A multi-label classification on topics of Indonesian news using K-Nearest Neighbor

To cite this article: Nikmah Isnaini et al 2019 J. Phys.: Conf. Ser. 1192 012027

View the article online for updates and enhancements.
A multi-label classification on topics of Indonesian news using K-Nearest Neighbor

Nikmah Isnaini1, Adiwijaya2, Mohamad Syahrul Mubarok3 and Muhammad Yuslan Abu Bakar4
1,2,3 School of Computing, Telkom University, Bandung, Indonesia
E-mail: 1 imaisna@students.telkomuniversity.ac.id,
2 adiwijaya@telkomuniversity.ac.id,
3 msyahrulmubarok.telkomuniversity.ac.id,
4 abubakar.yuslan@gmail.com

Abstract. News has become a basic human need along with technological and internet developments. This causes the process of disseminating information on the news that switched from print media to the digital era. Another problem that appears when classifying news is multi-label. Multi-label classification is different from single label classification. A single label classification will classify documents into one label only. While multi-label classification can group documents into more than one label. For example, news articles that discuss in detail the early detection of ovarian cancer with a bioinformatics approach may have more than one label such as health, bioinformatics, and women. In this paper, a classification model is developed that can identify classes in each multi-label news article using K-Nearest Neighbor. The advantages of K-Nearest Neighbor are algorithms that are very suitable for multi-label cases; even KNN can be superior to other classifiers. From the system created, the results of the value of system performance as measured by the size of the closeness are the comparison between Manhatten Distance, Euclidean Distance and Supremum Distance using the K = 11 parameters, resulting in a Hamming Loss value of 11.16%.

1. Introduction

News has become a basic human need along with the development of technology and the internet. The development of technology and the internet has led to the process of distributing information to the news from the way the delivery of the print media era to the digital era. News that is presented in the form of text on digital media is usually grouped based on its contents such as sports news, economics, science, and so on [22].

Another problem that arises when classifying news is multi-label. Multi-label classification differs from single label classification. The single label classification will classify a document into one label. While the multi-label classification can group documents into more than one label.

During this time, the classification of news documents only uses a single label. Multi-label classification differs from single label classification. Single label classification will classify a document into one document category only. Whereas multi-label classification can group documents into more than one category. For example, a news article in which detailed information on early detection of ovarian cancer with a bioinformatics approach can have more than one document label, namely: health, bioinformatics and women [4].

In this paper, we want to create a multi-label classification model on Indonesian language news topics using the K-Nearest Neighbor method. K-Nearest Neighbor is a supervised learning algorithm
where the results of the new test data are classified based on the majority of the category of the nearest K-neighbor. The KNN algorithm known as lazy learning classification, i.e. there is no learning process, algorithms that work by memorizing all training data and classifying the training data. KNN is very suitable for multi-label cases. For example, for assigning functions to genes based on expression profiles, some researchers find that KNN outperforms SVM, which is a much more sophisticated classification scheme [23].

2. Related work

2.1. Related research

Previous research that has conducted research related to the multi-label classification of text data, whether the data is news or not and use Indonesian language or not, Sigit Bagus Setiawan about “Classification of Indonesian News Topics using Weighted K-Nearest Neighbor” [18]. This study aims to create a system that can categorize every news in Indonesian in the class that should be. The choice of giving weight to the weighted k-nearest neighbor is quite influential on system performance. And the higher value of k will initially increase the percentage of truth in the classification; then if k is at the optimum point, then the percentage will tend to decrease.

Reynaldi Ananda Pane and Nanang Saiful Huda researching "A Multi-label Classification on Topics of Quranic Verses in English Translation using Multinomial Naive Bayes” [13]. In this research, the use of prior probability values that are purely obtained from training results produces a better classification performance than uniform probability values. It shows that the prior variation of each class is important for classifier performance.

Siti Nur Asiyah dan Kartika Fithriasari researching “Classification of Online News Using Support Vector Machine and K-Nearest Neighbor Methods” [7]. Both of these methods will be compared to find out the best classification accuracy. The results of this research that the linear kernel SVM and polynomial kernel produce the best classification accuracy is a polynomial kernel.

