INFORMATICS AND DIGITAL EXPERT (INDEX) - VOL. 6 NO. 2 (2024) 110-116
Terbit *online* pada laman web jurnal : http://e-journal.unper.ac.id
| ISSN (Print) 2775-2208 | ISSN (Online) 2715-0453 |

Comparison of Classification for Indonesian Language News Documents Using Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) Algorithms

Christian Sri Kusuma Aditya[1], Briansyah Setio Wiyono[2], Muh. Ridha Agam[3], Andhika Rezky Fadillah[4]

[1,2,3]*Informatics, Faculty of Engineering, University of Muhammadiyah Malang*
e-mail: christianskaditya@umm.ac.id[*1], brian@umm.ac.id[2], muhridhaagam@webmail.umm.ac.id[3], andhikarezky345@webmail.umm.ac.id[4]

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi : 06 Agustus 2024
Revisi Akhir : 05 November 2024
Diterbitkan *Online* : 30 November 2024

Kata Kunci :
Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), classification, news, indonesia, SMOTE

Korespondensi :
Telepon / Hp : +6271359659715
E-mail : **christianskaditya@umm.ac.id**

A B S T R A K

The development of online news has grown very fast. The high volume of text documents was triggered by activities from various news sources. Due to the large amount of news that is included on the website, sometimes the news is posted not according to its category which is most likely caused by human error. The grouping of online news is important for user convenience in searching for news according to its category. It need an intelligent system that can classify online news automatically. This research evaluates deep learning techniques using LSTM and RNN, and compared with the results obtained from previous studies, which used the NBC algorithm. To experiment the system, an Indonesia News Corpus with 7 different categories and total 2100 documents, collected by crawling online national news portals, is used. Due to the unbalanced number of class compositions or news categories, integration is also carried out SMOTE. The average empirical results show that the classification accuracy from RNN with SMOTE with an accuracy of 95.2% and followed by LSTM with SMOTE is 97.8%, both of which are able to outperform the NBC method with an accuracy of 73.2%.

1. INTRODUCTION

Advances in information technology in the last decade have brought major changes to the media, press and journalism industries. Internet technology has given birth to a new online media which is a threat to conventional media. The majority of information is now disseminated no longer through news in the form of print media such as newspapers [1]. The speed of news information owned by online news media is very high when compared to print media or other conventional media. Events that occur in the field can be uploaded directly in minutes or seconds [2]. A 2021 survey by UNESCO (United Nations Educational, Scientific, and Cultural Organization) reveals a continuing decline in the number of people who read news through traditional media, such as newspapers. In contrast, since 2010, the use of internet media, particularly online news portals, has been consistently rising each year [3], as illustrated in Figure 1.

Classification of news documents is the process of grouping and labeling based on certain topics or categories. By classifying news documents in a media or online news portal, users can easily navigate through relevant information according to their interests. Basically, the classification of news documents is done manually, but considering the increasing number of news documents available on the internet, it is very important to develop an automatic classification system by utilizing computer technology so that it can be more practical, fast and structured [4].

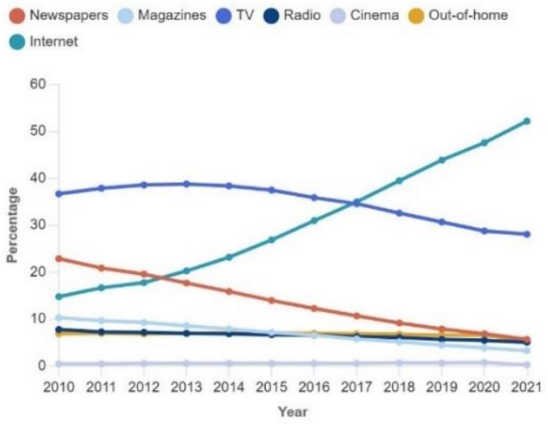

Figure 1. Graph of News Readers in Various Media (source: https://news.un.org/en/story/2022/03/1113702)

Several previous studies have carried out classification of news documents using various methods, including the use of the Naïve Bayes algorithm with the addition of bayesian boosting [5] which is capable of producing reliable classification models and often avoiding overfitting. This research was able to obtain results of accuracy, precision and recall of 72.1%. Furthermore, the application of bayesian boosting for labels that are polynomial has the effect of increasing the evaluation results by 1.3%.

