Using Adversarial Learning and Biterm Topic Model for an Effective Fake News Video Detection System on Heterogeneous Topics and Short Texts

HYEWON CHOI AND YOUNGJOONG KO
Department of Computer Science and Engineering, Sungkyunkwan University, Suwon 16417, South Korea
Corresponding author: Youngjoong Ko (youngjoong.ko@gmail.com; yjko@skku.edu)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korea Government (MSIT); in part by the Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques under Grant 2020-0-00368; in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2020R1A2C2100362; and in part by the Ministry of Science and Information and Communication Technology (ICT) (MSIT), South Korea, under the ICT Creative Consilience Program Supervised by the IITP under Grant IITP-2020-0-01821.

ABSTRACT Fake news videos are being actively produced and uploaded on YouTube to attract public attention. In this paper, we propose a topic-agnostic fake news video detection model based on adversarial learning and topic modeling. The proposed model estimates the topic distribution of a video using its title/description and comments by topic modeling and tries to identify the differences in stance by the topic distribution difference between title/description and comments. However, directly applying conventional topic models (e.g. LDA and PLSA) on such short texts may not work well. Therefore, we use BTM (Biterm Topic Model) that is robust even in short texts to estimate topic distribution. Then, it constructs an adversarial neural network to extract topic-agnostic features effectively. The proposed model can effectively detect topic changes for stance analysis and easily shifts among various topics. In this study, it achieves a 3.41%p greater F1-score than previous models used for fake news video detection.

INDEX TERMS Fake news video detection, topic modeling, adversarial neural network, YouTube.

I. INTRODUCTION
Recently, several studies have been conducted on detecting fake news videos uploaded on various video-sharing platforms such as YouTube. The topics of fake news videos characteristically change very quickly because their contents involve current societal issues. Conventional fake news detection studies that employ machine learning focus on the training data. However, fake news training data are highly volatile compared to those of other research areas. For example, if a fake news detection model is trained using the training data on fake news disseminated during the U.S. presidential election, it will not effectively detect fake news about COVID-19. Therefore, fake news video detection methods should extract the features of fake news without relying on particular topics to easily find any such videos with a new topic. Furthermore, fake news video detection methods should determine whether the topics in the same video are consistent.

In this study, we use two methods to detect fake news videos: (1) a topic modeling method for detecting topics in videos, and (2) an adversarial learning method for extracting features independent of the subject. Existing topic models such as LDA are often used to find hidden topics in a text corpus. However, it may not be proper to apply existing topic models like LDA to these short texts because the existing topic models capture the co-occurrence patterns of words in a document to find topics and it creates serious data sparsity problems when the document is short. To solve this problem, BTM (Biterm Topic Model) is used to estimate the topic distribution. The key idea of BTM is to learn topics over short texts based on the aggregated biters in the whole corpus to tackle the sparsity problem in single document. Specifically, we consider the whole corpus as a mixture of topics, where each biterm is drawn from a specific topic independently. Therefore, We extract the difference between the topic of comments and the title/description of the YouTube video by using BTM1 that is robust even in short sentences. Since the title/description and comments of YouTube videos usually consist of short sentences, it is very effective in
inferring topic distribution than the LDA (Latent Dirichlet Allocation) method [1], and the difference of the topic distributions between comments and the title/description is utilized to determine their stance distinction. Then, the stance difference of the videos will be reflected in encoding to improve the performance of fake news video detection. Three modules are constructed to proceed with adversarial learning [2], [3]: (1) a multimodal feature extractor, (2) a fake news video detector, and (3) a topic discriminator. The multimodal feature extractor extracts effective features from three types of data, namely video, title/description, and comments, and builds the encoding vector of a video. The fake news video detector is a classifier constructed with the features extracted by the multimodal feature extractor to determine whether the video is real or fake. The topic discriminator is a classifier constructed using the encoding vector of the input video received from the multimodal feature extractor to predict the topic distribution. Therefore, a loss function is designed to proceed with adversarial learning to ensure that the estimated topic distribution of the topic discriminator will not be the same as the distribution of the topics extracted by the BTM. Therefore, a loss function is designed to proceed with adversarial learning to ensure that the estimated topic distribution of the topic discriminator will not be the same as the distribution of the topics extracted by the BTM. Through this learning mechanism, the multimodal feature extractor can be trained to extract topic-agnostic features effectively to increase the performance of the fake news video detector while decreasing the performance of the topic discriminator. With these components, we can construct an effective fake news video detection model that adapts well to rapid topic changes by extracting topic-agnostic features when detecting fake news videos.

