International Journal of Data and Network Science 6 (2022). 347–354

Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

A combined deep learning model based on the ideal distance weighting method for fake news detection

Sarayut Gonwirat[a], Atchara Choompol[a*] and Narong Wichapa[b]

[a]Department of Computer and Automation Engineering, Faculty of Engineering and Industrial Technology, Kalasin University, Kalasin, 46000, Thailand
[b]Department of Industrial Engineering, Faculty of Engineering and Industrial Technology, Kalasin University, Kalasin, 46000, Thailand

CHRONICLE

Article history:
Received: October 25, 2021
Received in revised format: November 20, 2021
Accepted: January 5, 2022
Available online: January 5 2022

Keywords:
Fake news detection
Deep learning
Weighting method

ABSTRACT

Fake news has become a major problem affecting people, society, the economy and national security. This work proposes a combined deep learning model based on the ideal distance weighting method for fake news detection. The proposed model was validated on the ISOT and COVID-19 fake news datasets. Firstly, the ISOT and COVID-19 fake news datasets were collected. Secondly, the training-based models were used to provide accuracy values. After that, these values were transformed into criteria weights using the new ideal distance weighting method. Finally, the prediction value of the proposed model is calculated by the criteria weights. The results show that the proposed method is effective to distinguish the fake news datasets.

1. Introduction

With the rapid increase in communication technology, most people get news through social media, since it is a way to get news easily and quickly. News channels have used social media platforms for news distribution by providing members with the latest news in near real time. News media have evolved from newspapers and magazines to digital formats such as blogs, social media and other digital media formats (Ahmad, Yousaf, Yousaf, & Ahmad, 2020). These social media platforms are extremely efficient and helpful, enabling people to discuss and share ideas and debate on issues such as politics, education, business and health. Since the recipient can access the news easily and quickly, it has led some people to create distorted news for certain purposes, such as creating fake news to attack business competitors, disseminating news to hate individuals or organizations and clickbait (Aldwairi & Alwahedi, 2018). Fake news is one of the greatest threats to impact on commerce and politics all over the world. For example, the fake news about US president Barak Obama being injured in an explosion caused a loss in the stock market of US $130 billion. Another case of fake news campaigns that show a great impact is the escalation of tensions between Pakistan and India, which began with fake news reporting on the Balakot strike, and resulted in an enormous death toll of military personnel (Nasir et al., 2021; Popat et al., 2017; Zarrella & Marsh, 2016).

Currently, deep learning methods can be used to detect fake news on social media (Umer et al., 2020; Awan, 2020). The convolutional neural network (CNN) and recurrent neural networks (RNN) are two effective methods for natural language processing (NLP) tasks (Chung et al., 2014; He et al., 2015). In addition, selecting an optimizer tool is another important factor for improving the accuracy of a classification model. The Stochastic Gradient Descent (SGD), Root Mean Squared Propagation (RMSprop), Adaptive Moment Estimation (Adam), Nesterov-Accelerated Adaptive Moment Estimation (Nadam) and Adamax (Kingma & Ba, 2014; Ruder, 2016) are five popular optimizers that are widely used to change the weights and learning rate for reducing the losses. Many researchers have proven that a combination of deep learning models

* Corresponding author.
E-mail address: atchara.ch@ksu.ac.th (A. Choompol)

doi: 10.5267/j.ijdns.2022.1.003

can often perform better than any single model (Tan, Lim, & Cheah, 2014; Borovkova & Tsiamas, 2019; Yan, He, Zhang, & Xie, 2019; Rojarath & Songpan, 2021; Stab et al., 2018). Hence, the idea of the ensemble method is one of various ways to improve overall performance. There are several methods for calculating the objective weight, such as raking, cell-weighting, matching and propensity weighting. Each method has its advantages and disadvantages. The ideal distance weighting method is one popular method that is used to provide the objective weight. Therefore, this research presents a weighting method based on the ideal weighting method for calculating the weights of each model. The accuracy values obtained from training models can be transformed into criteria weights of each model using the ideal weighting method. Then multiply the obtained weights with the prediction values to calculate a new prediction.

