Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 6679661, 7 pages
https://doi.org/10.1155/2021/6679661

Research Article

Fraudulent News Headline Detection with Attention Mechanism

Hankun Liu,[1] **Daojing He**,[1] **and Sammy Chan**[2]

[1]*Software Engineering Institute, East China Normal University, Shanghai 200062, China*
[2]*Department of Electrical Engineering, City University of Hong Kong, Hong Kong 999077, China*

Correspondence should be addressed to Daojing He; djhe@sei.ecnu.edu.cn

Received 28 November 2020; Revised 3 February 2021; Accepted 27 February 2021; Published 15 March 2021

Academic Editor: Amparo Alonso-Betanzos

Copyright © 2021 Hankun Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

E-mail systems and online social media platforms are ideal places for news dissemination, but a serious problem is the spread of fraudulent news headlines. The previous method of detecting fraudulent news headlines was mainly laborious manual review. While the total number of news headlines goes as high as 1.48 million, manual review becomes practically infeasible. For news headline text data, attention mechanism has powerful processing capability. In this paper, we propose the models based on LSTM and attention layer, which fit the context of news headlines efficiently and can detect fraudulent news headlines quickly and accurately. Based on multi-head attention mechanism eschewing recurrent unit and reducing sequential computation, we build Mini-Transformer Deep Learning model to further improve the classification performance.

1. Introduction

With the rapid development of Internet, Internet security is suffering from various potential threats. The rise of Advanced Persistent Threat (APT) has caused traditional network defense systems to face increasingly severe challenges. According to statistics, social engineering is the main technique that attackers use to launch APT attacks, so it is of practical significance to research on defending against social engineering attacks. Cutting off the chains of attacks, detecting attacks and isolating attackers is the fastest and most effective method of defending against social engineering attacks.

Currently, in major cases of social engineering attacks, the essential operation of attackers to launch attacks is to distribute fraudulent news headlines on e-mail systems and online social media platforms, such as Instant Messaging services (e.g., QQ, WeChat, WhatsApp, Facebook Messenger, and Line) or microblogs (e.g., Twitter and Weibo). Some fraudulent news headlines often carry malicious links preset by attackers. Many curious users who see those news headlines would want to learn more about the detailed contents of those news by clicking directly on the malicious links, which leads to serious consequences, including personal privacy theft, account and password stealing, and even huge asset loss.

According to Symantec Internet Security Threat [1] (ISTR Volume 23), for the social engineering attacks on companies, 71.4% of targeted attacks in 2017 involved the use of spear-phishing e-mails. Therefore, the main vector of social engineering attacks to reach companies through their employees remains e-mail system.

Above all, it is of great importance to analyze and detect fraudulent news headlines, which has a profound impact on Internet security and the defense system against social engineering attacks.

In recent years, Deep Learning models, such as Long Short-Term Memory (LSTM) [2], attention layer [3] and Transformer [4], have demonstrated outstanding advantages in solving the problems of Natural Language Processing (NLP). In this paper, for the classification of news headline text data, we add one extra attention layer to the LSTM model and achieve a slight increase in accuracy. In addition, based on multi-head attention, we build Mini-Transformer without complex recurrent or convolutional neural networks to improve the classification performance (i.e., accuracy, precision, recall, and F1 score) dramatically.

2. Related Work

Although a considerable amount of literature has been published on Internet social engineering, the emerging security issues with e-mail systems and online social media platforms are still not addressed adequately. Moreover, since the operational principle of social engineering attacks has not been clearly revealed, it is difficult to construct an effective defense system.

Castillo et al. [5] raised the issue of fake information detection on Twitter. To examine newsworthy topics on Twitter, they evaluated various classification algorithms and analyzed four features (message, user, topic, and propagation). Automatic method was used to classify the credibility of Twitter messages and achieved high precision and recall.

Ma et al. [6] utilised Recurrent Neural Networks (RNN), including LSTM and Gated Recurrent Unit (GRU), to process massive text data. They proposed a novel method that learns continuous representations of microblog events for identifying rumors on Twitter and Weibo more quickly and accurately.

Guo et al. [7] investigated the relevant characteristics of social media and utilised attention mechanism to analyze the massive news and messages on the microblog. They designed an efficient classification scheme, which can detect rumors more accurately.