The major contributions of the fake news video detection system proposed in this paper are as follows: (1) The topic distribution of a news video is automatically estimated by the BTM, and the stance is distinguished using the difference of the topic distributions between the title/description and comments. Then, the stance distinction is reflected to improve performance. (2) topic-agnostic features are extracted using an adversarial learning method to adapt to rapid topic changes.

II. RELATED WORK
A. FAKE NEWS DETECTION
1) TEXT-BASED FAKE NEWS DETECTION
Many fake news detection methods that use only text to detect fake news have been proposed [4]–[6]. Xu et al. [4] used the term frequency-inverse document frequency (tf-idf) and topic modeling to detect fake news by focusing on the key terms of news articles. Baird et al. [5] achieved the highest performance in the first stage of the fake news challenge, which was a stance classification task, using a combination of deep convolutional neural networks and gradient-boosted decision trees containing lexical features. Hanselowski et al. [6] detected the stance of fake news using six hidden layers and five multi-layer perceptron (MLP)-based classifiers with manually produced features.

2) IMAGE- AND TEXT-BASED FAKE NEWS DETECTION
Studies have also been actively conducted on multi-modal fake news detection using news images and articles. Singhal et al. [7] proposed the SpotFake, which exhibited higher fake news detection performance than conventional detection models despite its simpler architecture. SpotFake comprises three modules. The first is a text feature extractor that uses a language model, namely the Bidirectional Encoder Representations from Transformers (BERT), to extract text features for each situation. The second is a visual feature extractor, and the third is a fusion module that combines the features obtained from the other two to form a single news feature vector.

Previous studies also used adversarial neural networks for fake news detection. Wang et al. [8] proposed a fake news detection model using an adversarial neural network to learn how to avoid extracting features related to one topic. However, because a fake news video may contain multiple topics, the discriminator proposed by Wang et al. [8] that predicts one topic only, cannot extract topic-agnostic features effectively. In this study, topic modeling is used to create the labels of the topic discriminator automatically and predict the distribution of more than one topic for more effective detection of fake news videos.

3) VIDEO- AND TEXT-BASED FAKE NEWS DETECTION
Previous studies also focused on the detection of fake news videos using video and text data. Recently, fake news video benchmark data for efficient detection of fake news videos were established and a model that can effectively detect fake news videos using the constructed benchmark data was proposed [9], [10]. Palod et al. [11] used simple functions based on machine learning models, such as naive Bayes, support vector machines, and decision tree, to detect fake news videos. For example, the percentages of “likes” and “dislikes” and the percentage of violent words expressed in a video title were considered. Simple features and long short-term memory (LSTM) networks were used to encode the comments of a video, and the encoding vectors of comments were linked to detect fake news videos.

III. PROPOSED METHOD
The proposed model uses BTM to estimate the topic distribution of a video and calculate the stance difference in a video automatically. Then, an adversarial learning method is applied to extract the topic-agnostic features for effectively detecting fake news videos. Figure 1 shows the process of the proposed model. The modules represented within the dotted lines are newly proposed and are implemented in the proposed method while the other modules are regarded as the baseline model.

A. BASELINE MODEL
1) COMMENTS ENCODING
All words in the comments are first converted into word segments using BERT and then encoded into vector
The k-th word segment vector of dimension d is denoted as \( x_k \in \mathbb{R}^d \). To encode the i-th comment, the comment is divided into word segments and a bidirectional LSTM is used to generate the comment embedding \( H^i = [h^i_{1}; h^i_{N}] \), where \( h^i_N \) is the vector representation of the i-th word segment with n word segments, \( h^i_{N} \) is n-th hidden vector from the forward direction of the LSTM, and \( h^i_{1} \) is 1-st hidden vector from the backward direction of the LSTM. In conventional fake news video detection studies, all comments existing in a video are simply averaged to create an overall comment embedding. However, the number of "likes" in each comment is exploited to create more effective embeddings than those created by conventional embedding methods. The number of "likes" in each comment is defined as \( C = \{c_1, c_2, \ldots, c_N\} \), where \( N \) is the number of comments contained in the video and \( c_i \) is the number of "likes" of the i-th comment. The overall comments embedding \( H \) of a video is generated as a weighted sum of the number of "likes" in a comment as follows:

\[
H = \sum_{i=1}^{N} \frac{c_i + 1}{c_{\text{total}} + N} H^i, \quad H \in \mathbb{R}^{2d},
\]

where \( c_{\text{total}} \) is the sum of the number of "likes" of all comments in a video; in equation (1), the Laplace estimation is applied by adding 1 to \( c_i \).

