Multimodal Fake News Analysis Based on Image–Text Similarity

Xichen Zhang, Sajjad Dadkhah, Alexander Gerald Weismann, Mohammad Amin Kanaani, and Ali A. Ghorbani, *Senior Member, IEEE*

***Abstract*— With the fast and extensive development of computer vision techniques, multimodal analyses are utilized more frequently for online fake news detection. To better understand the image–text relationship and its role in fake news detection, in this article, we proposed and evaluated four image–text similarities, namely, textual similarity, semantic similarity, contextual similarity, and post-training similarity. The textual and semantic similarities indicate the original image–text similarities in terms of the text information and image caption information. The contextual similarity reflects the image–text similarity in the format of meaningful named entities. The post-training similarity demonstrates how image–text similarity involves before and after a fake news detection model is trained. By evaluating the proposed similarity measurements on three real-world datasets, we find that fake news image–text similarity is higher than real news image–text similarity in most of the cases. Furthermore, the comparison of models' performance further validates the significance of visual information in online fake news detection. These findings may be considered as the fundamental logic to explain the original purpose of fake news creation and can be used as influential features for improving models' performance in the future.**

***Index Terms*— Fake news detection, image–text similarity, multimodal model.**

I. Introduction

NOWADAYS, social media services' extensive growth and development have witnessed the explosive availability and massive spread of digital information. Online social media, such as Twitter and Facebook, can facilitate creating and disseminating news, ideas, comments, interests, discussions, and personal opinions. In 2021, over 3.7 billion users used online social network services worldwide, and the number is projected to increase to 4.4 billion in 2025 [1]. Due to the inherent nature of fast transfer, easy access, and easy use, social media has become a powerful platform for users to consume online information regularly. Remarkably, the prevalence of smart devices and mobile equipment has enabled people to read news, share opinions, and post comments at anytime and anywhere. Consequently, the increasing popularity of social media services has changed how people communicate with each other, reshaping how information spreads on the Internet, and has had profound effects on people's daily lives. However, the unprecedented growth of social media services has created fertile ground for sowing low-credible information [2], [3]. This includes fake news, misinformation, rumors, biased political statements, fake reviews [4], and even misleading videos [4], [5]. For instance, a sharp increase in the quantity of fake news has accompanied the recent coronavirus pandemic, which has confused people in the health domain, swayed people's thoughts, and influenced people's judgments and behaviors [6], [7], [8]. Essentially, spreading false information on the Internet can lead to profound negative consequences. Because of this, the credibility and truthfulness of online information have been identified as significant issues that society must face and have received considerable attention in recent years [9], [10], [11].

Many papers focus on online fake news detection and identification [7], [12], [13], [14], [15], [16], [17], [18], [19]. With the unprecedented advancement and progression of computer vision techniques, visual information has become a critical component for identifying online misinformation. Recently, more and more studies have begun to pay attention to specific visual analyses and multimodal analyses of news content to verify its validity [20], [21], [22], [23], [24], [25]. The traditional fake news detection systems mainly utilize language-based news content for predicting the truthfulness of a piece of news. Unlike conventional works, multimodal approaches evaluate online information based on the news's textual information and corresponding visual information (i.e., images). It is alleged that humans can process and perceive visual information 60 000 times faster than text. Making it very important and valuable for cognitive and emotional influence [26]. Compared with pure textual data, visual-based data contain more intuitive insights and information about the news content. Hence, online news with visual information is more easily remembered and can be transmitted more quickly and extensively. Because of this, more and more online bad actors and fake news creators tend to generate misleading and hostile information based on not just textual information but also visual factors. Therefore, there is an immediate necessity for establishing a multimodal approach that can effectively discover the influential factors of fake news and comprehend the underlying relationships between textual and visual information in real and fake news.

Among all the existing multimodal studies, work [23] is the first work to exploit the similarity between the textual and the visual information in online articles for fake news detection. The authors assumed a considerable gap between text and

Manuscript received 20 July 2022; revised 21 December 2022 and 1 February 2023; accepted 6 February 2023. Date of publication 22 February 2023; date of current version 31 January 2024. *(Corresponding author: Sajjad Dadkhah.)*

The authors are with the Canadian Institute for Cybersecurity (CIC), Department of Computer Science, University of New Bruswick (UNB), Fredericton, NB E3B 5A3, Canada (e-mail: sdadkhah@unb.ca).

Digital Object Identifier 10.1109/TCSS.2023.3244068

graphic elements in fake news for the following two reasons. First, to attract people's attention, fake news creators prefer to use exaggerated pictures to spur emotion and give the reader a sense of drama. Second, fake news content is often created to mislead readers. Thus, it is not easy to find the corresponding images that match the news content exactly. As a result, the visual news content deviates far from its textual content. Similarly, Xue et al. [27] designed a multimodal approach for fake news detection based on the image–text similarity measurement. They also concluded that the mismatching between text and image is a fundamental feature for online fake news detection. Most of the existing works treat image–text similarity as an extra feature for training a multimodal fake news detection model. However, the current studies only present a few image–text mismatching cases as fake news examples. They do not give any intuition or statistical evaluations behind such scenarios, thus making the results difficult to reproduce for various types of online information. Understanding how the image–text similarity differs among fake news and real news is still missing. Furthermore, the existing studies incorporate advanced deep learning algorithms based on the image–text similarity feature and achieve the state-of-the-art detection performance. However, they are just black-box systems and are infeasible to apply in real-world applications since little systematic evidence has been studied to explain such models. Finally, most of the existing image–text similarity calculations are based on simple cosine similarity between textual vectors and visual vectors. They cannot provide any fundamental understanding and guidance regarding the relationship between news texts and images. There exists no study working on the evaluation and comparison of comprehensive image–text similarity calculation between real news and fake news. Aiming at filling this critical research gap, in this work, we propose a novel and practical study on evaluating the image–text similarity for both fake and real news. The significant contributions of this work can be summarized as follows.

1) We formally define the concepts of four image–text similarities, i.e., the textual similarity $\mathtt{Sim}_{\mathtt{text}}(t, v)$, the semantic similarity $\mathtt{Sim}_{\mathtt{sem}}(t, v)$, the contextual similarity $\mathtt{Sim}_{\mathtt{cont}}(t, v)$, and the post-training similarity $\mathtt{Sim}_{\mathtt{post}}(t, v)$. In addition, we provide the comprehensive and detailed information of how to calculate the abovementioned similarity measurements.
2) We propose a novel image caption generation approach. By utilizing the existing image captioning algorithm and the Google image search engine, we can produce an enhanced caption for a targeted image that contains more contextual and semantic meanings.
3) We conduct extensive experimental evaluations on the proposed similarity measurements on two real-world datasets. We find that in some certain circumstances, fake news image–text similarity is higher than that in real news. For example, the top fake news semantic similarities are commonly larger than the top real news semantic similarities and the same for the post-training similarities. The experimental results are essential and useful for researchers to understand the mechanisms of fake news creation and design effective detection and mitigation solutions.