2. Literature review

In this section, we have reviewed various techniques that have been proposed in the literature for fake news classification, a deep learning model and the ensemble learning method (Mohammad et al., 2016). Jindal and Liu (Jindal & Liu, 2008) first proposed research on spam reviews; they classified spam reviews into three categories, including unreal reviews, reviews on brands and irrelevant reviews. The user's behavior approach was offered to solve for filtering strategies and malicious behavior in networks. Aldwairi and Alwahedi (2018) proposed an effective approach to detect and filter websites that contain fake news or clickbait ads on websites. The experimental results show a 99.40% accuracy using a logistic classifier. Choudhary et al. (2021) offered a hybrid BerConvoNet model based on the concatenation of BERT and convolutional neural networks for fake news classification. The proposed model indicated that the BerConvoNet was powerful in identifying the truthfulness of news articles. Li et al. (Li, Cardie, & Li, 2013) offered an effective topic model based on Latent Dirichlet Allocation (LDA) for detecting deceptive reviews. Umer et al. (2020) proposed a deep learning architecture (CNN-LSTM) with Principle Component Analysis (PCA) and Chi-Squared for fake news stance detection. Thota et al. (2018) presented a neural network architecture to accurately predict fake news. The results show that the proposed model achieved an accuracy of 94.21%. Nasir et al. (2021) presented a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification. The proposed method was validated on two fake news datasets (ISOT and FA-KES). Based on the proposed results, the presented model is significantly better than other non-hybrid baseline methods. Xia et al. (2019) presented an ensemble algorithm of LSTM (eLSTM) to improve the outcome prediction in an intensive care unit. The proposed algorithm was tested with 18415 items of the Medical Information Mart for Intensive Care III (MIMIC-III) database. The results show that the proposed eLSTM algorithm is significantly better than other non-hybrid baseline approaches. Huang and Chen (Huang & Chen, 2020) proposed an ensemble learning model based on the Self-Adaptive Harmony Search (SAHS) algorithm for fake news detection. The proposed model was verified with state-of-the-art methods. The results show that the proposed method achieved a highest accuracy of 99.40%. Ahmad et al. (2020) presented a machine learning ensemble approach for automated classification of news articles. The proposed model was tested with 4 real world datasets. The results show that the proposed ensemble approach is better than individual approaches.

From the literature above, the idea of an ensemble learning method is to overcome the disadvantages of individual models, because the results of ensemble algorithms are significantly better than individual models. Although many methods were proposed to solve the fake news datasets, the CNN-LSTM model is one popular method to tackle fake news. The results of CNN-LSTM models in the above literature have shown that these models are powerful to discriminate the fake news datasets.

3. Proposed approach

This paper proposes a combined deep learning model based on the ideal distance weighting method for fake news detection. The five conventional models, the CNN-LSTM with SGD (M_1), CNN-LSTM with RMSprop (M_2), CNN-LSTM with Adam (M_3), CNN-LSTM with Nadam (M_4) and CNN-LSTM with Adamax (M_5), were utilized to develop four combined models, including the first two best models (M_6), the first three best models (M_7), the first four best models (M_8) and all models (M_9). In this paper, the calculation steps are as follows: (1) the ISOT fake news datasets (Ahmed et al., 2018) (Ahmed, Traore & Saad, 2017) and the COVID-19 fake news datasets (Koirala, 2020) were collected, (2) text pre-processing included data cleaning, removing stop wording and lemmatization, (3) training based models, (4) calculating the weights of each model using the ideal distance weighting method. Finally, the prediction value of the proposed model was calculated. Details of the proposed framework are shown in Fig.1

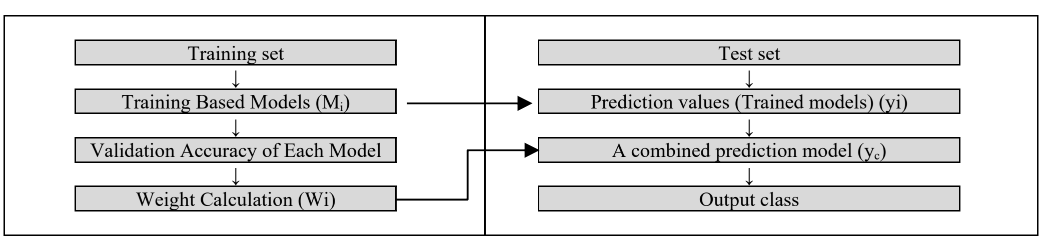

Fig. 1. The proposed framework.