2) TITLE/DESCRIPTION ENCODING
As the comments, the title/description of a video is converted into word segments using BERT. The Text-CNN proposed by Kim [13] uses multiple filters with various window sizes to extract a variety of features and generates the title/description encoding vector \( R \in \mathbb{R}^{2d} \).

3) VIDEO ENCODING
The pre-trained VGG-19 is utilized to extract visual features from video thumbnails and video frames [14]. The visual encoding vector of a video thumbnail is denoted as \( V_{\text{thumbnail}} \in \mathbb{R}^{2d} \), and the encoding vector of the i-th video frame is denoted as \( V_{\text{frame}(i)} \in \mathbb{R}^{2d} \). The cosine similarity score between \( V_{\text{thumbnail}} \) and \( V_{\text{frame}(i)} \) is calculated to extract the J frames that are most similar to \( V_{\text{thumbnail}} \). Then, \( V_{\text{thumbnail}} \) and the encoding vectors of the extracted J frames are concatenated to construct a two-dimensional matrix \( L \). \( V_{\text{thumbnail}} = V_{\text{frame}(J+1)} \) is added because the thumbnails of some videos are not extracted from the original video frames. For \( V_{\text{thumbnail}} \) and \( L \), the attention scores are calculated using equations (2) and (3), and the final video encoding vector \( V \) is calculated using equation (4):

\[
A_1 = V_{\text{thumbnail}} \cdot W_\text{Attention} \cdot L, \quad \text{(2)}
\]

\[
\alpha_i = \text{Softmax}(a_i), \quad \text{(3)}
\]

\[
V = \sum_{i=1}^{J+1} \alpha_i V_{\text{frame}(i)}, \quad \text{(4)}
\]

where \( A_1 = [a_1, a_2, \ldots, a_i, a_{J+1}] \) and \( V \in \mathbb{R}^{2d} \).

4) ENCODING VECTOR INTEGRATION
In previous studies, a linear combination is used to combine two feature vectors of text and video effectively [15], [16]. In this study, however, there are three encoding vectors (video vector \( V \), title/description vector \( R \), and comments vector \( H \)), and a linear combination is not suitable. Therefore, the following procedure is executed. First, the hyperparameter \( m_1 \) is learned to reflect the relationship between the video encoding vector \( V \) and the title/description encoding vector \( R \) to learn the encoding ratio as follows:

\[
m_1 = \sigma(w_g \cdot [V, R]), \quad \text{(5)}
\]

\[
P = m_1 \cdot V + (1 - m_1) \cdot R, \quad \text{(6)}
\]
where \( m_1 \) is a scalar value, and \( P \in \mathbb{R}^{2d} \) is an encoding vector containing the visual information of the title/description and video.

The same process is applied to combine \( H \), which includes the comment information of the video, with \( R \):

\[
m_2 = \sigma(w_y \cdot [H; R]),
\]

\[
E = m_2 \cdot H + (1 - m_2) \cdot R,
\]

where \( m_2 \) is a scalar value, and \( E \in \mathbb{R}^{2d} \). Then, \( P \) (equation (6)) and \( E \) (equation (8)) are concatenated and delivered to the fake news video detector to check whether the video is fake.

**B. STANCE DIFFERENCE ESTIMATION WITH TOPICS DISTRIBUTION**

The topics distribution estimation process in Figure 1 illustrates the method of estimating the comments’ topic distributions and the title/description by using the BTM based on Gibbs sampling. It is significant to infer topic distribution from the video because the proposed method infers the topic of the video by the topic modeling method. Therefore, we use the robust BTM even in short sentences to infer topic distribution. All \( M \) videos included in the training data are used to create title/description documents written by a video provider and the comment documents by readers. Then, after preprocessing a total of \( 2M \) documents, BTM estimates the distribution of \( K \) topics: the topic distribution \( \theta_{title/description} \) of the title/description documents, and the topic distribution \( \theta_{comments} \) of the comment documents.

