Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Education, Culture, Research and Technology with Second Grade (Sinta 2)
since year 2021 to 2026 according to the decree No. 158/E/KPT/2021

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 6 (2021) 1083 - 1089 ISSN Media Elektronik: 2580-0760

Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor

Nur Ghaniaviyanto Ramadhan
Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto
ghani@ittelkom-pwt.ac.id

Abstract

News is information disseminated by newspapers, radio, television, the internet, and other media. According to the survey results, there are many news titles from various topics spread on the internet. This of course makes newsreaders have difficulty when they want to find the desired news topic to read. These problems can be solved by grouping or so-called classification. The classification process is carried out of course by using a computerized process. This study aims to classify several news topics in Indonesian language using the KNN classification model and word2vec to convert words into vectors which aim to facilitate the classification process. The use of KNN in this study also determines the optimal K value to be used. In addition to using the classification model, this study also uses a word embedding-based model, namely word2vec. The results obtained using the word2vec and KNN models have an accuracy of 89.2% with a value of K=7. The word2vec and KNN models are also superior to the support vector machine, logistic regression, and random forest classification models.

Keywords: Indonesian, News Topic, Classification, KNN, Word2Vec

1. Introduction

News is a term that refers to information disseminated by newspapers, radio, television, the internet, and other media [1]. Hundreds of news articles are written every day on various online-based Indonesian news portals, due to the large number of news portals that switch to print media as electronic media that can be accessed online using the internet [2,3]. According to the Indonesian Digital Association (IDA), 96% of urban residents in Indonesia consume online information [4]. Meanwhile, a survey conducted by UC Browser in 2016 reported that 56.5% of internet users in Indonesia generally read 4-12 information articles per day [5]. According to the survey results in [2-5], there are lots of news headlines from various topics spread on the internet. This certainly raises a significant problem for newsreaders. News readers will have difficulty in finding a news topic that they want to read. These problems can be solved by grouping or so-called classification. The classification process is carried out of course by using a computerized process. Computerized classification proved to be more effective than manual classification [6].

Several previous studies have carried out the classification process of news text. The following are some examples of existing literature reviews for comparison of contributions. Research [3] raises the topic of how to form a classification of large Indonesian news data accurately using various computerized models such as Neural Network, SVM, Naïve Bayes, and KNN. In research [7] used Mutual Information (MI) as feature selection, for the classification method of Indonesian news text using Bayesian Network. Paper [8] discusses the classification of multilabel text that can group four labels from news articles with the proposed model of deep learning. In paper [9] made a multilabel classification model on Indonesian news topics using the K-Nearest Neighbor (KNN) method. Study [10] used the word insertion method Doc2vec, on the Turkish Text Classification 3600 dataset consisting of Turkish news texts classified based on deep learning.

Research [11] This research will apply the Porter Stemmer Enhancement algorithm in the stemming process and the Likelihood method for news classification by category and topic identification. A study [12] presents the implementation of multilabel classification using semantic features based on word2vec. In research [13] used word2vec to process Indonesian news headlines, the results of which were used to predict stock prices. Study [14] tested whether

Accepted by Editor: 07-10-2021 | Final Revision: 12-11-2021 | Online Publication: 30-12-2021

word2Vec can be used as input for deep learning in categorizing web news. Paper [15] discusses the special classification process for Indonesian sports news using the BM25 and KNN methods. Research [16] aims to classify Indonesian news titles based on positive-negative sentiments using the word2vec, LSTM, LSTM-CNN, and CNN-LSTM methods. Paper [17] discusses the multi-label classification using the Pseudo Nearest Neighbor Rule (PNNR) algorithm variant of the k-Nearest Neighbor (k-NNR) algorithm. Study [18] focused on the multilabel classification of Arabic text using the Bidirectional Long Short-Term Memory networks (BiLSTM) method, which showed superior results. Paper [19] implements a categorization model that uses a hybrid model consisting of BiLSTM and ANN which classifies news articles into selected topics using hypernyms and hyponyms of the words in them. In a study [20], the news headlines were classified using the NLP algorithm, namely LSTM. This study proposes three models to analyze the semantic similarity of Arabic question pairs using the XGBoost algorithm and word embedding [21].

