IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2021, accepted January 19, 2021, date of publication February 24, 2021, date of current version March 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062029

Learning to Detect Incongruence in News Headline and Body Text via a Graph Neural Network

SEUNGHYUN YOON[1], (Member, IEEE), KUNWOO PARK[2], MINWOO LEE[3], TAEGYUN KIM[4,5], MEEYOUNG CHA[4,5], (Member, IEEE), AND KYOMIN JUNG[3], (Member, IEEE)

[1]Adobe Research, San Jose, CA 95110, USA
[2]School of AI Convergence, Soongsil University, Seoul 06978, South Korea
[3]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[4]School of Computing, Korea Advanced Institute of Science and Technology (KAIST) , Daejeon 06978, South Korea
[5]Data Science Group, Institute for Basic Science, Daejeon 34126, South Korea

Corresponding author: Kyomin Jung (kjung@snu.ac.kr)

The work of Taegyun Kim and Meeyoung Cha was supported in part by the Institute for Basic Science in South Korea under Grant IBS-R029-C2, and in part by the Basic Science Research Program through the National Research Foundation of Korea under Grant NRF-2017R1E1A1A01076400. The work of Minwoo Lee and Kyomin Jung was supported by Samsung Electronics Company Ltd. under Grant IO201208-07852-01.

ABSTRACT This paper tackles the problem of detecting incongruities between headlines and body text, where a news headline is irrelevant or even in opposition to the information in its body. Our model, called the graph-based hierarchical dual encoder (GHDE), utilizes a graph neural network to efficiently learn the content similarity between news headlines and long body paragraphs. This paper also releases a million-item-scale dataset of incongruity labels that can be used for training. The experimental results show that the proposed graph-based neural network model outperforms previous state-of-the-art models by a substantial margin (5.3%) on the area under the receiver operating characteristic (AUROC) curve. Real-world experiments on recent news articles confirm that the trained model successfully detects headline incongruities. We discuss the implications of these findings for combating infodemics and news fatigue.

INDEX TERMS Graph neural network, headline incongruity, online misinformation.

I. INTRODUCTION

The volume of news content generated every day is surging [1]. In contrast to newspapers, which publish limited content each day, publishing articles online incurs little cost. Furthermore, some of these news articles (e.g., weather and financial reports) are written by automated algorithms [2], which further reduce the cost of news generation. To draw traffic to news articles among the plethora of competitors, some news media attempt to capture readers' attention by using news headlines unrelated to the main content. Such mismatches can be extremely harmful in an online environment, where readers usually skim headlines without consuming the content of the news articles [3]. Thus, misleading headlines potentially contribute to incorrect perceptions of events and inhibit their dissemination [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Moayad Aloqaily.

This study aims to tackle **the headline incongruity problem** [5], which involves determining whether the news headlines are unrelated to or distinct from the main parts of the full body text. Figure 1 illustrates an example in which based solely on the headline, a reader might expect to learn specific information related to the novel coronavirus; however, the body text contains an advertisement for a dietary supplement. The challenge is that many individual users will not notice the incongruity by simply reading the news headlines because the body text is revealed only after a click. Content incongruity is a growing problem that negatively impacts the news reading experience.

Researchers have proposed several practical approaches using deep learning to address the detection problem as a binary classification (i.e., incongruent or not) by determining the ground truth based on manual annotation. A recent method learns the characteristics of news headlines and body text jointly via a neural network [6]. However, there are two

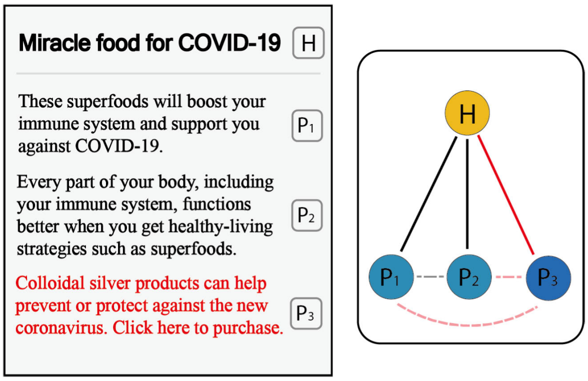


FIGURE 1. An example of an incongruent headline problem and its graph representation between the headline and body paragraphs. The red edge in the graph describes the incongruence between paragraph 3 and other texts.

critical challenges in these approaches. First, the existing models focus on learning the relationship between a short headline and lengthy body text that can reach thousands of words, posing challenges to efficient neural network-based learning due to the excessive lengths of news articles. Second, the lack of a large-scale dataset makes it difficult to train deep learning models, which have numerous parameters, to detect headline incongruities.

This paper presents a new method to tackle the headline incongruity problem: a graph-based hierarchical dual encoder (GHDE) that captures the textual relationship between a news headline and its body text of arbitrary size. It leverages the hierarchical nature of news articles by embedding the text content of the headline and body paragraphs as *nodes*. This approach is used to construct a graph in which headline nodes lie on one side and body paragraph nodes lie on the other. Then, we connect undirected edges between these nodes. The GHDE learns to compute edge weights between the headline and paragraph nodes and assigns a higher edge weight to the more relevant edges. Then, GHDE updates each node representation by aggregating information from its neighboring nodes. The iterative update process propagates the relevant information in paragraph nodes to the headline node, which is essential in determining content incongruities.

This work also presents a dataset generation method and makes a million-item-scale dataset available for future research. This dataset is currently the largest English dataset compiled for the headline incongruity problem. From the corpus of 7,127,692 English news articles published by 57 media outlets, our method iteratively matches two new stories to a similar topic and then combines their body paragraphs to create a synthetic news article with varying levels of incongruity.

The extensive experiments show that GHDE outperforms existing incongruity-detection models by a substantial margin (5.3%) on the AUROC metric (an improvement from 0.879 to 0.926). A study on real-world articles was conducted where crowdsourced workers were asked to annotate incongruous labels from recent news posts; then, GHDE was used to evaluate the incongruity. The results of this experiment demonstrate that the proposed method can be applied to incongruity detection in news articles in the wild. In fact, GHDE can successfully detect incongruence between headlines and body text even for unseen topics, such as health supplements for COVID-19, as demonstrated in Figure 1.

The remainder of this paper is organized as follows. Section II provides a brief review of the literature on headline incongruity detection and using graph neural networks with text. We propose an efficient automatic data generation method in Section III and introduce our newly created million-item-scale dataset for research in this field. In Section IV, we start by describing the baseline models considered in this paper, including the previous state-of-the-art neural network-based model and a recent BERT-based model. Next, we introduce the proposed model in detail. The experimental setup for model evaluation, a discussion of the result achieved by the various approaches, and empirical studies in the wild are presented in Section V. We conclude by discussing the implications of this study in the context of fighting against infodemics and news fatigue online. Finally, Section VI concludes the paper through a discussion on the limitations of this study and possible directions for future research on the news incongruence detection problem. The code and the data are available at https://github.com/minwhoo/detecting-incongruity-gnn.