Related research [6] for the purpose of classifying scientific sentences using the Recurrent Neural Network (RNN) algorithm. The use of RNN shows that the best learning rate results are obtained with the Stochastic

Gradient Descent (SGD) optimizer parameter with an accuracy value of 77.5% and a Loss value of 0.7%. RNN itself is a classification model in deep learning that is able to model sequential dependencies on input data and has a data memory containing previously generated information, this ability is due to the connection node (recurrent) in the model and processes input data sequentially [7]. RNNs are built to manage input sequences of varying lengths, making them ideal for applications like speech recognition, natural language processing, and time series analysis [8].

LSTM or Long Short-Term Memory is an updated model of the RNN model which is used to manage sequential data by storing the results of previous information [9]. LSTM is a development of the RNN method which has the ability to model long-term contexts and overcome the vanishing gradient problem that often occurs in RNNs. In another study aimed at classifying hoax news content [10], a comparison was made between training the data using a linear kernel Support Vector Machine (SVM) and LSTM with adam optimizer. LSTMs can handle variable length sequences and bidirectional input, which is useful for NLP tasks such as machine translation, text generation, and sentiment analysis. RNNs are simpler and faster to train than LSTMs, because they have fewer parameters and computations [11], whereas LSTM can learn more complex and long-range patterns. The use of the LSTM method in developing an Indonesian language news sentiment monitoring system based on content [12][13] is good which produces an accuracy value of 86.2%, with the amount of data used as many as 985 data which is divided into 886 training data and 99. The application of LSTM in this study uses the function early stopping to stop the data training process when an overfit occurs, while the batch size used is 128 with a total of 150 epochs.

Based on the description of the literature above, this study will develop a classification modeling with a deep learning approach in online Indonesian language news documents, as well as conducting a comparative test of the results of the accuracy between the use of the RNN and LSTM models. Due to the unbalanced number of class compositions or news categories, integration is also carried out with the Synthetic Minority Over-sampling Technique (SMOTE) [14], where SMOTE works by generating synthetic samples from the minority class by connecting close neighbors in the feature space [15]. Thus, this technique is able to create synthetic data that describes the variation and complexity of minority classes, thereby reducing bias and increasing overall classification accuracy.

2. RESEARCH METHODS

The flowchart of this research is carried out according to the plot shown in Figure 2.

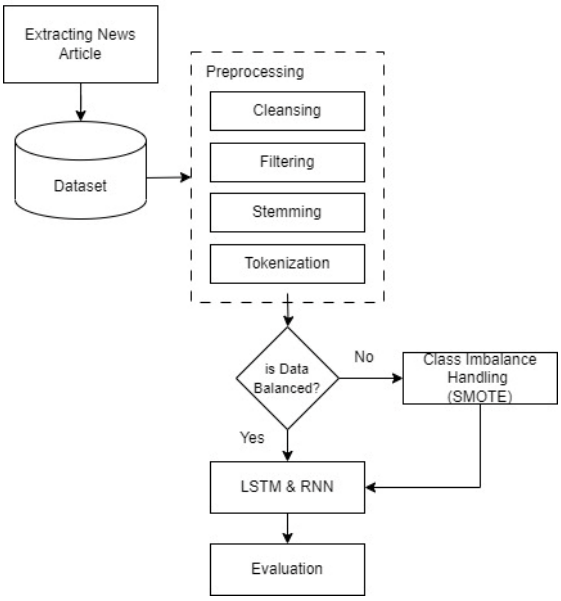

Figure 2. Stage of Research Method

2.1. Dataset

In the process of forming the dataset required in this study, the scraping technique was used. Where retrieval uses a technique that can retrieve index data containing title, link, category, and date, the value in XPATH on a kompas.com news website.

This research was conducted using a dataset from the results of the scraping process on one of the existing news websites, with a total of around 7610 data. Then a data selection process is carried out in order to get the most targeted 7 classes which in total is 2100 data. The total dataset will then be divided into 60% training data and 20% testing data and 20% validation data.

The following is the distribution or distribution of the number of news obtained for each class which can be seen in Table 2.