Song et al. [8] combined LSTM with attention mechanism and proposed a novel method of sentiment lexicon embedding for aspect-level sentiment analysis, which is better at representing sentiment word's semantic relationships to improve the sentiment classification performance.

Vaswani et al. [4] proposed a new network architecture (Transformer) based solely on attention mechanism, which is not only superior in machine translation quality but also more parallelizable so as to require significantly less time to train.

Our work focuses on the news headlines spread on e-mail systems and online social media platforms. We develop a set of models to detect massive fraudulent news headlines using LSTM and attention mechanism. To further improve the classification performance, we build Mini-Transformer, which consists of multi-head attention layers and fully connected dense layers rather than recurrent unit layers (i.e., LSTM layer and GRU layer).

3. Methodology

In this section, firstly, we briefly revisit LSTM [2]. Then, we present the formulations of attention layer proposed by Bahdanau et al. [3]. Finally, we show how we use multi-head attention mechanism to build Mini-Transformer.

3.1. Long Short-Term Memory (LSTM) Networks. LSTM is able to process variable-length input sequences by recursive operation [2]. With the ability to maintain the hidden states and fit the variations of contextual information in relevant time steps, LSTM is well-suited for classifying news headline text data.

Unlike the traditional Vanilla RNN unit whose hidden state is overwritten in each time step, LSTM unit maintains long memory cell state c_t in time step t. Given an input sequence $X = \{x_1, x_2, x_3, \ldots, x_{\text{len}}\}$ with length len, $\{x_t | 1 \le t \le \text{len}\}$ are real number vectors with dimension d_x, hidden state sequence $\{h_1, h_2, h_3, \ldots, h_{\text{len}}\}$ with length len, $\{h_t | 1 \le t \le \text{len}\}$ are real number vectors with dimension d_h, and long memory cell state sequence $\{c_1, c_2, c_3, \ldots, c_{\text{len}}\}$ with length len, $\{c_t | 1 \le t \le \text{len}\}$ are also real number vectors with dimension d_h. From $t = 1$ to len, the algorithm iterates as follows:

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot [\, h_{t-1}, x_t \,]\right), \\ i_t &= \sigma\left(W_i \cdot [\, h_{t-1}, x_t \,]\right), \\ \widetilde{c}_t &= \tanh\left(W_c \cdot [\, h_{t-1}, x_t \,]\right), \\ o_t &= \sigma\left(W_o \cdot [\, h_{t-1}, x_t \,]\right), \\ c_t &= f_t * c_{t-1} + i_t * \widetilde{c}_t, \\ h_t &= o_t * \tanh\left(c_t\right), \end{aligned} \tag{1}$$

where W_f, W_i, W_c, and W_o are weight matrices for forget gate, input gate, long memory cell, and output gate, respectively. The operator "$\cdot$" denotes the dot-product between the matrix and vector. The operator "$*$" denotes the element-wise multiplication (Hadamard product) between two vectors. $\sigma(\cdot)$ is the logistic sigmoid function, and $\tanh(\cdot)$ is the hyperbolic tangent function.

In LSTM unit, forget gate f_t controls the range of existing memory c_{t-1} removed from c_t, input gate i_t controls the range of new memory $\widetilde{c}_t$ added to c_t, and output gate o_t determines the amount of output memory. By removing part of the existing memory c_{t-1} and adding part of the new memory $\widetilde{c}_t$, long-term memory cell c_t is updated. LSTM unit is illustrated in Figure 1.

From $t = 1$ to len, after all iterative steps of the algorithm, last hidden state vector h_{len} is computed to generate real number output y via a fully connected dense layer whose activation is logistic sigmoid function.

3.2. Attention Layer. In 2014, Bahdanau et al. [3] introduced the attention mechanism to the NLP field for the first time and completed modeling, transduction, and alignment procedure on the machine translation task at the same time.

