Multi-label classification of Indonesian news topics using Pseudo Nearest Neighbor Rule

Reza Agung Pambudi 1, Adiwijaya 2 and Mohamad Syahrul Mubarok 3

1,2,3 School of Computing, Telkom University, Bandung

E-mail: 1 rezaagungp@students.telkomuniversity.ac.id, 2 adiwijaya@telkomuniversity.ac.id, 3 msyahrulmubarok@telkomuniversity.ac.id

Abstract. News is a form of text data that must be categorized to facilitate retrieval of information for the reader. One problem that arises when categorizing news is the many topics that news can discuss, which is known as a multi-label condition. To solve this problem, a system that can perform multi-label classification using a Pseudo Nearest Neighbor Rule (PNNR) algorithm—a variant of the k-Nearest Neighbor (k-NNR) algorithm—was developed in this study. This system yielded a cross-validation error of 0.1495, measured using the hamming loss method via Cosine proximity. From the experiment, it can be concluded that the performance of the PNNR algorithm is influenced by the type of proximity used and the number of nearest neighbors.

1. Introduction
A problem that can appear when categorizing news is the multi-label problem, a condition where a piece of news can contain more than one topic. When categorizing news, automation is needed because of the excessive amount of news that could result in too much time spent using human power. In addition, biasness could occur if one is not objective [3]. The increased amount of news released by companies can also impact the viability of the news, so a reliable and quick classification model is needed for categorizing news [12] to facilitate the retrieval of information for the reader [15]. The paragraph below shows an example of news with multi-label characteristics.

"Xiaomi resmi hadir di pasar Eropa dengan memasarkan dua handset andalannya, Xiaomi Mi A1 and Mi Mix 2 di Madrid, Spanyol. Di Spanyol, Mi A1 akan dibanderol seharga 299 euro atau sekitar Rp 3,6 jutaan. Sean-dgkan Mi Mix 2 dengan layar edge-to-edge akan dibanderol dengan harga 499 euro atau sekitar Rp 7,8 jutaan”-Kompas [14].

In the quoted news, two things are discussed: the first is the economy because the company in the news successfully opened up a new market; the second is technology because the company had released a new smartphone product. When classifying text using Natural Language Processing (NLP), data mining and machine learning need to be combined so the right patterns can be detected [1, 17]. Previous studies have used various methods to solve text classification problems. Nikhath, Subrahmanyam, and Vasavi [10] used the k-Nearest Neighbor Rule (k-NNR) method in their study.
Qin and Eang [16] solved the multi-label text classification problem using the Support Vector Machine method, while Sowmya, Chetan, and Srinivasa [1] used the Rochhio algorithm.

In this research, the Pseudo Nearest Neighbor rule was used to classify multi-label Indonesian news topics. The Pseudo Nearest Neighbor Rule (PNNR) is a supervised learning algorithm. It is also a variant of the k-Nearest Neighbor algorithm. Generally, PNNR is used to solve single-label problems. However, in this research, PNNR is used to solve a multi-label classification problem.

2. Related works

Many researcher has been doing research in the text classification field, such as in [18, 19, 20, 21]. Previous research on multi-label classification of text includes the work of Qin [16], who studied multi-label classification using SVM. The purpose of the research was to determine the performance of the SVM method in multi-label text classification where the total label for each data was two. From the SVM method used, the author concluded that the 1-a-rMC algorithm was more suitable for a small dataset with numerous classes, while the HSMC algorithm could be used for big datasets with numerous classes.

Nikath [10] investigated classification of multi-label text data using the k-NNR algorithm. The study focused on e-mail classification, which is a form of text classification application. The paper concluded that the k-NNR method is the easiest algorithm to modify to improve accuracy. However, the weakness of this method is the number of k in the k-NNR algorithm, which can greatly affect accuracy.

Afrianto [2] performed multi-label text classification for an Indonesian article. In his research, he compared the performance of two algorithms, namely the Neural Network and SVM algorithms. The results showed that classification using a Neural Network and WordNet yielded better performance than using SVM. However, the Neural Network required a longer running time than SVM.