Table 1. Distribution of the Amount of Data to the Category

Categories	Amount of data
BOLA	328
HYPE	250
MONEY	196
NEWS	331
OTOMOTIF	344
REGIONAL	312
TREN	339

2.2. Preprocessing

Text preprocessing is a stage or process of refining text where previously unstructured data becomes structured data. The preprocessing stage consists of several stages, namely: cleansing, filtering, stemming,

tokenization [16]. In the cleansing process, various contents that have URLs listed in the news content are deleted. This is done in order to maintain the semantic value that exists in each sentence. In this process there is also case folding which changes letters from uppercase to lowercase [17], an example of the cleaning process can be seen in Table 2.

Table 2. Cleansing Process

Input	Output
www.kompas.com Duta Besar Amerika Serikat (AS) untuk Indonesia	duta besar amerika serikat (as) untuk indonesia

In the process of filtering or stopword removal is done by using a "sastrawi" library from python which will identify various kinds of words that are not needed so that they will be deleted and make sentences have a higher classification value because there is no need to process words that are not needed [18], an example of the filtering process can be seen in Table 3.

Table 3. Filtering Process

Input	Output
jakpro secara tegas menyatakan lahan yang dimanfaatkan oleh ruko merupakan aset pt jakpro	jakpro secara tegas menyatakan lahan dimanfaatkan ruko merupakan aset pt jakpro

Stemming is a stage in text processing which aims to find the basic words of the original words that appear in a text. The stemming algorithm used in this study uses Enhanced Confix Stripping [19][20], an example of the stemming process can be seen in Table 4.

Table 4. Stemming Process

Input	Output
berita tentang pernyataan pt jakarta propertindo jakpro menyebut lahan yang dicaplok ruko	berita tentang nyata pt jakarta propertindo jakpro sebut lahan caplok ruko

Tokenization involves converting words in sentences into integers for machine analysis of the data [21][22]. One hot encoding is a method used to represent categorical variables as numerical values in machine learning models. In this technique, each word (including symbols) in the text data is transformed into a vector. Each word is represented by a unique one hot vector. An example of the stemming process can be found in Table 5.

Table 5. Tokenization Process

Input	Output
Berita tentang pernyataan PT Jakarta Propertindo	[31, 229, 4, 230, 89, 231]

2.3. SMOTE

After the preprocessing stage, the next step is to balance the data distribution across classes. A fundamental oversampling technique involves randomly duplicating data in the minority class, known as random oversampling. However, this approach often leads to overfitting because the classifier encounters the same information multiple times. To address this issue, the synthetic minority oversampling technique (SMOTE) is recommended. SMOTE works by generating artificial data to increase the amount of minority class data, making it equal to the majority class [23]. This is achieved by creating synthetic samples based on the k-nearest neighbors. For categorical variables, the Value Difference Metric (VDM) formula is used to calculate distances between examples in the minority class [24]. The impact of this data distribution adjustment is illustrated in Figure 3.

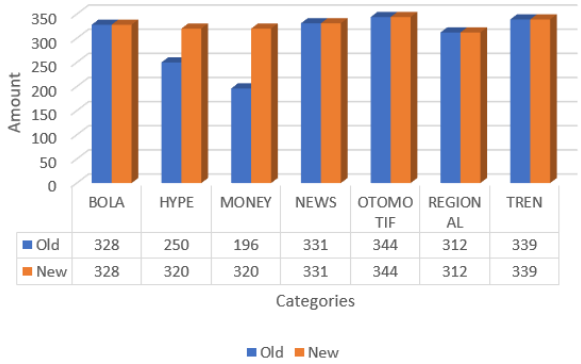

Figure 3. SMOTE Result

2.4. Recurrent Neural Network (RNN)

RNN include feedback loops within their recurrent layers, allowing them to retain information over time. However, training RNNs to address issues that involve learning long-term temporal relationships can be challenging. This difficulty arises because the gradient of the loss function diminishes exponentially as time progresses, a phenomenon known as the vanishing gradient problem.

$h(t)$ in RNN maintains a hidden state at time t, serving as the network's memory. $h(t)$, equation 8, is calculated based on the current input and the previous time step's hidden state.

$$h(t) = g(Wx_t + Uh_{t-1}) \quad (8)$$

where g is a non-linear neuron activation function, W and U are a weight matarices, x_t are updated neural inputs, and h_{t-1} is the hidden state if previous timestep.