II. RELATED WORKS

A. MACHINE LEARNING FOR HEADLINE INCONGRUITY

Incongruity between news heading and body content is a common type of misinformation on the Internet [7]. In digital environments, people are less likely to read full news stories; they tend to only peruse the news headlines. Such news reading habits aggravate the harm caused by misleading (incongruent) headlines [4]. Several machine learning techniques have been proposed to tackle this challenge. The main challenge is that this field of study still lacks large-scale realistic datasets; consequently, many of the existing studies relied on relatively small datasets of manually annotated data. In terms of data complexity, the best-known model utilizes an attention-based hierarchical dual encoder to process the long body paragraphs common in news articlesefficiently [8].

Headline incongruity detection is also related to the *stance detection problem*, which aims at identifying the stance of specified claims against a reference text. The similarity of stance detection to our task is that both require a model to investigate relationships between a short claim and a long article. The Fake News Challenge 2019 was held to promote the development of methods for stance detection, and many of the teams utilized deep learning models (e.g., [9]). The winning model was an XGBoost [10] model based on hand-designed features. An unsupervised learning technique was introduced to detect the stance of users in social media [11]. Most recently, a study proposed a method that detects headline incongruity via a semantic matching framework between the original and synthetically generated headlines [12].

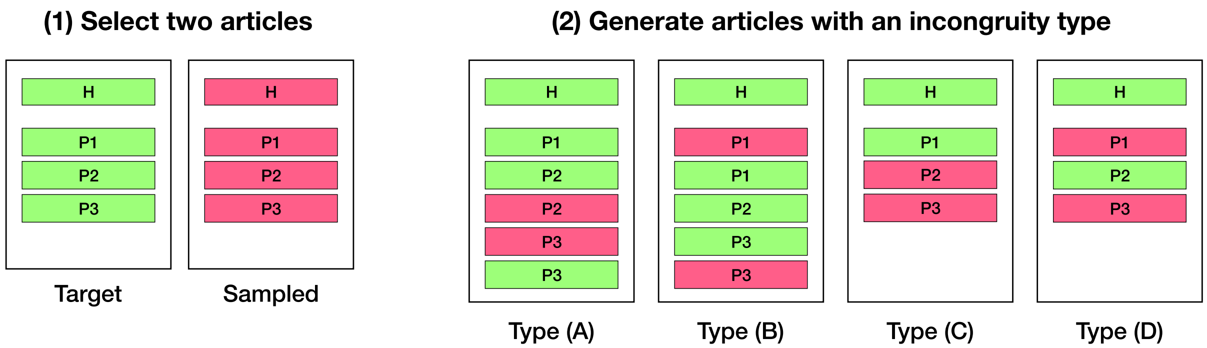


FIGURE 2. An illustration of the generation process of news articles with incongruent headlines (H: Headline, P: Paragraph).

B. GRAPH NEURAL NETWORKS FOR TEXT

A graph neural network (GNN) utilizes graph-like structural information to either explicitly or implicitly represent data [13]. Several methods can embed network information. For example, Veličković *et al.* employed an attention mechanism to aggregate node vectors by learning the importance of each edge [14]. Palm *et al.* adopted a recurrent node updating approach to capture changes in information across time [15].

A GNN embeds relational information in textual data. Thanks to its unique architecture, such information is propagated into neighboring nodes during the training process. Hence, GNN models can perform reasoning regarding their nodes and edges, which is challenging for more standard architectures such as recurrent neural network (RNN) and convolutional neural network (CNN) models. For example, GNNs have excelled at question-answering [16]–[18], relation extraction [19], and knowledge base completion [20] tasks.

C. FAKE NEWS DETECTION AND GNNs

The problem of fake news detection has been actively studied over the past several years. To estimate the truthfulness of claims and further enable the detection of fake news, researchers have relied on resources available on fact-checking websites such as politifact.com [21]. The Liar dataset is one representative example of such a resource; it comprises 12.8 K political statements with veracity labels on a 6-point scale [22]. That study also showed that a CNN can achieve a reasonable performance using only text. A recent study suggested an approach that learns and constructs discourse-level structures from articles to detect false claims [23]. The active prevention of fake news was addressed by detecting early-stage diffusion [24], a blockchain proof-of-authority protocol [25], and by developing misinformation reputing measures [26].

Most recently, a handful of studies have proposed using GNNs to detect fake news. The researchers implemented a model named Fake Detector that learns the representations of news articles, creators and subjects simultaneously through a gated graph neural network [27]. A follow-up study proposed a hierarchical attention mechanism that learns the importance of each node and a schema for fake news detection [28]. Yet another study proposed aggregating information by considering content characteristics, sharing behaviors, and social connections through a graph neural network [29].

Based on recent GNN developments, this study presents a GNN-based model to address the headline incongruity problem. We will introduce how to define the nodes and how to learn edge weights for this task.

III. DATA GENERATION

There are two main challenges in detecting incongruities between headlines and body text: (*i*) the lack of large training datasets and (*ii*) the length of news stories. This section focuses on the first challenge by presenting a rule-based approach to generate news articles with incongruity. We will address the second challenge in the coming section as well.

While previous studies manually annotated the ground truth [6], [30], it is almost impossible to apply a manual method to datasets consisting of millions of news articles. Therefore, we propose an alternative approach that instead generates news articles with incongruous headlines automatically. This process starts with an extensive collection of real news stories. For each news article in a randomly chosen set of "target" news stories that we wish to manipulate, we replace the body text of each target article with paragraphs from a different news article, again chosen from the remaining news corpus (which we call a "sampled" article). Here, the assumption is that the seed target article's headline and body text are consistent with regard to the news content.

The seed corpus of real news stories comes from Real-News [31], which consists of 32,797,763 English news articles published over multiple years. Following the guidelines from the Media Bias/Fact Check [32], we consider only 7,127,692 of these news articles written by listed trustworthy media outlets as the seed corpus, because untrustworthy news sources may already share incongruent headings. Based on 1,000 of the news articles sampled from the corpus, we manually confirmed that trustworthy media are unlikely to publish incongruent headlines.

Figure 2 illustrates the process through which the dataset of incongruent labels (i.e., "positive" labels) is built. The figure shows one selected ‘target’ and one ‘sampled’ news story. We generate two types of datasets: one where the ‘sample’ news stories are randomly chosen (i.e., Random dataset) and one in which ‘sample’ news stories are chosen to contain the similar news stories (i.e., Similar dataset). Then, paragraphs from the ‘sample’ news stories are mixed into the ‘target’ news story. The number of swapped paragraphs is randomly determined and ranges from 1 to the number of sampled paragraphs, which causes the incongruity difficulty to vary.