LSTM layer needs to return all hidden states $\{h_t | 1 \le t \le \text{len}\}$ as the input of attention layer. In attention layer, attention weight scores $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{\text{len}}\}$ are computed with v_α, W_α, and input sequence $\{h_1, h_2, h_3, \ldots, h_{\text{len}}\}$. $\{\alpha_i | 1 \le i \le \text{len}\}$ are real numbers reflecting the importance of each state h_i. As trainable parameters, v_α is a real number vector with dimension d_{attn}, and W_α is a real number matrix with shape (d_{attn}, d_h). From $i = 1$ to len, the algorithm iterates as follows:

$$\widetilde{\alpha}_i = v_\alpha^T \cdot \tanh\left(W_\alpha \cdot h_i\right), \tag{2}$$

where v_α^T is the transpose of v_α.

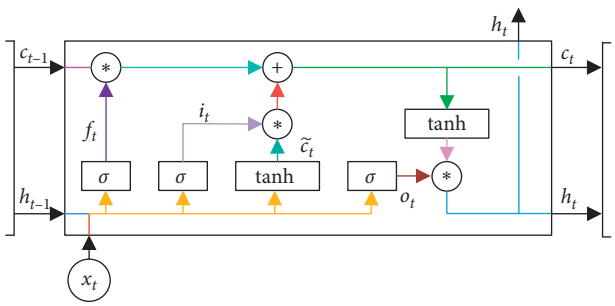

Figure 1: An illustration of LSTM unit.

For normalization that $\text{Sum}(\{\alpha_i|1 \le i \le \text{len}\}) = 1$, softmax function is called to generate α_i, i.e., $\{\alpha_i|1 \le i \le \text{len}\} = \text{Softmax}(\{\tilde{\alpha}_i|1 \le i \le \text{len}\})$. From $i = 1$ to len, the algorithm iterates as follows:

$$\alpha_i = \frac{\exp(\tilde{\alpha}_i)}{\sum_{j=1}^{\text{len}} \exp(\tilde{\alpha}_j)}, \tag{3}$$

where $\exp(.)$ is the exponential function.

We take a weighted sum of all states $\{h_i|1 \le i \le \text{len}\}$ as computing an expected state h_{sum} with dimension d_h, which is similar to h_{len}. The formula reads as follows:

$$h_{\text{sum}} = \sum_{i=1}^{\text{len}} \alpha_i h_i. \tag{4}$$

Weighted sum state h_{sum} is computed to generate real number output y via fully connected dense layer whose activation is logistic sigmoid function.

3.3. Multi-Head Attention. In 2017, Vaswani et al. [4] introduced the multi-head attention mechanism, which consists of several attention heads running in parallel. Then, they built the Transformer without any recurrence or convolution to improve the machine translation quality. In addition, the Transformer is more parallelizable so as to require significantly less time to train.

In this paper, we propose a simplified Transformer, called Mini-Transformer, for the classification of news headline text data. Mini-Transformer is composed of multi-head attention layers and eschews recurrence or convolution.

For single-head dot-product attention, given an input sequence $X = \{x_1, x_2, x_3, \ldots, x_{\text{len}}\}$ with length len, where $\{x_t|1 \le t \le \text{len}\}$ are real number vectors with dimension d_x, we generate Q (Query), K (Key), and V (Value) with trainable parameter matrices, W_q, W_k, and W_v. They are all real number matrices with shape (d_{head}, d_x), where d_{head} denotes dimension of attention head. The formulas are as follows:

$$\begin{aligned} q_i &= W_q \cdot x_i, Q = \{q_i|\, 1 \le i \le \text{len}\} = \text{query}\left(W_q, X\right), \\ k_i &= W_k \cdot x_i, K = \{k_i|\, 1 \le i \le \text{len}\} = \text{key}\left(W_k, X\right), \\ v_i &= W_v \cdot x_i, V = \{v_i|\, 1 \le i \le \text{len}\} = \text{value}\left(W_v, X\right). \end{aligned} \tag{5}$$

After generating Q, K, and V with shape (len, d_{head}), we compute single dot-product attention head as follows:

$$\text{head} = \text{Attention}(Q, K, V) = \text{Softmax}\left(Q \cdot K^T\right) \cdot V. \tag{6}$$