Sowmya [1] researched multi-label text classification using the Rochhio algorithm, where two algorithms, k-NN and Rochhio, were run and the results compared. The paper also showed that the Rochhio algorithm is more efficient and accurate than the k-NNR algorithm. Furthermore, the Rochhio algorithm was easier to parallelize.

Mishnu, Zaman, and Sadia [13] conducted research using a supervised learning model for text classification. The paper aimed to determine the best model classification from a few supervised learning algorithms. The research concluded that the best algorithm with the best performance for text classification was the Neural Network using a Backpropagation algorithm. This algorithm gave the best accuracy but required the longest running time for the training process.

3. PNNR for multi-label data

The input dataset text used in this system is Indonesian news crawled from selected sites. The maximum total label for each data is three. The data example can be seen in table 1. The system flowchart for this study is presented in figure 1.

| ID | News content | Label | URL |
|----|--------------|-------|-----|
| 1  | <news content> | Economy, Education | <link> |
| 2  | <news content> | Politics, Law, Economy | <link> |
| 3  | <news content> | Defense and Security, Entertainment | <link> |
| 4  | <news content> | Sports, Technology | <link> |
| 5  | <news content> | Sports, Culture | <link> |

The first process in the proposed system is preprocessing. Input data text will perform tokenizing, which is the process of separating a text document into words, where separation is based on punctuation or spaces. After that, the stopwords or words that do not represent the content are omitted.
Next, the stemming process is performed to minimize the form of grammatical differences for a word and to cripple the inflection and derivatives of a word [2, 7]. The final phase of preprocessing is indexing, which is the calculation of the appearance of a word, so that data can be represented as a vector. The results from this step are similar to that of table 2. Features A to I are variables that represent a term (word).

![System flowchart.](image)

**Figure 1.** System flowchart.

**Table 2.** Data vector for input.

| ID | Attribute/Feature | Label                              |
|----|-------------------|------------------------------------|
|    | A | B | C | D | E | F | G | H | I |                      |
| 1  | 5 | 6 | 7 | 8 | 9 | 0 | 0 | 0 | 0 | Economy, Education   |
| 2  | 2 | 0 | 0 | 0 | 2 | 2 | 9 | 8 | 0 | Politics, Law, Economy |
| 3  | 0 | 0 | 0 | 0 | 0 | 9 | 8 | 7 | 0 | Defense and Security, Entertainment |
| 4  | 8 | 0 | 0 | 7 | 0 | 0 | 0 | 5 | 0 | Sports, Technology   |
| 5  | 0 | 3 | 0 | 6 | 7 | 8 | 0 | 9 | 0 | Sports, Culture      |

After preprocessing the data, the next process is feature extraction using the TF-IDF (Term Frequency Inverse Document) method, the higher the TF-IDF score for a word, the more relevant the word is to the document [8]. After that, a k-fold cross-validation is performed, where the data is split into two types i.e. training data and testing data. Before classification, label-based transformation is performed [4]. The purpose of data label transformation from multi-label to single label is to associate a multi-label class with its instances. Then, the testing data is predicted using the Pseudo Nearest Neighbor algorithm, where the data is classified based on the surrounding class of data that has the closest distance to the testing data distance [17, 11, 5], calculated using eq. (1):

$$d_{total,i} = \sum_{j=1}^{K} u_j d_{ij}$$  \hspace{1cm} (1)

Where k is the parameter score k, $u_j$ is the weight of the j neighbor, and $d_{ij}$ is the proximity score between number i training data, which is also a neighbor of the number j testing data that will be predicted. The score for $u_j$ is the weight of the neighbor number j calculated using eq. (2):

$$u_j = \frac{1}{j}$$  \hspace{1cm} (2)
The variable \( j \) is the \( j \)-th serial number of the total parameters that \( k \) has specified. The value of \( d_{ij} \) is the proximity value between training data and testing data. Suppose we have training data (A) and testing data (B) with a number of attributes (n). Calculations using the Euclidean, Manhattan, and Cosine proximity is written sequentially as Equations (3), (4), and (5), respectively.