2.5. Long Short Term Memory (LSTM)

LSTM is an advancement of the RNN deep learning technique, offering the benefit of handling relatively long sequences of data. It addresses the issues of vanishing and exploding gradients by incorporating new gates, like input and forget gates, which enhance control over the gradient flow and improve the maintenance of "long-range dependencies." In LSTM,

the issue of long-range dependencies in RNNs is managed by increasing the number of repeating layers.

An LSTM unit comprises three main parts known as gates, which control the flow of information into and out of the memory cell. The first gate is the Forget gate (Equation 1), followed by the Input gate (Equation 2), and the last is the Output gate (Equation 3). An LSTM unit or its gates can be compared to a layer of neurons in a traditional feedforward neural network, where each neuron has a hidden layer and a current state [25].

In an LSTM neural network cell, the primary function is to decide whether to keep or discard information from the previous time step. The forget gate is represented by the following equation.

Forget Gate

$$f_t = \sigma\left(x_t * U_f + H_{t-1} * W_f\right) \quad (1)$$

Input Gate

$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i) \quad (2)$$

Output Gate

$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o) \quad (3)$$

x_t: input to the current timestamp.
U_f: weight associated with the input
U_i: weight matrix of input
U_o: weight matrix of output
H_{t-1}: hidden state of the previous timestamp
W_f: weight matrix associated with the hidden state
σ: sigmoid

Afterwards, a sigmoid function is applied, which transforms f_t into a value between 0 and 1. This f_t is then multiplied by the cell state from the previous time step, as illustrated below. The equation from sigmoid function can be seen at equation 4 and 5.

$$C_{t-1} * f_t = 0 \ldots if\ f_t = 0\ (forget\ everything) \quad (4)$$

$$C_{t-1} * f_t = C_{t-1} \ldots if\ f_t = 1\ (forget\ nothing) \quad (5)$$

The new piece of information, represented by equation 6, that must be incorporated into the cell state depends on the hidden state from the previous time step $t-1$ and the input x at the current time step t. The activation function used is tanh. As a result of the tanh function, the new information will range between -1 and 1. If the value of N_t is negative, the information is subtracted from the cell state; if it is positive, the information is added to the cell state at the current time step.

$$N_t = tanh(x_t * U_c + H_{t-1} * W_c)\ (new\ information) \quad (6)$$

However, the N_t won't be added directly to the cell state. Here comes the updated equation 7.

$$C_t = f_t * C_{t-1} + i_t * N_t\ (updating\ cell\ state) \quad (7)$$

Here, C_{t-1} is the cell state at the current timestamp, and the others are the values we have calculated previously.

The primary distinction between RNN and LSTM lies in their ability to retain information over long durations. LSTM has the upper hand over RNN because it can manage and preserve information in memory for extended periods, unlike RNN [26]. Figure 4 provides a visual comparison between standard RNN and LSTM.

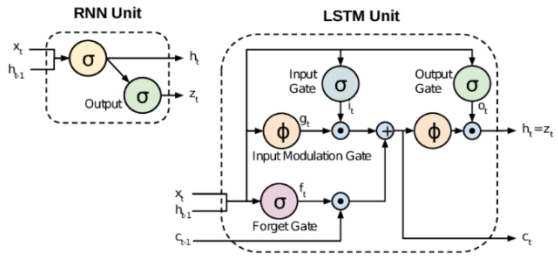

Figure 4. RNN vs LSTM

2.6. Evaluation

The confusion matrix is a tool used in predictive analytics to show and compare actual or true values with those predicted by a model. It helps generate evaluation metrics such as Precision (described in equation 9), Recall (described in equation 10), Accuracy (described in equation 11), and F1-Score (described in equation 12). The matrix includes four key values: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (10)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (11)$$

$$\text{F1} - \text{Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (12)$$

3. RESULTS AND DISCUSSIONS

3.1. Classification Results with RNN

In the first trial scenario, after going through the preprocessing stage and balancing the number of datasets, then proceed to RNN modeling with the softmax activation parameter, the adam optimizer and a dropout rate value of 0.2, and implemented early stopping to avoid overfitting which easily occurs in the RNN. The accuracy results obtained are 95.2% while the validation results are at 71.1% which can be seen in Figure 5, and in Figure 6 is the Confusion matrix from RNN modeling.

