Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

News text classification using long-term short memory (LSTM) algorithm

Indra Triyadi[1], Budi Prasetiyo[2], Tiara Lailatul Nikmah[3]
[1,2,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jan 18, 2023
Revised Feb 26, 2023
Accepted Mar 6, 2023

Keywords:

Text classification
LSTM
News text

ABSTRACT

Over the past few years, the classification of texts has become increasingly important. Because knowledge is now available to users through various sources namely electronic media, digital media, print media, and many more. One of them is the development of so much news every day. LSTM is one of the algorithms of deep learning methods that can classify a text. This research proves for the LSTM algorithm on the classification of news text sentences. The data used is the news text from the Kaggle data center set i.e. aggregator news data. The results of the LSTM experiment from 10 epochs obtained with an accuracy value of 93,15% on the classification of texts into four categories, namely entertainment, bussines, science, and health.

Corresponding Author:

Tiara Lailatul Nikmah,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang 50229, Indonesia.
Email: tiaralaila21@gmail.com

1. INTRODUCTION

Text classification is the process of grouping a text into a specified category. Text categorization of the application of the arrangement with the aim of: D x C ◊ {T, F}, D is part of the text and C is the defined group of categories [1]. A frequently updated News Site will create a large amount of news information. Text classification can be as an alternative to analyzing news texts by defining news types [2], [3]. Text classification can also make it easier for readers to obtain news from the large amount of news information available.

Currently, the process of classifying a text is facilitated by the use of a computer so that it is more efficient than done manually. In addition, the use of computers increases efficiency and minimizes errors [4]. Nowadays text classification is very popular using machine learning. Many studies have been conducted in classifying text using machine learning [5]–[11]. The completion of news text classification has now grown so much using various algorithms from machine learning and deep learning [12]–[21]. The use of machine learning methods in the application of news text classification includes the Naïve Bayes algorithm [22]–[25], TF-IDF [26], [27] and SVD [28]–[30]. As well as the use of deep learning methods among LSTM, CNN, MLP algorithms [31]–[34].

Based on the reference [35] the application of LSTM, SVM and RF algorithms for classifying LSTM Javanese-language text expressions obtained the highest accuracy with 92%. Reference [36] the accuracy result from LSTM got 91.9% for social media sentiment analysis. There are deep learning algorithms that can be used to classify news texts. Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) algorithms are two algorithms of the deep learning method as alternatives commonly used to recognize related data [37], [28]. The LSTM algorithm is an upgrade of the RNN algorithm [38]. Because there is a weakness in RNN before, there is a weakness, namely the flexibility of RNN memory which cannot predict if a word is stored in long-term memory [39], [40]. From these references can be obtained the LSTM algorithm has good accuracy in the classification of texts.

In this study, managing data from Kaggle sources, namely news aggregator data. The data will go through a preprocessing process, namely shuffle, one hot encoding and tokenizer. Then the data will be classified using the Long Short Term Memory (LSTM) algorithm with sequential layers and adam optimizer. This research will classify news into four categories, namely entertainment, bussines, science, and health. Testing will be carried out using several epoch counts to get the best accuracy. The accuracy of the model is measured using metric accuracy and loss.

2. METHOD

2.1. Data and data sources

The managed data is a data set of "news aggregator dataset" obtained from Kaggle. The data was taken from a web aggregator between March 10, 2014 and August 10, 2014. The data is a table consisting of 8 columns and 423,000 rows. The data is seen in Table 1 and Figure 1 as follows.

Table 1. Dataset attributes

Atribute	**Description**
ID	the numeric ID of the article
Title	the headline of the article
Url	the URL of the article
Publisher	the publisher of the article
Category	the category of the news item; one of: -- *b* : business -- *t* : science and technology -- *e* : entertainment -- *m* : health
Story	alphanumeric ID of the news story that the article discusses
Hostname	hostname where the article was posted
Timestamp	approximate timestamp of the article's publication, given in Unix time (seconds since midnight on Jan 1, 1970)

.