After this process is completed, an equal number of articles are sampled from the remaining news pool to include congruent data (i.e., those with a “negative” label). The final dataset consists of 1,366,025 news articles with a balanced distribution between incongruity labels, mixing types (i.e., Types in the figure) and the number of swapped paragraphs.

Headline similarity is measured by the Euclidean distance of the fastText embeddings pre-trained on the WikiNews corpus [33]. To avoid selecting sample news stories that are not incongruent with the target article (e.g., stories reporting the same event), we filter out news stories published in a similar period. We apply a maximum threshold for the similarity measure to control the incongruity difficulty of the generated dataset. We use an efficient implementation of the similarity search [34], which consumes approximately 3 hours on a server equipped with a 32-core Intel Xeon CPU to find similar articles for more than 2 million target articles.

The data generation methods from the existing work insert sampled paragraphs into a target article [8], leading to longer news stories, depicted as Type A and Type B in Figure 2. However, such a change to the article length can be mistakenly learned by the detector as a trivial feature for the detection task. Therefore, it is crucial to maintain a length distribution of news stories similar to that of the original distribution. The existing generation methods also do not consider textual similarities between the target and sample articles, resulting in trivial topic differences. Because ordinary news articles cover a single topic, this inconsistency could induce the machine learning models to focus on body text patterns rather than on understanding the relationship between headline and body text. Compared to the previous approaches (Types A and B), our approach (Types C and D) can generate news articles with headline incongruities while preventing the detection models from learning the artifacts produced by data generation.

The dataset includes labels specifying whether each paragraph originates from the sampled article; we exploit thee paragraph labels to dynamically represent news articles as a graph structure. More details of the final dataset are described in Table 1. The dataset constructed using random sampling exhibits a similar data distribution in terms of word counts, as shown in Table 2. When splitting the dataset into training, development, and test sets, we ensured that they do not have an overlapping period with one another to prevent the models from unintentionally focusing on topical patterns.

TABLE 1. Data characteristics. H and B indicate headline and body text, respectively.

	Train	Dev	Test
Number of data	1,347,097	9,493	9,435
Avg. word counts (H)	11.73	12.83	12.98
Min. word counts (H)	3	3	3
Max. word counts (H)	56	35	35
Avg. word counts (B)	765.25	793.79	715.52
Min. word counts (B)	20	63	29
Max. word counts (B)	27,597	7,173	11,597

TABLE 2. Data characteristics of the random dataset. H and B indicate headline and body text, respectively.

	Train	Dev	Test
Number of data	1,360,095	9,478	9,395
Avg. word counts (H)	11.04	12.59	12.70
Min. word counts (H)	3	3	3
Max. word counts (H)	57	35	35
Avg. word counts (B)	760.61	794.92	709.94
Min. word counts (B)	20	49	22
Max. word counts (B)	27,362	5,918	11,964

IV. METHODS

Our objective is to detect whether a news headline is incongruent to any subset of body text. Formally, we consider the detection task as one in which each news article is provided as a tuple (H, P), where H is the headline, and P is a set of paragraphs comprising the body text. Each paragraph $p_i \in P$ is a sequence of words that may consist of one or more sentences. Our goal is to determine a binary incongruity label y. Paragraph-level incongruity labels $Y_P = \{y_1, \ldots, y_{|P|}\}$ are available as additional supervision during training.

We first review the learning approaches that have been proposed to detect headline incongruity. We then present a new graph-based neural network model that embeds the relationship information between a headline and its corresponding body text.

A. BASELINE APPROACHES

We discuss four prominent baseline approaches.

1) XGBoost

XGBoost, which implements gradient boosted decision trees, is a well-recognized and fast algorithm for classification tasks [10]. We adopted XGBoost as a representative baseline because it was used in the winning model for the stance detection challenge in news headlines [35]. Here, given a news headline and the text body content, the task was to assign the news headline’s stance label to one of the following: agree, disagree, discuss, or unrelated. Using an incongruity label instead, we implemented the winning model from this challenge by extracting a feature set consisting of TF-IDF vectors based on word occurrences. Singular values decomposed from these vectors indicate word-vector similarities between a headline and its corresponding body text. We call this model **XGB**.

2) ATTENTIVE HIERARCHICAL DUAL ENCODER

Among the available approaches for the headline incongruity problem is the attentive hierarchical dual encoder (**AHDE**), which has a two-level hierarchy of recurrent neural networks [8]. This model utilizes paragraph structure to address the arbitrarily long sizes of news article. It returns $\mathbf{u}_H$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_{|P|}\}$, which correspond to a headline and the paragraphs of the associated body text. An attention mechanism is applied to the headline's hidden states and the paragraphs to learn the importance of each paragraph and detect incongruity in its relationship with the headline. This model is the current state-of-the-art. The vector $\mathbf{h}_B$, which is the context vector for the entire body text, is calculated as follows:

$$\begin{aligned} s_i &= \mathbf{v}^\intercal \tanh(\mathbf{W}_u^B \mathbf{u}_i + \mathbf{W}_u^H \mathbf{u}_H), \\ \mathbf{a}_i &= \exp(\mathbf{s}_i)/\textstyle\sum_j \exp(\mathbf{s}_j), \\ \mathbf{h}_B &= \textstyle\sum_i \mathbf{a}_i \mathbf{u}_i, \end{aligned} \tag{1}$$

where i is the paragraph index. The output probability of the headline and body being incongruent is computed by

$$\hat{y} = \sigma(\mathbf{h}_H{}^\intercal W \mathbf{h}_B + b), \tag{2}$$

where W and b are trainable weights, and $\mathbf{h}_H$ is $\mathbf{u}_H$.

3) BERT-BASED DUAL ENCODER

BERT is a transformer network that was pretrained for a masked language model and with a next-sentence prediction objective [36]. The pretrained network provides a fixed-dimensional representation for each input token by jointly conditioning the left and right contexts from the previous layers. We input a headline and its corresponding body text and retrieve h_H and h_B by mean-pooling the hidden vectors of the last layer, respectively.[1] The output probability is calculated by Equation (2). Using the BERT-based model as a backbone, we train the BDE model while freezing the weights of the pretrained BERT network due to the lack of computational resources. We call this model **BDE**. As another baseline, we also measure the next-sentence prediction performance of **BERT**.

