Journal of KIIT. Vol. 21, No. 4, pp. 45-56, Apr. 30, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.4.45

A Pooled RNN-based Deep Learning Model based on Data Augmentation for Clickbait Detection

Jeong-Jae Kim*[1], Sang-Min Park*[2], and Byung-Won On*[3]

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by Korean Government through the Ministry of Science and ICT (MSIT) under Grant NRF-2022R1A2C1011404

Abstract

Recently, the fake news detection problem is one of the most urgent issues in the data engineering. In this paper, we propose two novel approaches to solve the clickbait detection problem, the most essential problem among various subproblems in the fake news detection. We first propose a new deep learning model that works based on Recurrent Neural Network (RNN) with multi-layers of Bi-LSTM, max-pooling layers, and fully-connected layers. Then, to improve the accuracy of deep learning models, we propose a novel method of pseudo-generating the large-scale but yet high-quality training data. We show that the pseudo-generated training data almost match those made by human evaluators. The proposed deep learning model improves the accuracy of the state-of-the-art (SOTA) methods by 36% and our new approach of automatically generating training data significantly boosts up the accuracy of the deep learning model. To the best of our knowledge, our effort is the first study of proposing a new deep learning model and pseudo-generation of training data.

요 약

최근 가짜 뉴스 탐지 문제는 데이터 공학에서 가장 시급한 문제 중 하나이다. 본 논문에서는 가짜 뉴스 탐지 문제 중 가장 본질적인 문제인 낚시성 기사 탐지 문제를 해결하기 위해 두 가지의 새로운 접근 방법을 제안한다. 먼저, RNN 기반의 Bi-LSTM 다중 계층들, max-pooling 계층들, 그리고 fully-connected 계층들로 구성된 딥러닝 모델을 제안한다. 또한, 딥러닝 모델의 정확도를 향상하기 위해 대용량, 고품질의 학습 데이터를 자동으로 생성하는 데이터 증강 알고리즘을 제안한다. 제안된 알고리즘으로 생성된 학습 데이터가 인간 평가자가 만든 데이터와 거의 일치함을 보인다. 제안된 딥러닝 모델은 기존 주요 방안에 비해 36% 정확도 향상시키며, 학습 데이터를 자동으로 생성하는 새로운 접근 방식은 딥러닝 모델의 정확도를 크게 높인다. 이러한 제안 방안은 낚시성 기사 감지를 위해 현재까지 시도되지 않은 새로운 연구이다.

Keywords
clickbait detection, deep learning, data augmentation

* School of Software, Kunsan National University
(*[3] Corresponding author)
- ORCID[1]: https://orcid.org/0000-0003-2412-2822
- ORCID[2]: https://orcid.org/0000-0003-4730-1997
- ORCID[3]: https://orcid.org/0000-0001-9881-5336

· Received: Mar. 06, 2023, Revised: Apr. 03, 2023, Accepted: Apr. 06, 2023
· Corresponding Author: Byung-Won On
School of Software, Kunsan National University, 558, Deahak-ro, Gunsan, Jeollabuk-do, Korea
Tel.: +82-63-469-8913, Email: bwon@kunsan.ac.kr

Ⅰ. Introduction

US President Donald J. Trump wrote to Tweeter on December 10, 2016, six weeks before his presidency, that reports by CNN had been ridiculous and untrue - FAKE NEWS![1] This tweet has sparked controversy over whether CNN broadcasters are producing fake news and has triggered the worldwide issue of "what is really fake news?" Through active use of social network services in recent years, fake news, a story made for the purpose of deception, is spreading rapidly, causing frequent social disruption and becoming a global problem. For instance, the total number of comments on the top-20 fake news distributed through Facebook over the three months of the US presidential election campaign was 8.7 million, while the number of comments on the top-20 major news media was 7.36 million[2]. This fact clearly shows that fake news has more influential than news articles issued by major existing news media. According to the research report of Hyundai Research Institute in 2017, the domestic economic impact of fake news was estimated to be close to 30 billion dollars per year, assuming that the proportion of fake news among all news was 1%. As such, fake news has become one of urgent issues to be addressed as a national and social issue, such as threatening democracy around the world and fostering social division.

Clickbait detection is one of the major issues in the big category of fake news detection. Clickbait takes ad revenue based on the number of clicks that lead to clicks on links that are not interesting or worthwhile for readers. This is causing social disruption due to the spread of fake news and rising as a global problem. For example, most of Google's revenue comes from Google AdSense, which monetize websites. However, there is an abuse that relies solely on the revenue of the ad to induce the reader to click on the link regardless of the value of the content or the quality of the information. If these occur frequently, users can no longer trust information on the web because many news articles may often contain exaggerated advertising, sexual, or spam content.