The paper [22] has used text mining techniques to analyze ancient and modern English. We have introduced the Common-Words Counting algorithm and vector processing using TF-IDF. Study [23] presents an overview of concepts, application of search and answer (SQA), and issues regarding text mining for surah Qur'an (ITQ) with tokenization and stemming techniques. Research [24] aims to improve the Indonesian language stemming algorithm which is suitable for Indonesian text data with slang from social media. Research [25] created a computational environment that allows for the mining of the Qur'anic text, which aims to facilitate people to understand each verse in the Qur'an. The classification method used is SVM, Naïve Bayes, KNN, and J48. In paper [26] tries to measure the ability of the algorithm by applying it to text classification. Paper [26] aims to compare the exact modeling of Deep Learning Neural Network results with two other commonly used algorithms, namely Naïve Bayes and Support Vector Machine (SVM).

The study [27] aims to detect cyberbullies based on text and user credibility analysis and inform them about the dangers of cyberbullying using SVM and KNN methods. Research [28] aims to find the right algorithm to automatically classify a news article in Indonesian using the Naïve Bayes and SVM methods, the dataset comes from the website www.cnnindonesia.com. Research [29] developed an Indonesian hoax filter based on text vector representation based on Term Frequency and Document Frequency as well as SVC classification techniques. Study [30] uses a sentiment classification system which includes several steps such as text preprocessing, feature extraction, and SVM classification. Paper [31] experiment uses text classification to predict personality based on text written by Twitter users. The languages used are English and Indonesian. The classification method applied is Naive Bayes, K-Nearest Neighbors, and Support Vector Machine.

Word2Vec is an efficient tool for transferring words to distribution representation and transferring words into vectors in K dimensions [32]. The word2vec concept in [13] is a cluster that measures the cluster's proximity of words to each other. The advantage of the Word2Vec model is that it can reduce dimensions efficiently and contains a lot of semantic meaning [32]. In general, the form of word2vec can be seen in Figure 1.

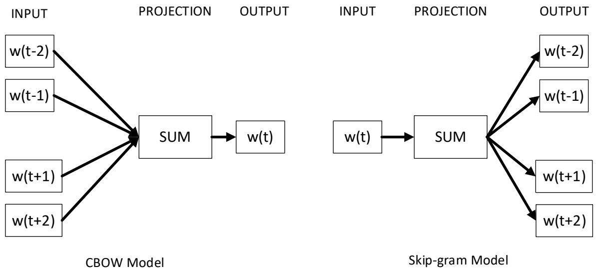

Figure 1. CBOW and Skip-gram Model [33]

The Word2Vec model trains words based on the idea of a distributed representation. It uses two types of models, namely the CBOW model and the Skip-gram model (Figure 1). The CBOW model uses the w(t) word context to predict the current word, and the skip-gram model uses the w(t) word to predict its context [32]. While the notion of KNN is an instance-based method of lazy learning that does not have an offline training phase [34]. Its main computation is an online assessment of training documents given a test document to find the k nearest neighbors [34].

So, this research aims to classify several Indonesian news topics using the word2vec model and K-Nearest Neighbor (KNN). In table 1 is a comparison of contributions with previous studies.

Table 1. Comparison of Previous Study

Authors	Word2Vec	Methods
[9]	No	KNN
[12]	Yes	Semantics Feature
[13]	Yes	Neural Network
[14]	Yes	Deep Learning
[16]	Yes	LSTM, CNN
This Study	Yes	KNN

2. Research Method

Figure 2 is the proposed design for this research.

2.1. Dataset

The first process is carried out by collecting data on various Indonesian news topics through the website [34].

In Table 2 are the topics used in this study along with the number of documents used in each topic.

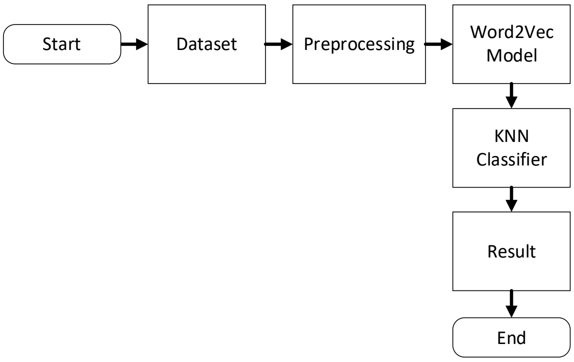

Figure 2. Design System

Table 2. Dataset News Topics

No	Topics	Number of Document
1	Covid 19	10
2	Finance	10
3	Political	10
4	Social	10

The topics used in this research are the most discussed today in Indonesia.