Because there are three types of data to be encoded in the proposed model, the data should be effectively encoded according to the characteristics of each video. Among the data types, comments have been used as a key feature for detecting fake news in many studies [17]. The topic distribution difference, or stance difference, is commonly large between the title/description and comments of fake news videos. Therefore, we increase the encoding ratio of the comments when the difference between two estimated distributions of the comments and title/description is large.

We measure the stance difference \( s \) between the topic distribution of the comments by the readers and the topic distribution of the title/description by the video provider using the Jensen–Shannon divergence as follows [18]:

\[
s = JS\!D(\theta_{title/description}, \theta_{comments}).
\]

Min-max normalization is applied to normalize the score of each video using minimum and maximum scores for all training videos:

\[
s_{normalized} = \frac{s - \min_{s_j \in S}s_j}{\max_{s_j \in S}s_j - \min_{s_j \in S}s_j} \in [0, 1],
\]

where \( s_{normalized} \) is the normalized stance difference, \( S \) is the set of \( s \) of all videos contained in the training data. The calculated \( s_{normalized} \) is applied to equation (7) instead of \( m_2 \) to determine the encoding ratio of \( H \). If the stance difference is large, the encoding ratio of the comments increases; otherwise, the encoding ratio of the title/description increases as follows:

\[
E = s_{normalized} \cdot H + (1 - s_{normalized}) \cdot R.
\]

Finally, \( E \) is concatenated with \( P \) of equation (6), and the results are used as an input in the fake news video detector.

**C. ADVERSARIAL NEURAL NETWORK**

1) FAKE NEWS VIDEO DETECTOR

The fake news video detector is constructed with MLP. The detection loss \( L_y \) is calculated using the cross-entropy as follows:

\[
L_y = -\mathbb{E}_{y \sim (VC, y)}[y \log(d_{fake}) + (1 - y) \log(1 - d_{fake})]
\]

where \( d_{fake} \) is the output of the fake news video detector, \( Y \) is the label set that indicates whether the video is fake, and \( VC \) is the set of all videos.

2) TOPIC DISCRIMINATOR

The topic discriminator consists of the gradient reversal layer (GRL) and the MLP. We use the GRL proposed by Ganin et al. [19], which works as an identity function in the forward step and delivers the result of multiplying the gradient by \(-\lambda\) to the previous layer. We create the multimodal features \( O \) by concatenating all three features to be provided to the topic discriminator:

\[
O = [H; R; V],
\]

The topic discriminator predicts the distribution of \( K \) topics based on the input \( O \). The loss function \( L_{td} \) of the topic discriminator can be calculated as follows:

\[
L_{td} = -\mathbb{E}_{y \sim (VC, Y_{td})}[\sum_{i=1}^{K} y_i \log(\hat{y}_i)],
\]

where, \( Y_{td} \) is the topic distribution of the title/description estimated by the BTM, \( \hat{y}_i \) is the output of the topic discriminator, and the probability of the \( i \)-th topic in the topic distribution. The final loss function \( L_{final} \) of our proposed model can then be expressed as follows:

\[
L_{final} = L_y - \lambda L_{td}
\]

Then, the proposed model can be updated as follows:

\[
\theta_f = \theta_f - \eta \left( \frac{\partial L_{final}}{\partial \theta_f} - \frac{\lambda}{\partial \theta_f} \right)
\]

To stabilize the training process, we reduce the learning rate \( \eta \) as follows [19]:

\[
\eta' = \frac{\eta}{(1 + \alpha \cdot X)^\beta}
\]

where \( \eta' \) is the reduced learning rate, \( \alpha = 10, \beta = 0.65 \), and \( X \) changes linearly between 0 and 1 according to the training progress.
Table 1 shows the data distribution of the three datasets used in the experiments. The Fake Video Corpus (FVC) dataset, which was developed in the In VIdeo Veritas (InVID) project [9], is composed of fake and real news videos of various topics, including politics, sports, and accidents. The Volunteer Annotated Video Dataset (VAVD) is a publicly available fake news video dataset developed by Palod et al. [11], and the Misleading YouTube Video Corpus (MYVC) was collected and annotated for this study. First, we collected fake news and real news from the popular fact-checking websites, Snopes, Fact Check-CNN, and PolitiFact. Next, we collected fake and real news from YouTube to augment the fake news video data. Two researchers evaluated the collected video data. The obtained Cohen’s kappa coefficient was 0.89, indicating that the constructed video data could be considered reliable. Finally, we combined all three datasets to create a full dataset after removing the duplicate videos.