Technically, we view clickbait detection to a typical Big Data problem because clickbait articles essentially have four attributes of Big Data - (1) Large *volumes* of clickbait articles are generated from a variety of sources, including online sites, social media, and mobile messengers; (2) Real-time processing (*velocity*) is required to block the rapid spread of clickbait articles; (3) Clickbait articles are one of typical unstructured data (*variety*); and (4) Clickbait articles are filtered out to improve the quality of all news data (*veracity*). In fact, through these four Vs, Big Data is defined by Gartner and IBM. Obviously, it is time-consuming for human evaluators to manually identify clickbait articles among large volumes of news articles. On the other hand, deep learning models on behalf of human evaluators can detect a lot of clickbait articles automatically, whereas the accuracy of the deep learning models to classify clickbait articles is relatively low, compared to humans. In this work, we focus on improving the accuracy of the deep learning models used for detecting clickbait articles among massive news data.

In the clickbait detection problem, there are two different approaches to avoid the overfitting problem that degrades the accuracy of the deep learning models including FNN, CNN, and RNN[3]. The first approach is to reduce the complexity of the deep learning models. The other is to use large-scale training data. In the former approach, deep learning models themselves have high complexity with a number of hidden units and weight parameters. The accuracy would be decreased if we reduced the complexity of the models. On the other hand, our work focuses on the latter approach. As a good example, the accuracy of Google translation services

has recently improved significantly because a sequence-to-sequence model, one of RNN-based deep learning models, is trained with a huge number of training data including one billion pairs of Korean sentences that match English ones. However, it is non-trivial to obtain large-scale training data in which each news article has its class label indicating that it is clickbait or not. Even though we collect many news articles automatically, human evaluators must manually check each news article to determine whether it is really clickbait or not. In this way, collecting large-scale training data by human judgement is not possible.

To solve the above problem, we first propose a new deep learning model that works based on Recurrent Neural Network(RNN) with multi-layers of Bi-LSTM, max-pooling layers, and fully-connected layers. Then, to more improve the accuracy of the proposed deep learning model, we propose a novel algorithm of *pseudo*-generating the large-scale but yet high-quality training data, thereby considerably improving the accuracy of the existing deep learning models. We also show that the pseudo-generated training data almost match those made by human evaluators. To the best of our knowledge, this is the first study of proposing the deep learning model that solves the clickbait detection problem, gathering the training data automatically.

To evaluate the proposed method, we compare it to SOTA methods such as (1) SOLAT in the SWEN[4], (2) Athene(UKP Lab)[5], and (3) UCL Machine Reading[6]. Our experimental result shows that the proposed method outperforms the SOTA methods in the clickbait detection problem.

The contributions of the proposed method are as follows:

- Recently, the fake news detection problem is one of the most urgent issues in data engineering. In this article, we propose two novel approaches to solve the clickbait detection problem, the most essential problem among various subproblems in fake news detection. One is a new deep learning model that detects clickbait articles well and has the highest accuracy compared to the SOTA methods. The other is a new approach of *automatically* generating large-scale training data, given a collection of news articles. To the best of our knowledge, our solution suggests a new methodology that has not existed in the clickbait detection problem.
- Unlike the SOTA methods, the proposed deep learning model makes use of only word embedding features. Because the SOTA methods use various features in addition to word embedding features, their accuracy is not high. In general, collecting various features is a time-consuming and labor-intensive task and some features are not directly related to solving the clickbait detection problem. In our deep learning model, such hand-crafted features are not necessary.
- The proposed deep learning model improves the accuracy of the SOTA methods by 36% and our new approach of automatically generating training data significantly boosts up the accuracy of the deep learning model. Thus, if the proposed approaches can be applied to real applications, it will be useful for blocking news articles containing ad and spam.

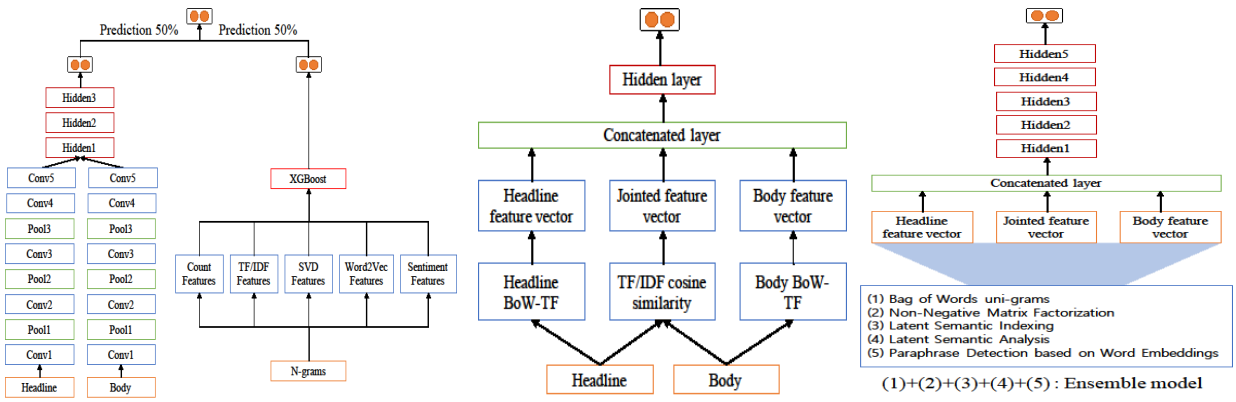
The rest of this paper consists of the followings: In Section II, we discuss related work. In Section III, we formally define the clickbait detection problem. Furthermore, we present the details of the proposed methods in Section IV. Subsequently, Section V shows our experimental set-up and results. Concluding remarks and future works follow in Section VI.