2.2. Preprocessing

This stage is carried out by cleaning sentences if they have characters such as (!?<’") and others. After these characters are removed then the word2vec process is carried out.

2.3. Word2Vec Model

At this stage, the word2Vec process will be carried out which will change the news titles on each topic into x and y vectors. The word2vec process can be seen in Figure 3.

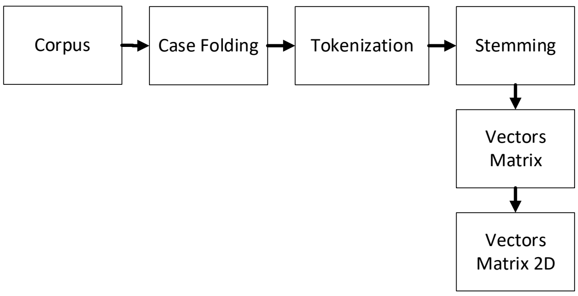

Figure 3. Word2Vec Process

The process starts from a corpus containing a collection of texts. Then case folding is done to change all letters in the corpus to only letters a-z which are accepted, in other words, other than letters are omitted (Table 4). Next, the process of dividing the text or sentence into certain parts is called tokenization. The last step is the process of finding a basic word from a word by removing the affix on the word. So that it will produce a vector matrix value that can be used for classification. Table 3 is an example of the form of news data used.

Table 3. Example of Dataset

News Title	Label	Length
Keuangan DKI sangat terpengaruh jika skenario terburuk pandemi terjadi. DKI's finances are greatly affected if the worst scenario of a pandemic occurs.	Finance	9
Anosmia atau Hilangnya Daya Penciuman Pasien COVID-19 dan Cara Penyembuhannya. Anosmia or loss of smell in COVID-19 patients and how to cure it.	Covid 19	10
Masyarakat Sayangkan Baliho Tokoh Politik Bertebaran di Jalanan. The Community Regrets Billboards of Political Figures Scattered on the Streets.	Political	8
Menteri Sosial melelang nasi goreng buatannya untuk amal. The Minister of Social Affairs auctions his fried rice for charity.	Social	8

Table 4 is a news title that has gone through the cleaning and case folding process. In this process the sentence in the news title will be cleaned of special characters and the sentence will be changed to lower case for each letter.

Table 4. Dataset Cleaning

News	Label
keuangan dki sangat terpengaruh jika skenario terburuk pandemi terjadi	Finance
anosmia atau hilangnya daya penciuman pasien dan cara penyembuhannya	Covid 19
masyarakat sayangkan baliho tokoh politik bertebaran di jalanan	Political
menteri sosial melelang nasi goreng buatannya untuk amal	Social

DOI: https://doi.org/10.29207/resti.v5i6.3547

Table 5 is an example of a dataset that has been tokenized. Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens [36]. The purpose of tokenization is the exploration of words in a sentence. The results are in the form of a list of tokens that will be input for further processing such as parsing or text mining.

Table 5. Dataset Tokenizations

News	Label
"keuangan", "dki", "sangat", "terpengaruh", "jika", "scenario", "terburuk", "pandemic", "terjadi"	Finance
"anosmia", "atau", "hilangnya", "daya", "penciuman", "pasien", "dan", "cara", "penyembuhannya"	Covid 19
"masyarakat", "sayangkan", "baliho", "tokoh", "politik", "bertebaran", "di" "jalanan",	Political
"Menteri", "social", "melelang", "nasi" "goreng", "buatannya" "untuk", "amal"	Social

The tokenization results can be seen in table 5 where the news title turns into a collection of several words that were originally one sentence. The results of the tokenization will then be transformed from words to stemming process. the same root.