	ID	TITLE	URL	PUBLISHER	CATEGORY	STORY		
0	1	Fed official says weak data caused by weather,...	http://www.latimes.com/business/money/la-fi-mo...	Los Angeles Times	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.latimes.com	1394470370698
1	2	Fed's Charles Plosser sees high bar for change...	http://www.livemint.com/Politics/H2EvwJSK2VE6O...	Livemint	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.livemint.com	1394470371207
2	3	US open: Stocks fall after Fed official hints ...	http://www.ifamagazine.com/news/us-open-stocks...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371550
3	4	Fed risks falling 'behind the curve', Charles...	http://www.ifamagazine.com/news/fed-risks-fall...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371793

Figure 1. Data table

2.2. Research steps

The research steps are divided into three stages of Preprocessing, Modeling, Evaluation with the sequence of stages found in Figure 2.

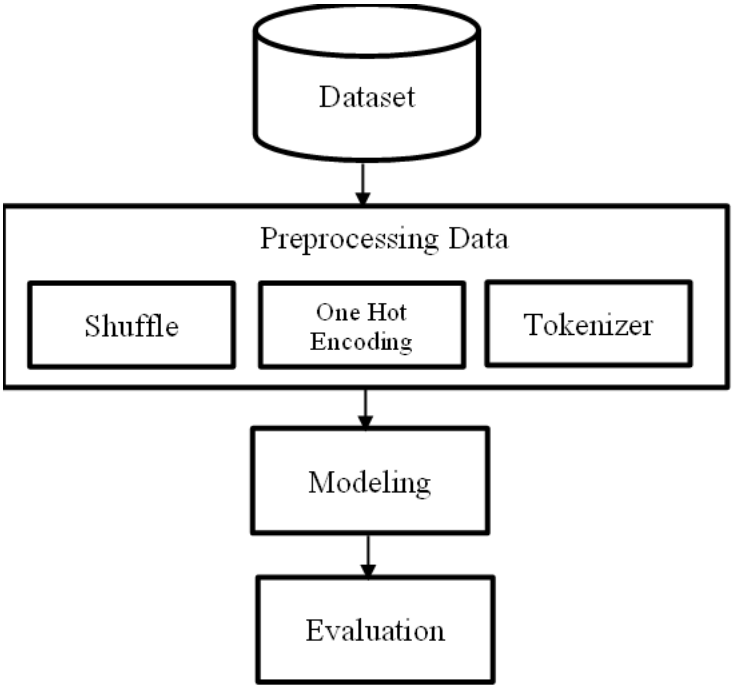

Figure 2. Research steps

2.2.1. Shuffle

Shuffle is used to not bias the data sequence during the data acquisition process [41]. So it is necessary to shuffle rows against the data set. So it is necessary to shuffle rows against the data set. Data shuffling is performed before training the model [42]. This aims to minimize data variants, generalize data well and make the model able to study the data well so as to reduce overfitting on the model.

2.2.2. One-hot encoding

One-Hot Encoding is an alternative process used for multi-class classification problems [43]–[45]. The one hot encoding process represents the category model data into a binary vector that has integer values of 1 and 0 [46], [47]. So that the data of each category class must be converted into an integer value using the One-Hot Encoding process. In this study, four classes were determined, namely e, b, t, m. The class converted to an integer obtains the number 0 for 'e', the number 1 for 'b', the number 2 for 't' and the number 3 for 'm'. The result of one hot encoding of the dataset process in Figure 3.

```
36244     0
165528    3
152141    3
131116    2
108964    2
157191    3
75932     1
142147    3
83870     1
127815    2
Name: LABEL, dtype: int64
[[1. 0. 0. 0.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [0. 0. 1. 0.]
 [0. 0. 1. 0.]
 [0. 0. 0. 1.]
 [0. 1. 0. 0.]
 [0. 0. 0. 1.]
 [0. 1. 0. 0.]
 [0. 0. 1. 0.]]
'\n [1. 0. 0. 0.] e\n [0. 1. 0. 0.] b\n [0. 0. 1. 0.] t\n [0. 0. 0. 1.] m\n'
```

Figure 3. One hot encoding result

2.2.3. Tokenizer

At this stage the process checks all the text in the data and cuts the text into a set of tokens and/or sentences. In the tokenizer, the removal of all punctuation marks is also carried out, symbols such as'!" #$%&()*+,-./:;<=>?@[\]^_`{|} ~' [48]. The tokenizer in this study used num_words parameter set to 8000 and max_len 130 so that 51806 tokens were obtained which were retrieved in the data. The tekonizer process uses the text_to_sequences method.