B. PROPOSED APPROACH

The existing approaches compute a similarity score between the headline and body text, and many of these methods suffer from performance degradation due to the increased content complexity that occurs when an article is too long. AHDE, the state-of-the-art model, utilizes a hierarchical structure to cope with long news stories and abstract content at the paragraph level. We also exploit this hierarchical structure by considering the headline and paragraphs as analysis units. We further utilize graph-based learning to better detect incongruities by learning the importance of each paragraph in an end-to-end manner. The proposed model is a graph-based hierarchical dual encoder (**GHDE**) that computes the headline incongruity probability of a news article in four steps, as illustrated in Figure 3. It first computes a node representation of each headline and paragraph using a hierarchical RNN structure. The headline node and each paragraph node are paired to compute a matching score, which is considered as an edge weight for those nodes. After the graph is completed using the previous steps, the graph neural network propagates information between nodes to examine the article's incongruity. The final step fuses the updated information from each node and outputs the incongruity predictions. We describe this model in more detail in the following section.

1) THE HIERARCHICAL NODE ENCODING STEP

The GHDE constructs an undirected graph $G = (V, E)$ for each news article that represents its innate structure, which is then used to train a graph neural network. V is the set of nodes comprising the headlines and each paragraph of the news content. An edge in E is formed between the headline and each paragraph, resulting in a total of $|E| = |P|$ edges.

A hierarchical dual encoder layer learns the initial node representation using a two-level hierarchy. To encode a headline into a fixed-size vector, a gated recurrent unit (GRU)-based RNN takes word sequences as input. The final hidden state of the RNN's $\mathbf{h}_{head}$ corresponds to the headline's representation. For the body text, a GRU-based RNN learns the word sequence of each paragraph and takes the last hidden state of the RNN as the representation of each paragraph: $\{\mathbf{h}_1, \cdots, \mathbf{h}_{|P|}\}$. The GRU-based bidirectional RNN then learns the paragraph representation from the first level of the RNN and the context-aware paragraph representation $\{\tilde{\mathbf{h}}_1, \cdots, \tilde{\mathbf{h}}_{|P|}\}$.

2) THE EDGE LEARNING STEP

The next step is to learn the edge weights of the input graph G to prevent detrimental smoothing of the node representation between the congruent and incongruent paragraphs during GNN propagation. A bilinear operation with sigmoid nonlinearity σ computes an edge weight e_i between the news headline and the i-th paragraph:

$$e_i = \sigma(\mathbf{h}_{head}{}^\intercal W_E \tilde{\mathbf{h}}_i + b_E), \tag{3}$$

where W_E and b_E are trainable weights. The use of the sigmoid function bounds the edge weight to a value between zero and one; these weights plays a mask role when the features are aggregated in the GNN. We add supervision to the edge weights by using the paragraph congruity value of $1 - y_i$ as a label during the cross-entropy loss, where y_i indicates whether a paragraph originates from another article. This edge-level supervision enables the GHDE to assign high weights to congruent paragraphs and low weights to incongruent paragraphs; thus, it helps congruent paragraphs

[1] We take the average of hidden vectors that correspond to valid tokens (other than special tokens such as [CLS], [SEP], and [PAD]). We utilize the mean-pool operation instead of using the hidden vector corresponding to the first special token [CLS] based on the results of comparison experiments in [37].

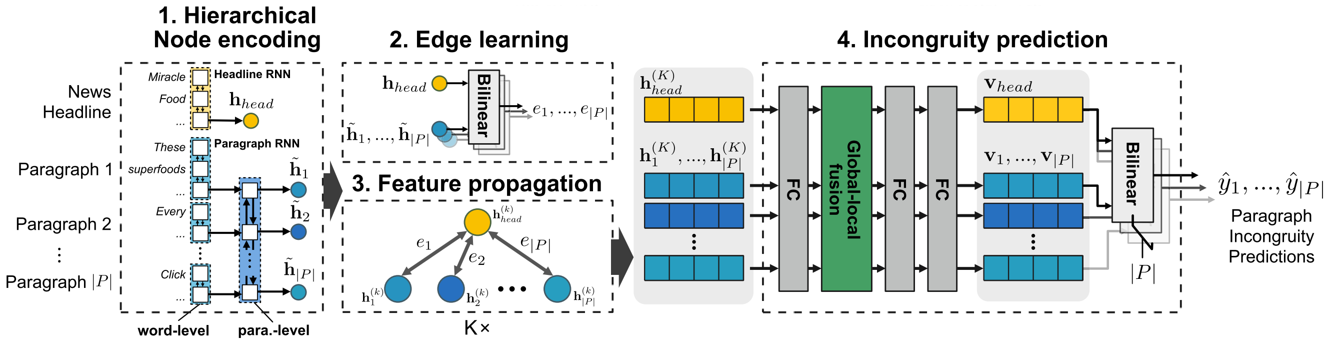


FIGURE 3. An overview of the GHDE (graph-based hierarchical dual encoder) model. The first hierarchical node-encoding step computes the initial hidden representations for the news headline and each paragraph of an article. In the second, edge-learning step, the model computes an edge weight between each paragraph and the headline. The computed edge weights are used to update hidden representations using the GNN during the feature propagation step. The final step computes paragraph-level incongruity scores from the updated hidden representations.

propagate more information to the headline node than can the incongruent paragraphs alone from the propagation step.

The following loss helps in learning weights such that that the edges of congruent paragraphs are retained, while the edges of incongruent paragraphs are masked.

$$\mathcal{L}_{edge} = -\sum_i (1 - y_i)\log(e_i) + y_i \log(1 - e_i). \tag{4}$$

3) THE FEATURE PROPAGATION STEP

The third step is to propagate the node features into the neighboring nodes through the pre-defined graph structure and the trainable edge weights from the GNN framework. GHDE employs an edge-weighted variant of the graph convolutional network (GCN) aggregation function from [13]:

$$\mathbf{z}_i^{(k)} = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{e_{ij}}{\sqrt{\tilde{d}_i \tilde{d}_j}} \mathbf{h}_j^{(k)}, \tag{5}$$

where $\mathbf{z}_i^{(k)}$ is the information propagated to the i-th node from the corresponding set of neighbor nodes $\mathcal{N}(i)$, e_{ij} is the edge weight, and $\tilde{d}_i$ is the degree of the i-th node in the augmented graph with self-loops. The edge weights for the self-loops e_{ii} are set to 1. After feature aggregation, a nonlinear transformation is applied to the resulted outputs as follows:

$$\mathbf{h}_i^{(k+1)} = \mathrm{ReLU}(W_G^{(k)} \mathbf{z}_i^{(k)} + \mathbf{b}_G^{(k)}), \tag{6}$$

where $W_G^{(k)} and \mathbf{b}_G^{(k)}$ are trainable weights. The graph propagation layer is iterated for k times with residual connections.

4) THE INCONGRUITY PREDICTION STEP

The final step predicts the incongruity scores of news articles; this is equivalent to the graph classification task in GNN. To fuse the global-level graph representation with the local-level node representation, GHDE adapts a fusion block as proposed in [38]. It concatenates the node embedding outputs from every GNN layer and passes the embedding through a fully-connected (FC) layer. It then concatenates each node embedding with the max-pooled and sum-pooled representations of the node embeddings in G.