IV. EXPERIMENTAL SETTINGS

A. YouTube DATASETS

Table 1 shows the data distribution of the three datasets used in the experiments. The Fake Video Corpus (FVC) dataset, which was developed in the In VIdeo Veritas (InVID) project [9], is composed of fake and real news videos of various topics, including politics, sports, and accidents. The Volunteer Annotated Video Dataset (VAVD) is a publicly available fake news video dataset developed by Palod et al. [11], and the Misleading YouTube Video Corpus (MYVC) was collected and annotated for this study. First, we collected fake news and real news from the popular fact-checking websites, Snopes, Fact Check-CNN, and PolitiFact. Next, we collected fake and real news from YouTube to augment the fake news video data. Two researchers evaluated the collected video data. The obtained Cohen’s kappa coefficient was 0.89, indicating that the constructed video data could be considered reliable. Finally, we combined all three datasets to create a full dataset after removing the duplicate videos.

B. IMPLEMENTATION DETAILS

Excluding the modules marked by dotted lines in Figure 1, to evaluate the performance of our proposed model we denoted the model as FANVM\(_{\text{Base}}\) (FAke News Video Model) and set it as the baseline system in our experiments; the proposed model was denoted as FANVM\(_{\text{Our}}\). The experiment was implemented using Pytorch in Ubuntu 16.04 LTS with a TITAN X GPU. The hidden state size of the LSTM was set to 300, and the size of the BERT was set to 768. Three filters with sizes of 3, 4, and 5 were considered for the CNN, each of which consisted of 14 kernels. Furthermore, a drop-out method was applied to the fully connected layer to prevent over-fitting. The drop-out ratio was set to 0.5. The proposed model was trained using Adam with a learning rate of 0.0001 [20]. The optimal parameters of BTM were determined using perplexity and coherence. The number of topics was set to 15, and the number of iterations was set to 50. The model that achieved the highest F1-score in the validation set was selected for the final test. Identical hyper-parameter settings were used in all experiments unless otherwise specified.

V. RESULTS AND DISCUSSION

A. EFFECT OF STANCE DIFFERENCE ESTIMATION

Table 2 shows the F1-score achieved by each module in the proposed model. The F1-score increased by 2.00%\(^\text{p}\) when the topic distribution stance distinction was added to the FANVM\(_{\text{Base}}\). We analyzed why the performance was improved when comments were encoded with a high ratio when detecting fake news videos. In the comments encoding in Figure 1, we used the number of “likes” to assign weight to comment embedding. Put differently, the comment embedding’s influence is enhanced by a large number of “likes” in the comments. From the comments of the fake and real news videos, we found that those with many “likes” in the fake news videos were mostly ones that fact-checked the corresponding fake news videos or ones that rejected incitement. In fake news video detection, when the stance difference is large, i.e., when an input video has a high probability of being fake news, highly weighted comment embeddings can increase the importance of the abovementioned comments with many “likes”.

B. EFFECT OF ADVERSARIAL NEURAL NETWORK

We applied two methods in the adversarial neural network experiments. One method, called the Label method, uses a one-hot representation of the topics of videos, and the other method, called the Distribution method, uses a probability distribution of the topics of videos. The Label method estimates the topic distribution in the title/description in Figure 1 and sets the topic with the highest probability among the distribution of \(K\) topics as the ground truth label of the topic discriminator. Conversely, the Distribution method sets the distribution of \(K\) topics of the title/description as the ground truth labels of the topic discriminator to perform the training. When the adversarial neural network was added to FANVM\(_{\text{Base}}\) the Label method improved the F1-score by 2.45%\(^{p}\) while the Distribution method improved the F1-score by 3.39%\(^{p}\) indicating that the extraction of topic-agnostic features based on the topic distribution and the adversarial neural network is more effective in detecting fake news videos. This is attributed to a single fake news video commonly containing various topics. Finally, when both the stance difference and adversarial neural network were applied to FANVM\(_{\text{Base}}\), a 3.92%\(^{p}\) increase in the F1-score was obtained.