II. Related Work

In this section, we review the state-of-the-art work on fake news multimodal detection. Zhou et al. [23] deployed a multimodal fake news detection system based on similarity-aware evaluations between visual and textual information. Deep learning algorithms are adopted to extract textual and visual features from news articles. Then, the relationship between image features and text features is investigated across modalities. Finally, the image–text similarity measurement and the textual and visual representation are jointly learned by the model to predict fake news. Xue et al. [27] presented a multimodal consistency neural network model for fake news identification. Their model is based on the following five components: 1) a textual feature extraction module; 2) a visual semantic feature extraction module; 3) a visual tampering feature extraction module, 4) a similarity measurement module, and 5) a multimodal fusion module. In this way, the proposed model can assess the consistency between data in different modalities and capture social media information's overall nature. Wang et al. [20] studied the problem of how to identify fake news on newly emerged events. They designed an end-to-end framework called event adversarial neural network. It can extract event-invariant features from online news articles and assist in early fake news detection on upcoming events. The proposed framework has three essential components: a multimodal feature extractor, a fake news detector, and an event discriminator. The textual and visual features cooperate with the fake news detector and the event discriminator to learn distinguishing representations for detecting online fake news. Jin et al. [28] introduced an end-to-end online rumor detection framework based on the recurrent neural network with an attention mechanism. In their framework, the visual features are incorporated with both textual features and social context features to train the long short-term memory (LSTM) network jointly. Khattar et al. [22] described a multimodal framework for detecting online misinformation based on a bimodal variational autoencoder. The proposed system can learn probabilistic latent variable representations by optimizing a bound on the observed data's marginal likelihood and identifying fake news depending on both textual and visual information. An LSTM network is used to extract textual representations, while VGG-19 is used to extract visual information. Qian et al. [29] demonstrated a hierarchical multimodal contextual attention approach for misinformation identification on social networks. BERT and ResNet are used to learn textual and visual representations, respectively. The proposed model collaboratively trains the multimodal contextual information and the hierarchical semantics of news content, which can fuse both intermodality and intramodality relationships of online fake news. Finally, a hierarchical encoding network is designed to estimate the rich hierarchical semantics of news articles. In summary, almost all of the existing works on fake news multimodal detection aim at enhancing and improving the model performance by introducing the visual information from the news. There are some other related

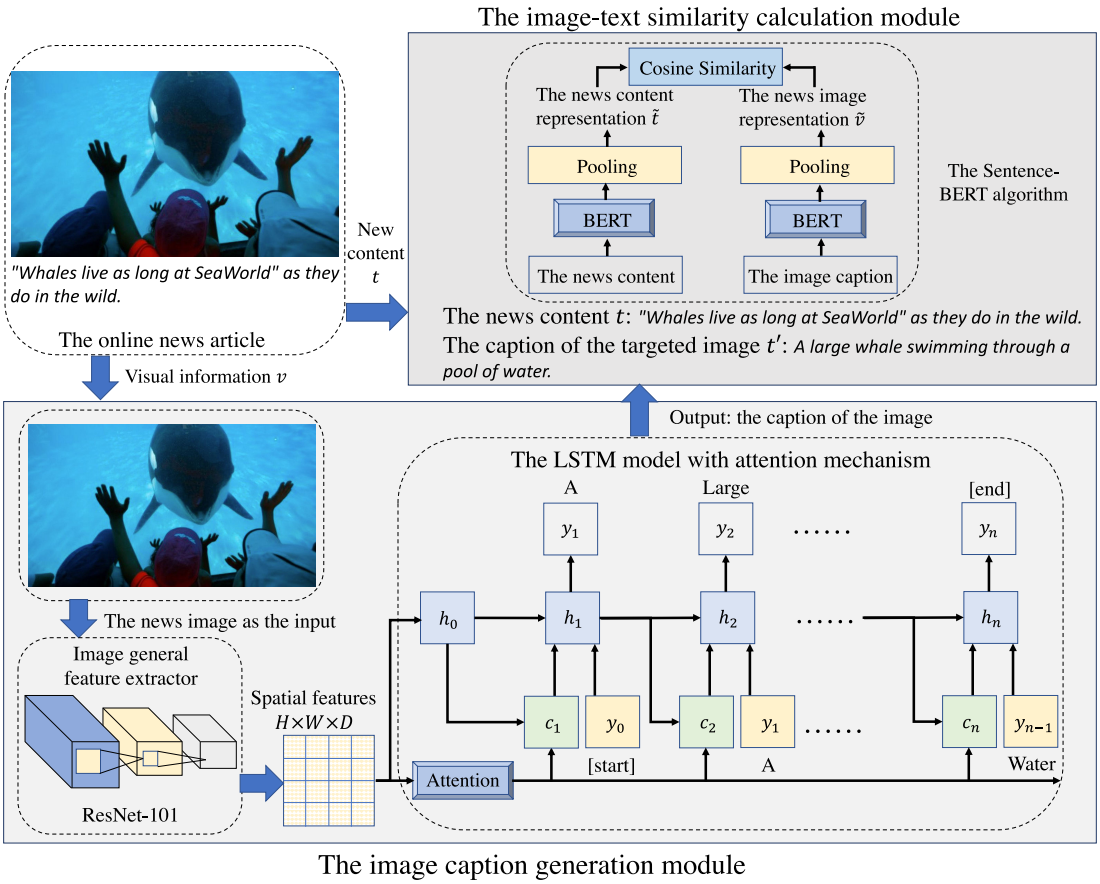


Fig. 1. Overall illustration of image captioning algorithm used in this work.

works on multimodal fake news detection by image captoning [30] and web source credibility analysis [31]. In summary, most of the works aim at improving model's performance by introducing news visual information in different ways. Although the comparison shows that multimodal fake news detection model outperforms single model, the reasons of such improvement are not well explained. There is no reasonable explanation on the connection between their assumption (e.g., the visual information deviates far from its textual content) and their results (e.g., the multimodal performance is better).

III. Methodology

In this section, we mainly discuss the details of the proposed scheme for calculating the image–text similarity for online news articles, e.g., textual similarity, contextual similarity, semantic similarity, and post-training similarity.

A. Image–Text Textual Similarity

In this section, given a news article $\mathcal{N} = (t, v)$, we define the image–text textual similarity as the relationship between textual information extracted from both images and news content, which is denoted by $\texttt{Sim}_{\texttt{text}}(t, v)$. The overall workflow of the image–text textual similarity analyses is shown in Fig. 1. The news in the example is stated by the SeaWorld theme park on March 24, 2015, in an advertisement, which PolitiFact later reported as a half-true news.[1] In particular, there are two modules to compute image–text textual similarity, namely, the image caption generation module `CapGen` and the image–text similarity calculation module `TextSimCal`.

1) *`CapGen`:* In the image caption generation module, the LSTM model [32] with attention mechanism in [33] is used to extract the image caption. Given the news image v as the input information, the `CapGen` module outputs a caption t' as a sequence of encoded words

$$t' = \{y_1, y_2, \ldots, y_n\} \tag{1}$$

where for $i \in [1, n]$, $y_i \in \mathbb{R}^k$ is an encoded work, k is the size of the vocabulary, and n is the length of the generated caption. The details of `CapGen` are discussed as follows. First, same as in [33], we implement the convolutional neural network (CNN) ResNet101 (101 layers) [34] to extract the CNN spatial features. Following the experiments in [35], each image is encoded with the final convolutional layer of the ResNet101 architecture. After applying the spatially adaptive max-pooling operation, the output of each image is a set of l feature vectors $\texttt{Vec}_l$, i.e.,

$$\texttt{Vec}_l = \texttt{ResNet101}(v) \tag{2}$$

where `ResNet101` represents the CNN-based feature extractor model and v is the input image. $\texttt{Vec}_l$ contains l different annotation vectors, and each vector $\texttt{vec}_i$ for $i \in [1, d]$ is a d-dimensional representation associated

[1] https://www.politifact.com/factchecks/2015/mar/24/seaworld/seaworld-says-their-whales-live-long-wild-whales-d/

with a part of the image, i.e.,

$$\texttt{Vec}_l = \{\texttt{vec}_1, \texttt{vec}_2, \ldots, \texttt{vec}_l\} \quad \forall \texttt{vec}_i \in \mathbb{R}^d. \quad (3)$$