Asriyanti Indah Pratiwi was researching "In the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis” [14]. In the research, how to choose the optimal features and classification. Handle a large number of features and provide better classification/grouping of sentiments. The proposed method reduces more than 90% of unnecessary features while the proposed classification scheme reaches 96% accuracy of grouping sentiments.

Another case regarding classification, namely "spect-based Sentiment Analysis to Review Products Using Naive Bayes” [12]. This research aims to analyze and extract sentiment polarity in product reviews based on certain aspects of the product. It can be concluded that the Naive Bayes classification works well for aspect-based sentiment analysis with the best F1-Measure of 78.12%. Then “Classification of hadith into the positive suggestion, negative suggestion, and information” [5]. This research aims to classify the translation of hadith languages into three categories using the machine learning approach and produced SVM with a linear kernel reaching the best F1-Score of 0.88.

“An implementation of support vector machine on sentiment classification of movie reviews” [24] research aims to classify sentiments on film review documents using SVM. Based on the research, the conclusion is the more data is used as training data, F1-Score results are better to classify.

2.2. News and text classification

News, according to Big Indonesian Dictionary (KBBI) is a story or description of a warm event or event. News must be in accordance with the reality, not made up (fictitious), and the latest / latest. The news is one way to communicate through important, latest, and interesting events. We can find news anywhere like in newspapers, magazines, internet, television, radio, even in school mading. News must contain elements of 5W + 1H (What, Who, When, Where, Why, and How) so that the reader can find out more about an event. In English "news" is called news [17].
Text Classification is the process of determining the category of a text document to the characteristics of the text [16] [8] [11] [10] [6]. The main problem in text classification depends on words that appear in documents that affect features or dimensions so that the frequency of words affects the results of the classification. Before the classification process is required the transformation of text data into vector documents. Figure 1 is the data transformation process.

![Figure 1. Flowchart transforming text data into vector data.](image)

This is an explanation of figure 1. Input data in the form of text or documents, need to be preprocessed. This stage aims to clean the noise dataset to be ready for use in the next stage [13]. The first step is case folding is the process of cleaning data where all sentences are converted into lowercase letters [21]. Tokenization is the process of dividing text in sentences in a dataset into single word pieces. Stopword Removal is helped words such as person pronouns, conjunctions, and words that have no function in the sentence will be removed [19]. The last step of the preprocessing is stemming, where is the process of getting the basic words from a word in a sentence by separating the prefix and suffix from each word.

2.3. Multi-label classification

Text classification is part of supervised learning the purpose of classifying data based on its label. Each data can be grouped in one label, several labels or without labels. There are two types of classification models namely single-label only has a label length = 1 and multi-label, that has more than two label length which is the focus of this paper. Previous research on multi-label has been carried out by Min-Ling Zhang [25]. The approach used in multi-label cases is one of the transformation problems. The problem transformation approach handles multi-label issues into single labels by describing the problem of multi-label learning into independent binary classification problems, where each binary classification problem corresponds to the label that might be labeled as binary relevance [25].

TF-IDF process is a feature extraction process that is used to transform string attributes into weights. TF-IDF value is seen from the number of words that appear in a document. Term frequency is the frequency of the appearance of a term in the document concerned. The greater the number of occurrences of a term (high TF) in the document, the greater the weight or the greater suitability value [9]. The calculation can be seen in equation 1.

\[
\text{TF} = 1 + \log(f_{td})
\]

Where \( f_{td} \) is term frequency (t) on a document (d). While Inverse Document Frequency is the relationship between the availability of a term in all documents. The smaller the number of documents containing the terms in question, the greater the IDF value [9]. The calculation can be seen in equation 2.
\[ IDF = \log \frac{N}{df_j} \]  

Where \( N \) is the number of all documents in the collection while \( df_j \) is the number of documents containing term \( t_j \). The weight of the term is calculated using the TF-IDF size in equation 3.

\[ W = TF \cdot IDF \]  

Where \( tf \) is the appearance of the term of each document, and is the weight of the document against the word or the weight of the key against the document.