The output node embeddings of the fusion layer are passed through two FC layers to compute the news headline representation $\mathbf{v}_{head}$ and paragraph representations $\{\mathbf{v}_1, \cdots, \mathbf{v}_{|P|}\}$. At this point, GHDE can determine an incongruity label for each paragraph in a news article based on a bilinear operation:

$$\hat{y}_i = \sigma(\mathbf{v}_{head}{}^{\intercal} W_B \mathbf{v}_i + b_B), \tag{7}$$

where σ is the sigmoid nonlinear activation function and W_B and b_B are the learned model parameters. The paragraph-level incongruity scores $\{\hat{y}_1, \cdots, \hat{y}_{|P|}\}$ are merged to determine the article-level incongruity score $\hat{y}$ by taking the maximum of the paragraph-level scores:

$$\hat{y} = \max\{\hat{y}_1, \ldots, \hat{y}_{|P|}\}. \tag{8}$$

GHDE is trained in an end-to-end manner to minimize the following loss:

$$\begin{aligned} \mathcal{L}_{article} &= CE(\hat{y}, y) \\ \mathcal{L}_{edge} &= \textstyle\sum_{i=1}^{|P|} CE(e_i, 1 - y_i) \\ \mathcal{L} &= \mathcal{L}_{article} + \lambda \mathcal{L}_{edge} \end{aligned} \tag{9}$$

where y is the incongruity label of the input news article, CE is the cross-entropy loss, and $\mathcal{L}_{article}$ and $\mathcal{L}_{edge}$ are the loss for the article incongruity prediction and the edge weight, respectively. λ is a hyperparameter for adjusting the tradeoff.

V. PERFORMANCE EVALUATION

A. DETECTION ON THE GENERATED DATASET

We conducted classification experiments to compare the newly proposed GHDE model with baseline methods in terms of accuracy and the area under the receiver operating characteristic (AUROC) curve. We report the average value of all the results after running the experiments five times with distinct seeds.

For AHDE, we employ two single-layer GRUs with 200 hidden units for the word-level RNN and another two single-layer bidirectional GRUs with 100 hidden units for the paragraph-level RNN. For regularization, we apply dropout at ratios of 0.7 and 0.9 for the word-level RNN and paragraph-level RNN, respectively. We used the Adam optimizer with norm gradient clipping at a threshold of 1 [39]. We used $\text{BERT}_{\text{base}}$ for BDE, which includes 12 transformer layers and 12 attention heads and outputs hidden vectors with 768 dimensions. The model is trained using the AdamW optimizer with the learning rate set to 0.001.

For GHDE, we utilize a single-layer GRU with 200 hidden units to encode a headline and each paragraph of the corresponding body text, and use a single-layer bidirectional GRU with 100 hidden units for the paragraph-level RNN. The number of GNN layers, K, is set to 3 with 200 hidden units in each layer. The hidden unit dimensions of the FC layers applied after feature propagation on the graphs are 200, 200, and 100, respectively. The model is trained using the Adam optimizer with a batch size of 120, and gradient clipping is applied with a threshold of 1.0. We decay the learning rate every three epochs starting from an initial learning rate of 0.001 at a decay rate of 0.1. The tradeoff hyperparameter λ for edge loss is set to 0.1.

For all the models that include an embedding layer, we initialize the layer using the pre-trained GloVe embedding matrix [40]. The vocabulary size of the embedding matrix is determined by the number of words that occur at least eight times in the training dataset. All the hyperparameters are optimized on the development set based on more than twenty trials. The dataset and the implementation details for the empirical results will be available via a public web repository.[2] For the experiments, we use a computer equipped with an Intel(R) Core(TM) i7-6850K CPU (3.60 GHz) and a GeForce GTX 1080 Ti GPU. The software environments are Python 3.6 and PyTorch 1.2.0. The total number of trainable parameters in GHDE is 1,214,702. A single GHDE training run takes approximately 3 hours and 18 minutes. The resulting accuracy and AUROC scores on the validation set were 0.8561 and 0.9326 on the Similar dataset and 0.9560 and 0.9860 on the Random dataset.

Table 3 displays the model performances when applied to detect headlines incongruencies on two datasets: Similar (where the target and sampled articles have similar topics) and Random (where the target and sampled articles are random matches with no constraint on the topic of the sampled article compared to the topic of the target article). Other than the method for selecting the sampled news articles, the generation processes for these two datasets are identical. We measure the performance on each different test set; consequently a high performance value does not imply the superiority of a sampling method for detection.

From these results, we make two observations. First, GHDE achieves the highest accuracies on the Similar and Random datasets, 0.852 and 0.959, respectively. The next best algorithm is AHDE, which reaches 0.799 and 0.922, respectively. Both models embed the hierarchical structures

[2]The pointer to the repository will be placed here after the review process.

TABLE 3. Experimental results of headline incongruity predictions on two datasets: Similar and Random. The top scores for each comparison set are highlighted in bold text.

Model	Similar Dataset		Random Dataset	
	Accuracy	AUROC	Accuracy	AUROC
GHDE	**0.852**	**0.928**	**0.959**	**0.989**
AHDE	0.799	0.879	0.922	0.971
XGB	0.700	0.776	0.687	0.756
BDE	0.654	0.712	0.720	0.799
BERT	0.510	0.487	0.512	0.561

TABLE 4. Ablation results of the GHDE model with varying levels of supervision and different graph structures.

	Accuracy	AUROC
Supervision		
Article-level	0.838	0.916
Paragraph-level	0.832	0.923
Paragraph-level + Edge loss	0.846	0.927
Graph structure		
+ Inter-paragraph edges	0.832	0.921
+ Fully-connected edges	0.827	0.917
Article-level supervision + Edge loss	**0.852**	**0.928**

of news articles; however, our graph-based neural network further exploits the news headlines and the unique structures of body paragraphs. The coherence values between a headline and a body paragraph and across body paragraphs are learned as edge weights of the graph-like structure. Second, all four models exhibit better performances on the Random dataset than on the Similar dataset. This suggests that identifying incongruent articles generated by random sampling is easier, yet it does not answer the question of which type of data better represents incongruent news articles in the real world.

B. ABLATION EXPERIMENTS

To test the individual components of the GHDE model, we conducted an ablation study by examining the performances of models after removing each model component. Table 4 shows the ablation results.