\[
d(A;B) = \sum_{i=1}^{n} (A_i - B_i)^2
\]

(3)

\[
d(A;B) = \sum_{i=1}^{n} A_i B_{ij}
\]

(4)

\[
\cos(q) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}
\]

(5)

When the multi-label testing data that has undergone the process of classification is done spawning, the PNNR algorithm will count the class in the dataset. The function of each PNNR in spawn is to perform binary classifications for the testing data. This includes the \( n \)-th class, as shown in fig. 2.

![Figure 2. Multi-label classification process using PNNR.](image)

The output of the model is a set of binary numbers from class 1 to \( n \)-th class e.g., if the value of \( n \) is five, then, the set of predicted results of data 1,0,0,0,1 is obtained, where the labels for the testing data will be one and five. At the end of the system process, the model is evaluated using a hamming loss method. The smaller the value of the hamming loss, the better the model performance [9, 6].

4. Results and analysis
The main purpose of system testing is to classify the news into existing classes. To achieve good performance from the system, the parameters of the nearest neighbor number (k) and type of proximity are tested. The minimum and maximum tested parameter values for \( k \) are two and four, respectively, because the distribution of the amount of data labeled other than label 13 on one of the most partitions is four. Meanwhile, the type of proximity to be tested is the Euclidean, Manhattan, and Cosine...
proximities. Table 3 presents the values hamming loss obtained from the result of k parameter test and types of proximity used.

**Table 3.** The value of hamming loss for testing k and proximity parameters.

| Partition | k | Proximity   | Euclid | Manhattan | Cosine |
|-----------|---|-------------|--------|-----------|--------|
| Partition A | 2 | 0.1815 | 0.1772 | 0.1538 |
| Partition B | 2 | 0.1668 | 0.1832 | 0.1461 |
| Partition A | 3 | 0.1815 | 0.1789 | 0.1538 |
| Partition B | 3 | 0.1685 | 0.1590 | 0.1452 |
| Partition A | 4 | 0.1815 | 0.1798 | 0.1538 |
| Partition B | 4 | 0.1824 | 0.1590 | 0.1478 |

From table 3, we can obtain the cross-validation error using the hamming loss metrics for each value k and the type of proximity tested. Table 4 describes the values of the cross-validation error test result.

**Table 4.** Value of cross-validation error using hamming loss.

| Proximity | Hamming loss score |
|-----------|--------------------|
|           | k = 2 | k = 3 | k = 4 |
| Euclidean | 0.1742 | 0.1750 | 0.1819 |
| Manhattan | 0.1802 | 0.1690 | 0.1694 |
| Cosine    | 0.1500 | 0.1495 | 0.1508 |

Based on table 4, the performance of the PNNR algorithm yields a hamming loss value that is different for every kind of proximity and k parameter tested. The table is used to find the mean value of hamming loss. The best parameters to be tested are shown in figure 3.

**Figure 3.** Mean value of hamming loss based on proximity.

In figure 3, the PNNR algorithm uses the Cosine proximity similarity to give the smallest hamming loss value compared to the Euclidean and Manhattan proximities, which means that the Cosine proximity similarity gives higher accuracy because it gives less score miss-classification.
Based on figure 4, when testing the parameters, the number of nearest neighbors from 2 to 4 was used. The smallest hamming loss value was achieved with a nearest neighbor number of 3. When testing, the number of nearest neighbors greater than 4 was not selected because of the uneven distribution of data on both partitions, so it was still possible to use smaller hamming loss values to test the value of the number of nearest neighbors.

![Mean hamming loss based on parameters k](image)

Figure 4. Mean value of hamming loss based on parameters of the nearest neighbor number (k).