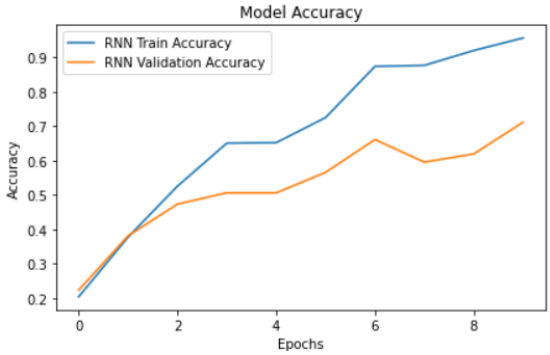

Figure 5. Accuracy Chart on RNN

Although in RNN modeling Adam optimizer and early stopping have been added to help **prevent** overfitting and improve the generalization performance of the model, it can be seen that the validation results are still below the training results. RNN have the disadvantage of being computationally expensive to train, especially when dealing with long sequences. This is because the network has to process each input in sequence.

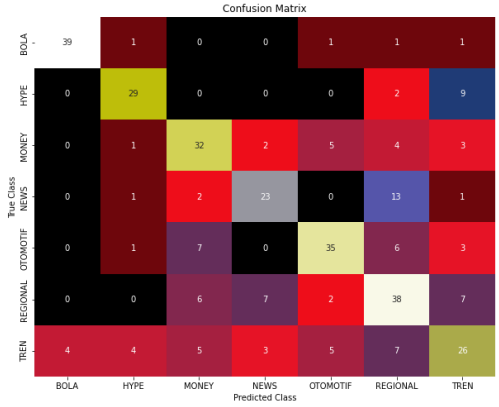

Figure 6. Confusion Matrix on RNN

For comparison results in the form of precision, recall, F1-Score and accuracy with RNN modeling for each category can be seen in table 6.

Table 6. RNN Results for Each Category

Categories	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
Bola	98.1	99.2	99.3	99.2
Hype	99.3	98.3	99.1	97.4
Money	96.5	95.3	96.3	95.2
News	92.4	95.7	94.5	95.2
Otomotif	94.4	96.5	95.3	95.1
Regional	92.3	88.3	90.5	93.3
Tren	96.8	95.2	96.6	91.2
Avg	95.4	95.3	95.3	**95.2**

3.2. Classification Results with LSTM

In the second trial scenario, LSTM modeling was carried out with the same activation function parameters and dropout values as in the first trial scenario. The accuracy results obtained are 97.8% while the validation results are at 92.8%. The results of the train can be seen in Figure 7.

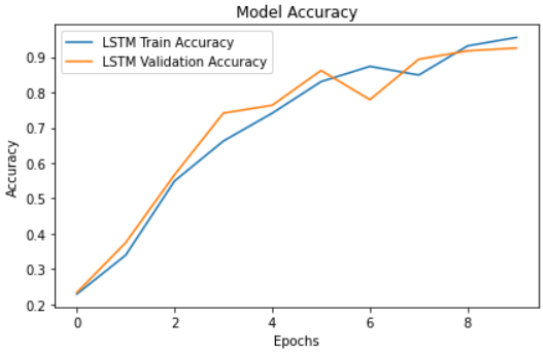

Figure 7. Accuracy Chart on LSTM

The confusion matrix of the LSTM modeling which is a representation of the classification performance results can be seen in Figure 8, while the results of the comparison of data testing for each category with the LSTM modeling can be seen in Table 7.