2.2.4. Modelling

In modeling the LSTM algorithm using softmax function activation with the number of 4 neurons with Adam (Adaptive moment estimation) optimization. Adam is a combination of RMSprop, adaptive learning rate and momentum. Adam works by changing the accumulation Gradient into Weighted Moving Average [49]. Then for evaluation using accuracy and loss categorical_crosentropy metrics to find out the loss value. The parameters used in the training process are batch sizes 256, emb_dim 128 and epoch 10 with the application seen in the model summary in Figure 4.

```
((135000, 130), (135000, 4), (45000, 130), (45000, 4))
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 130, 128)          1024000

 spatial_dropout1d (SpatialD  (None, 130, 128)         0
 ropout1D)

 lstm (LSTM)                 (None, 64)                49408

 dense (Dense)               (None, 4)                 260

=================================================================
```

Figure 4. Model summary

.

2.2.5. Evaluation

In this study, the model evaluation used accuracy and loss metrics. Loss uses 'categorical_crossentropy' as it is a classification of many classes. Accuracy is a measure of the proportion of correct data predictions based on the total amount of data [50]. The accuracy calculation formula is given in Equation (1).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

3. RESULTS AND DISCUSSIONS

The results of the test scenarios shown are presented in Table 1 as follows:

Epoch	Acc	Loss
2	87,78%	0.3495
4	91,03%	0.2597
6	92,15%	0.2261
8	92,66%	0.2100
10	93,15%	0.1968

Table 1. Test Results

Table 1 is the result of testing the LSTM algorithm with batch size parameters of 256, emb_dim of 128 and epochs of 10 obtained accuracy results increasing and losses decreasing. So that an accuracy test set of 93,15% with a loss of 0.1968 was obtained.