The method used in a multi-label classification on topics of Indonesian news is K-Nearest Neighbor. K-Nearest Neighbor is that \( k \) is a variable that shows the number of closest neighbors that affect other data. K-nearest neighbor is a supervised learning algorithm where the results of the new test data are classified based on the majority of the category of the nearest K-neighbor. In this case, the distance calculation uses Minkowski Distance, which includes Manhattan Distance, Euclidean Distance, and Supremum Distance [3].

### 2.4. Proximity measure

This paper does not focus on analyzing the best measure of similarity, because it depends on certain problems. However, focusing on the comparison of various distance measures that can explain the performance of each method used. The Proximity Measure method used is Minkowski Distance, which includes Manhattan Distance, Euclidean Distance, and Supremum Distance. Explanations and formulas are as follows.

**Manhattan Distance**

Manhattan Distance also referred to as "city block distance" is the amount of distance from all objects [20]. For two data points \( x_i \) and \( y_j \) in p-space dimensions, Manhattan Distance between these points is defined as follows:

\[ d(x_i; x_j) = \sum_{n=1}^{p} |x_{in} - x_{jn}| \]  

\( x_{in} \) : i test data on the nth variable  
\( x_{jn} \) : j training data on the nth variable  
\( d(x_i; x_j) \) : Manhattan distance  
\( p \) : data dimension

**Euclidean Distance**

Euclidean Distance is the most commonly used distance measurement on numerical data [20]. For two data points \( x_i \) and \( y_j \) in p-space dimensions, Euclidean Distance between these points is defined as follows.

\[ d(x_i; x_j) = \sqrt{\sum_{n=1}^{p} (x_{in} - x_{jn})^2} \]  

\( x_{in} \) : i test data on the nth variable  
\( x_{jn} \) : j training data on the nth variable  
\( d(x_i; x_j) \) : Euclidean distance  
\( p \) : data dimension
Supremum Distance
Supremum Distance is also called "Maximum Distance". Defined as the maximum value of the distance of its attributes [20]. For two data points \( x_i \) and \( y_j \) in \( p \)-space dimensions, the Supremum Distance between points is defined as follows.

\[
d(x_i; x_j) = \max(j|x_{i1} - x_{j1}|; \ldots; j|x_{ip} - x_{jp}|)
\]

\( x_{in} \): i test data on the nth variable
\( x_{jn} \): j training data on the nth variable
\( d(x_i; x_j) \): Supremum distance
\( p \): data dimension

2.5. K-Fold cross validation
K-fold is one of the popular Cross Validation methods by folding data as much as k and repeating the experiment as much as k. In K-Fold Cross Validation, data is partitioned into the same segment [15], where one segment is used as training data, and the other segment is used as testing data. The difference is, in this method, every data contained in the dataset must be used as training data and also testing data. With Cross Validation all data will be divided into k groups. With the size of the number of datasets divided by the number of k that will be used [13].

2.6. Hamming Loss
Hamming Loss is used to evaluating the performance of the system in multi-label classification. This process works by comparing the actual output with the output released on the system, then calculating each error and divided by the number of labels and data. The smaller the value of Hamming Loss, the better the classification model built, and vice versa. The Hamming Loss formula can be seen below.

\[
\text{Hamming Loss} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} \left[ \sum_{q} y_{j}^{(i)} y_{j}^{(i)} \right]
\]

\( N \): the amount of data analyzed
\( L \): the number of labels on the data analyzed
\( y_{j}^{(i)} \): output label
\( y_{j}^{(i)} \): target label

3. System design
The input data used in this system is Indonesian news text data taken from certain sites. There are thirteen labels that will be classified by the system. Before entering the classification stage, preprocessing is carried out which has been described in figure 1. Table 2 is an explanation of a number of labels from label 1 to label 13, and in table 5 is an example of the dataset used.
### Table 1. Class label of Indonesian news.

| Class Label | Explanation |
|-------------|-------------|
| 1           | Politic     |
| 2           | Law         |
| 3           | Economy     |
| 4           | Social      |
| 5           | Culture     |
| 6           | Technology  |
| 7           | Life Style  |
| 8           | Sport       |
| 9           | Entertainment |
| 10          | Education   |
| 11          | Defense     |
| 12          | Health      |
| 13          | Others      |

