research has been running in accordance with the objectives of the study or not, the evaluation of this study was carried out with the calculation of accuracy, precision, recall, and F1-Score of the classification results displayed on the confusion matrix. By using formula of recall(5), precision(6),accuracy (7) and F1-Score(8), the results from results of the classification that has been done are shown in this table as follow:

Table 8. Evaluation Results

Accuracy	93.75%
Precision	94.09%
Recall	93.56%
F1-Score	93.71%

Referring to Table 8, it can be seen that each evaluation result from Classification using KNN. From these results obtained accuracy of 93.75%, Precision of 94.09%, Recall of 93.56%, and F1-Score of 93.71%.

4. Conclusion

Classification results obtained and presented in Table 7 confusion matrix is based on the results of classification by the system with the accuracy of 93.75%, Precision of 94.09%, Recall of 93.56%, and F1-Score of 93.71% for K-Nearest neighbor classification with value of k is k=7. The

occurrence of errors or failures in the classification due to the similarity between the words composing a class with other classes as well as the wording of each data used as an example in the automotive and money classes, the wording of the automotive news topics similar to the wording of the news money topics, thus causing errors in the classification, the number of datasets used can also affect the results of the classification, the more the amount of data used allows to improve the classification results.

References

[1] L. Septiani and Y. Sibaroni, “Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia,” *J. Linguist. Komputasional*, 2019, doi: 10.26418/jlk.v2i2.23.

[2] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2006. doi: DOI: 10.1017/CBO9780511546914.

[3] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. 2008. doi: 10.1017/cbo9780511809071.

[4] D. Ariadi and K. Fithriasari, “Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer,” *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.

[5] P. Putra, A. M. H. Pardede, and S. Syahputra, “Analisis Metode K-Nearest Neighbour (Knn) Dalam Klasifikasi Data Iris Bunga,” *J. Tek. Inform. Kaputama*, vol. 6, no. 1, pp. 297–305, 2022.

[6] A. P. Permana, K. Ainiyah, and K. F. H. Holle, “Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 3, pp. 178–188, 2021, doi: 10.14421/jiska.2021.6.3.178-188.

[7] D. N. D. Iriantoro, C. Dewi, and D. Fitriani, “Klasifikasi pada Penyakit Dental Caries Menggunakan Gabungan K-Nearest Neighbor dan Algoritme Genetika,” *J. Pengemb. Teknol. Inf. dan Ilmu Komputer; Vol 2 No 8*, Sep. 2017, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1715

[8] A. F. Hidayatullah, “Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016, doi: 10.14421/jiska.2016.11-01.

[9] A. K. Nikhath, K. Subrahmanyam, and R. Vasavi, “Building a K-Nearest Neighbor Classifier for Text Categorization,” *Int. J. Comput. Sci. Inf. Technol.*, 2016.

[10] Dwi Hartanti, Kusrini, and E. L. Taufiq, “PENERAPAN NAÏVE BAYES DALAMS PREDIKSI KETERCAPAIAN NILAI KRITERIA KETUNTASAN MINIMAL SISWA JUSIKOM PRIMA (Junal Sistem Informasi Ilmu Komputer Prima),” *Jusikom Prima*, vol. 2, no. 1, pp. 15–22, 2018.