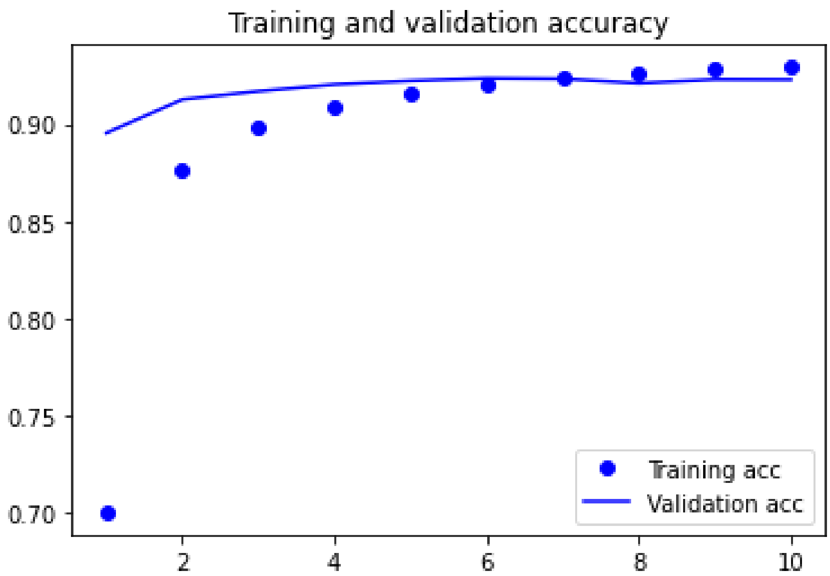

Figure 5. LSTM Train trial accuracy results

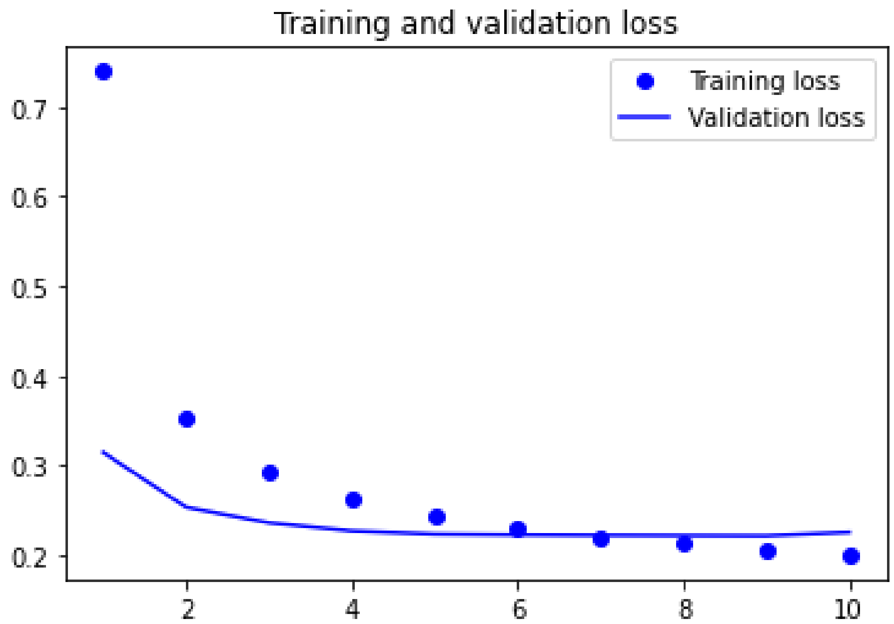

Figure 6. LSTM loss trial results

Testing the application to a sentence with labeling 'entertaiment' , 'bussiness', 'science/tech', 'health' can be seen in Figure 5 and Figure 6.

```
txt = ["For the last few years, text mining has been gaining significant importance. Since Knowledge is now available to users through variety
seq = tokenizer.texts_to_sequences(txt)
padded = pad_sequences(seq, maxlen=max_len)
pred = model.predict(padded)
labels = ['entertainment', 'bussiness', 'science/tech', 'health']
print(pred, labels[np.argmax(pred)])
```

```
/1 [==============================] - 0s 51ms/step
[0.00193299 0.0021199  0.991197   0.0047502 ]] science/tech
```

Figure 7. Sentence application

The results of the test of applying a sentence successfully classifying are seen in Figure 7. Testing of the LSTM algorithm on the application of effective text classification with a high degree of accuracy. This is also confirmed in research Putra et al [35], Astari et al [36] the use of the LSTM algorithm can be as an alternative to the classification of texts, especially news texts.

4. CONCLUSION

In this study, the LSTM algorithm on the data was classified into four categories, namely entertainment, bussines, science, and health. The results obtained by doing as much as 10 times the epoch of potential accuracy on the data with a high accuracy value of 93,15%. This strengthens the classification of news texts using deep learning methods with the LSTM algorithm effective as an alternative used in text classification.

REFERENCES

[1] A. D. Arifin, I. Arieshanti, and A. Z. Arifin, "Implementasi algoritma k-nearest neighbor yang berdasarkan one pass clustering untuk kategorisasi teks," *ITS, Surabaya*, pp. 1–7, 2012.

[2] A. Y. Rofiqi, "Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat," *J. Simantec*, vol. 6, no. 1, 2017.

[3] R. Hartono, Y. Wibisono, and R. A. Sukamto, "Damropa (Damage Roads Patrol): Aplikasi Pendeteksi Jalan Rusak Memanfaatkan Accelerometer pada Smartphone," *OSF Prepr.*, 2017, doi: https://doi.org/10.31219/osf.io/yekpr.

[4] A. Rizaldy and H. A. Santoso, "Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents," in *2017 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2017, pp. 227–232.

.

[5] W. B. Trihanto, R. Arifudin, and M. A. Muslim, "Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Naive Bayes Classifier," *Sci. J. Informatics*, vol. 4, no. 2, pp. 179–190, 2017, doi: 10.15294/sji.v4i2.11876.

[6] N. P. Ririanti and A. Purwinarko, "Implementation of Support Vector Machine Algorithm with Correlation-Based Feature Selection and Term Frequency Inverse Document Frequency for Sentiment Analysis Review Hotel," *Sci. J. Informatics*, vol. 8, no. 2, pp. 297–303, 2021, doi: 10.15294/sji.v8i2.29992.