### Table 2. Dataset of Indonesian news and label.

| No | News Article                                                                                                                                                                                                 | Class Label |
|----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|
| 1  | The dynamics of dismantling the 2019 presidential candidate pair continues. Whoever is a floating figure can be juxtaposed. Moreover, a number of continually emerging are considered alternative candidates.                                    | 1           |
| 2  | Today (30/11) Butet Kartarejasa will present 138 visual works at the exhibition at the National Gallery, Jakarta. The visual works take the painted media. This way of art is deliberately done by Butet to conduct social criticism of the community and the government. | 5,9         |
| 3  | The Vice President of the Republic of Indonesia, Jusuf Kalla delivered a written warning to the Minister of Youth and Sports Imam Nahrawi to allocate more funds to prepare for a number of sports ahead of the 2018 Asian Games. | 1,3,8       |
| ...| ...                                                                                                                                                                                                          | ...         |
| 178| Red plate cement producer, PT Semen Indonesia Tbk (SMGR) decided to distribute profits or dividends of 40 percent of the total net profit of Rp 2.01 trillion                                                                 | 1           |

### 4. Proposed system design

Researchers proposed a system design to classify multi-label of Indonesian news using K-Nearest Neighbor. To build this system design, there are several processes including preprocessing, TF-IDF, K-Fold Cross Validation, KNN model, and Model Evaluation. Figure 2 is a proposed design.
Based on figure 2, system flow in general in this study starts from a dataset in the form of text going through the preprocessing stage, where there are four stages in it namely case folding, tokenization, stopword removal, and stemming. Explanation of the preprocessing stage is found in point 2.3. The next process is the feature extraction using the TF-IDF (Term Frequency-Inverse Document Frequency) method which aims to weight the vector data so that it becomes a matrix [1] [2]. The data that has been weighted, then carried out K-Fold Cross Validation, is the process of dividing the dataset into training data and test data. For each training data fold is used to create a classification model. Test data is data used to evaluate the performance of the model. The test data is predicted using the K-Nearest Neighbor algorithm by classifying 13 classes in the dataset into binary classifications. Calculation of the distance used is Manhattan Distance, Euclidean Distance, and Supremum Distance. The output in the system is a set of binary numbers from the 1st label to the 13th label. After that the set of predictions was obtained, for example, the 19th news of the results of the prediction of the 5th and 7th classes, namely 0,0,0,0,1,0,1,0,0,0,0,0,0. After that, the system will evaluate its performance using the Hamming Loss method.

5. Results and discussion
5.1. Analysis of test results
The goal of system testing is to classify the news into 13 existing labels. The system built has three main processes, namely preprocessing, feature extraction, and classification. To get optimal performance from the system, testing of k and proximity parameters is carried out.

5.1.1. Class label testing.
How to build a K-Nearest Neighbor model to solve a multi-label classification on topics of Indonesian news using K-Nearest Neighbor. At this point, explain the design of the K-Nearest Neighbor model that is how the model predicts how many labels will be tested and what data is included in any label. In this case, there are 13 labels. The K-Nearest Neighbor model can predict the number of labels and includes any labels, namely KNN to recognize label 1 or not, KNN to recognize label 2 or not, to KNN to identify label 13 or not. Check one by one the testing data for all training data. After that, the label is changed to binary to calculate the dominant label. Then the binary dominant is calculated (dominant 0 or 1). The number of K-Nearest Neighbor models follows a number of labels on training
data. In table 5 is the test results of the dataset used, as discussed in point 2.4 the approach used for classification is the binary relevance of the problem transformation method.