The above dot-product single-head attention outputs a real number matrix with shape (len, d_{head}). For multi-head attention, we employ n_{head} parallel attention heads. Due to the reduced dimension of each head (d_{head}), the total computational cost is about the same as that of single-head attention with full dimensionality, but multi-head attention is more parallelizable for GPU to train. The formulas are as follows:

$$\begin{aligned} \text{head}_j &= \text{Attention}\left(Q_j, K_j, V_j\right) \\ &= \text{Softmax}\left(Q_j \cdot K_j^T\right) \cdot V_j \\ &= \text{Softmax}\left(\text{query}\left(W_{qj}, X\right) \cdot \left(\text{key}\left(W_{kj}, X\right)\right)^T\right) \\ &\quad \cdot \text{value}\left(W_{vj}, X\right), \\ \text{multi-head} &= \text{Concat}\left(\left\{\text{head}_j|1 \le j \le n_{\text{head}}\right\}\right). \end{aligned} \tag{7}$$

Finally, multi-head attention outputs a real number matrix with shape (len, $n_{\text{head}} \times d_{\text{head}}$), as depicted in Figure 2.

4. Fraudulent News Headline Detection

Our work focuses on classifying massive news headlines data into fraudulent class (label 1) and true class (label 0). There are three Deep Learning models based on LSTM, LSTM with attention layer, and Mini-Transformer, respectively. The proposed scheme consists of labeled data source, text data preprocessing and training, and test and evaluation of Deep Learning models.

4.1. Scheme Flow Chart. The flow chart of our proposed scheme for fraudulent news headline detection is shown in Figure 3.

4.2. Labeled Dataset. Three data sources are used in this paper. All of them are publicly available at Kaggle, the world's largest data science community [9].

If the length of a news headline is greater than or equal to 7, the news headline would be considered as valid news headline data. For balanced sampling, there are a total of 1,481,814 news headlines, including 736,009 items with label 1 and 745,805 items with label 0.

Fraudulent news headline dataset is The Examiner - Spam Clickbait Catalog [10]. Original source is the pseudo news site, The Examiner. At a certain point, the site was the 10th largest site on mobile and was attracting twenty million unique visitors per month. However, The Examiner no longer exists at present, Kaggle keeps the last record. Our work focuses on the fraudulent news headlines from January 1, 2013, to December 31, 2015, a total of 736,009 fraudulent news headlines (with a class label of 1).

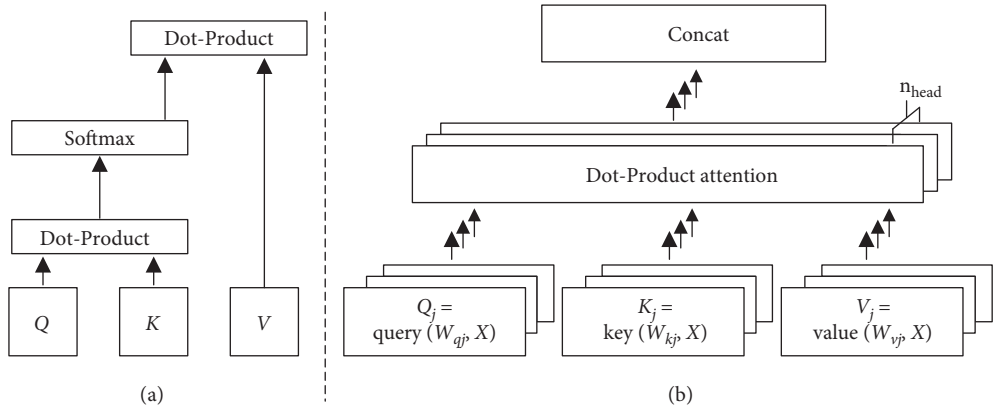

Figure 2: (a) Dot-Product attention. (b) Multi-Head attention consists of parallel attention heads.

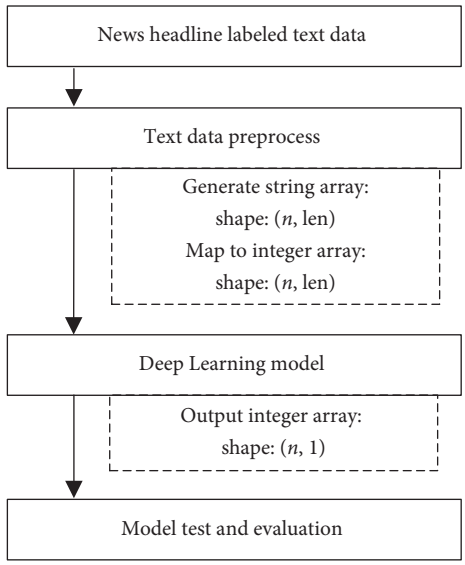

Figure 3: Fraudulent news headline detection scheme.