C. COMPARISON WITH EXISTING SYSTEMS

Table 3 shows the F1-scores obtained with the FANVM\(_{\text{Our}}\), MVAE [21], SpotFake [7], and UcNet [11]. For accurate performance comparison, we conducted the experiments in the same experimental environment as a conventional fake video dataset specifications.

### TABLE 1. YouTube video dataset specifications.

| Dataset       | Fake | Real |
|---------------|------|------|
| FVC           | 1,301| 1,204|
| VAVD          | 117  | 331  |
| MYVC          | 902  | 903  |
| Full dataset  | 2,201| 2,421|

### TABLE 2. F1-score comparison (%) of our proposed models on the full dataset.

| Model                                      | Full dataset  |
|--------------------------------------------|---------------|
| FANVM\(_{\text{Base}}\)                    | 83.09         |
| FANVM\(_{\text{Base}}\) w/ Stance difference estimation | 85.09 (+2.00) |
| FANVM\(_{\text{Base}}\) w/ Adversarial neural network (Label) | 85.54 (+2.45) |
| FANVM\(_{\text{Base}}\) w/ Adversarial neural network (Distribution) | 86.48 (+3.39) |
| FANVM\(_{\text{Base}}\) (=FANVM\(_{\text{Our}}\)) w/ Stance difference estimation w/ Adversarial neural network (Distribution) | 87.01 (+3.92) |
H. Choi, Y. Ko: Using Adversarial Learning and BTM for Effective Fake News Video Detection System

TABLE 3. Overall F1-score comparison (%) among models on the different datasets.

| Models       | FVC   | VADV  | MVVC  | Full dataset |
|--------------|-------|-------|-------|--------------|
| UcNet [11]   | 74.22 | 75.22 | 70.14 | 74.10        |
| MVAE [21]    | 81.37 | 80.88 | 81.22 | 80.56        |
| SpotFake [7] | 84.88 | 83.34 | 82.31 | 83.60        |
| FANVM_{Our}  | 86.91 (±2.03) | 86.94 (±3.60) | 88.12 (±5.81) | 87.01 (±3.41) |

FIGURE 2. Comments and title/description encoding ratio based on the full dataset.

D. COMPARISON OF ENCODING RATIO

Figure 2 shows the encoding ratio of comments (H) and title/description (R) obtained on the full dataset using equation (11). For fake news videos, the encoding ratio of the comments is 74% on average. In contrast, only 29% of real news videos are encoded. These results indicate that when our proposed stance difference estimation of topic distribution is used to detect a fake news video, the encoding ratio should be adjusted dynamically depending on whether the news video is real or fake.

E. TOPIC DISCRIMINATOR PERFORMANCE ANALYSIS

In section III.B, the target topic distribution of the topic discriminator is the distribution of title/description. Through comparisons between the performance of FANVM_{Our} on each dataset, we verified that our selection, title/description, is more effective than comments. Figure 3 shows the F1-scores of FANVM_{Our} when a different target topic distribution is used. As a result, FANVM_{Our} including the target topic distribution of title/description achieved better performance on all the datasets than FANVM_{Our} with that of comments. These results indicate that the topic distribution of title/description is more effective in detecting fake videos because the title/description was written by the video provider but the comments were made by many readers, that is, the comments generally contain diverse topics regardless of video contents.

F. COMPARISON OF F1-SCORE ACCORDING TO TOPIC MODELING METHODS

Figure 4 shows the F1-scores of FANVM_{Our} according to the topic modeling method. As a result of the experiment, the model inferring topic distribution using BTM had higher performance in all datasets than using the LDA. Since the proposed model computes the stance between the title/description and comments by the topic distribution and utilizes the topic distribution of the title/description as a topic label for an adversarial neural network, it is very important to extract the topic distribution from the video. The title/description and comments of videos mostly consist of short sentences. In particular, the FVC data showed a greater...
Table 4 shows the F1-scores of the FANVM method. More effective in detecting fake news videos than the LDA method is because the title/description and comments on topic modeling was used to dynamically adjust the encoding ratio of the comments to determine whether a video is real or fake. Second, to detect video topic changes effectively, we proposed an adversarial neural network framework that is adaptable to new topics along with a multi-modal feature extractor that can extract topic-agnostic features effectively. Furthermore, we presented the FANVM and the MYVC datasets collected by our group. The experiments on three datasets collected from YouTube demonstrated that the proposed model outperforms state-of-the-art models.