In terms of supervision, article-level supervision and edge loss indicate $\mathcal{L}_{article}$ and $\mathcal{L}_{edge}$ in Eq. 8, respectively. Paragraph-level supervision indicates the cross-entropy loss between the i-th paragraph incongruity label y_i and the paragraph-level incongruity prediction $\hat{y}_i$ averaged over all the paragraphs in an article. Training the model with article-level supervision alone outperforms the state-of-the-art (i.e., AHDE) by an accuracy margin of 0.038 and an AUROC margin of 0.032. Paragraph-level supervision further improves the AUROC value, but reduces the accuracy. The model that combines article-level supervision with edge loss achieves the best performance, which suggests that dynamic edge updating is a crucial aspect of for detecting headline incongruity through a graph neural network.

We also investigated the benefit of the graph structure itself by training a GHDE model with augmentation in which

TABLE 5. Instruction used for educating annotators in Amazon Mechanical Turk. We further provided the annotators with specific examples, as shown in Figures 4 and 5.

[Overview]

In this task, you are supposed to read random news articles.
Please read the article carefully and help us determine whether the headline is incongruent with the main body text.
We consider a headline incongruent when it makes claims that are unrelated to or distinct from the story.

[Instructions]

1. Read each news article carefully; the article consists of a headline along with its corresponding body text.
2. Give us your thoughts regarding whether each headline is incongruent with its body text.

What is an incongruent headline?

"A headline that does not accurately represent the story in the body text"
- Type 1: A headline makes claims that only partially represent the story in the body text
- Type 2: A headline that is distinct from the main story in the body text

(Type 1: Partial representation)

The following article introduces multiple stories in the body text. However, the headline describes only a partial story; it fails to cover the multiple stories embodied in the body text. We consider this type of 'briefing' article incongruent because of the mismatch between headline and body text.

(Type 2: Incorrect representation)

In the following article, the headline promises to provide specific benefits of wearing masks to fight against COVID-19, yet the body text does not explain the benefits of wearing masks; it includes only a direction to wear a mask and introduces an advertisement.

inter-paragraph edges connect the edges between each pair of consecutive paragraphs, and fully-connected edges connect all the possible combinations of paragraph pairs. Here, we utilize paragraph-level supervision for a fair comparison. The results show that the additional connections between paragraphs are redundant; adding the inter-paragraph edges resulted in a negligible performance difference, while adding the fully-connected edges decreases the performance accuracy of 0.005 and the AUROC of 0.006. We suspect that the additional edges may cause detrimental smoothing effects between the features corresponding to congruent nodes and those corresponding to nodes of incongruent paragraphs during the feature propagation step, making each node feature less discriminative.

C. DETECTION ON REAL NEWS ARTICLES

To test whether the trained model can identify incongruous headlines in the wild, we conducted experiments on Amazon Mechanical Turk using real articles where the body text was not manipulated by any generation method. Through iterative rounds, we asked Turkers to label the following kinds of incongruity based on the definitions of the literature [5], [7], [8]: (1) when the headline only partially supports the claims of the main article, or (2) when the headline does not represent the body text.

Table 5 shows the instructions given to the annotators in the crowdsourcing task. We asked the crowd workers to read the provided news articles carefully and mark their decisions as to whether each article has an incongruent headline. We provided three types of incongruent headline examples and one congruent headline example to help the workers decide. We further asked the workers to classify the incongruent type of each news article. The task included an optional question so that the workers could detail the reasoning leading to their decisions.

The evaluation experiment involves newly gathered news stories that were not used during the training phase. We gathered 63,271 English news articles from news media outlets known for their biased political orientations and active use of clickbait [30], [32]: FoxNews, BuzzFeed, and The Huffington Post. In addition to 500 randomly sampled articles, assuming that an article's prior probability of being incongruity is low, we included the top 40 articles in terms of prediction scores from each of the five models. We assigned at least ten Turkers to each article to annotate incongruity based on the three criteria above. We aggregated the responses by majority


'Start Here': Trump demands to meet whistleblower as Schiff confirms testimony agreement

1. Whistleblower testimony

President Donald Trump said he wants to meet the whistleblower whose anonymous complaint spurred an impeachment probe.

…

2. Hong Kong chaos

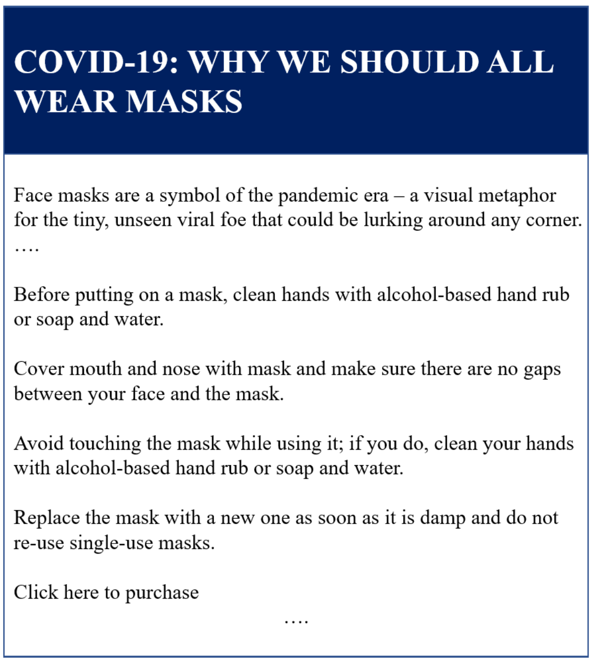
There were violent clashes between pro-democracy demonstrators and police in the streets of Hong Kong over the weekend.

…

3. Child arrests

The two 6-year-old students in Orlando, Florida, whose arrests sparked nationwide outrage and the firing of a school resource officer, are counted among thousands of child arrests in the U.S. annually, raising questions about how school officials discipline misbehavior.

FIGURE 4. Example of an incongruent headline (Type 1: Partial representation).


COVID-19: WHY WE SHOULD ALL WEAR MASKS

Face masks are a symbol of the pandemic era – a visual metaphor for the tiny, unseen viral foe that could be lurking around any corner.

….

Before putting on a mask, clean hands with alcohol-based hand rub or soap and water.

Cover mouth and nose with mask and make sure there are no gaps between your face and the mask.

Avoid touching the mask while using it; if you do, clean your hands with alcohol-based hand rub or soap and water.

Replace the mask with a new one as soon as it is damp and do not re-use single-use masks.

Click here to purchase

….

FIGURE 5. Example of an incongruent headline (Type 2: Incorrect representation).

voting and assigned an incongruity classification when an article received 7 or higher out of 10 votes.

Figure 6 presents the model performances evaluated by the annotated labels. We account for bias in real-world experiments and report the unweighted accuracy (UA) or each class's average accuracy. GHDE trained on the Similar dataset achieves the best performance, achieving an accuracy of 0.760 and an AUROC of 0.784. As was apparent from Table 3, the Similar dataset is more challenging than the Random dataset. However, the real-world evaluation results suggest that the proposed data generation method represents headline incongruity better, enabling a model to be trained that can effectively capture headline incongruity problems in the real world. The models trained on the Random dataset result in poor performance in the wild: most of the models achieved UA scores similar to or lower than 0.6. This result is likely due to the training on the Random dataset, which may induce the models to learn trivial features of topical differences.