True news headline datasets are A Million News Headlines [11] and News Category Dataset [12], a total of 745,805 true news headlines (with a class label of 0).

For A Million News Headlines, the original source is Australian Broadcasting Corporation. It includes the entire corpus of articles published by the ABC news website. With a volume of two hundred articles per day and a good focus on international news, every event of significance has been captured. It contains a total of 577,264 true news headlines from February 19, 2003, to December 31, 2019.

For News Category Dataset, the original source is HuffPost. Each news headline has a corresponding category (e.g., parenting, style and beauty, entertainment, wellness, and politics). It contains a total of 168,541 true news headlines from January 28, 2012, to May 26, 2018.

4.3. Text Data Preprocessing. We preprocess the original labeled news headline text data, including deleting repeated news headlines, removing unnecessary English symbols (i.e., ( ) ' " , . ?: - ! #), removing redundant space characters, NLTK Lemmatization [13], truncating the news headlines that are too long, padding the news headlines that are too short, and converting uppercase letters to lowercase, etc.

The data with time order is generally called sequence. In this paper, news headline text data are typical sequences. The representation of news headlines is a two-dimensional string array with shape (n, len), where $n = 1,481,814$ is the total number of news headlines, and len is the maximum length of news headlines. For example, the two-dimensional string array can be as follows:

(i) ['tom', 'act', '*a*', 'cat', '*o*', ...] 0
(ii) ['jerry', 'act', '*a*', 'mouse', '*e*', ...] 0
(iii) ['goofy', 'act', 'the' , 'sanguine', 'dog', ...] 0
(iv) ['jerry', 'act', 'the' , 'hypothetical', 'cat', ...] 1,

where label 0 denotes true class and label 1 denotes fraudulent class.

We calculate the frequency of each English word in the two-dimensional string array, so as to identify high-frequency words and generate a high-frequency word dictionary. Significantly, in the procedure of generating the high-frequency word dictionary, our work mainly focuses on ignoring the extremely short words, marking stopwords [14] uniformly with tag 1 and marking low-frequency words uniformly with tag 2. For example, the word dictionary can be as follows:

(i) Stopword: '*a*' ->1, 'the' ->1, ...
(ii) Low-Frequency word: 'sanguine' ->2, 'hypothetical' ->2, ...
(iii) High-Frequency word: 'act' ->3, 'cat' ->4, 'jerry' ->5, 'dog' ->6, 'goofy' ->7, 'mouse'->8, 'tom' ->9, ...,

where '*a*' and 'the' are stopwords marked uniformly with tag 1, 'sanguine' and 'hypothetical' are low-frequency words marked uniformly with tag 2.

The original news headline is composed of several words; to facilitate the training and test of Deep Learning model, we map each word string to the corresponding integer based on the generated word dictionary, thus the news headline two-dimensional string array can be converted to a two-dimensional integer array with shape (n, len), e.g., the two-dimensional integer array can be as follows:

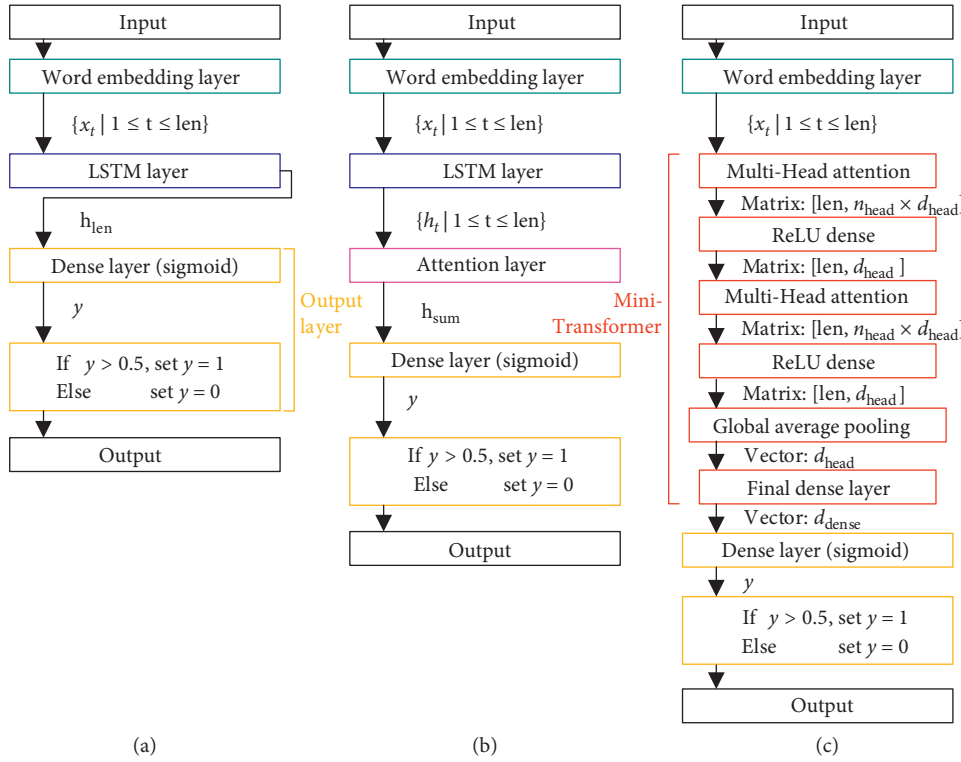

Figure 4: (a) LSTM. (b) LSTM with attention layer. (c) Mini-Transformer.

(i) [9, 3, 1, 4, . . .] 0
(ii) [5, 3, 1, 8, . . .] 0
(iii) [7, 3, 1, 2, 6, . . .] 0
(iv) [5, 3, 1, 2, 4, . . .] 1,

where 'o' and 'e' are extremely short and unnecessary words (only one letter) that have been ignored, tag 1 denotes all stopwords, tag 2 denotes all low-frequency words, and tags which are greater than or equal to 3 denote corresponding high-frequency words.

4.4. Deep Learning Model Structure. In this section, we propose three Deep Learning models, all of them contain word embedding layers and output layers; their structures are shown in Figure 4.

Word2Vec [15] provides a simple and effective method for vectorized representation of words, which can be employed in word embedding task. In word embedding layer, each integer in news headline two-dimensional integer array will be converted to real number vector with dimension d_x. Eventually, the two-dimensional integer array will be converted to a real number array with shape (n, len, d_x), i.e., each piece of news headline text data will be converted to word vector input sequence $X = \{x_1, x_2, x_3, \ldots, x_{len}\}$.

In output layer, vector h_{len}, vector h_{sum}, or the vector with dimension d_{dense} returned from final dense layer in Mini-Transformer will be converted to real number y via dense layer whose activation is logistic sigmoid function. If y is greater than threshold 0.5, it will be set to 1, else it will be set to 0.

In Mini-Transformer, we employ two layers of multi-head attention sublayer and fully connected dense sublayer without bias. It is worth noting that the activation function of dense sublayer is Rectified Linear Unit [16] (ReLU), but final dense layer has no activation function.

5. Experimental Settings and Results

If the frequency of a word is greater than or equal to 140 times, the word will be considered as high-frequency word; from word frequency statistics, the total number of high-frequency words is 7,996, so the length of word dictionary is 7,998, including low-frequency words and stopwords.

To configure the Deep Learning model for training, in tf.keras.Model.compile, we set that optimizer = Adam (learning_rate = 0.0002), loss = BinaryCrossentropy().

For word embedding layer, dimension of word vectors $\{x_t | 1 \le t \le \text{len}\}$ (d_x) is 25 and the maximum length of news headlines (len) is 15. For LSTM and attention layer, dimension of hidden states $\{h_t | 1 \le t \le \text{len}\}$ (d_h) is 16 and d_{attn} is 32.