[7] U. I. Larasati, M. A. Muslim, R. Arifudin, and A. Alamsyah, "Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis," *Sci. J. Informatics*, vol. 6, no. 1, pp. 138–149, 2019.

[8] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, and A. S. Bahri, "Comparison of LSTM , SVM , and Naive Bayes for Classifying Sexual Harassment Tweets," *J. Soft Comput. Explor.*, vol. 3, no. 2, pp. 131–137, 2022, doi: https://doi.org/10.52465/joscex.v3i2.85.

[9] Sulistiana and M. A. Muslim, "Support Vector Machine (SVM) Optimization Using Grid Search and Unigram to Improve E-Commerce Review Accuracy," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 8–15, 2020.

[10] A. Falasari and M. A. Muslim, "Optimize Naïve Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency For Amazon Review Sentiment Analysis," *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 31–36, 2022, doi: 10.52465/joscex.v3i1.68.

[11] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 1–7, 2020.

[12] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese news text classification based on machine learning algorithm," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2018, vol. 2, pp. 48–51.

[13] P. Barberá, A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler, "Automated text classification of news articles: A practical guide," *Polit. Anal.*, vol. 29, no. 1, pp. 19–42, 2021.

[14] S. Kaur and N. K. Khiva, "Online news classification using deep learning technique," *Int. Res. J. Eng. Technol.*, vol. 3, no. 10, pp. 558–563, 2016.

[15] L. Deping, W. Hongjuan, L. Mengyang, and L. Pei, "News text classification based on Bidirectional Encoder Representation from Transformers," in *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, 2021, pp. 137–140. doi: 10.1109/CAIBDA53561.2021.00036.

[16] Y. Zhu, "Research on News Text Classification Based on Deep Learning Convolutional Neural Network," *Wirel. Commun. Mob. Comput.*, vol. 2021, p. 1508150, 2021, doi: 10.1155/2021/1508150.

[17] N. Sun and C. Du, "News Text Classification Method and Simulation Based on the Hybrid Deep Learning Model," *Complexity*, vol. 2021, p. 8064579, 2021, doi: 10.1155/2021/8064579.

[18] W. Zhao, L. Zhu, M. Wang, X. Zhang, and J. Zhang, "WTL-CNN: a news text classification method of convolutional neural network based on weighted word embedding," *Conn. Sci.*, vol. 34, no. 1, pp. 2291–2312, 2022, doi: 10.1080/09540091.2022.2117274.

[19] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 2018, pp. 890–893. doi: 10.1109/ITME.2018.00199.

[20] M. Shopon, "Bidirectional LSTM with Attention Mechanism for Automatic Bangla News Categorization in Terms of News Captions," in *Electronic Systems and Intelligent Computing*, 2020, pp. 763–773.

[21] R. Saputra, A. Waworuntu, and A. Rusli, "Classification of Indonesian News using LSTM-RNN Method," in *2021 6th International Conference on New Media Studies (CONMEDIA)*, 2021, pp. 72–77. doi: 10.1109/CONMEDIA53104.2021.9617187.

[22] F. Wang, X. Deng, and L. Hou, "Chinese News Text Multi Classification Based on Naive Bayes Algorithm," in *Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control*, 2018, pp. 1–5. doi: 10.1145/3284557.3284704.

[23] Y. Ying, T. N. Mursitama, Shidarta, and Lohansen, "Effectiveness of the News Text Classification Test Using the Naïve Bayes' Classification Text Mining Method," *J. Phys. Conf. Ser.*, vol. 1764, no. 1, p. 12105, Feb. 2021, doi: 10.1088/1742-6596/1764/1/012105.

[24] Q. Wang, H. Xu, and Y. Li, "Classification of News Texts Based on Bayes Algorithm," in *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 2022, pp. 1288–1291. doi: 10.1145/3501409.3501636.

[25] U. Parida, M. Nayak, and A. K. Nayak, "News Text Categorization using Random Forest and Naïve Bayes," in *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, 2021, pp. 1–4. doi: 10.1109/ODICON50556.2021.9428925.

[26] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, 2016, pp. 112–116.