After that, the LSTM model with attention mechanism is utilized to create the image caption for each input v. Instead of using the static spatial feature extracted from the image as in [33], we implement the dynamic approach to reweigh the CNN-based spatial features with attention mechanism in LSTM such that a different image annotation vector $\texttt{vec}_i$ for $i \in [1, l]$ can be focused at each time step [35], [36]. The LSTM can produce a sequence of words (as the caption) one at a time based on a context vector $\texttt{vec}_i$, the previous hidden state H, and the previous generated words Y

$$\begin{cases} H = \{h_1, h_2, \ldots, h_n\} \\ Y = \{y_1, y_2, \ldots, y_n\} \end{cases} \quad (4)$$

where n is the number of words (time steps) in the generated caption. By involving the attention-based image feature to the cell node of the LSTM algorithm, all the parameters are optimized by the following equations:

$$\begin{cases} \text{Encoded Word:} \\ y_{\text{ts}} = \texttt{LE}(w_{\text{ts}-1}) \quad \forall ts \le 1, \quad w_0 = \text{BOS} \\ \text{Input Gate:} \\ i_{\text{ts}} = \sigma\left(W_{(i,y)} y_{\text{ts}} + W_{(i,h)} h_{\text{ts}-1} + b_i\right) \\ \text{Forget Gate:} \\ f_{\text{ts}} = \sigma\left(W_{(f,y)} y_{\text{ts}} + W_{(f,h)} h_{\text{ts}-1} + b_f\right) \\ \text{Output Gate:} \\ o_{\text{ts}} = \sigma\left(W_{(o,y)} y_{\text{ts}} + W_{(o,h)h_{\text{ts}-1}} + b_o\right) \\ \text{Memory State:} \\ c_{\text{ts}} = i_{\text{ts}} \bigodot \phi\Big(W^{\otimes}_{(c,y)} y_{\text{ts}} + W^{\otimes}_{(c,I)} I_{\text{ts}} \\ \qquad\qquad + W^{\otimes}_{(c,h)} h_{\text{ts}-1} + b^{\otimes}_z\Big) + f_t \bigodot c_{\text{ts}-1} \\ \text{Hidden State: } h_{\text{ts}} = o_{\text{ts}} \bigodot \tanh(c_{\text{ts}}). \end{cases} \quad (5)$$

Here, each sentence is marked with a Beginning Of Sentence (BOS) token and an End Of Sentence (EOS) token at the beginning and end of it. LE represents a linear embedding function and y_{ts} represents the encoded word. $i_{\text{ts}}, f_{\text{ts}}, o_{\text{ts}}, c_{\text{ts}}$, and h_{ts} are the input, forget, output, memory, and hidden states at time step ts in the LSTM algorithm, respectively. Specifically, I_{ts} is the attention-derived image feature based on a specific annotation vector $\texttt{vec}_i$ for $i \in [1, l]$. $W_{(*,*)}$ and b_* are the learned weight matrices and biases, respectively. ϕ is a maxout nonlinearity with two units ($\bigotimes$ represents the units), $\bigodot$ denotes the Hadamard product, and σ is the sigmoid function. At the end of each time stamp t, the LSTM algorithm will produce a softmax distribution to get the next word $y_{\text{ts}+1}$. It is estimated that in language processing tasks, a current word y_{ts} is assumed conditionally dependent on the preceding $ts-1$ historical words [37]. Therefore, the log probability of a sentence $(y_1, y_2, \ldots, y_{\text{ts}-1})$ can be defined as

$$\log p(y_1, y_2, \ldots, y_T) = \sum_{\text{ts}=1}^{T} \log p(y_{\text{ts}} | y_1, y_2, \ldots, y_{\text{ts}-1}). \quad (6)$$

Consequently, the cross-entropy loss used in optimizing the parameters in LSTM can be defined as

$$L(\theta) = -\sum_{\text{ts}=1}^{T} p(y_{\text{ts}} | y_1, y_2, \ldots, y_{\text{ts}-1}) \quad (7)$$

where $p(y_{\text{ts}} | y_1, y_2, \ldots, y_{\text{ts}-1})$ is given by the softmax distribution and θ represents the parameters of the model.

2) *TextSimCal:* After getting the image caption as the textual information of the news, the Sentence-BERT (SEBRT) algorithm [38] is used in the image–text similarity calculation module to convert both the generated image caption and the news content into fixed-length numeric vectors. In our work, each sentence is projected to a 768-D dense vector space. Finally, the cosine similarity measurement is used to calculate the similarity between the image caption vector and the news text vector.

In summary, the textual similarity $\texttt{Sim}_{\texttt{text}}(t, v)$ is calculated as follows:

$$\begin{aligned} t' &= \texttt{LSTM}(v) \\ \texttt{Vec}_t &= \texttt{SBERT}(t) \\ \texttt{Vec}_{t'} &= \texttt{SBERT}(t') \\ \texttt{Sim}_{\texttt{text}}(t, v) &= \frac{\texttt{Vec}_t \cdot \texttt{Vec}_{t'}}{||\texttt{Vec}_t|| \times ||\texttt{Vec}_{t'}||} \end{aligned} \quad (8)$$

where LSTM and SBERT represent the LSTM model and SBERT model mentioned above, respectively.

B. Image–Text Semantic Similarity

The caption generated by the LSTM model with the attention mechanism can provide us with some high-level and general information about a targeted image. However, it contains limited contextual and semantic information about the image. For example, Fig. 2 shows three examples of news images and their corresponding original generated captions. Literally speaking, the generated caption ideologically aligns the image content by describing the persons' behaviors and their interactions. However, it does not provide any contextual and semantic news information. By just reading the image, readers have no idea about the news content. Therefore, comparing the similarity between the generated caption and the news content cannot reveal the underlying relationship of the image–text similarity in fake news creation. The evaluation results based on such similarities may be biased and not accurate.

To address this problem, in this section, by combining the Google image search engine and the image caption generation module discussed in Section III-A, we propose a novel approach to generate an enhanced caption of a news image,

Original Caption: "a crowd of people standing around each other."

Original Caption: "a group of people standing outside of a building."

Original Caption: "A group of people standing next to each other. "

Fig. 2. Examples of the original generated captions and enhanced captions for some news images from the PolitiFact website.

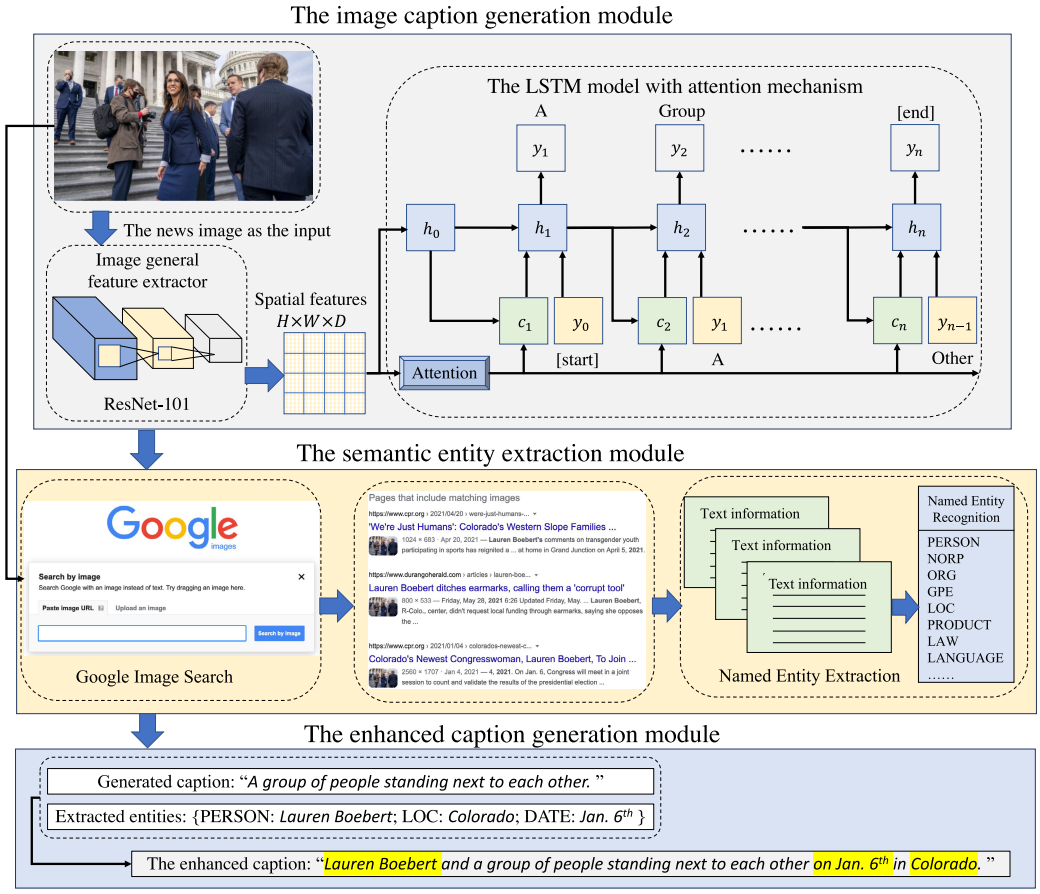


Fig. 3. Overall illustration of enhanced image captioning algorithm used in this work.

namely, the semantic caption. The overall framework contains three modules: 1) the image caption generation module `CapGen`; 2) the semantic entity extraction module `SenExt`; and 3) the enhanced caption generation module `EnhGen`. The overall workflow of the proposed enhanced caption generation framework is shown in Fig. 3. The image example in Fig. 3 is from the news "Colorado's Newest Congresswoman, Lauren Boebert, to Join Electoral College Objection" in CPR News.[2] The details of each module are discussed as follows.