Table 3. News classification results.

| Document to | Label | Binary |
|-------------|-------|--------|
| 39          | 4     | 0 0 0 1 0 0 0 0 0 0 0 0 0 |
| 37          | 1,4   | 1 0 0 1 0 0 0 0 0 0 0 0 0 |
| 168         | 3     | 0 0 1 0 0 0 0 0 0 0 0 0 0 |
| 176         | 1,9,8 | 1 0 0 0 0 0 1 1 0 0 0 0 |
| 91          | 1,1,1 | 1 0 0 0 0 0 0 0 0 1 0 0 |
| 146         | 2,4   | 0 1 0 1 0 0 0 0 0 0 0 0 0 |
| 137         | 3,8   | 0 0 1 0 0 0 0 1 0 0 0 0 |
| 73          | 6,9   | 0 0 0 0 0 1 0 0 1 0 0 0 |
| 148         | 3,5,6 | 0 0 1 0 1 1 0 0 0 0 0 0 |
| 40          | 2,3,9 | 0 1 1 0 0 0 0 0 1 0 0 0 |
| . . .       | . . . | . . . |
| 128         | 4,7   | 0 0 0 1 0 0 1 0 0 0 0 0 |

Based on the table above, it was found that the results of testing the system from the Indonesian language news classification were able to classify news into more than one label. From 178 data, divided into five fold. In the table above, the test data is folded 1 (1st document to 36th document) and the training data is fold 2 to fold 5. One news document is tested one by one against all training data; then the system will classify its automatic.

5.1.2. Testing parameters k.
The researcher assumes that the difference in k parameters and the difference in proximity measures used will affect the system performance. This is caused by the parameter k which affects the distance between the data to the system being built.

The following is a test table to determine optimal accuracy using the proposed k parameter. For the parameter value k used is between 5 to 23, while the proximity to be tested is Manhattan, Euclidean, and Supremum.

Table 4. Comparison of Hamming Loss value with k value on KNN.

| Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme | Testing Scheme |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Fold           | K=5            | K=7            | K=9            | K=11           | K=13           | K=15           | K=17           | K=19           | K=21           | K=23           |
| 1              | 0.1261         | 0.1090         | 0.1154         | 0.1154         | 0.1132         | 0.1154         | 0.1197         | 0.1218         | 0.1218         | 0.1218         |
| 2              | 0.1154         | 0.1175         | 0.1218         | 0.1239         | 0.1282         | 0.1282         | 0.1282         | 0.1282         | 0.1282         | 0.1282         |
| 3              | 0.1197         | 0.1111         | 0.1068         | 0.1068         | 0.1132         | 0.1111         | 0.1132         | 0.1132         | 0.1132         | 0.1132         |
| 4              | 0.0983         | 0.0940         | 0.1046         | 0.0940         | 0.0983         | 0.1026         | 0.1047         | 0.1047         | 0.1047         | 0.1047         |
| 5              | 0.1154         | 0.1154         | 0.1111         | 0.1176         | 0.1176         | 0.1176         | 0.1154         | 0.1154         | 0.1154         | 0.1154         |
| Mean           | 0.1150         | 0.1094         | 0.1119         | 0.1116         | 0.1141         | 0.1150         | 0.1158         | 0.1167         | 0.1167         | 0.1167         |

Based on the observations above, it was found that the minimum hamming loss value in parameter k = 11 was 0.1116. The division of the dataset, in this case, is 4: 1, which is 4 fold for training data and 1 fold for testing data. The range of observed k values is from 5 to 23. Apparently, after observing, the optimal k value for the case of Indonesian language news classification is 11. From the results of this experiment can be analyzed to determine the selection of k values to be important because it will affect the performance of the KNN algorithm. The value of k is too small, then the classification results will be more affected by noise. On the other hand, if the k value is too high it will reduce the effect of noise on the classification [23]. Therefore, the value of k is influenced by the
proximity between neighbors in Indonesian news data. The system recommendation for further development is to expand the Indonesian language dataset because with more datasets it is expected to add references from KNN and improve system performance.

5.1.3. Proximity measure testing.
After getting the optimal k parameter that has been tested in test 1, it is obtained k by 11. In this section, it will compare the proximity measure used to measure the proximity between data with the selected parameters. The distance measurement used is Manhattan Distance, Euclidean Distance, and Supremum Distance.