II. Related Work

For clickbait detection, Mengxi et al. integrated knowledge graph with deep learning techniques to achieve explainability[8]. Dimpas et al. presented a simple method based on only Bi-directional LSTM for determining if it is clickbait from Filipino and English

news articles[9]. Our experimental results show that our proposed model is better than Bi-directional LSTM. We will discuss the detailed results in the experimental validation section. In Wei and Wan work, an ambiguous headline is a headline whose meaning is unclear relative to that of the content of the story and a misleading headline is a headline whose meaning differs from that of the content of the story[10]. The definition of their clickbait detection problems is different from our problem that is defined as deliberately clickbait detection. Especially, the main contribution of Wei and Wan is to exploit surface-level linguistic features collected by human annotators. In general, since this task is time-consuming and labor-intensive, it is difficult to collect large-sized label dataset, where some labels are biased. Thus, this hand-crafted feature engineering limits to improving accuracy of learning algorithms. Because of these reasons, we excluded Wei and Wan's methods as baseline method. Instead, we compared the proposed method with the three clickbait methods below.

Firstly, the SOLAT in the SWEN proposed the ensemble method between a deep neural network and decision tree models. This model predicts either true or fake news, based on an weighted average between the two models. As illustrated in Figure 1 (a), two deep neural networks are made for news headlines and bodies independently. Each deep neural network is composed of three hidden layers on top of five convolutional and max-pooling layers. Each word (w) of a headline(e.g., 'local') is transformed to k-dimensional word vector using word2vec. For example, the first dimension of the word 'local' is 0.012 and the second dimension is 0.141. The k-dimensions indicate the similarity scores between w and each of k words related to w in the vocabulary of the corpus. k headline vectors are entered as the deep neural network for headlines. Through the five convolutional and max-pooling layers, the most discriminative word features are selected from k·(# of words) in the headline. This is, the objective of this process is to abstract the word features of the headline. In the same way, the deep neural network for news bodies is trained to abstract the word features of the news body. Then, the three hidden layers are used to find the optimal parameter values through back propagation. If the model predicts incorrectly, the weight values among nodes in different hidden layers are readjusted to minimize the errors between the actual values and the predicted values. In most cases, this deep learning model works well but another method called gradient-boosted decision tree method(XGBoost) is proposed to improve the accuracy of the deep learning model. XGBoost is an ensemble of weak prediction models - decision trees.



(a) SOLAT in SWEN (b) UCL machine reading (c) Athene(UKP lab)

Fig. 1. SOTA clickbait detection methods

The feature set for XGBoost is Count, TF/IDF, SVD, Word2Vec, and Sentiment. The count feature is # of uni-grams, bi-grams, and tri-grams that appear simultaneously in the news headline and body. TF/IDF is the normalized value of the count feature, quantifying the importance of the uni-grams, bi-grams, and tri-grams. Through SVD, top-k important uni-grams, bi-grams, and tri-grams are selected. The word2vec feature is the similarity scores between w (one of uni-grams, bi-grams, or tri-grams) and each of k words related to w in the vocabulary of the corpus. The sentiment feature includes the polarity values (positive, neutral, or negative) of the uni-grams, bi-grams, and tri-grams. The experimental results show that XGBoost supplements the weakness of the deep learning model. The reason is: Some words in a news article are so ambiguous that the news article is classified incorrectly. It turns out that such words are less commonly used in the corpus. The count and TF/IDF features of XGBoost can detect such words easily and prevent some sort of misclassification.

Secondly, the models in Figure 1(b) and (c) are quite simple, compared to the deep learning model presented by SOLAT in the SWEN. Athene proposed five-layers FNN model to train the headline vector, the body vector, and the joint vector. They uses embedding of the nouns and verbs of the headline and the body as word features. And they also uses features such as Non-Negative Matrix Factorization, Latent Semantic Indexing, Latent Semantic Analysis to derive latent meanings through the matrix. Each of the above five features is trained independently and composed as an ensemble model.

Thirdly, Athene's model is based on deep feed-forward neural network, while UCL's model is based on shallow feed-forward neural network. The UCL uses BoW-TF and the cosine similarity between headline and body. This model is concatenated to learn features and has a hidden layer. Unlike SOLAT's model, the two models do not use convolutional and pooling layers to abstract word features. Besides, the models are close each other in that the input of the models includes joint feature vectors.