3.1 Dataset and Text pre-processing

In this paper, we used the ISOT fake news datasets (Ahmed et al., 2018; Augenstein et al., 2018; Ahmed et al., 2017) and the COVID-19 fake news datasets (Koirala, 2020) for testing the proposed algorithm. The ISOT fake news datasets include two categories, true and fake. The true articles were collected from the Reuters website, and the fake ones from various sources flagged as fake sources by Wikipedia and from Politifact. The datasets comprise the full body of each article, the title, date and topic. The main article topics are "politics" and "world news" and the dates fall between 2016 and 2017. The COVID-19 fake news dataset was collected using Webhose.io and was manually labelled, within the interval of December 2019 - July 2020. It consists of 3 subcategories of news: fake news, true news and partially fake news. For the fake classification, both partially fake news and fake news were labelled 0 and true news was labelled 1. Due to data being text datasets (Unstructured format), they must be transformed to structure their data format. In this paper, the pre-processing is as follows: (1) data cleaning, where URLs, punctuation and special characters are removed from the text, (2) stop words (like is, a, an, the) are removed from the dataset using a dictionary-based technique, (3) the lemmatization process brings back multiple forms of the same word to their common root, e.g., 'coming', 'comes' into 'come'

3.2 Training based models

After obtaining the structure data format, the dataset is divided into 2 sets, including a training set of 70% and a testing set of 30%. This paper proposes the conventional deep learning models, including M_1, M_2, M_3, M_4 and M_5. The architecture of the CNN-LSTM model is shown in Table 1.

Table 1
The architecture of the CNN-LSTM model

CNN-LSTM	
Input (300,100)	Conv1D (3,64, ReLU)
Conv1D (3,32,ReLU)	MaxPooling (2)
Conv1D (3,32, ReLU)	LSTM (128)
MaxPooling (2)	Dense (1,sigmoid)

In the first layer, the training data is transformed to an embedding matrix using the pre-trained embedding layer. The second and third layers are the one-dimensional CNN layer (Conv1D) for extraction of local features using 32 filters of size 3. In this model, the activation function is defined as Rectified Linear Unit (ReLU). In the fourth layer, the large feature vectors generated using CNN are pooled by feeding them into a MaxPooling1D layer with a window size of 2. In the fifth layer, the one-dimensional CNN layer (Conv1D) uses 64 filters of size 3. The sixth layer is a MaxPooling1D layer with a window size of 2. In the next layer, the pooled feature maps are fed into the LSTM layer, and the dimension of the output is set to 128. Finally, the trained feature vectors are classified using a Dense layer that shrinks the output space dimension to 1, which corresponds to the classification label (real or fake). This layer applies the Sigmoid activation function.

3.3 Weight calculation

In this paper, the dataset is spilt into training and test subsets (70-30% split). In the training dataset, the dataset of 30% is used for validation. After that, the conventional models will be tested *n* times. The accuracy values will be used for determining the performance of each model. The accuracy values can be defined as

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100 \tag{1}$$

where *TP* is the number of cases correctly identified as true news. *TN* is the number of cases correctly identified as fake news. *FP* is the number of cases incorrectly identified as true news. *FN* is the number of cases incorrectly identified as fake news.

Finally, the accuracy values of each model will be used to evaluate the criteria weights using ideal distance weight. Details are shown in Eq. (2),

$$w_j = \frac{s_j^-}{(s_j^- + s_j^+)} \tag{2}$$

$$s_j^- = \sum_{i=1}^{m} ac_{ij} - n \cdot ac^-, \; \forall j = 1,2,3,\ldots,n \tag{3}$$

$$s_j^+ = n \cdot ac^+ - \sum_{i=1}^{m} ac_{ij}, \forall j = 1,2,3,\ldots,n \tag{4}$$

where ac^{-} is a minimum accuracy value of all models.
ac^{+} is a maximum accuracy value of all models.

After obtaining the weights of each model, the weights will be normalized as in Equation (5)

$$w_j^* = \frac{w_j}{\sum_{i=1}^{n} w_j} \tag{5}$$

3.4 A combined prediction models

The prediction value of proposed models (y_c) is obtained by multiplying the weight value of each model (w_j^*) by the corresponding prediction of each model (y_j) using Equation (6).

$$y_c = \sum_{j=1}^{n} w_j^* \cdot y_j \tag{6}$$

4. Evaluation results

4.1 Data collection

This work uses the ISOT fake news dataset (Ahmed, Traore, & Saad, 2018) and the COVID-19 fake news dataset (Koirala, 2020). The ISOT dataset consists of 45,000 news articles and the COVID-19 dataset consists of 3,117 articles. The characteristics of the datasets are shown in Table 2 and Table 3.