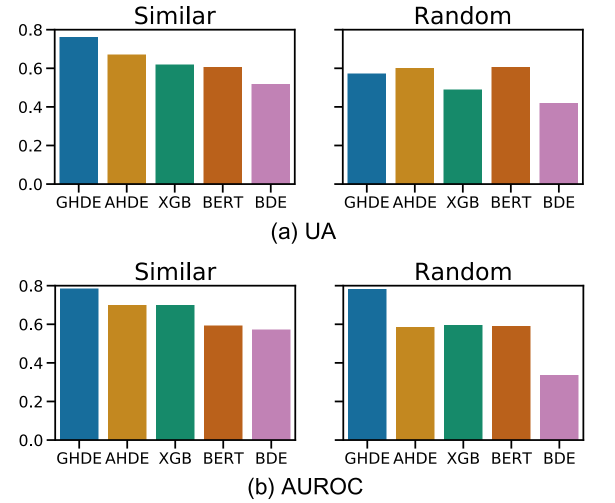


FIGURE 6. Human evaluation results measured on real-world articles.

Compared to the performances measured on the synthetic test set (see Table 3), the performance values on the real-world evaluation are slightly lower. This reduction calls for future studies to develop a more robust detection model and a more realistic data generation method for the headline incongruity problem.

VI. DISCUSSION

News headlines are known to play a crucial role in news selection in online media [41]. According to the Pew Research Center, most U.S. adults (62%) are unlikely to click and read a full news story; instead, they prefer to consume news in aggregated forms via news headlines blurbs [3]. Twitter also announced that they plan to introduce a new design to urge users to click a link before retweeting it because many users do not read the main content [42]. Consequently, when the short headline text does not accurately represent the main content, it can mislead and adversely affect the entire news reading experiences [43]. Therefore, detecting incongruence between headlines and article bodies is both a timely and important aspect for minimizing the negative consequences of potential misinformation.

This paper demonstrated the use of a graph neural network to solve the headline incongruity problem. We found that the hierarchical nature of news posts (i.e., composed of a single headline and multiple paragraphs that are semantically closely related) lends itself well to a graph structure. Therefore, content incongruity can be learned based on low edge weights between hijacked paragraphs and the headline and low edge weights with other paragraphs. The real-world case study confirms that the model is topic-independent (i.e., it can be applied to previously unseen topics such as breaking news).

The solid performances achieved in the crowdsourced experiments suggest that the data generation method contributes to training models that can detect misleading news headlines. Nonetheless, through manual annotations, we observed a few false-positive cases in which a model misidentified a coherent article containing an incongruent headline. For example, one article that covered multiple issues in the main text belonged to this case. According to the definition of headline incongruity, the model correctly predicted the label, but such a "briefing" article does not mislead readers by presenting incorrect information.

Our findings highlight the need for future studies to improve the data generation method and build a training dataset that better represents headline incongruity in the wild. This study did not edit the headline to make it incongruent with the main content, which is a challenging task even for humans. In newsrooms, editors are typically responsible for crafting the headlines of news articles written by reporters. One could address this challenging task by developing a generative model that produces an incongruent headline by inputting pairs of congruent headlines and body text through a generative neural model [44]. The line of research on controlled generation and text style transfer could be used to generate synthetic datasets for headline incongruity [45], [46].

A methodology to discover the relational information among sentences and paragraphs can help in understanding long news articles. A graph neural network is a reliable choice for such cases because the graph embedding propagates information between nodes (given that a node represents a sentence or a paragraph). Many successful studies have adopted graph-based models to tackle NLP tasks such as question answering, document understanding, and other text-related tasks [16], [47], [48]. These studies propose different graph topologies for the learning task, where the graph topology determines the path that allows information to flow between nodes.

In addition to graph-based neural networks, the headline incongruity problem could benefit from other approaches. One such approach would be to use pretrained models such as BERT, which was compared in this study in rudimentary form due to its computational load. Future works could fine-tune the pretrained weights and improve on the transformer layer. For example, in GHDE, we utilize the hierarchical dual encoder (HDE) block to embed the node information corresponding to the headline and paragraphs, but it is possible to use transformer layers instead of an RNN-based block. Another direction might be to directly encode the entire text without adopting a hierarchical model architecture. The previous BERT and its variant models possess limitations in that they can address a maximum token length of only 512. Using recent technologies such as Transformer-XL [49] and Longformer [50], it would be possible to overcome this limitation and explore different ways of computing node representations.

Furthermore, the proposed GHDE can be applied to other applications that require content understanding, such as document summarization, detecting reasoning sentences for question-answering systems, and possibly understanding multimodal (i.e., text, image) contents.

VII. CONCLUSION

In this paper, we studied the detection of news articles that feature headline and body text incongruity, which is an important type of misinformation. Inspired by the hierarchical nature of news articles, we propose a graph-based hierarchical dual encoder (GHDE) that facilitates information flows between headlines and paragraphs to aid in incongruity detection. The evaluation experiments suggest that the proposed approach successfully identifies such misinformation with high accuracy. We hope this study contributes to the construction of more credible online environments for news consumption.

ACKNOWLEDGMENT

(Seunghyun Yoon and Kunwoo Park contributed equally to this work.)

REFERENCES

[1] The Atlantic. (2016). *How Many Stories do Newspapers Publish Per Day?* Accessed: Sep. 21, 2020. [Online]. Available: https://bit.ly/3clqqai
[2] T. Montal and Z. Reich, "I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism," *Digit. J.*, vol. 5, no. 7, pp. 829–849, 2017.
[3] E. S. Matsa and K. Eva, "News use across social media platforms," Pew Res. Center, Washington, DC, USA, Tech. Rep., 2018. [Online]. Available: https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
[4] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, "The effects of subtle misinformation in news headlines," *J. Exp. Psychol., Appl.*, vol. 20, no. 4, p. 323, 2014.
[5] S. Chesney, M. Liakata, M. Poesio, and M. Purver, "Incongruent headlines: Yet another way to mislead your readers," in *Proc. EMNLP Workshop*, 2017, pp. 56–61.
[6] W. Wei and X. Wan, "Learning to identify ambiguous and misleading news headlines," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 4172–4178.
[7] L. Lagerwerf, C. Timmerman, and A. Bosschaert, "Incongruity in news headings: Readers' choices and resulting cognitions," *J. Pract.*, vol. 10, no. 6, pp. 782–804, 2016.
[8] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha, and K. Jung, "Detecting incongruity between news headline and body text via a deep hierarchical encoder," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 791–800.
[9] B. Riedel, I. Augenstein, G. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," *CoRR*, vol. abs/1707.03264, pp. 1–6, Jul. 2017.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[11] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised user stance detection on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 141–152.
[12] R. Mishra, P. Yadav, R. Calizzano, and M. Leippold, "MuSeM: Detecting incongruent news headlines using mutual attentive semantic matching," 2020, *arXiv:2010.03617*. [Online]. Available: http://arxiv.org/abs/2010.03617