1) CapGen: The image caption generation module `GapGen` is same as we discussed in Section III-A. In a nutshell, given a targeted image v, `CapGen` generates its title t'.

2) SemExt: In the semantic entity extraction module, inspired by a recent related work [44], we incorporate the Google image search into the existing image caption generation module to produce a novel caption with more semantic meanings and understandings of the new content. Specifically, we first define the named entity and then discuss the semantic information extraction procedure.

Definition 1 (Named Entity): A named entity represents a collective name of real-world objects in information retrieval, typically named entities such as person, location, organization, product, and date.

The commonly used named entities and their corresponding descriptions are listed in Table I. Next, given a targeted news image v, we search v in the Google Image Search engine and crawl all the related news websites that contain the same or similar images. We consider all the crawled news websites as v's contextual information in history. After that, by using the name entity extraction techniques, we extract a set of significant named entities $\mathbb{E}^v$ in the searched articles as the

[2]https://coloradosun.com/2020/12/24/lauren-boebert-electoral-college-challenge/

TABLE I
MOST COMMONLY USED NAMED ENTITIES

Entity Name	Description of Each Entity
PERSON	People, including fictional
NORP	Nationalities, religious or political groups
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Geopolitical entities like countries, cities, states.
LOC	Non-GPE locations, like mountain ranges, etc.
EVENT	Hurricanes, battles, wars, sports events, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day

semantic entities of the targeted image v. An example of the explanation of $\mathbb{E}^v$ is discussed as follows.

Example 1: Given the image v in Fig. 3, we first upload v in Google Image Search to crawl all the matched news websites that contain the same or similar images. Then, we extract a set of three entities from the news websites, they are $\mathbb{E}^v = \{\text{PERSON}^v, \text{DATE}^v, \text{GPE}^v\}$, where $\text{PERSON}^v = \{$"Lauren Boebert," "Ben Goldey," "Diane Mitsch Bush," "Bruce Leshan," ...$\}$, $\text{DATE}^v = \{$"Jan 6th," "Jan 4th," "Feb 5th"$\}$, and $\text{GPE}^v = \{$"Colorado," "Washington," "D.C.," "America," "Rifle," "Columbia," ...$\}$.

It is worth noting that each type of entity may include many items that belong to it. For instance, the PERSON entity contains a different person's name, i.e., "Lauren Boebert," "Ben Goldey," "Diane Mitsch Bush," and so on. The importance of each item can be ranked by its frequency of occurrence in all the matched news. For each named entity, its items are ranked by the frequency of occurrence in descending order. We assume that the more frequent an item repeats, the more important and relevant it is to the target image v.

3) EnhGen: In this step, we generate the enhanced caption $\tilde{t}'$ of the targeted image v based on the previously generated caption t' and the contextual entity set $\mathbb{E}^v = \{\text{PERSON}^v, \text{DATE}^v, \text{GPE}^v\}$. Specifically, we first manually define some particular words in t', which indicates the individual personal subjects of the sentence t', e.g., "woman," "man," "person," "people," "kid," "old woman," and "old man." These words are called individual subject indicators and are denoted as $\texttt{ind}_{t'}$. Next, we go through all the items in the PERSON^v entity set. If this entity set is not empty, we select the first item (most frequent item) in PERSON^v entity set and then replaced it with the subject indicator $\texttt{ind}_{t'}$. Finally, we extract the first items in the DATE^v entity set and GPE^v entity set and add them at the end of the sentence. After the modification on t', we can get an enhanced caption $\tilde{t}'$ that contains more contextual information about the targeted image. In particular, we also manually define some group personal subject indicators $\texttt{ind}^g_{t'}$, e.g., "a group of people" and "many people." If $\texttt{ind}^g_{t'}$ appears in t', then we just add the first item in PERSON^v entity set at the front of $\texttt{ind}^g_{t'}$. If neither $\texttt{ind}_{t'}$ or $\texttt{ind}^g_{t'}$ shows in t', then we just add the first items in the Data^v entity and GPE^v entity at the end of the sentence.

Example 2: Given the image in Fig. 3 as an example, the generated caption t' is: "A group of people standing next to each other." In the extracted entity group, the first items in PERSON^v entity set, DATE^v entity set, and GPE^v entity set are "Lauren Boebert," "Jan 6th," and "Colorado," respectively. Therefore, the enhanced caption $\tilde{t}'$ is "Lauren Boebert and a group of people standing next to each other on January 6 in Colorado."

Given a news article $\mathcal{N} = t, v$, we define the image–text semantic similarity as the relationship between the news content t and the enhanced caption $\tilde{t}'$, which is denoted by $\texttt{Sim}_{\texttt{sem}}(t, v)$. After generating the enhanced caption $\tilde{t}'$, we can calculate the semantic similarity $\texttt{Sim}_{\texttt{sem}}(t, v)$ based on the SBERT algorithm discussed in Section III-A.

In summary, the semantic similarity $\texttt{Sim}_{\texttt{text}}(t, v)$ is calculated as follows:

$$\begin{aligned} t' &= \texttt{LSTM}(v) \\ \tilde{t}' &= \texttt{EnhGen}(t') \\ \texttt{Vec}_t &= \texttt{SBERT}(t) \\ \texttt{Vec}_{\tilde{t}'} &= \texttt{SBERT}(\tilde{t}') \\ \texttt{Sim}_{\texttt{sem}}(t, v) &= \frac{\texttt{Vec}_t \cdot \texttt{Vec}_{\tilde{t}'}}{||\texttt{Vec}_t|| \times ||\texttt{Vec}_{\tilde{t}'}||}. \end{aligned} \tag{9}$$

C. Image–Text Contextual Similarity

In this section, we first define a novel image–text similarity called contextual similarity, which is denoted as $\texttt{Sim}_{\texttt{cont}}(t, v)$. Next, we mainly describe how to calculate a news article's $\texttt{Sim}_{\texttt{cont}}(t, v)$ based on the Jaccard similarity.

Basically speaking, for a given news article $\mathcal{N} = (t, v)$, we can generate its contextual entity set for visual information based on Section III-B, which is denoted as $\mathbb{E}^v$, where $\mathbb{E}^v = \{\text{PERSON}^v, \text{GPE}^v, \text{DATE}^v, \text{TIME}^v, \text{NORP}^v, \text{LOC}^v, \text{ORG}^v\}$. In addition, it is also easy to generate the contextual entity set for the text information, which is denoted as $\mathbb{E}^t$, where $\mathbb{E}^t = \{\text{PERSON}^t, \text{GPE}^t, \text{DATE}^t, \text{TIME}^t, \text{NORP}^t, \text{LOC}^t, \text{ORG}^t\}$. Here, for two groups A and B, $A \cap B$ means the number of common elements in A and B, whereas $A \cup B$ is the number of total unique elements in A and B. The image–text contextual similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$ is the set of six Jaccard similarities mentioned above, i.e., $\texttt{Sim}_{\texttt{cont}}(t, v) = \{J_{\text{PERSON}}, J_{\text{GPE}}, J_{\text{DATE}}, J_{\text{TIME}}, J_{\text{NORP}}, J_{\text{LOC}}\}$. Essentially, the contextual similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$ indicates the similarity of the context information between the news text and the news image in terms of different meaningful entities and can be considered as a practical image–text similarity measure for fake news evaluation. Consequently, for the corresponding entity set in $\mathbb{E}^v$ and $\mathbb{E}^t$, we calculate the following seven Jaccard similarity:

$$J_{\text{PERSON}} = \frac{\left|\text{PERSON}^t \cap \text{PERSON}^v\right|}{\left|\text{PERSON}^t \cup \text{PERSON}^v\right|}$$

$$J_{\text{GPE}} = \frac{\left|\text{GPE}^t \cap \text{GPE}^v\right|}{\left|\text{GPE}^t \cup \text{GPE}^v\right|}$$

$$J_{\text{DATE}} = \frac{\left|\text{DATE}^t \cap \text{DATE}^v\right|}{\left|\text{DATE}^t \cup \text{DATE}^v\right|}$$

$$J_{\text{TIME}} = \frac{\left|\text{TIME}^t \cap \text{TIME}^v\right|}{\left|\text{TIME}^t \cup \text{TIME}^v\right|}$$

$$J_{\text{NORP}} = \frac{\left|\text{NORP}^t \cap \text{NORP}^v\right|}{\left|\text{NORP}^t \cup \text{NORP}^v\right|}$$
$$J_{\text{LOC}} = \frac{\left|\text{LOC}^t \cap \text{LOC}^v\right|}{\left|\text{LOC}^t \cup \text{LOC}^v\right|}$$
$$J_{\text{ORG}} = \frac{\left|\text{ORG}^t \cap \text{ORG}^v\right|}{\left|\text{ORG}^t \cup \text{ORG}^v\right|}. \tag{10}$$

D. Image–Text Post-Training Similarity

Up to now, we have proposed three different image–text similarities: textual similarity $\texttt{Sim}_{\texttt{text}}(t, v)$, sematic similarity $\texttt{Sim}_{\texttt{sem}}(t, v)$, and contextual similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$. All three similarities are calculated based on the information extracted from the original image/text or some related knowledge for the image/text (e.g., the entities). In this section, we proposed a novel image–text similarity called post-training similarity, which is denoted as $\texttt{Sim}_{\texttt{post}}(t, v)$. Basically speaking, we first train ViLBERT [45], a state-of-the-art visiolinguistic algorithm, as the multimodal fake news detection classifier. ViLBERT introduces a separate training process for vision and language with co-attention transformer layers, which can integrate different modalities to provide cross-modal interactions and connections at different depths. After the model is fully trained, we extract the hidden image/text representations from ViLBERT and calculate their cosine similarity. Hereafter, we describe the intuition behinds $\texttt{Sim}_{\texttt{post}}(t, v)$. Notably, instead of examining the original image–text similarity for a given news article $\mathcal{N}$, we deploy a multimodal fake news classifier, which can learn influential and representative features for both image v and text t. In specific, given a news article $\mathcal{N} = (v, t)$, Faster R-CNN [46] with ResNet-101 backbone is pretrained on the Visual Genome dataset [47] and is used to extract a sequence of image region features. Then, v and t are represented as a set of the regional features $v_1, \ldots, v_{\mathcal{T}}$ and a series of tokens $w_1, \ldots, w_T$, respectively. Similar to BERT, we add two special tokens to the beginning of vision and linguistic features. The inputs for ViLBERT are the image–text pairs $\{\texttt{IMG}, v_1, \ldots, v_{\mathcal{T}}, \texttt{CLS}, w_1, \ldots, w_{\mathcal{T}}, \texttt{SEP}\}$ and outputs are $h_{\texttt{IMG}}, h_{v1}, \ldots, h_{v\mathcal{T}}, h_{\texttt{CLS}}, h_{w1}, \ldots, h_{wT}$. We take $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$, the corresponding vectors to the IMG and CLS tokens, as the holistic representations of the image and text inputs. Finally, the cosine similarity between $h_{\texttt{IMG}}$ and $h_{\texttt{CLS}}$ is used as the post-similarity, i.e.,

$$\texttt{Sim}_{\texttt{post}}(t, v) = \frac{h_{\texttt{IMG}} \cdot h_{\texttt{CLS}}}{||h_{\texttt{IMG}}|| \times ||h_{\texttt{CLS}}||}. \tag{11}$$

Different from the previously defined similarities, the post-training similarity $\texttt{Sim}_{\texttt{post}}(t, v)$ considers how the text feature t and image feature v are processed in model training and optimization. Therefore, by computing the cosine similarity between the representative features $h_{\texttt{CLS}}$ and $h_{\texttt{IMG}}$, we can understand how the image–text similarity is evolving for distinguishing between fake news and real news in the fake news detection model.

IV. Experiment Evaluation

In this section, we evaluate the comparison of different similarity measurements between real news and fake news and then perform a multimodal approach for fake news detection.

A. Dataset

In this work, we use three real-world datasets to evaluate the proposed similarity measurements for fake news analyses.

1) The first dataset is the famous Buzzfeed news dataset.[3] The Buzzfeed news collection comprises Facebook posts from nine news agencies from September 19 to September 23, 2016. Every social post is fact-checked by five journalists and labeled as either fake or real. The significant features in the Buzzfeed dataset are ID, title, news text, source publisher or author, image URL, and truthfulness label. After eliminating the news that their image URLs are not available, we get in total 100 real news and 100 fake news in our experimental evaluation.
2) The second dataset is the Medieval Twitter dataset [48], which is a benchmark data corpus for analyzing multimedia usage and detecting fake multimedia posts on social media. Each tweet in this dataset consists of text, image/video, or social contextual information. After removing all the tweets without images and filtering out the tweets whose image URLs are no longer available, we finally got 258 real tweet posts and 906 fake tweets.
3) Due to the issue of the existing dataset (e.g., the number of available image links is limited), we built a smart web crawler to collect online fake news for multimodal analysis. In specific, we crawled news from PolitiFact, one of the most popular fact-checking websites. After we get the original news statement from PolitiFact, we follow their image URL to download the corresponding news images. Finally, we get 1589 real news and 1976 fake news. This dataset is referred as CIC dataset.

B. Evaluation of the Enhanced Caption Generation

As we discussed above, in this work, we proposed an enhanced caption generation module in Section III-B by incorporating the traditional image caption generation algorithm and Google image search engine. This module is the fundamental building block for assessing image–text semantic similarity $\texttt{Sim}_{\texttt{sem}}(t, v)$. This section evaluates the reasonableness and intuitive performance of the enhanced caption generation approach.

Fig. 4 shows some comparisons examples between the enhanced captions generated by the improved captioning approach and the original captions generated by the RNN model with the attention mechanism for the Buzzfeed data. From Fig. 4, we can see that the enhanced caption can provide more contextual information about the targeted image. For example, in Fig. 4 [(1), (3), and (8)], the proposed enhanced captioning algorithm can accurately identify the persons in the images, which are "Hillary Clinton," "Joe Biden," and "Jan Schakowsky," respectively. Furthermore, in Fig. 4 [(6) and (7)],

[3]https://www.kaggle.com/sohamohajeri/buzzfeed-news-analysis-and-classification

(1) Caption: "a woman in a blue shirt is smiling."
Enhanced Caption: "Hillary Clinton in a blue shirt is smiling in 2020 in Washington."

(2) Caption: "two men standing next to each other on a stage."
Enhanced Caption: "two men standing next to each other on a stage in March 2015 in NY prison."

(3) Caption: "a man in a suit holding a camera."
Enhanced Caption: "Joe Biden in a suit holding a camera."

(4) Caption: "a group of people standing next to each other."
Enhanced Caption: "Television presenter and a group of people standing next to each other."

(5) Caption: "a crowd of people standing around each other."
Enhanced Caption: "Protesters and a crowd of people standing around each other today in America."

(6) Caption: "a man in a red jacket holding an umbrella.."
Enhanced Caption: "a man in a red jacket holding an umbrella on Wednesday in Kansas City."

(7) Caption: "a group of people standing outside of a building."
Enhanced Caption: "a group of people standing outside of a building on Wednesday in Wuhan City."