Table 5. Comparison of Hamming Loss values on proximity measure.

| Fold | Manhattan Distance | Euclidean Distance | Supremum Distance |
|------|--------------------|--------------------|-------------------|
| 1    | 0.1154             | 0.1175             | 0.1239            |
| 2    | 0.1239             | 0.1282             | 0.1346            |
| 3    | 0.1068             | 0.1175             | 0.1389            |
| 4    | 0.0940             | 0.1303             | 0.1175            |
| 5    | 0.1176             | 0.1448             | 0.1267            |
| Mean | 0.1116             | 0.1277             | 0.1283            |

Table 5 displays a comparison of proximity measure using the parameter k = 11. Based on table 6, you can see the comparison between Manhattan Distance, Euclidean Distance, and Supremum Distance. When compared with the Supremum Distance which produces the minimum distance but the worst performance, it is due to the very small distance of the data, it will be affected by noise, resulting in a performance of 0.1283. When compared with Euclidean Distance which produces a greater distance than the Supremum Distance, but it is still affected by noise, resulting in a performance of 0.1277 which causes the performance of Hamming Loss not too good when the parameter K = 11. The results show that the minimum Hamming Loss average is 0.1116 with the k = 11 parameters calculated using Manhattan Distance. With the results of calculating a distance that is much greater than Euclidean Distance and Supremum Distance. The greater the distance generated between data, the better the Hamming Loss performance will be produced. As explained in point 2.9, that Hamming Loss works by comparing the actual output with the output issued to the system, and calculating each error then divided by the number of labels and data. Distribution of datasets will also affect system performance. With a dataset of 178 data, the division is 4: 1, which is 4 fold for training data and 1 fold for testing data. The more data/references in the training/learning process, the better the classification process/the better the system performance.

6. Conclusion
Based on the results of the testing and analysis obtained, the K-Nearest Neighbor method can be used to a multi-label classification on topics of Indonesian news. In the tests that have been done, there is no data that after being classified does not have a label or has a label that is more than 3. The method used to measure the distance is Minkowski Distance, which includes Manhattan Distance, Euclidean Distance, and Supremum Distance. For the case of multi-label classification, the performance measurement of the system uses Hamming Loss. Of the three types of proximity measure used, Manhattan Distance provides the best performance compared to Euclidean Distance and Supremum Distance. The range of k values tested is 5-23. Using the parameter k = 11 produces a Hamming Loss value of 0.1116. This happens because it is influenced by the proximity of the neighbors between the news text data. To determine the selection of k values and the type of proximity measure used, it is important because it will affect the performance of the K-NN algorithm. The value of k and the result
of proximity is too small, then the classification results will be more affected by noise. On the other hand, if the k value is too high it will reduce the effect of noise on the classification. The system recommendation for further development is to expand the Indonesian language dataset because with more datasets it is expected to add references from K-NN and improve system performance. Use of libraries that have more vocabulary for Stopword Removal and Stemming than Sastrawi also using feature selection to reduce some unimportant features.