Table 2
The characteristics of the ISOT fake news dataset

Label of News	***Number of articles***	***Article type***	***Articles***
True	21,415	World news	10,145
		Politics news	11,272
Fake	23,481	Government news	1,570
		Middle east	778
		Left news	4,459
		Politics	6,841
		News	9,050

Table 3
The characteristics of the COVID-19 fake news dataset

Label of News	***Articles***
True	2,060
Fake	1,057

4.2 Results of training performance

This paper aimed to formulate the new models from five conventional models, including CNN-LSTM with SGD (M_1), CNN-LSTM with Rmsprop (M_2), CNN-LSTM with Adam (M_3), CNN-LSTM with Nadam (M_4) and CNN-LSTM with Adamax (M_5). These models were validated with 15 epochs for each model. The validation accuracy of each conventional model on the ISOT dataset of each epoch during the training process for evaluating ensemble models is shown in Table 4 and Table 5.

Table 4
The Validation Accuracy of Conventional Models on the ISOT fake news dataset

Epoch(i)	Accuracy of each model (j)				
	M_1	M_2	M_3	M_4	M_5
1	0.528	0.806	0.949	0.847	0.916
2	0.528	0.858	0.983	0.974	0.974
3	0.528	0.974	0.989	0.977	0.979
4	0.730	0.982	0.953	0.981	0.984
5	0.785	0.978	0.986	0.956	0.986
6	0.813	0.988	0.986	0.985	0.887
7	0.832	0.991	0.984	0.987	0.918
8	0.821	0.992	0.992	0.986	0.971
9	0.847	0.990	0.990	0.987	0.987
10	0.861	0.991	0.987	0.988	0.987
11	0.847	0.984	0.990	0.989	0.983
12	0.857	0.989	0.989	0.987	0.987
13	0.890	0.989	0.989	0.984	0.988
14	0.871	0.990	0.991	0.988	0.986
15	0.915	0.992	0.991	0.988	0.988
$\sum_{i=1}^{m} ac_{ij}$	11.654	14.494	14.749	14.604	14.520

After obtaining the validation accuracy of each model, the accuracy values were used to generate w_j using Eq. (2) and Eq. (3). For example, a combined 2 models (M_2 and M_3) was selected based on the top two maximum accuracy values.

$$w_1 = \frac{s_1^-}{(s_1^- + s_1^+)}$$

where ac^- = min(0.806, 0.858,…,0.991) = 0.806, ac^+ = max(0.806, 0.858,…,0.991) = 0.992 and n =15.

s_1^- = (14.494 - (15×0.806)) = 2.398, s_1^+ = ((15×0.992) - 14.494) = 0.385

w_1 = 2.398/(2.398+0.385) = 0.862

With the same calculation steps, w_2 was 0.954. After that, the w_j^* were obtained using Eq. (5) For example,

$$w_1^* = \frac{w_1}{w_1 + w_2}$$

w_1^* = 0.862/(0. 862 + 0.954) = 0.475

With the same calculation steps, w_2^* was 0.525.

The weights of each model were taken into Eq. (6): y_c = 0.475 y_1 + 0.525 y_2 . Finally, the new prediction model was tested with the testing set. The performances of each model on the ISOT fake news dataset and the COVID-19 fake news dataset are shown in Table 6 and Table 7.

Table 6
The performance of each model on ISOT fake news dataset

Model	Accuracy
CNN-LSTM-SGD (M_1)	0.9176
CNN-LSTM-Rmsprop (M_2)	0.9928
CNN-LSTM-Adam (M_3)	0.9925
CNN-LSTM-Nadam (M_4)	0.9927
CNN-LSTM-Adamax (M_5)	0.9887
A combination of 2 models (M_6)	0.9941
A combination of 3 models (M_7)	0.9956*
A combination of 4 models (M_8)	0.9952
A combination of 5 models (M_9)	0.9907