After tokenization, the next process is stemming. Stemming is a process without variations of the word form into a representative general form [36]. For example, the word: "*hilangnya*" can be reduced to a general representation of "*hilang*". This process is widely used in offering texts for information retrieval (IR) based on the assumption that asking questions with a presentation implies an interest in the document containing the wording of the presentation and being presented. Table 6 is a representative result of the stemming process.

Table 6. Stemming Dataset

News	Label
"uang" "dki" "sangat" "pengaruh" "jika" "skenario" "buruk" "pandemi" "jadi"	Finance
"anosmia" "atau" "hilang" "daya" "cium" "pasien" "dan" "cara" "sembuh"	Covid 19
"masyarakat" "sayang" "baliho" "tokoh" "politik" "tebar" "di" "jalan"	Political
"Menteri" "social" "lelang" "nasi" "goreng" "buatan" "untuk" "amal"	Social

In the stemming process, there are usually 2 possible errors that occur, namely over-stemming and under-stemming [32]. Over-stemming is when two words with different stems come from the same root. This is also known as a false positive. Under-stemming is when the two words must not come from the same root.

After the stemming process is complete, then the next step is to change the word form into a vector. Table 7 is a vector representation of the results of each word.

Table 7. Word Matrix Vector

Word No	Feature Number			
	1	2	….	300
1	-0.00121	-0.00109	….	0.000314
2	-0.00073	0.000408	….	0.001027
3	0.000255	-0.00155	….	-0.00016
4	0.000016	0.000533	….	0.001333
5	0.000068	0.000937	….	0.000775
6	-0.00044	-0.00056	….	0.000112
….	…..	…..	…..	…..
286	0.000516	-0.00123	….	0.000299

Furthermore, the vector results obtained in table 6 will be reduced to 2 dimensions (x and y) to facilitate the visualization. The technique used is T-Distributed Stochastic Neighbor Embedding (T-SNE), T-SNE is a dimension reduction technique used to represent high-dimensional datasets in two- or three-dimensional low-dimensional spaces so that we can visualize them [37]. The TSNE formula can be seen in (1) [37].

$$Pj|i = \frac{\exp\left(\dfrac{-\left||xi - xj|\right|^2}{2\sigma\dfrac{2}{i}}\right)}{\sum k \neq i \ \exp\left(\dfrac{-\left||xi - xj|\right|^2}{2\sigma\dfrac{2}{i}}\right)} \tag{1}$$

Where $Pj|i$ is a conditional probability. σi is the Gaussian variance centered on the data point xi. For the low-dimensional counterparts yi and yj of the high-dimensional data points xi and xj, it is possible to calculate similar conditional probabilities. This research set the cube of the Gaussian variance used in the calculation of the conditional probability $q\ j|i$ to $1\sqrt{2}$. Formula (2) [32] is used to calculate the low dimension.

$$Pj|i = \frac{(1 + -\left||yi - yj|\right|^2)^{-1}}{\sum k \neq i \ (1 + -\left||yi - yj|\right|^2)^{-1}} \tag{2}$$

This study uses a single degree of freedom, because it has good properties, namely $(1 + -\left||yi - yj|\right|^2)^{-1}$ approaches the inverse square law for large pairwise distances *yi yj* on the dimension low. Formula (3) [29] is used to calculate gradient descent.

DOI: https://doi.org/10.29207/resti.v5i6.3547

$$\frac{\delta C}{\delta yi} = 4\sum_{i}(Pij - Qij)(yi - yj)(1 + ||yi - yj||^2)^{-1} \quad (3)$$

Where this study uses the gradient between two low-dimensional data points yi and yj as a function of the paired Euclidean distance in a high-dimensional and low-dimensional space that is, as a function of xi xj and yi yj. same root.

Table 8 is a 2D vector-matrix obtained from the reduced dimensions in the vector-matrix table 7.

Table 8. Word Matrix Vector 2D

Word No	Feature Number X	Y
1	-1.54038	-8.67283
2	-8.08947	1.229208
3	1.966984	-2.24732
4	-3.55171	-4.86175
5	2.871634	0.588452
6	-5.10687	-3.58582
….	….	….
286	-4.19933	-2.40809

The vectors produced in this study are the values of x and y which interpret a 2D vector. So, it can be concluded that the variable x in word2vec is the independent variable, while the variable y is the dependent variable. For example, it can be seen in Table 9 is the result obtained from a sentence Preparing learning models in the digitalization era "*Mempersiapkan model pembelajaran pada era digitalisasi*".