[27] A. A. Khan, S. Jamwal, and M. M. Sepehri, "Applying Data Mining to Customer Churn Prediction in an Internet Service Provider," *Int. J. Comput. Appl.*, vol. 9, no. 7, pp. 8–14, 2010, doi: 10.5120/1400-1889.

[28] I. A. Kandhro *et al.*, "Classification of Sindhi Headline News Documents based on TF-IDF Text Analysis Scheme," *Indian J. Sci. Technol.*, vol. 12, no. 33, pp. 1–10, 2019.

[29] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018.

[30] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and others, "News article text classification in Indonesian language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017.

[31] X. Li and H. Ning, "Chinese text classification based on hybrid model of CNN and LSTM," in *Proceedings of the 3rd International Conference on Data Science and Information Technology*, 2020, pp. 129–134.

[32] X. She and D. Zhang, "Text classification based on hybrid CNN-LSTM hybrid model," in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, 2018, vol. 2, pp. 185–189.

[33] G. Nergız, Y. Safali, E. Avaroğlu, and S. Erdoğan, "Classification of Turkish News Content by Deep Learning Based LSTM Using Fasttext Model," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–6. doi: 10.1109/IDAP.2019.8875949.

[34] M. Zhang, "Applications of deep learning in news text classification," *Sci. Program.*, vol. 2021, p. 9, 2021, doi: https://doi.org/10.1155/2021/6095354.

[35] O. V. Putra, A. Musthafa, and K. P. Wibowo, "Klasifikasi Ekspresi Teks Berbahasa Jawa Menggunakan Algoritma Long Short Term Memory," *Komputika J. Sist. Komput.*, vol. 10, no. 2, pp. 137–143, 2021.

[36] Y. yuli Astari, A. Afiyati, and S. W. Rozaqi, "Analisis Sentimen Multi-Class pada Sosial Media menggunakan metode Long

Short-Term Memory (LSTM)," *J. Linguist. Komputasional*, vol. 4, no. 1, pp. 8–12, 2021.
[37] Y. Widhiyasana, T. Semiawan, I. G. A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 4, pp. 354–361, 2021.
[38] F. Qian and X. Chen, "Stock prediction based on LSTM under different stability," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2019, pp. 483–486.
[39] F. Landi, L. Baraldi, M. Cornia, and R. Cucchiara, "Working memory connections for LSTM," *Neural Networks*, vol. 144, pp. 334–341, 2021.
[40] Y. Huang, X. Dai, Q. Wang, and D. Zhou, "A hybrid model for carbon price forecasting using GARCH and long short-term memory network," *Appl. Energy*, vol. 285, p. 116485, 2021.
[41] S. Al Faraby and A. Romadhony, "Pengaruh Distribusi Panjang Data Teks pada Klasifikasi: Sebuah Studi Awal," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 3, pp. 1501–1508, 2022.
[42] T. T. Nguyen *et al.*, "Why globally re-shuffle? Revisiting data shuffling in large scale deep learning," in *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2022, pp. 1085–1096.
[43] C.-H. Chen, P.-H. Lin, J.-G. Hsieh, S.-L. Cheng, and J.-H. Jeng, "Robust multi-class classification using linearly scored categorical cross-entropy," in *2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*, 2020, pp. 200–203.
[44] M. N. Rizaldi, A. Adiwijaya, and S. Al Faraby, "Klasifikasi Argument Pada Teks dengan Menggunakan Metode Multinomial Logistic Regression Terhadap Kasus Pemindahan Ibu Kota Indonesia di Twitter," *J. Media Inform. Budidarma*, vol. 4, no. 4, pp. 904–913, 2020.
[45] P. Rodr\'\iguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image Vis. Comput.*, vol. 75, pp. 21–31, 2018.
[46] P. Arsi, L. N. Hidayati, and A. Nurhakim, "Komparasi Model Klasifikasi Sentimen Issue Vaksin Covid-19 Berbasis Platform Instagram," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 1, pp. 459–466, 2022.
[47] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing." 2018.
[48] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *J. Khatulistiwa Inform.*, vol. 7, no. 1, 2019.
[49] S. Bera and V. K. Shrivastava, "Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2664–2683, 2020.
[50] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Informatics*, vol. 36, pp. 82–93, 2019.

.