(8) Caption: "a man and a woman standing next to each other."
Enhanced Caption: "a man and Jan Schakowsky standing next to each other on November 3rd in U.S."

Fig. 4. Some examples of the comparisons between the enhanced captions and the original captions for news images.

TABLE II
COMPARISON BETWEEN THE PROPOSED IMAGE CAPTION ENHANCE APPROACH WITH THE STATE-OF-THE-ART METHODS

Model	RNN	CNN	LSTM	BERT	Enhanced Caption	PERSON	DATE	ORG	GPE	LOC	Year
Our Approach		x	x	x	x	x	x	x	x	x	2022
Show, Attend and Tell [39]		x	x								2015
Mind's Eye [40]	x	x									2015
gLSTM [41]	x		x								2015
Self Critical Sequence Training [35]		x	x								2017
Stimulus Driven [42]		x	x								2019
SAFE [23]		x									2020
NICNDA [43]		x	x								2020

our approach can accurately detect the locations of the images, which are "Wuhan City" and "Kansas City," respectively. Furthermore, we compare our proposed image caption enhancement approach with the state-of-the-art approaches. The comparison results are shown in Table II. It is very clear to show that the most distinguishing feature of our method is the "enhanced component" in the result. Instead of only outputting general information describing an image, our method can accurately depict the essential named entities, such as "person," "date," "organization," and "location," which are significant for readers to understand the basic information in the news images and identify the mismatching between the news texts and its corresponding visual information. In addition, it is worth noting that our major purpose is to generate captions for the news images, which can be used for understanding the news generation mechanism and comparing the image–text similarity. Therefore, we simply deploy the pretrained model for image caption generation in our experiment.

C. Evaluation of the Extracted Entities in News Images

To evaluate the extracted entities in news images for calculating the contextual similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$, we present several typical examples in Fig. 5. In particular, we compare two real news with two fake news. Given each image v, we enumerate the common entities extracted from both v and t. For example, in the first real news example, the common GPE entities extracted from both v and t can be represented as $|\text{GPE}^t \cap \text{GPE}^v| = \{$"Estonia," "Sochi," "Russia"$\}$. From Fig. 5, first, we can see that the commonly extracted entities can provide us with additional information about the image content and context. Rather than visually interpreting the image, the extracted entities can convey relevant and essential knowledge for the images in terms of important aspects, such as a person, location, and organization. Furthermore, based on the definition of $\texttt{Sim}_{\texttt{cont}}(v, t)$, the commonly extracted entities from v and t extend the concept of image–text similarity to a new level. In other words, instead of only focusing on the information that is directly related to the text/image content, we concentrate on the contextual background for computing the similarity. The entity-based Jaccard measurement indicates the vital contextual fact on image–text similarity and can be used as meaningful evidence for detecting and evaluating online fake news.

D. Evaluation of Different Image–Text Similarities

In this section, we calculate four similarity measurements for the three real-world datasets, and the experimental results are shown as follows.

Fig. 5. Examples of common entities extracted from texts and images for both real news and fake news.

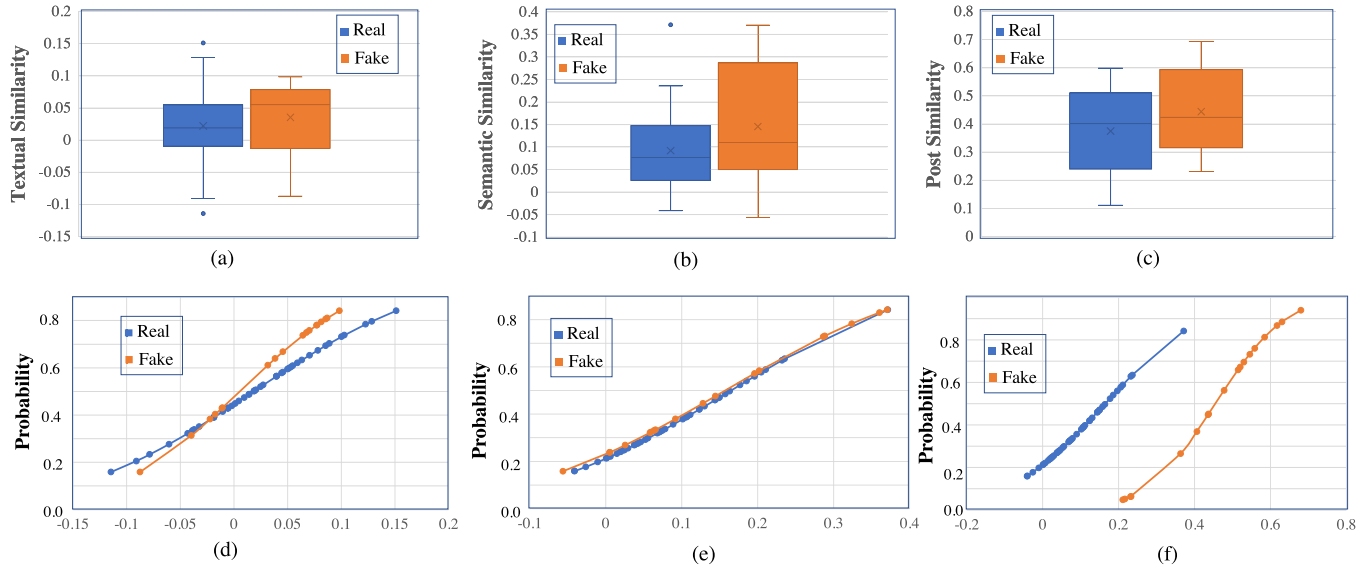


Fig. 6. Statistic evaluation of different similarities for the Buzzfeed dataset. The p-values of the Mann–Whitney test for textual similarity, semantic similarity, and post similarity between real and fake news are 3.62e-09, 1.71e-09, and 2.66e-09, respectively, which concludes the image–text similarity in the Buzzfeed dataset between real and fake news significantly different from each other. (a) Buzzfeed dataset for textual similarity. (b) Buzzfeed dataset for semantic similarity. (c) Buzzfeed dataset for post similarity. (d) Buzzfeed dataset for textual similarity. (e) Buzzfeed dataset for semantic similarity. (f) Buzzfeed dataset for post similarity.

Figs. 6–8 show the comprehensive statistical results of different similarity measurements between real news and fake news for the Buzzfeed dataset, Medieval dataset, and the CIC dataset, respectively. First, it is easy to find that the semantic similarity $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ is higher than textual similarity $\mathtt{Sim}_{\mathtt{text}}(t, v)$ in all the datasets [e.g., compare Figs. 6(a)–8(a) with Figs. 6(b)–8(b)]. The reason is that $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ is calculated based on the enhanced image caption which contains more contextual and background information. Consequently, the semantic similarity between two input sentences is higher. Furthermore, it is interesting to find that both $\mathtt{Sim}_{\mathtt{text}}(t, v)$ and $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ for fake news are higher than that in the real news in all the three datasets. For instance, there are more fake news whose $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ values are in the range of $[0.3, 0.4]$ than the real news in the Buzzfeed dataset. In addition, the fake news's smallest value of $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ in the CIC dataset is larger than real news's largest value [as seen in Fig. 8(b)]. This conclusion is diametrically opposed with the existing multimodal fake news studies, where they claimed that the image–text similarity is higher in real news than fake news [23], [27]. In particular, from Fig. 6(b), we can see that compared with the real news, there are more fake news whose semantic similarity $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ are larger than 0.3. There is no obvious pattern, which can be found for the post-training similarity $\mathtt{Sim}_{\mathtt{post}}(t, v)$. The cumulative distribution function (cdf) curve in Figs. 6–8 also validates the conclusion we discussed, where the image–text similarity in fake news is higher than that in real news for most of the cases.