III. Problem Definition

In this section, we formally define the clickbait detection problem which is one of the important problems in fake news detection, referring to Table 1.

Table 1. Notation terms

Term	Description
n_i	the i-th news articles in the corpus
$H(n_i)$	True headline of n_i
$H^f(n_i)$	Fake headline of n_i
$B(n_i)$	Body of n_i
w_i	The i-th word in $H(n_i)$ or $B(n_i)$
s_i	The i-th sentence of $B(n_i)$
$K(B(n_i))$	A key sentence of $B(n_i)$
$f_t(s_i)$	Weight of s_i by $\sum_i TFIDF(w_i \in s_i)$
$f_c(f_t(H(n_i)), f_t(s_i))$	Cosine sim. score b\w $f_t(H(n_i)$ and $f_t(s_i)$
$f_n(s_i)$	Random shuffled list of nouns from s_i
$f_w(s_i)$	Word embedding(fastText) of s_i

A news article n_i is composed of news headline $H(n_i)$ and news content $B(n_i)$. The ad and non-text data are removed in the news content so the output of $B(n_i)$ is the list of words in the news content. The order of the words is not also considered in the problem. In this set-up, a fake headline news H_f (n_i) is defined as $\exists n_i$ such that $H(n_i) \neq B(n_i)$. Given a news article n_i as input, the goal of AI-based models is to automatically classify n_i to either $H(n_i)$ or $H_f(n_i)$, where $H(n_i)$ and $H_f(n_i)$ are the true and fake headline news, respectively. This classification problem is also known as the stance detection problem. As already discussed in Section II, most solutions to fake news detection are focused on social media, and the features used are user profile information, context, network structure, content, and

multi-modal features. However, relatively the only content-based clickbait detection problem with news data is rarely studied until now [7]. In this work, we focus on the content-based clickbait detection problem with news data.

As described above, the goal of our research is to automatically determine if the title and content of a given news article are consistent or not. By detecting fishing titles or deliberately distorted titles, we can prevent disinformation from being transmitted due to news headlines. In our work, the fake headline news that corresponds to five levels of difficulty are classified by domain experts. Table 2 shows one of typical examples in the clickbait detection problem.

The *very easy* fake headline news is the level at which the revision of the title can be found in the article "Matching Vocabulary in a Sentence"; The *easy* fake headline news is the level at which the modification of the title can be found in the article "Changing a vocabulary within a sentence";

Table 2. An example of the clickbait detection problem

True headline
British TV channel BBC predicted victory of 'Leave' in the Brexit referendum
Body
British TV channel BBC predicted victory of 'Leave' in the Brexit referendum on June 24. The United Kingdom held a referendum on whether the UK leave the EU on June 23. It's currently ballots from 82 percent of precincts counted. 'Leave' was leading with 52 percent, while 'Remain' trailed with 48 percent.
Fake headline
Very easy: 'Leave' was leading with 65 percent in the Brexit referendum **Easy**: The EU predicted victory of 'Leave' in the Brexit referendum **Normal**: British TV channel BBC won the prediction of the Brexit referendum **Difficult**: BBC predicted that the UK will not leave the EU **Very difficult**: 'Remain' led 'Leave' by a margin of 48 - 52 percent with 82 percent of precincts reporting

The *normal* fake headline news is the level at which the correction of the title can be found in the article "Change sentence in one sentence"; The *difficult* fake headline news refers to the level of finding the revision of the title as "sentence inference in one sentence"; The *very difficult* fake headline news refers to the level at which the revision of the title can be found in the content of the article as "sentence reasoning more than two sentences."

IV. Proposed Model

In the previous section, we explained the main SOTA models, all of which are based on deep learning models like FNN and CNN. However, we claim that such methods do not work well because they do not consider the sequence of words per sentence. In the language model, the following word are generated under the influence of the previous word[11]. For example, in a sentence "I eat XX", the probability of generating the word XX is directly affected by the previous word 'eat'. The probability of generating 'steel' as XX is very low, while that of generating 'apple' is very high. Since news headlines and bodies consist of sentences, the order of words per sentence should be considered in deep learning models. To handle this point, we propose a new deep learning model that is the hybrid of news headline and body models. As shown in Fig. 2, each model is composed of four hidden layers, where the first and third layers are Bi-LSTM (Bidirectional Long Short-Term Memory), and the second and the fourth layers are max-pooling layers. By Bi-LSTM, the generation of the current word is influenced by both previous and next words. Through max-pooling layers, word features are abstracted to a few features as the most discriminative features. The fully-connected layers are used to concatenate the headline and body models. The input of the proposed model is word2vec of words and the output of the model is the probability values of true and fake headline to a news article.