In Mini-Transformer, for multi-head attention sublayer, number of dot-product attention heads (n_{head}) is 8 and dimension of that (d_{head}) is 64, and for final dense layer, d_{dense} is 16, which is the same as d_h.

After shuffled, 80% of original labeled dataset is split into the training set and 20% is split into the test set for cross-validation; for batch training [17], we combine

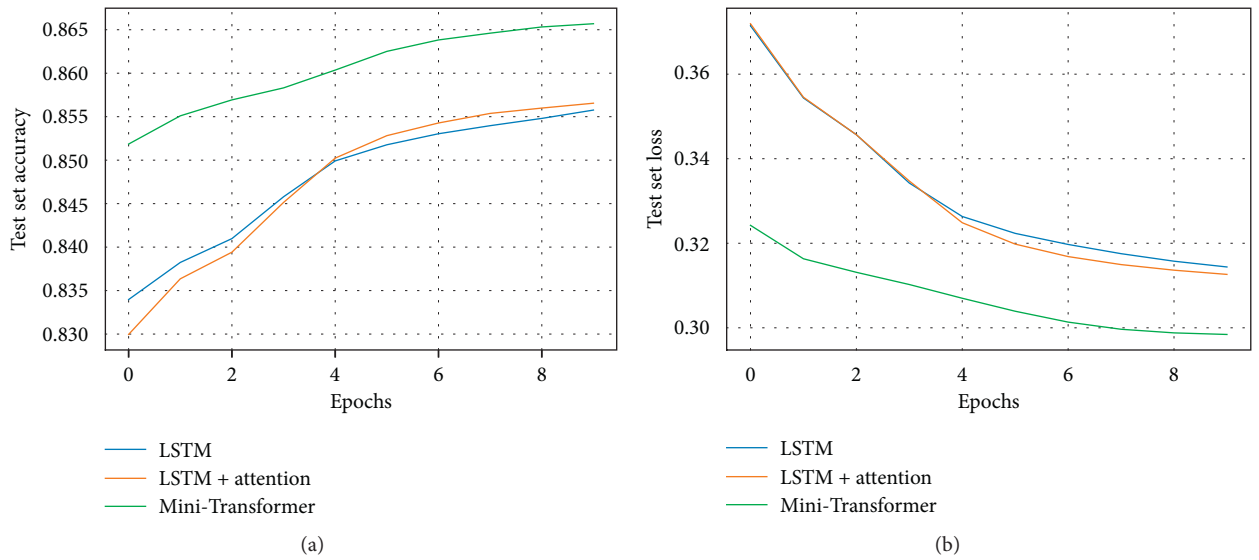

Figure 5: (a) Test set accuracy. (b) Test set loss.

Table 1: Accuracy, precision, recall, and F1 score.

Model	Accuracy (%)	Class	Precision (%)	Recall (%)	F1 score
Logistic regression	63.2167	Fraudulent	69.7941	45.7810	0.5529
		True	60.0367	80.4344	0.6875
Linear SVM	62.7882	Fraudulent	69.5541	44.6495	0.5439
		True	59.6196	80.7000	0.6858
Random forest	71.6277	Fraudulent	71.2205	71.9842	0.7160
		True	72.0385	71.2757	0.7166
LSTM	85.5761	Fraudulent	83.7513	88.0529	0.8585
		True	87.5720	83.1304	0.8529
LSTM + attention	85.6551	Fraudulent	83.8343	88.1208	0.8592
		True	87.6456	83.2202	0.8538
Mini-Transformer	86.5692	Fraudulent	84.6380	89.1490	0.8683
		True	88.6894	84.0216	0.8629

consecutive news headlines of this text dataset into batches, batch size is set to 768.

For LSTM, LSTM with attention layer and Mini-Transformer, test set accuracy and loss curves in 10 epochs are shown in Figure 5; accuracy, precision, recall, and F1 score are shown in Table 1.

For further comparison with the models from [18], we conducted several contrast experiments by employing three regular Machine Learning models: logistic regression [19], linear support vector machine (Linear SVM) [20], and random forest [21].

For logistic regression, primal formulation is implemented with liblinear solver (dual = false). For Linear SVM, the algorithm is selected to solve primal optimization problem (dual = false). For random forest, the minimum number of samples required to split an internal node is 50 (min_samples_split = 50). Other hyper-parameters are default from scikit-learn.