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907
[14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.
[15] R. Palm, U. Paquet, and O. Winther, "Recurrent relational networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3368–3378.
[16] N. De Cao, W. Aziz, and I. Titov, "Question answering by reasoning across documents with graph convolutional networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 2306–2317.
[17] L. Song, Z. Wang, M. Yu, Y. Zhang, R. Florian, and D. Gildea, "Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks," 2018, *arXiv:1809.02040*. [Online]. Available: http://arxiv.org/abs/1809.02040
[18] Y. Xiao, Y. Qu, L. Qiu, H. Zhou, L. Li, W. Zhang, and Y. Yu, "Dynamically fused graph network for multi-hop reasoning," 2019, *arXiv:1905.06933*. [Online]. Available: http://arxiv.org/abs/1905.06933
[19] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2205–2215.
[20] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 593–607. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-93417-4_38
[21] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proc. ACL Workshop Lang. Technol. Comput. Social Sci.*, 2014, pp. 18–22.
[22] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 422–426.
[23] H. Karimi and J. Tang, "Learning hierarchical discourse-level structure for fake news detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 3432–3442.
[24] Y. Liu and Y.-F.-B. Wu, "FNED: A deep network for fake news early detection on social media," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–33, Jun. 2020.
[25] Q. Chen, G. Srivastava, R. M. Parizi, M. Aloqaily, and I. A. Ridhawi, "An incentive-aware blockchain-based solution for Internet of fake media things," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102370.
[26] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 5, pp. 1159–1167, Oct. 2020.
[27] J. Zhang, B. Dong, and P. S. Yu, "FakeDetector: Effective fake news detection with deep diffusive neural network," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 1826–1829.
[28] Y. Ren and J. Zhang, "Fake news detection on news-oriented heterogeneous information networks through hierarchical graph attention," 2020, *arXiv:2002.04397*. [Online]. Available: http://arxiv.org/abs/2002.04397
[29] S. Chandra, P. Mishra, H. Yannakoudakis, M. Nimishakavi, M. Saeidi, and E. Shutova, "Graph-based modeling of online communities for fake news detection," 2020, *arXiv:2008.06274*. [Online]. Available: http://arxiv.org/abs/2008.06274
[30] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 9–16.
[31] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9051–9062.
[32] Latest Media Bias Fact Check. (2015). *Media Bias/Fact Check*. Accessed: Sep. 21, 2020. [Online]. Available: https://mediabiasfactcheck.com
[33] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–4.
[34] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," 2017, *arXiv:1702.08734*. [Online]. Available: http://arxiv.org/abs/1702.08734
[35] C. Talos. (2017). *Fake News Challenge—Team SOLAT IN THE SWEN*. Accessed: Sep. 21, 2020. [Online]. Available: https://github.com/Cisco-Talos/fnc-1
[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.
[37] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3973–3983.
[38] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, p. 146, 2019.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[40] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[41] M. Y. Almoqbel, D. Y. Wohn, R. A. Hayes, and M. Cha, "Understanding facebook news post comment reading and reacting behavior through political extremism and cultural orientation," *Comput. Hum. Behav.*, vol. 100, pp. 118–126, Nov. 2019.
[42] T. Verge. (2020). *Twitter is Bringing its 'Read Before You Retweet' Prompt to All Users*. Accessed: Sep. 26, 2020. [Online]. Available: https://bit.ly/2EEhD6Y
[43] J. Reis, F. Benevenuto, P. V. de Melo, R. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," in *Proc. ICWSM*, 2015, pp. 1–10.
[44] K. Lopyrev, "Generating news headlines with recurrent neural networks," 2015, *arXiv:1512.01712*. [Online]. Available: http://arxiv.org/abs/1512.01712
[45] J. Ficler and Y. Goldberg, "Controlling linguistic style aspects in neural language generation," in *Proc. Workshop Stylistic Variation*, 2017, pp. 94–104.
[46] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6830–6841.
[47] G. Nikolentzos, A. Tixier, and M. Vazirgiannis, "Message passing attention networks for document understanding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 8544–8551.
[48] H. Linmei, T. Yang, C. Shi, H. Ji, and X. Li, "Heterogeneous graph attention networks for semi-supervised short text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4823–4832.
[49] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2978–2988.
[50] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*. [Online]. Available: http://arxiv.org/abs/2004.05150

SEUNGHYUN YOON (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Seoul National University, in 2020. He is currently a Research Scientist at Adobe Research, San Jose, USA. His research interests include machine learning and natural language processing (NLP), focusing on question-answering systems and learning language representations for NLP tasks.

KUNWOO PARK received the Ph.D. degree in web science from the School of Computing, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2018. He was a Postdoctoral Researcher with KAIST, the Qatar Computing Research Institute, and UCLA. He is currently an Assistant Professor with the School of AI Convergence, Soongsil University, South Korea. His research interest includes the detecting misinformation and social bias from online media through natural language processing (NLP) and multi-modal methods.

MINWOO LEE received the B.S. degree in integrated technology from Yonsei University, South Korea, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Seoul National University, South Korea. His current research interests include natural language processing (NLP) and graph neural networks.

TAEGYUN KIM is currently pursuing the bachelor's degree with the School of Computing, Korea Advanced Institute of Science and Technology (KAIST), South Korea. His research interest includes applying natural language processing techniques to social science problems.

MEEYOUNG CHA (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2008. From 2008 to 2010, she was a Postdoctoral Researcher with the Max Planck Institute for Software Systems, Germany. Since 2010, she has been a Faculty Member with the School of Computing and the Graduate School of Culture Technology, KAIST. She is currently a Chief Investigator at the Institute for Basic Science, South Korea. Her research interests include data science, information science, and computational social science with an emphasis on modeling socially relevant information propagation processes.

KYOMIN JUNG (Member, IEEE) received the B.S. degree in mathematics from Seoul National University, Seoul, South Korea, in 2003, and the Ph.D. degree in mathematics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009. From 2009 to 2013, he was an Assistant Professor with the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST). Since 2016, he has been an Assistant Professor and then an Associate Professor with the Department of Electrical and Computer Engineering, Seoul National University (SNU), where he is also an Adjunct Professor with the Department of Mathematical Sciences. He is currently working with the Department of Electrical and Computer Engineering, ASRI, SNU. His research interests include natural language processing, deep learning and applications, data analysis, and web services.