As we already discussed in the introduction section,

we also propose a novel method that improves the accuracy of existing deep learning models by **pseudo-generating** *large-scale* but yet *high-quality* training data. It is well known that all deep learning models work well if the size of the training data is increasing. In the clickbait detection problem, to pseudo-generate training data, we propose Algorithm 1.

Given a news article n_i, Algorithm 1 generates a fake headline. First, compute the similarity score between $H(n_i)$ and each of sentences in the news body and then mark the sentence with the highest score as the key sentence $K(B(n_i))$. Next, extract nouns from $H(n_i)$ and $K(B(n_i))$. Finally, generate a fake news headline to exchange the noun of $H(n_i)$ with the noun of $K(B(n_i))$, where the nouns are chosen at random. In this way, Algorithm 1 can automatically generate a large amount of training data for deep learning models.

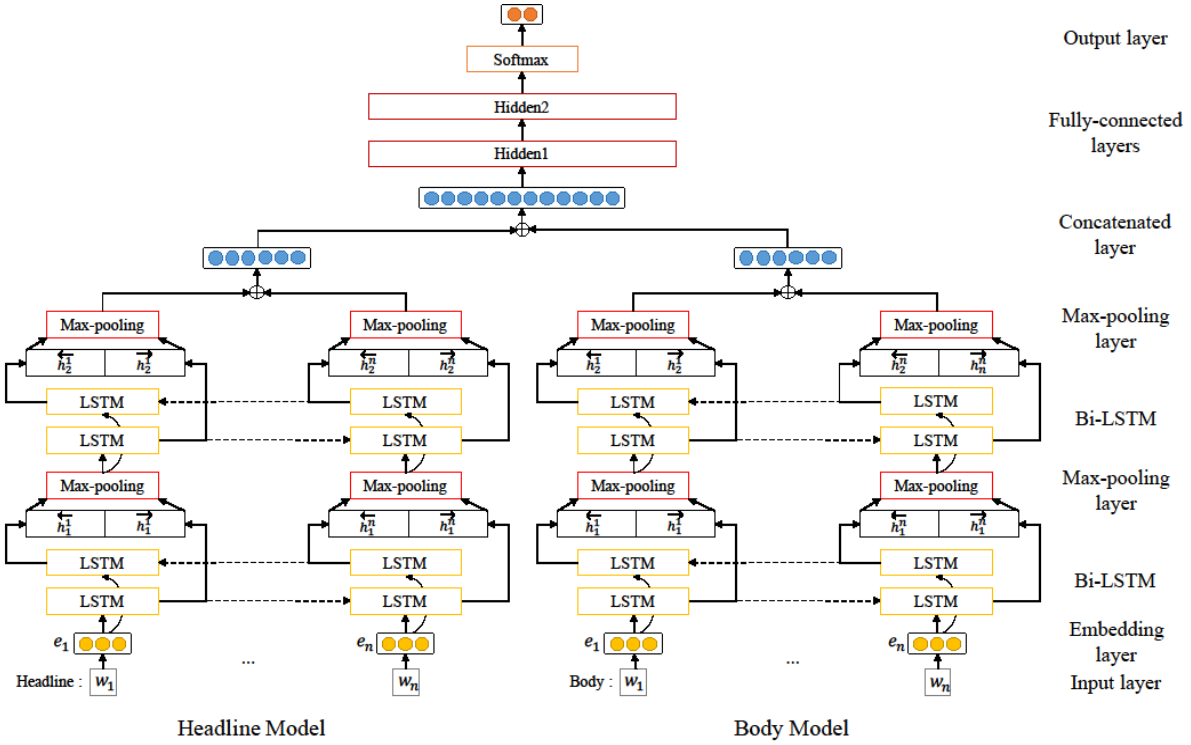


Fig. 2. Proposed model

Table 3. Algorithm 1: Pseudo-generating large-scale but yet high-quality training data for deep learning models

$n_i = \{H(n_i), B(n_i)\}$;
$H(n_i) = \{w_1,...,w_h\}$;
$B(n_i) = \{s_1,...,s_b\}$;
for $s_i \in B(n_i)$ do
  $f_c(f_t(H(n_i))), f_t(s_i))$;
$K(B(n_i))$ =The sentence with the largest cosine sim. score by f_c is a key sentence== $\{w_1,...,w_k\}$;
for $h_n \in f_n(H(n_i))$ do
  for $k_n \in f_n(K(B(n_i))$ do
    if $h_n \neq k_n$ then
      $H^f(n_i) = Replace\ h_n$ in $H(n_i)$ $with\ k_n$;
      $v \rightarrow$ vector$< real, f_w(H(n_i)), f_w(B(n_i)) >$ in the training set;
      $v \rightarrow$ vector$< fake, f_w(H^f(n_i)), f_w(B(n_i)) >$ in the training set;

V. Experimental Validation

5.1 Experimental set-up

In the previous section, we described the proposed pseudo-generating algorithm for collecting large-scale but yet high-quality training data for deep learning models. Now we introduce the process of evaluating main deep learning models based on the proposed method, comparing to the existing SOTA models. In the experiment, we used 49,997 news articles as training data and 350 news articles as test set. Table 4 shows the brief characteristics of the data sets.