Table 9. Word2Vec Result

No	Word	x	y
1	*Mempersiapkan* Preparing	-26.0291	-337.468
2	Model	-274.054	-67.5149
3	*Pembelajaran* Learning	307.3602	-185.014
4	*Pada* In this	-93.9533	251.8026
5	Era	265.3956	179.1799
6	*Digitalisasi* Digitalization	35.7445	-31.7941

2.4. K-Nearest Neighbor (KNN)

This study uses a supervised learning classification model, namely K-Nearest Neighbor (KNN) to see the results of the accuracy of the classification of news topics. This algorithm works by finding the most optimal K value or the closest value to the results. To measure similarity efficient in KNN, use the following formula (4), (5), (6), and (7) [34]:

$$Sim(d1, d2) = \frac{d1.d2}{||d1||_2 ||d2||_2} \quad (4)$$

Where d1 and d2 are vector documents used.

Furthermore, each neighbor is given a weight using the similarity in each neighbor to d0, as shown in formula (5).

$$score(d0, C_i) = \sum_{dj \in KNN(d0)} Sim(d0, dj)\delta(dj, Ci) \quad (5)$$

Where KNN(d) is the closest K-neighbor set from document d0. δ(dj,Ci) stands for classification for dj documents related to class Ci. Formula 6 is a derivative of the formula δ(dj,Ci).

$$\delta(dj, Ci) = \begin{Bmatrix} 1\ dj\ \in Ci \\ 0\ dj \notin Ci \end{Bmatrix} \quad (6)$$

Finally, to decide on KNN, can use formula (7).

$$C = \arg max_{ci}\left(score(d0, C_i)\right) \quad (7)$$

The KNN algorithm was chosen because it proved to be capable of not only being used for text classification, but also being used in leaf image classification [38].

3. Result and Discussion

To calculate the accuracy results obtained can use formula (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Table 10 is the result of classification using KNN with several K values experiments.

Table 10. Classification Result

K-value	Accuracy %
3	87.1
5	88.2
7	**89.2**
10	88.2
15	85
17	86
20	84
23	81.7
27	80.6
Mean	85.5

It can be seen from the results in table 10, the most optimal K value used for this research case is 7. The selection of the right K value in the KNN model is very important and affects the classification results. The results also show that if the value of K is getting smaller then the accuracy is increasing, otherwise, if the value of K is getting bigger then the accuracy is decreasing.

Figure 4 is a plot for the confusion matrix generated using KNN with a value of K=7. The confusion matrix can be used as additional material for the analysis of the classification results, for example, how many values are predicted true and actually true on label 2 or how many values are predicted to be wrong but are actually true to label 3. The values in the confusion matrix greatly affect the accuracy results.

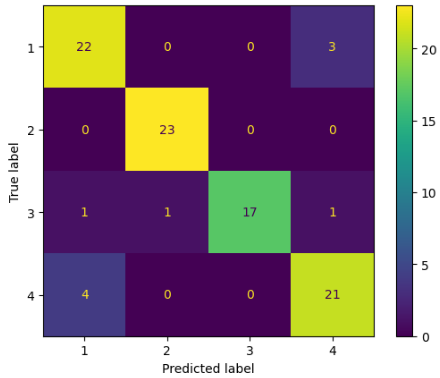

Figure 4. Confusion Matrix K=7

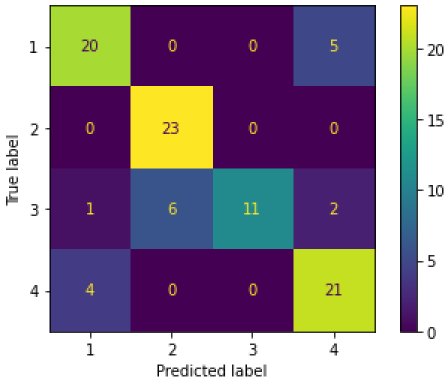

Figure 5. Confusion Matrix K=27

Figure 5 shows the confusion matrix K=27. This proves that it is true that if the value of K is greater, there will be an error in the classification. the value of 11 on the predicted label 3 and the actual label 3 tends to be less likely to guess at that label compared to the value of K=7 (Figure 4).