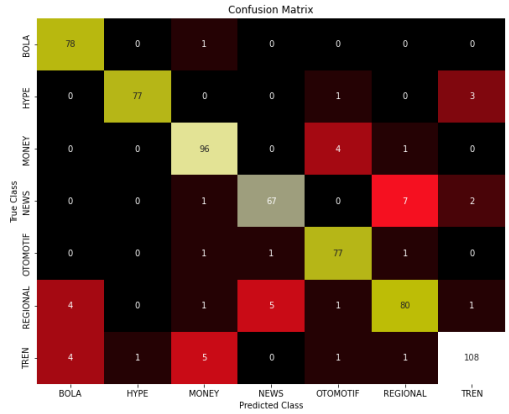

Figure 8. Confusion Matrix on LSTM

Table 7. LSTM Results for Each Category

Categories	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
Bola	95.3	100	97.2	98.5
Hype	100	98.3	99.3	99.2
Money	100	98.4	99.4	97.7
News	98.5	94.6	96.2	97.4
Otomotif	98.2	99.7	98.1	98.4
Regional	93.4	94.2	94.7	96.5
Tren	98.3	98.3	98.2	97.1
Avg	**97.8**	**97.5**	**97.2**	**97.8**

scenario, it can be seen that the training results from RNN vs LSRM have a difference that is not too different at 2.6%. Meanwhile, in the comparison of validation results, the difference in value comparison is quite large, 21.7%. This proves that the ability of the LSTM model to recognize and classify new data is better.

3.3. Effects of Applying SMOTE

In the third scenario, testing is carried out to see the effect of using SMOTE due to the imbalance in the distribution of the number of datasets owned.

A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. One way to solve this problem is to oversample the examples in the minority class. Table 8 shows that the average level of accuracy in both LSTM and RNN modeling has increased in accuracy by 10.2%.

Table 8. Comparison of the Effects of Using SMOTE

Model	Accuracy (%)
RNN without SMOTE	82.5
RNN with SMOTE	95.2
LSTM without SMOTE	88.2
LSTM with SMOTE	97.8

3.4. Comparison of the Results of the Proposed Modeling Design with Previous Research

From the several test scenarios above, it can be seen that when compared with the results of experiments in previous studies with the same dataset, the accuracy results using LSTM and RNN are able to outperform using the Naïve Bayes method which has been optimized with bayesian boosting and Information Gain (IG) feature selection. The following is in table 9 for details on the comparison of accuracy results with several machine learning models.

Table 9. Comparison of Accuracy Results with Previous Research

Model Machine Learning	Accuracy (%)
NBC [2]	72.3
IG-NBC [2]	69.5
Bayesian Boosting-NBC [2]	73.2
IG-*Bayesian Boosting-NBC* [2]	73.2
RNN with SMOTE	95.2
LSTM with SMOTE	97.8

4. CONCLUSION

The results of the research that has been done show that the application of deep learning with the Long Short Term Memory (LSTM) and Recurrent Neural Network (RNN) models is able to classify news documents in Indonesian quite effectively.

The LSTM model obtains an accuracy of 96.3% on training data and 71.8% on validation data. While the RNN model achieves an accuracy of 95.2% on the training data and 66.1% on the validation data. Both of these methods are able to achieve high values of precision, recall, and F1-score, which indicates that these two models have good performance in news text classification compared to reference journals [14] which use the NBC model.

In this case it also shows that the SMOTE method can increase the value of accuracy in modeling designs that are developed in the form of unbalanced datasets.

Thus, it can be concluded that LSTM and RNN have great potential to be used in various other text classification applications, not only news text classification.

Suggestions for future research, where there is still room to improve performance, especially on validation result values due to overfitting problems, several efforts can be made including increasing the number of datasets (oversampling), performing feature selection, and also normalizing or feature scaling at the preprocessing stage.

ACKNOWLEDGEMENTS

This manuscript is based on research supported by the Universitas Muhammadiyah Malang. The authors would also express their gratitude for the UMM Informatics Laboratory, who have supported the implementation of this research.

DAFTAR PUSTAKA

[1] Prihantoro, E., & Fitriani, D. R. (2015). Modalitas dalam teks berita media online. Prosiding PESAT, 6.

[2] Kencana, W. H., Situmeang, I. V. O., Meisyanti, M., Rahmawati, K. J., & Nugroho, H. (2022). Penggunaan Media Sosial dalam Portal Berita Online. IKRA-ITH HUMANIORA: Jurnal Sosial Dan Humaniora, 6(2), 136-145.

[3] https://news.un.org/en/story/2022/03/1113702

[4] Raphael, M. M. T., Hafizh, M. K., Damasyifa, F. A., Setiawan, S. R., Putra, P. R. B., & Yudistira, N. DETEKSI HOAKS PADA BERITA LOKAL INDONESIA MENGGUNAKAN MODEL BERBASIS RECURRENT NEURAL NETWORK.