Table 4. Data characteristics

Data set	True	False
Training set	36,545	13,427
Validation set	1,827	671
Test set	175	175

The class(true or fake) of all news articles in both training and test sets were manually labeled by domain experts. To generate fake news, the news article changed the headline to fake by human evaluators. The generated test set consists of easy problems such as vocabulary matching and change problems, and difficult problems requiring human-level interpretation and reasoning. The validation data set was evaluated at each training step of the model by applying 5% of the training data set. In the preprocessing step, we normalized the text to lower case after removing advertising texts and images. Then we removed the stop words[12] from all news articles and converted words into root forms through stemming software[13].

We implemented the proposed method in addition to Bi-LSTM, max-pooling layer, and fully-connected layer in Python and Tensorflow[14].

As an activation function, ReLU was used for all layers except the output layer to which softmax function was applied. We used cross entropy as loss function in addition to Adam optimizer for carrying out backward propagation of errors. Our model showed the best performance when learning rate is 0.00001 and batch size is 35. The length of the embedding layer is fixed at 300, padding for small, and 300 words for long. Each Bi-LSTM has 256 units; the layer size is 300; and the same as the embedding length. Each max-pooling layer has a kernel size of 8, a stride of 4, and padding and the length of each fully-connected layer is 1,024. For weight initialization of the model, we used the truncated normal method[15].

Our proposed model was trained for pseudo-generating large-scale but yet high-quality training data by applying Algorithm 1 to the collected news articles. In the training process, it is selected randomly as mini batch in training data set and the model evaluated the validation data set every 10 mini batches, and stored it as the final model when it has the highest accuracy. We generated a neural word embedding model by training the collected news articles in the skip-gram model of fastText[16]. The trained neural word embedding model expresses each word in 300 dimensions and size of the vocabulary is 727,503. Each model was in standalone executed in a high-performance workstation server with Intel Xeon 3.6GHz CPU with eight cores, 24GB RAM, 2TB HDD, and TITAN-X GPU with 3,072 CUDA cores, 12GB RAM, and 7Gbps memory clock.

5.2 Experimental result

5.2.1 Results of various experiments in clickbait detection

In the introduction section, we viewed clickbait detection to a typical Big Data problem. The deep learning models are better than the traditional classification models such as SVM and Random Forest in the Big Data problem. Thus, we propose a deep learning model based on RNN with multilayers of Bi-LSTM, max-pooling layers and fully-connected layers considering the sequence of words per sentence.

In the first experiment, we compared the main SOTA models to assess the excellence of the proposed method in clickbait detection. We evaluated 49,972 training sets and 350 test sets for evaluation of various models. The clickbait detection is a fifty-fifty chance, because it detects true or fake headlines. However, the test set consists of the difficulty levels as shown in Table 2, which can not be solved by simple word search, rule base, and morphological analysis. We note that the test set is very difficult data because it requires human interpreting in difficulty and above.

Table 5 summarizes the accuracy of different types of methods and Table 6 summarizes accuracies of fake news headlines in different levels.

Table 5. Accuracy of various deep learning models

Methods	Accuracy
Athene	0.4886
SOLAT	0.5514
UCL	0.5171
Bi-LSTM	0.5057
Bi-LSTM+POOL	0.5486
Bi-LSTM+FNN	0.7257
Bi-LSTM+POOL+FNN	0.7543
Bi-LSTM+Algo. 1	0.9743
Bi-LSTM+FNN+Algo. 1	0.9886
Bi-LSTM+POOL+Algo. 1	0.9942
Bi-LSTM+POOL+FNN+Algo. 1	1.0

Athene's model is trained in deep feed-forward neural networks with the title and body features such as word embedding, latent semantic analysis features and joint features. The model has the lowest accuracy of 0.4846. In terms of difficulty levels, easy ("changing a vocabulary within a sentence") and normal("changing sentence in one sentence") are more accurate. It is observed that uni-gram and matrix operation due to the change of position of words or sentences in latent semantic analysis affected easy and normal clickbait detection. Because Athene learns about word-based latent semantic analysis features and embedding in FNN models, it does a good job of word-based analysis, but performance of context understanding is very low.

UCL is trained with TF/IDF features which indicates how important a word is in a particular set of documents when there are various document groups on shallow feed-forward neural networks. The accuracy of the model is 0.5171, and the difficulty levels show the highest performance with difficulty ("sentence inference in one sentence"). Fake news headlines in difficulty have completely different sentence structures than true news headlines, so it is observed that the TF/IDF attributes influence the search for non-words in the article. UCL is difficult to solve clickbait detection because the model is sallow and uses word-based features similar to Athene.