In future studies, we will focus on utilizing the subtitles and speech in videos to detect fake news videos more effectively.

VI. CONCLUSION AND FUTURE WORK

We proposed two effective techniques for detecting fake news videos. First, the estimation of stance difference between title/description and comments on topic modeling was used to dynamically adjust the encoding ratio of the comments to determine whether a video is real or fake. Second, to detect video topic changes effectively, we proposed an adversarial neural network framework that is adaptable to new topics along with a multi-modal feature extractor that can extract topic-agnostic features effectively. Furthermore, we presented the FANVM and the MYVC datasets collected by our group. The experiments on three datasets collected from YouTube demonstrated that the proposed model outperforms state-of-the-art models.

In future studies, we will focus on utilizing the subtitles and speech in videos to detect fake news videos more effectively.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, M. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 2096–2096, May 2015.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014, arXiv:1406.2661.

[4] K. Xu, F. Wang, H. Wang, and B. Yang, “Detecting fake news over online social media via domain reputations and content understanding,” Tsinghua Sci. Technol., vol. 25, no. 1, pp. 20–27, Feb. 2020.

[5] S. Baird, D. Sibley, and Y. Pan, “Talos targets disinformation with fake news challenge victory,” Fake News Challenge, [Online]. Available: https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html

[6] A. Hanselowski, P. Avinash, B. Schiller, and F. Caspelherr, “Description of the system developed by team Athene in the FNC-1,” Fake News Challenge, [Online]. Available: https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf

[7] S. Singhal, R. R. Shah, T. Chakraborty, P. Kamaraguru, and S. Satoh, “SpotFake: A multi-modal framework for fake news detection,” in Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM), Sep. 2019, pp. 39–47.

[8] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, “EANN: Event adversarial neural networks for multi-modal fake news detection,” in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2018, pp. 849–857.

[9] O. Papadopoulo, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, “A corpus of debunked and verified user-generated videos,” Online Inf. Rev., vol. 43, no. 1, pp. 72–88, Feb. 2019.

[10] O. Papadopoulo, M. Zampoglou, S. Papadopoulos, Y. Kompatsiaris, and D. Teyssou, “Invid fake video corpus v2.0 (version 2.0),” Fake News Dataset.

[11] P. Palod, A. Patwari, S. Babay, S. Bagchi, and P. Goyal, “Misleading metadata detection on YouTube,” in Proc. Eur. Conf. Inf. Retr., Jan. 2019, pp. 140–147.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[13] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, arXiv:1408.5882.
[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[15] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, “Online video recommendation based on multimodal fusion and relevance feedback,” in Proc. 6th ACM Int. Conf. Image Video Retr. (CIVR), 2007, pp. 73–80.

[16] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, “Multimodal residual learning for visual QA,” 2016, arXiv:1606.01455.

[17] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, “Web video verification using contextual cues,” in Proc. 2nd Workshop Multimedia Forensics Secur., Jun. 2017, pp. 6–10.

[18] B. Fuglede and F. Topsoe, “Jensen–Shannon divergence and Hilbert space embedding,” in Proc. Int. Symp. Inf. Theory (ISIT), Jun. 2004, p. 31.

[19] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.

[20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, “MVAE: Multimodal variational autoencoder for fake news detection,” in Proc. World Wide Web Conf., May 2019, pp. 2915–2921.

HYEWON CHOI is currently pursuing the M.S. degree with the Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, South Korea. His research interests include natural language processing, fake news detection, information retrieval, machine learning (deep neural networks), and big data analysis.

YOUNGJOONG KO received the Ph.D. degree from the Department of Computer Science, Sogang University, Seoul, South Korea, in 2003. He worked at LG-EDS, from 1996 to 1997. He worked with the Computational Linguistics and Information Processing Laboratory (CLIP), University of Maryland, College Park, as a Visiting Scholar, from 2011 to 2012. In 2019, he moved Sungkyunkwan University, Suwon. Since 2004, he has been joined a Faculty Member of Dong-A University, Busan, where he leaded the Intelligent System Laboratory, Department of Computer Engineering. His research interests include natural language processing, machine learning (deep neural networks), spoken dialogue systems, question answering systems, knowledge graph engineering, information retrieval, and big data analysis.

***