Table 11 is the result of a comparison using other classification models.

Table 11. Comparison to Other Methods

Classification Methods	Accuracy %
Support Vector Machine	68.8
Logistic Regression	86.9
Random Forest	87
KNN (K=7)	**89.2**

The comparison model used consists of a support vector machine which is based on finding the largest margin (hyperplane), logistic regsression based on regression values, and tree-based random forest. So, based on the results obtained in table 11, the selection of the KNN model used in this study is correct. This can be seen from the results of comparisons using other machine learning models, where KNN with a value of K=7 is still superior. This proves that the algorithm used in this study can produce high classification accuracy.

4. Conclusion

This research aims to classify several Indonesian news topics using the word2vec model and K-Nearest Neighbor (KNN). So, based on the results of the experiments and analysis carried out, it is concluded that the word2vec and KNN models are a combination that can be used in the case of text-based multilabel classification. The selection of the K value in the KNN model also affects the classification results. The value of K used would be better to use a small value. The results of the KNN accuracy are superior to the support vector machine, logistic regression, and random forest models. For further research, you can use more news topics, use news topics from other languages, and use word embedding algorithms and other classifications.

Acknowledgements

The authors declare no conflict interest. The authors would like to thank Institut Teknologi Telkom Purwokerto for funding this research.

References

[1] Li-Juan, Zhu, et al. "A classification method of Vietnamese news events based on maximum entropy model." *2015 34th Chinese Control Conference (CCC)*. IEEE, 2015.

[2] Rizaldy, Adhy, and Heru Agus Santoso. "Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents." *2017 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, 2017.

[3] Irham, Lalu Gias, Adiwijaya Adiwijaya, and Untari Novia Wisesty. "Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 3.4. 284-292. 2019

[4] www.ida.or.id. [Online]. Available: www.ida.or.id. [Accessed: September 2021]

[5] www.pcplus.co.id. [Online]. Available: www.pcplus.co.id/. [Accessed: September 2021]

[6] Dadgar, Seyyed Mohammad Hossein, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification." *2016 IEEE International Conference on Engineering and Technology (ICETECH)*. IEEE, 2016.

[7] Nurfikri, Fahmi Salman, and Mohamad Syahrul Mubarok. "News topic classification using mutual information and bayesian network." *2018 6th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2018.

[8] Sari, Winda Kurnia, Dian Palupi Rini, and Reza Firsandaya Malik. "Multilabel Classification for News Article Using Long Short-Term Memory." *Sriwijaya Journal of Informatics and Applications* 1.1. 2020.

[9] Isnaini, Nikmah, Mohamad Syahrul Mubarok, and Muhammad Yuslan Abu Bakar. "A multi-label classification on topics of Indonesian news using K-Nearest Neighbor." *Journal of Physics: Conference Series*. Vol. 1192. No. 1. IOP Publishing, 2019.

[10] Doğru, Hasibe Büşra, et al. "Comparative Analysis of Deep Learning and Traditional Machine Learning Models for Turkish Text Classification."