Table 6. Accuracy of fake news detection in different levels(%)

Methods	Very easy	Easy	Normal	Hard	Very hard
Athene	57.5	65	67.3	56	40
SOLAT	70	77.5	70	60	55
UCL	65	67.5	63.3	68	45
Bi-LSTM	76.7	65	56.7	84	55
Bi-LSTM+POOL	71.7	57.5	56.7	84	55
Bi-LSTM+FNN	66.7	82.5	63.3	88	65
Bi-LSTM+POOL+FNN	68.3	82.5	73.3	92	70
Bi-LSTM+Algo. 1	89.3	97.5	100	92	85
Bi-LSTM+FNN+Algo. 1	100	100	100	100	100
Bi-LSTM+POOL+Algo. 1	100	97.5	100	100	100
Bi-LSTM+POOL+FNN+Algo. 1	100	100	100	100	100

SOLAT's deep learning model is trained with the abstract word features of the headline and SOLAT's XGBoost is trained with count, TF/IDF, SVD, Word2Vec, and sentiment features. The predictions of these two models were reflected in half. The accuracy of SOLAT is 0.5514, which is higher than both Athene and UCL. In SOLAT, Easy has the highest accuracy. It is observed that XGBoost's n-gram features, the feature extraction, information abstraction through the convolution layers, and the pooling layers affect the word change detection. Similar to the proposed model, SOLAT's deep-running model learns feature representation with the Convolutional Neural Network and reduces the loss of classification through a fully-connected layer, but does not take into account the word sequence. The results of the SOTA models in the test set are the same as accidental results and have only word-based analysis. In Table 6, the SOTA models show that the detection ability decreases as the difficulty level increases. Also, Very Difficulty is the lowest accuracy of all three models, and Athene and UCL are less than half. The higher the degree of difficulty, the more contextual interpretation is required, which is difficult to detect with word feature analysis or latent semantic analysis.

In order to demonstrate the analysis and superiority of the proposed model, we conducted the experiments in Table 6, lines 4 to 7. We use RNN-based Bi-LSTM to train the sequence of words per sentence. Since LSTM learns about storing, maintaining, ejecting, and deleting information, it solves vanishing gradient problem of RNN. In addition, Bi-LSTM introduces forward and backward concepts to compensate for loss of information according to the direction of LSTM. The Bi-LSTM+POOL model layered the Bi-LSTM layer and the Max-pooling layer as shown in Figure 2.

The Max-pooling layer is to reduce the spatial size of the output of Bi-LSTM and is to abstract the most distinctive features of the word feature. As a result, Bi-LSTM+POOL is improved by 8.4% compared to Bi-LSTM. The Bi-LSTM+FNN model is a specialized model for information learning in the merging process of the Headline model and the Body model. Rather than detecting true and fake with a simple linear technique method, the model solves complex nonlinear detection problems by stacking two fully-connected layers. The accuracy of Bi-LSTM+FNN is significantly improved to 0.7257. The proposed model is Bi-LSTM+POOL+FNN model. It is repeated twice as the sequence learning of Bi-LSTM and the abstraction of Max-pooling layers for the headline and the body respectively. And the model solved nonlinear detection problem with two fully-connected layers for information learning in merging process of two models. Except for the Bi-LSTM model, the performance of fake headline detection of all models is high. As a result, our model performs well in feature representation with Bi-LSTM and max-pooling layer to consider the sequence of words per sentence and reduces the loss of clickbait detection through the fully-connected layer.

In the next section, we compared the main existing models to the proposed model according to various sizes of learning data for the performance evaluation of Algorithm 1.

5.2.2 Validation of effectiveness of algorithm 1

As discussed in the introduction section, we focus on how to obtain large amounts of data to avoid the overfitting problem in the deep learning model. The cost and time of crawling and preprocessing are expensive to obtain high quality large data. In this article, we propose an algorithm to generate large-scale but yet high-quality training data, and the approach of generating fake news in train set rather than analyzing fake news to solve fake news problems. It is possible to generate a large number of fake news headlines by attempting to create a text and

headline composed of real news through Algorithm 1. In Table 6, the accuracies of the deep learning models with Algorithm 1 are over 0.97, and the proposed model correctly detected all test sets. The proposed model, Bi-LSTM+POOL+FNN, correctly detected all test sets. Fake headline detection of all models except Bi-LSTM model shows high performance. All models except the Bi-LSTM model have an accuracy of over 0.97 at all difficulty levels. In general, collecting various features is a time-consuming and labor-intensive task and some features of the existing methods are not directly related to solving the clickbait detection problem. In our deep learning model, such hand-crafted features are not necessary. SOLAT, UCL, and the proposed model show that the accuracy of the test set increases as the size of the training data increases. It is observed that the accuracy of the proposed model increases exponentially with the larger data size. Through the above experiments, we can obtain high quality learning data through Algorithm 1, which proves that the accuracy increases in proportion to the size of the data. As a result, the proposed deep learning model improves the accuracy of the main existing deep learning models by 36% and Algorithm 1 significantly boosts up the accuracy of the deep learning model.