Table 7
The performance of each model on the COVID-19 fake news dataset

Model	Accuracy
CNN-LSTM-SGD (M_1)	0.6709
CNN-LSTM-Rmsprop (M_2)	0.7254
CNN-LSTM-Adam (M_3)	0.7201
CNN-LSTM-Nadam (M_4)	0.7233
CNN-LSTM-Adamax (M_5)	0.6998
A combination of 2 models (M_6)	**0.7372***
A combination of 3 models (M_7)	**0.7372***
A combination of 4 models (M_8)	**0.7361**
A combination of 5 models (M_9)	**0.7329**

As seen in Table 6 and Table 7 based on comparisons with based models, the solutions of the proposed model can achieve maximum accuracy for the datasets. Clearly, the proposed weighting method based on the ideal weighting method is simple but powerful. The results of both datasets show that the proposed weighting method is most effective when applied to a combination of three models (M_7). However, application of the proposed model should be tested with more datasets, and we believe that the ideas of the proposed model can be applied to integrate with other algorithms to enhance the validity of the research output further.

5. Conclusions

This paper presents combined deep learning models based on the ideal distance weighting method for solving the fake news detection problem. The proposed method was tested with ISOT fake news and COVID-19 fake news datasets. Firstly, text pre-processing including data cleaning, removing stop wording and lemmatization was performed. Secondly, training-based models were tested. Next, the criteria weights of each model were calculated using the ideal distance weighting method. Finally, the prediction values of each model were evaluated. The results have shown that the proposed model can obtain the maximum accuracy.

For future research, the limitations of this paper lie in that only the ISOT fake news and the COVID-19 fake news dataset were studied. Application of the proposed model should be tested with more datasets of classification problems to enhance the validity of the research output further. In addition, we believe that the ideas of the proposed model can be used to integrate with other algorithms for improving the validity of research output further.

Acknowledgement

The authors are very grateful to the Department of Computer and Automation Engineering, Kalasin University for supporting this research. Finally, the authors would like to thank the reviewers for their valuable comments and recommendations which enabled the improvement of the quality of this paper.

References

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, 2020, pp.1-11.

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, *1*(1), 1-15.

Aldwairi, M., & Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, *141*(1), 215–222.

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. *ArXiv Preprint ArXiv:1606.05464*.

Awan, M. J. (2020). Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory. Article in *International Journal of Emerging Technologies in Learning (IJET)*, *11*(5), 209–212.

Borovkova, S., & Tsiamas, I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal of Forecasting*, *38*(6), 600–619.

Choudhary, M., Chouhan, S. S., Pilli, E. S., & Vipparthi, S. K. (2021). BerConvoNet: A deep learning framework for fake news classification. *Applied Soft Computing*, *110*, 107614.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv Preprint ArXiv:1412.3555*.

He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1576–1586).

Huang, Y.-F., & Chen, P.-H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, *159*, 113584.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Koirala, A. (2020). COVID-19 Fake News Classification with Deep Learning. *Preprint*.

Li, J., Cardie, C., & Li, S. (2013). Topicspam: a topic-model based approach for spam detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 217–221).

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 31–41).

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), pp.100007.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1003–1012).

Rojarath, A., & Songpan, W. (2021). Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*, *51*, 4908–4932.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *ArXiv Preprint ArXiv:1609.04747*.

Stab, C., Miller, T., & Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *ArXiv Preprint ArXiv:1802.05758*.

Tan, C. J., Lim, C. P., & Cheah, Y. (2014). A multi-objective evolutionary algorithm-based ensemble optimizer for feature selection and classification with neural network models. *Neurocomputing*, *125,* 217–228.

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: a deep learning approach. *SMU Data Science Review*, *1*(3), 10.

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B.-W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). *IEEE Access*, *8*, 156695–156706.

Xia, J., Pan, S., Zhu, M., Cai, G., Yan, M., Su, Q., … Ning, G. (2019). A long short-term memory ensemble approach for improving the outcome prediction in intensive care unit. *Computational and Mathematical Methods in Medicine*, 2019, pp.1-10.

Yan, X., He, F., Zhang, Y., & Xie, X. (2019). An optimizer ensemble algorithm and its application to image registration. *Integrated Computer-Aided Engineering*, *26*(4), 311–327.

Zarrella, G., & Marsh, A. (2016). Mitre at semeval-2016 task 6: Transfer learning for stance detection. *ArXiv Preprint ArXiv:1606.03784*.