In addition, from Figs. 6 to 8, we can see the evolvement of the image–text similarity before and after a fake news detection model is trained. In particular, the textual similarity $\mathtt{Sim}_{\mathtt{text}}(t, v)$ and the semantic similarity $\mathtt{Sim}_{\mathtt{sem}}(t, v)$

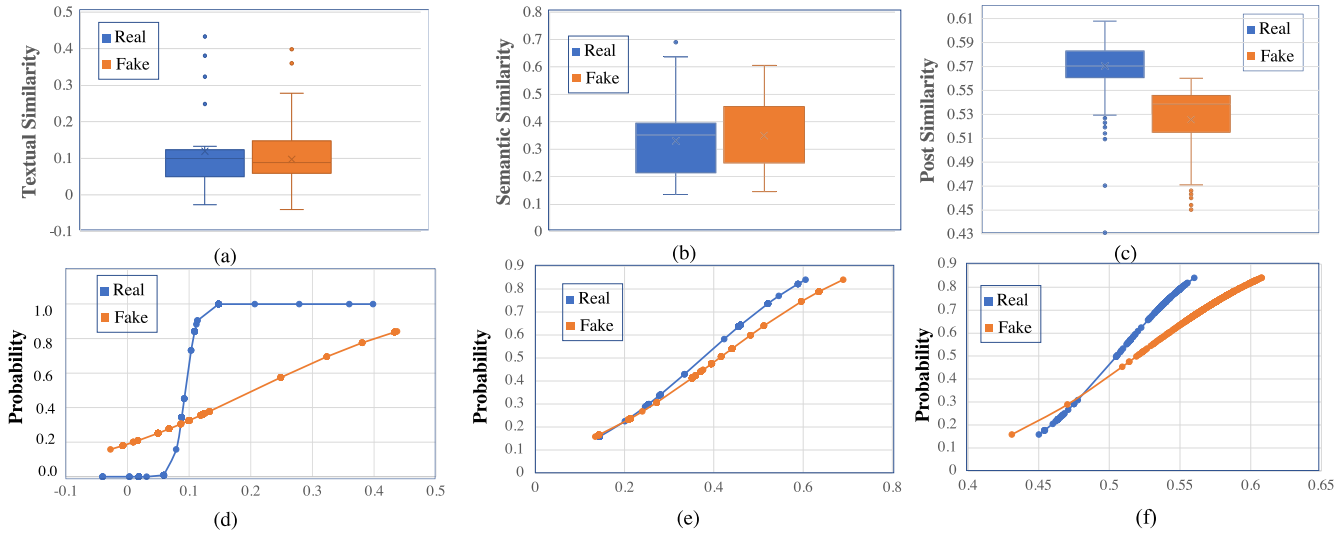


Fig. 7. Statistic evaluation of different similarities for the Medieval dataset. The p-values of the Mann–Whitney test for textual similarity, semantic similarity, and post similarity between real and fake news are 1.71e-12, 1.59e-11, and 1.82e-13, respectively, which concludes the image–text similarity in the Medieval dataset between real and fake news significantly different from each other. (a) Medieval dataset for textual similarity. (b) Medieval dataset for semantic similarity. (c) Medieval dataset for post similarity. (d) Medieval dataset for textual similarity. (e) Medieval dataset for semantic similarity.

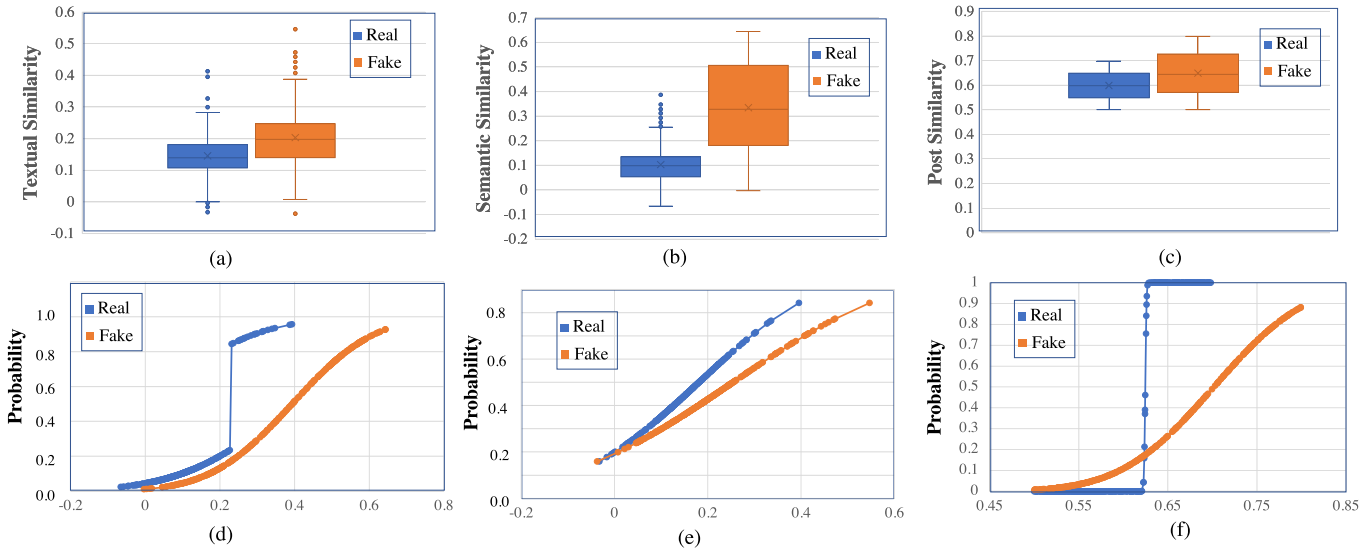


Fig. 8. Statistic evaluation of different similarities for the CIC dataset. The p-values of the Mann–Whitney test for textual similarity, semantic similarity, and post similarity between real and fake news are 3.24e-17, 2.05e-9, and 5.78e-12, respectively, which concludes the image–text similarity in the CIC dataset between real and fake news significantly different from each other. (a) CIC dataset for textual similarity. (b) CIC dataset for semantic similarity. (c) CIC dataset for post similarity. (d) CIC dataset for textual similarity. (e) CIC dataset for semantic similarity. (f) CIC dataset for post similarity.

represent the image–text similarity before training, while the post-training similarity $\mathtt{Sim}_{\mathtt{post}}(t, v)$ indicates the image–text similarity after training. In particular, the fake news image–text similarity is very similar to that in the real news [see Figs. 6(a)–8(a)]. However, after the model is fully trained and optimized, the difference turns out to be more obvious, with fake news being larger than real news in the Buzzfeed and CIC datasets. The effects of model processing on fake news image–text similarity are bigger than real news image similarity. After multiple training rounds, our ViLBERT model finds that making the fake news image–text similarity higher can positively impact the model performance.

Fig. 9 shows the average Jaccard measurement for both real news and fake news regarding different entities for both Medieval and CIC datasets. From 9 (a), we can see that

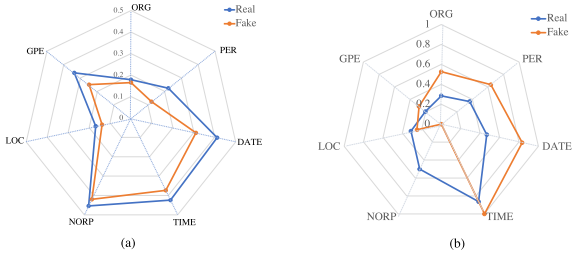


Fig. 9. Radar plot of average Jaccard similarities on (a) Medieval and (b) CIC datasets for both fake news and real news.

real news contain more entities such as DATE, GPE, PER, and TIME. By exhibiting more contextual and event-based descriptions such as time, date, person, and geopolitical entities, real news is more likely to deliver the basic background information and facts regarding the news. On the other hand,