VI. Concluding Remark and Future Work

Recently, the fake news detection problem is one of the most urgent issues in data engineering. In this article, we propose two novel approaches to solve the clickbait detection problem, the most essential problem among various subproblems in fake news detection. Specifically, we propose a novel deep learning model that abstracts the word sequence in news title and body for detecting clickbait articles well. Furthermore, we propose a brand-new algorithm for automatically pseudo-generating large-scale but high-quality training data, given a collection of news articles. Our experimental results show that the proposed deep learning model improves the accuracy of the existing deep learning models by 36% and our new approach of automatically generating training data significantly boosts up the accuracy of the deep learning model.

Our future research direction is to expand the proposed models to another subproblems in the fake news detection problem. One of the typical problems is to detect news content that is not related to the context of news. In other words, detecting mixed advertising or fishing content in news texts prevents exposure to unintended information. In detail, we will attempt to three cases. In the first case, a news article includes some paragraphs of other field in the article's overall content. In the second case, it includes some sentences of the other field in the overall article. In the last case, it includes some paragraphs of the same field in the entire article.

References

[1] Donald J. trump on twitter, https://twitter.com/realdonaldtrump/status/807588632877998081?lang=en [Accessed: Dec. 01, 2016]

[2] This analysis shows how viral fake election news stories outperformed real news on facebook, https://www.buzzfeed.com/craigsilverman/viralfake-election-news-outperformed-real-news-on-facebook?utm_term=.sgr3XNKWw.rl4xGwy6L. [Accessed: Nov. 01, 2016]

[3] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning(adaptive computation and machine learning series)", MIT Press, 2016.

[4] Fake news challenge, http://www.fakenewschallenge.org. [Accessed: Jun. 01, 2017]

[5] M. Balmas, "When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism", Communications research, Vol. 41, No. 3. pp. 430-454, Jul. 2012. https://doi.org/10.1177/0093650212453600.

[6] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news", ASIST, pp. 1-4, Nov. 2015. https://doi.org/10.1002/pra2.2015.145052010082.

[7] Computational Solutions against Fake News: AI vs. DB Approaches, https://john.cs.olemiss.edu/nhassan/file/tutorial-aaai18-ai.pdf. [Accessed: Aug. 29, 2018]

[8] M. Zhou, W. Xu, W. Z. Zhang, and Q. Jiang, "Leverage Knowledge Graph and GCN for fine-grained-level Clickbait Detection", World Wide Web, Mar 2022. https://doi.org/10.1007/s11280-022-01032-3.

[9] P. Dimpas, R. Po, and M. Sabellano, "Filipino and English Clickbait Detection using a Long Short Term Memory Recurrent Neural Network", 2017 International Conference on Asian Language Processing(IALP), Singapore, Dec. 2017. https://doi.org/10.1109/IALP.2017.8300597.

[10] W. Wei and X. Wan, "Learning to identify ambiguous and misleading news headlines", IJCAI, May 2017. https://doi.org/10.24963/ijcai.2017%2F583.

[11] Y. Kim and J. Gim, "A Study on Knowledge Embedding Method for Extending Contextual Information of Words", The Journal of Korean Institute of Information Technology, Vol. 20, No. 11. pp. 29-38. Oct. 2022. https://doi.org/10.14801/jkiit.2022.20.11.29.

[12] Stop word, https://en.wikipedia.org/wiki/Stop_word [accessed: Mar. 24, 2023]

[13] The Porter Stemming Algorithm, https://tartarus.org/martin/PorterStemmer/index.html [accessed: Mar. 24, 2023]

[14] Tensortflow, https://www.tensorflow.org [accessed: Mar. 24, 2023]

[15] Y. Lecun, L. Bottou, G. B. Orr, and K. R. Muller, "Efficient backprop", In Neural Networks: Tricks of the trade, 2002.

[16] fastText, https://fasttext.cc [accessed: Mar. 24, 2023]

저자소개

Jeong-Jae Kim

2020 ~ present : Graduate School in Cognitive Science, Yonsei University

Research Interests : Data Mining, Natural Language Processing, Artificial Intelligence

Sang-Min Park

2020 : Master of Software Convergence Engineering, Kunsan National University

2021 ~ present : Senior Researcher, Saltlux AI Labs, Inc.

Research Interests : Natural Language Processing, Large Language Model, Generative Model

Byung-Won On

2007 : PhD, Pennsylvania State University

2008 ~ 2009 : Post-doctoral Researcher, University of British Columbia

2010 : Senior Research Engineer, Advanced Digital Sciences Center

2011 ~ 2014 : Senior Researcher, Advanced Institute of Convergence Technology

2014 ~ present : Professor, School of Software, Kunsan National University

Research Interests : Data Mining, Natural Language Processing, Artificial Intelligence, Reinforcement Learning