[5] Prakoso, B. S., Rosiyadi, D., Utama, H. S., & Aridarma, D. (2019). Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 3(2), 227-232.

[6] Firmansyah, M. R., Ilyas, R., & Kasyidi, F. (2020, September). Klasifikasi Kalimat Ilmiah Menggunakan Recurrent Neural Network. In Prosiding Industrial Research Workshop and National Seminar (Vol. 11, No. 1, pp. 488-495).

[7] Rais, I. L., & Jondri, J. (2020). Klasifikasi Data Kuesioner dengan Metode Recurrent Neural Network. eProceedings of Engineering, 7(1).

[8] Ivanedra, K., & Mustikasari, M. (2019). Implementasi Metode Recurrent Neural Network Pada Text Summarization Dengan Teknik Abstraktif. J. Teknol. Inf. dan Ilmu Komput, 6(4), 377.

[9] Pakpahan, J. A., Panjaitan, Y. C., Amalia, J., & Pakpahan, M. B. (2022). Model Klasifikasi Berita Palsu Menggunakan Bidirectional LSTM dan Word2vec sebagai Vektorisasi. JATISI (Jurnal Teknik Informatika dan Sistem Informasi), 9(4), 3319-3331.

[10] Aditya, C. S. K., Wicaksono, G. W., & Hilman Abi Sarwan, H. (2023). Sentiment Analysis of the 2024 Presidential Candidates Using SMOTE and Long Short Term Memory. Jurnal Informatika Universitas Pamulang, 8(2), 279-286.

[11] Tannady, S. M. N., Setiabudi, D. H., & Tjondrowiguno, A. N. (2022). Penerapan Long-Short Term Memory dengan Word2Vec Model untuk Mendeteksi Hoax dan Clickbait News pada Berita Online di Indonesia. Jurnal Infra, 10(2), 28-34.

[12] Widhiyasana, Y., Semiawan, T., Mudzakir, I. G. A., & Noor, M. R. (2021). Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia. Jurnal Nasional Teknik Elektro dan Teknologi Informasi| Vol, 10(4).

[13] Liliana, D. Y., Hikmah, N. N., & Harjono, M. (2021). PENGEMBANGAN SISTEM PEMANTAUAN SENTIMEN BERITA BERBAHASA INDONESIA BERDASARKAN KONTEN DENGAN LONG SHORT-TERM MEMORY. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 8(5).

[14] Mualfah, D., Fadila, W., & Firdaus, R. (2022). Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest. Jurnal CoSciTech (Computer Science and Information Technology), 3(2), 107-113.

[15] Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Information sciences, 501, 118-135.

[16] Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis, 26(2), 168-189.

[17] Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1), 114-146.

[18] Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

[19] Winarti, T., Kerami, J., & Arief, S. (2017). Determining term on text document clustering using algorithm of enhanced confix stripping stemming. Int. J. Comput. Appl, 157(9), 8-13.

[20] Arifin, A. Z., Mahendra, I. P. A. K., & Ciptaningtyas, H. T. (2009, August). Enhanced confix stripping stemmer and ants algorithm for classifying news document in indonesian language. In The International Conference on Information & Communication Technology and Systems (Vol. 5, pp. 149-158).

[21] A. Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. Journal of Open Source Software, 3(23), 655.

[22] Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenization tools-an analysis. Advanced Computational Intelligence: An International Journal (ACII), 3(1), 37-47.

[23] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

[24] Barro, R. A., Sulvianti, I. D., & Afendi, F. M. (2013). Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. Xplore: Journal of Statistics, 1(1).

[25] Cahyo, P Winar., & U,S Aesyi. (2023). Perbandingan LSTM dengan Support Vector Machine dan Multinomial Naive Bayes pada Klasifikasi Kategori Hoax. Jurnal Transformatika. Vol 20(2). pp 23-29.

[26] Ivanedra, Kasyfi., & M, Mustikasari. (2018). Implementasi Metode Recurrent Neural Network pada Text Summarization dengan Teknik Abstraktif. Jurnal Teknologi Informasi dan Ilmu Komputer. Vol 6(4). hal 377-382. 10.25126/jtiik.201961067.