DOI: https://doi.org/10.29207/resti.v5i6.3547

[11] Rukmi, Alvida Mustika, Devi Andriyani, and Imam Mukhlas. "Identification of topics in News Articles Using Algorithm of Porter Stemmer Enhancement and Likelihood Classifier." *Journal of Physics: Conference Series*. Vol. 1490. No. 1. IOP Publishing, 2020.
[12] Rahmawati, Dyah, and Masayu Leylia Khodra. "Word2vec semantic representation in multilabel classification for Indonesian news article." *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. IEEE, 2016.
[13] Ramadhan, Nur Ghaniaviyanto, and Imelda Atastina. "Neural Network on Stock Prediction using the Stock Prices Feature and Indonesian Financial News Titles." *International Journal on Information and Communication Technology (IJoICT)* 7.1. 54-63. 2021
[14] Kato, Ryoma, and Hiroyuki Goto. "Categorization of web news documents using word2vec and deep learning." *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia*. 2016.
[15] Septrinas, Enggar, and Arief Andy Soebroto Indriati. "Klasifikasi Berita Olahraga Berbahasa Indonesia Menggunakan Metode BM25 dan K-Nearest Neighbor." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN* 2548 964X. 2020.
[16] Hermanto, Dedi Tri, Arief Setyanto, and Emha Taufiq Luthfi. "Algoritma LSTM-CNN untuk Binary Klasifikasi dengan Word2vec pada Media Online." *Creative Information Technology Journal* 8.1. 64-77. 2021
[17] Kakulapati, V., and S. Mahender Reddy. "Multimodal Detection of COVID-19 Fake News and Public Behavior Analysis—Machine Learning Prospective." *Intelligent Healthcare*. Springer, Cham. 225-241. 2021.
[18] AlBatayha, Duha. "Multi-Topic Labelling Classification Based on LSTM." *2021 12th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2021.
[19] Sanagavarapu, Sowmya, Sashank Sridhar, and S. Chitrakala. "News Categorization using Hybrid BiLSTM-ANN Model with Feature Engineering." *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021.
[20] Bhuiyan, Md Rafiuzzaman, et al. "An Approach for Bengali News Headline Classification Using LSTM." *Emerging Technologies in Data Mining and Information Security*. Springer, Singapore. 299-308. 2021.
[21] Hammad, Mahmoud, et al. "Using deep learning models for learning semantic text similarity of Arabic questions." *International Journal of Electrical & Computer Engineering (2088-8708)* 11.4. 2021.
[22] Alam, Saqib, and Nianmin Yao. "Big data analytics, text mining and modern english language." Journal of Grid Computing 17.2 (2019): 357-366.
[23] Putra, Syopiansyah Jaya, Teddy Mantoro, and Muhamad Nur Gunawan. "Text mining for Indonesian translation of the Quran: A systematic review." 2017 International Conference on Computing, Engineering, and Design (ICCED). IEEE, 2017.
[24] Maylawati, Dian Sa'adillah, et al. "An improved of stemming algorithm for mining indonesian text with slang on social media." 2018 6th International Conference on Cyber and IT Service Management (CITSM). IEEE, 2018.
[25] Hidayat, Rahmat, and Sekar Minati. "Comparative Analysis of Text Mining Classification Algorithms for English and Indonesian Qur'an Translation." IJID (International Journal on Informatics for Development) 8.1. 47-51. 2019.
[26] Mariel, Wahyu Calvin Frans, Siti Mariyah, and Setia Pramana. "Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text." *Journal of Physics: Conference Series*. Vol. 971. No. 1. IOP Publishing, 2018.
[27] Nurrahmi, Hani, and Dade Nurjanah. "Indonesian twitter cyberbullying detection using text classification and user credibility." 2018 International Conference on Information and Communications Technology (ICOIACT). IEEE, 2018.
[28] Wongso, Rini, et al. "News article text classification in Indonesian language." Procedia Computer Science 116. 137-143. 2017.
[29] Prasetijo, Agung B., et al. "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD." 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, 2017.
[30] Naradhipa, Aqsath Rasyid, and Ayu Purwarianti. "Sentiment classification for Indonesian message in social media." 2012 International Conference on Cloud Computing and Social Networking (ICCCSN). IEEE, 2012.
[31] Pratama, Bayu Yudha, and Riyanarto Sarno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM." 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE, 2015.
[32] Wu, Chunzi, and Bai Wang. "Extracting topics based on Word2Vec and improved Jaccard similarity coefficient." *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2017.
[33] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*. 2013.
[34] Tan, Songbo. "An effective refinement strategy for KNN text classifier." *Expert Systems with Applications* 30.2. 290-298. 2006.
[35] www.id.news.search.yahoo.com. [Online]. Available: www.id.news.search.yahoo.com/. [Accessed: October 2021].
[36] Kannan, Subbu, et al. "Preprocessing techniques for text mining." *International Journal of Computer Science & Communication Networks* 5.1. 7-16. 2014.
[37] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.11, 2008.
[38] Ahmad, Andani, and Abdul Latief. "Perbandingan Metode KNN Dan LBPH Pada Klasifikasi Daun Herbal." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5.3. 557-564. 2021.

DOI: https://doi.org/10.29207/resti.v5i6.3547