References

[1] Adiwijaya Aplikasi Matriks dan Ruang Vektor Graha Ilmu Yogyakarta 2014
[2] Adiwijaya Matematika dan Aplikasinya Alfabeta Bandung 2016
[3] Adiwijaya Aulia Masyithah Nur Mubarok M Syahrul Novia W Untari and Nhita Fhira A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronunciation classification system In Information and Communication Technology (ICoIC7) 2017 5th International Conference on pages 1–5 IEEE 2017
[4] Arianto Rio Bayu and Kurniawati Lisa Yuli Kategorisasi dokumen teks secara multi label menggunakan fuzzy c-means dan k-nearest neighbors pada artikel berbahasa indonesia JUTI: Jurnal Ilmmiah Teknologi Informasi 11(1):23–28 2013
[5] Faraby Said Al Jasin Eliza Rachmawati and Kusumaningrum Andina Classification of hadith into positive suggestion, negative suggestion, and information In Journal of Physics: Conference Series volume 971 page 012046 IOP Publishing 2018
[6] Arifin Azmi Hafizha Rahman Zainal Mubarok Muhammad Syahrul and Adiwijaya Adiwijaya. Learning struktur bayesian networks menggunakan novel modified binary differential evolution pada klasifikasi data In Indonesia Symposium on Computing (IndoSC) 2016 2016
[7] Asiyah Siti Nur and Fithriansari Kartika Klasifikasi berita online menggunakan metode support vector machine dan k-nearest neighbor Jurnal Sains dan Seni ITS 5(2) 2016
[8] Firmansyah Achmad Fauzi Bagus and Pramana Setia Ensemble based gustafson kessel fuzzy clustering In Journal of Data Science and Its Applications (JDSA) volume 1 pages 1–9 Journal of Data Science and Its Applications 2018
[9] INFORMATIKALOGI Pembobotan kata atau term weighting tf-idf 2018
[10] Izzaty Al Mira Khonsa Mubarok M Syahrul Huda Nanang Saiful and Adiwijaya Adiwijaya A multi-label classification on topics of quranic verses in english translation using tree augmented naïve bayes In Information and Communication Technology (ICoICT) 2018 6th International Conference on volume 5 pages 1556–1561 IEEE 2018
[11] Manuel Bram and Tricahyono Dodie Classifying electronic word of mouth and competitive position in online game industry In Journal of Data Science and Its Applications (JDSA) volume 1 pages 20–27 Journal of Data Science and Its Applications 2018
[12] Mubarok Mohamad Syahrul Adiwijaya and Aldhi Muhammad Dwi Aspect-based sentiment analysis to review products using naïve bayes In AIP Conference Proceedings volume 1867 page 020060 AIP Publishing 2017

[13] Pane Reynaldi Ananda Mubarok Mohamad Syahrul Huda Nanang Saiful and Adiwijaya Adiwijaya A multi-lable classification on topics of quranic verses in english translation using multinomial naive bayes In Information and Communication Technology (ICoICT. IEEE) 2018 6th International Conference on IEEE 2018

[14] Pratiwi Asriyanti Indah et al On the feature selection and classification based on information gain for document sentiment analysis Applied Computational Intelligence and Soft Computing 2018 2018

[15] Refaeilzadeh Payam Tang Lei and Liu Huan Cross-validation 2008

[16] Ridok Achmad and Latifah Retnani Klasifikasi teks bahasa indonesia pada corpus tak seimbang menggunakan nwknn Proceedings Konferensi Nasional Sistem dan Informatika (KNS&I) 2015

[17] Sasrawan Hedi Natural language processing (NLP) 2016

[18] Setiawan Sigit Bagus Adiwijaya Adiwijaya and Mubarok Mohamad Syahrul Klasifikasi topik berita berbahasa indonesia menggunakan weighted k-nearest neighbor eProceedings of Engineering 5(1) 2018

[19] Sitompul Desi Adiwijaya Adiwijaya and Mubarok Mohamad Syahrul Analisis sentimen level kalimat pada ulasan produk menggunakan bayesian networks eProceedings of Engineering 4(3) 2017

[20] Supianto Ahmad Afif Pengenalan pola klasterisasi data 2014

[21] Halteren Hans van Syntactic wordclass tagging volume 9 Springer Science & Business Media 1999

[22] Widodo Pangestu Putra Januar Adi Afiadi Suwanto Arifin Agus Zainal and Herumurti Darlis Klasifikasi kategori dokumen berita berbahasa indonesia dengan metode kategorisasi multi-label berbasis domain specific ontology Jurnal Ilmiah Teknologi Informasi Terapan 2(2) 2016

[23] Wu Xindong Kumar Vipin Quinlan J Ross Ghosh Joydeep Yang Qiang Motoda Hiroshi McLachlan Geoffrey J Ng Angus Liu Bing Philip S Yu et al Top 10 algorithms in data mining. Knowledge and information systems 14(1):1–37 2008

[24] Yulietha IM Faraby SA Widyaningtyas WC et al An implementation of support vector machine on sentiment classification of movie reviews In Journal of Physics: Conference Series volume 971 page 012056 IOP Publishing 2018

[25] Zhang Min-Ling Li Yu-Kun Liu Xu-Ying and Geng Xin Binary relevance for multi-label learning: an overview. Frontiers of Computer Science pages 1–12 2018