From Table 1, three regular Machine Learning models do not achieve good classification results, this may be because they are too simplistic to process massive news headline data.

Compared with LSTM, Mini-Transformer achieves an obvious accuracy improvement in classification performance (0.9%–1.0%). However, LSTM with attention layer achieves a slight accuracy improvement in classification performance (<0.1%); this may be because the maximum length of news headlines (len) is so short that general attention layer cannot play a sufficient role in reflecting the importance of all hidden states returned from LSTM layer.

6. Conclusion

Existing work has not focused on fraudulent news headline detection. In this paper, we have compared the classification performance of mainstream LSTM network and general attention mechanism for fraudulent news headline detection using massive news headline data, which is helpful for the research on the defense system against social engineering attacks.

In addition, according to relevant experience, we have built a more advanced Deep Learning model, Mini-

Transformer, which further improves the classification performance.

There is still room to optimize the proposed Deep Learning model. For future work, we can employ Bidirectional Encoder Representations from Transformers (BERT) as NLP pre-training method. Additionally, adversarial training and virtual adversarial training may be beneficial to improving the classification performance.

Data Availability

The data used to support the findings in this study are available from The Examiner-Spam Clickbait Catalog|Kaggle: https://www.kaggle.com/therohk/examine-the-examiner, A Million News Headlines|Kaggle: https://www.kaggle.com/therohk/million-headlines, and News Category Dataset|Kaggle: https://www.kaggle.com/rmisra/news-category-dataset.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Key R&D Program of China (2017YFB0802805 and 2017YFB0801701), the National Natural Science Foundation of China (Grant No. U1936120), and the Basic Research Program of State Grid Shanghai Municipal Electric Power Company (52094019007F).

References

[1] *Symantec Internet Security Threat Report*, Symantec Corporation, 2018, https://docs.broadcom.com/doc/istr-23-03-2018-en.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, San Diego, CA, USA, May 2015.

[4] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 5998–6008, Montréal, Canada, December 2017.

[5] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, pp. 675–684, Hyderabad, India, March 2011.

[6] J. Ma, W. Gao, P. Mitra et al., "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 3818–3824, Palo Alto, CA, USA, July 2016.

[7] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 943–951, Turin, Italy, October 2018.

[8] M. Song, H. Park, and K.-S. Shin, "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean," *Information Processing & Management*, vol. 56, no. 3, pp. 637–653, 2019.

[9] Kaggle: Your Machine Learning and Data Science Community, https://www.kaggle.com.

[10] The Examiner - Spam Clickbait Catalog | Kaggle, 6 Years of Crowd Sourced Journalism, Rohit Kulkarni, https://www.kaggle.com/therohk/examine-the-examiner.

[11] *A Million News Headlines | Kaggle*, News headlines published over a period of 18 Years, Rohit Kulkarni, https://www.kaggle.com/therohk/million-headlines.

[12] *News Category Dataset | Kaggle*, Identify the type of news based on headlines and short descriptions, Rishabh Misra, https://www.kaggle.com/rmisra/news-category-dataset.

[13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc., Sebastopol, CA, USA, 2009.

[14] *Default English stopwords list from Ranks NL Webmaster Tools*, https://www.ranks.nl/stopwords.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 3111–3119, Lake Tahoe, NV, USA, December 2013.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 315–323, Lauderdale, FL, USA, April 2011.

[17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: generalization gap and sharp minima," in *Proceedings of the 5th International Conference on Learning Representations*, Palais des Congrès Neptune, Toulon, France, April 2017.

[18] K. Rajesh, A. Kumar, and R. Kadu, "Fraudulent news detection using machine learning approaches," in *Proceedings of the 2019 Global Conference for Advancement in Technology*, Bangalore, India, October 2019.

[19] J. S. Cramer, *The Origins of Logistic Regression*Tinbergen Institute, Amsterdam, Netherlands, 2002.

[20] P. S. Bradley and O. L. Mangasarian, "Massive data discrimination via linear support vector machines," *Optimization Methods and Software*, vol. 13, no. 1, pp. 1–10, 2000.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.