5. Conclusion
Based on the results of testing and analysis obtained, the Pseudo Nearest Neighbor rule (PNNR) algorithm can be used to perform text classification of multi-label Indonesian news. The PNNR algorithm value for parameter k and type of proximity used affected the performance of the algorithm. Of the three types of proximity tested, the Cosine proximity provided the best performance compared to the Manhattan and Euclidian proximities. Future research should utilize addition of data so that the distribution of data can be partitioned in a balanced manner and the value for the variable parameter k, which is tested, will be larger. Future research should also use a library with more vocabulary.

References
[1] C B J Sownya and Srinivasa Larage scale multi-label text classification of a hierarchical dataset using rocchio algorithm International Conference on Computational Systems and Information Systems for Sustainable Solutions 2016
[2] A R Bayu and L Y Kurniati Kategorisasi teks secara multi-label menggunakan fuzzy c-means dan k-nearest neighbors pada artikel berbahasa Indonesia Jutti 2013
[3] A C Charu A survey of text classification algorithm Springer 2012
[4] C de Andre and F Alex A review paper on algorithms used for text classification Research Gate 2009
[5] Dr Suyanto Data Mining Untuk Klasifikasi dan Klasterisasi Data Informatika Bandung 2017
[6] M J Grodzicki Rafał and W Lipo Improved multi-label classification with neural networks Springer 2008
[7] K Gurmeet and B Karan News classification using neural networks Communications on Applied Electronics 2016
[8] R Juan Using tf-idf to determine word relevance in document queries Rutgers University
[9] Z Min-Ling and Z Zhi-Hua Multi-label neural networks with application functional genomics and text categorization IEEE 2006
[10] S k Nikhath A Kousar and V R Building a k-nearest neighbor classifier for text categorization
International Journal of Computer Science and Information Technologies (IJCSIT) 2016
[11] B Nitin and Vandana Survey of nearest neighbor techniques International Journal Computer Science and Information Security 2010
[12] A I Pratiwi and Adiwijaya On the feature selection and classification based on information gain for document sentiment analysis Appl Computational Intell Soft Computing 28(8):1–5 2018
[13] M Z Sadia and R M S Performance analysis of supervised machine learning algorithms for text classification International Conference of Computer and Information Technology 2016
[14] P K Wahyunda Xiaomi resmi masuk pasar eropa dengan mi a1 dan mix 2 http://tekno.kompas.com/read/2017/11/08/20330007/xiaomi-resmi-masuk-pasar-eropa-dengan-mi-a1-dan-mi-mix-2 2017 Online Access on 9 November 2017
[15] P A J Widodo Pangestu A Z A Afiadi Suwanto and H Darlis Klasifikasi kategori dokumen berita berbahasa indonesia dengan metode kategorisasi multilabel berbasis domain spesific ontology Jurnal Ilmiah Teknologi Terapan 2016
[16] Q Yu-ping and W Xiu-kun Study on multi-label text classification based on svm sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009
[17] Y Y Zeng Young and Z Liang Pseudo nearest neighbor rule for pattern classification Elsevier 2009
[18] Al Faraby S Jasin E R R and Kusumaningrum A 2018 March Classification of hadith into positive suggestion, negative suggestion, and information In Journal of Physics: Conference Series (Vol 971 No 1 p 012046) IOP Publishing
[19] Mubarok M S Adiwijaya and Aldhi M D 2017 Aspect-based sentiment analysis to review products using Naive Bayes In AIP Conference Proceedings (Vol 1867 No 1 p 020060) AIP Publishing
[20] Pane Reynaldi Ananda Mubarok Mohamad Syahrul Huda Nanang Saiful Adiwijaya 2018 A Multi-label Classification on Topics of Quranic Verses in English Translation using Multinomial Naive Bayes In 2018 6th International Conference on Information and Communication Technology (ICoICT) IEEE
[21] Izzaty Al Mira Khonsa Mubarok Mohamad Syahrul Huda Nanang Saiful Adiwijaya 2018 A Multi-label Classification on Topics of Quranic Verses in English Translation Using Tree Augmented Naive Bayes In 2018 6th International Conference on Information and Communication Technology (ICoICT) IEEE