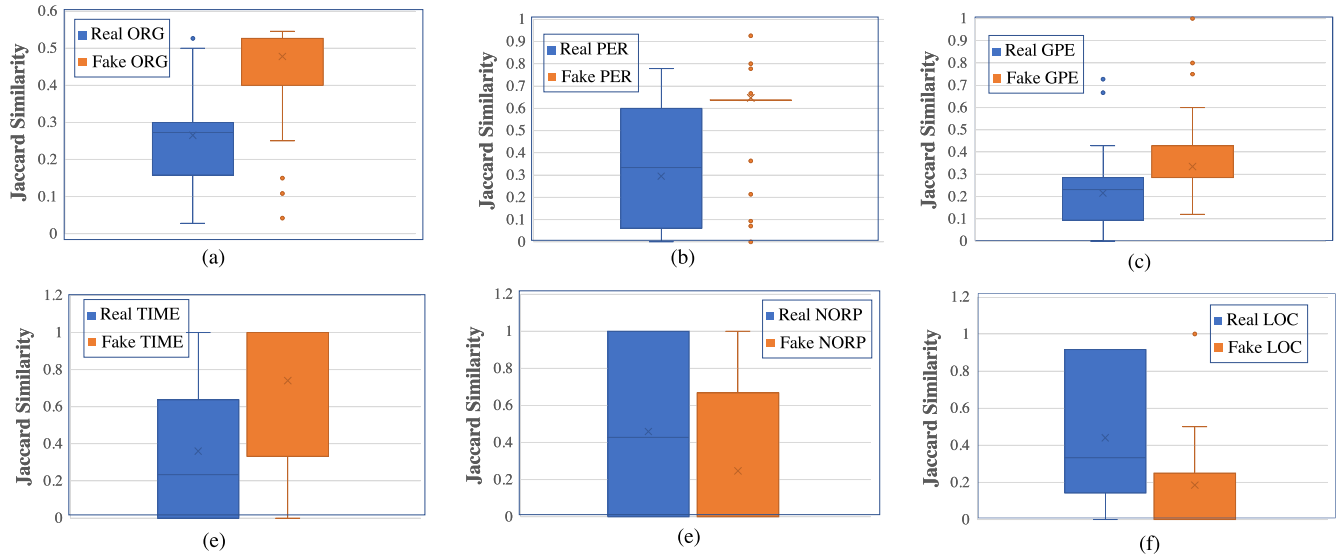


Fig. 10. Statistic evaluation of Jaccard similarity for the CIC dataset. (a) CIC dataset for ORG similarity. (b) CIC dataset for PER similarity. (c) CIC dataset for GPE similarity. (d) CIC dataset for TIME similarity. (e) CIC dataset for NORP similarity. (f) CIC dataset for LOC similarity.

(1) Title of the news: "9 times Ted Cruz insulted Donald Trump before endorsing him."
Enhanced Caption: "Ted Cruz holding up his hands in the air."
Semantic Similarity: 0.37
Label: True

(2) Title of the news: "Trump puzzling pitch to black voters"
Enhanced Caption: "Trump and a group of people standing next to each other."
Semantic Similarity: 0.24
Label: True

(3) Title of the news: "The mystery of Trumps Man in Moscow: Carter Page"
Enhanced Caption: "Cater Page in a suit and tie holding a microphone on 2016 in Russia."
Semantic Similarity: 0.23
Label: True

(4) Title of the news: "New York bomb suspect sued police!"
Enhanced Caption: "Bombing and a group of people standing around a green truck on Monday in New Jersey."
Semantic Similarity: 0.38
Label: False

(5) Title of the news: "Bill and Hillary Clinton have a major aids Scandal!"
Enhanced Caption: "First lady Hillary Clinton and a person standing next to each other in Washington, 2016."
Semantic Similarity: 0.32
Label: False

(6) Title of the news: "Exposed: Soros funding fake veterans PAC to take Trump down."
Enhanced Caption: "George Soros and a couple of men standing next to each other recently."
Semantic Similarity: 0.29
Label: False

Fig. 11. Examples of the semantic similarity $\mathtt{Sim}_{\mathtt{sem}}(t, v)$ for both real news and fake news.

with less entity-based information, fake news seems to be more emotional and sensational in distributing the message. However, in 9 (b), we see that fake news outperform real news in terms of the number of matched GPE, ORG, PER, DATE, and TIME entities. In other words, most of the fake news collected by PolitiFact tend to be similar to real news regarding the named entities and terminologies discussed in the news.

Fig. 10 shows the detailed comparison of different entity similarities between real news and fake news in the CIC dataset. Similarly, we can see that fake news contextual similarity $\mathtt{Sim}_{\mathtt{cont}}(t, v)$ is higher than the contextual similarity

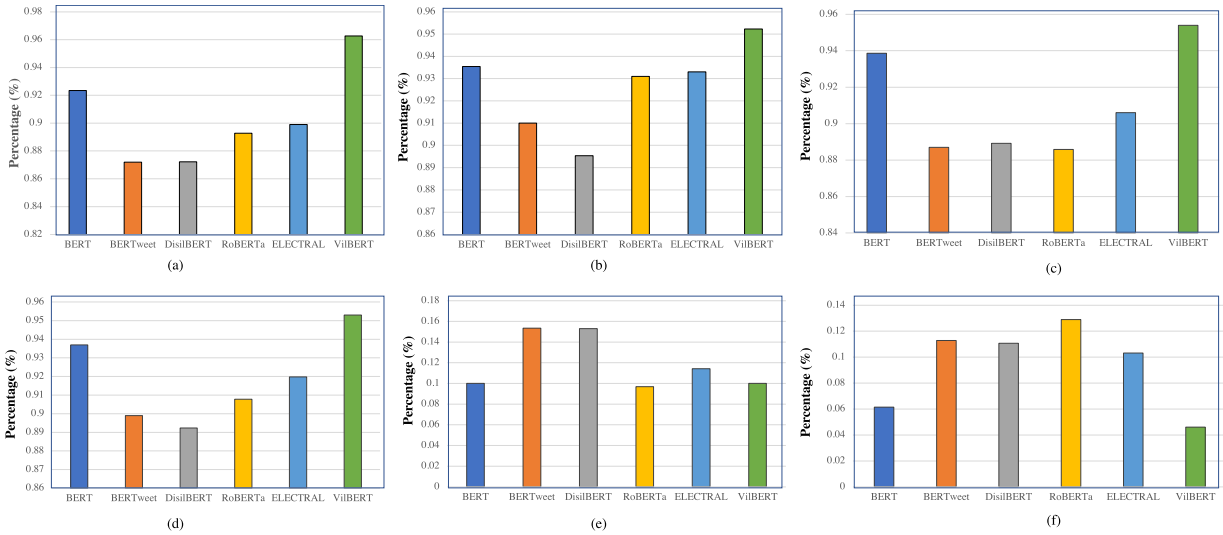


Fig. 12. Performance evaluation between different models on CIC dataset. (a) Accuracy for CIC dataset. (b) Precision for CIC dataset. (c) Recall for CIC dataset. (d) F-1 for CIC dataset. (e) FP for CIC dataset. (f) FN for CIC dataset.

$\texttt{Sim}_{\texttt{cont}}(t, v)$ in real news in terms of ORG, PER, GPE, and TIME entities.

In order to further investigate this interesting discovery, we explore the top-3 fake news and real news in BuzzFeed in terms of the image–text semantic similarity $\texttt{Sim}_{\texttt{sem}}(t, v)$. In Fig. 11, for each image example, we showed the news title (i.e., news content t), the enhanced image caption $\tilde{t}'$, the semantic similarity score $\texttt{Sim}_{\texttt{sem}}(t, v)$, and the label of the news (i.e., fake or real). We can see that the $\texttt{Sim}_{\texttt{sem}}(t, v)$ value for fake news is higher than that for real news in those examples. In contrast to earlier findings, we find that in some cases of fake news, the news images are more relevant to the news content in terms of semantical matching. Rather than using exaggerated pictures, the fake news creators tend to select as content-related figures as possible. Therefore, in the fake news generation, the image–text similarity is high (even higher than real news) to mislead the readers. The content-related image is a key component in fake news such that a reader cannot tell the truthfulness of the information by simply reading the image.

E. Multimodal Fake News Detection

In this section, we conduct multiple fake news detection models for comparing the detection performance with text-only approaches and the multimodal approach. In specific, we deploy the following models, which represent the state-of-the-art text-only methodologies: BERT [49], BERTweet [50], DistilBERT [51], RoBERTa [52], and ELECTRAL [53], whereas we use ViLBERT [45] algorithm as the multimodel for fake news identification. Table III shows the details of the detection performance among different models. Fig. 12(a)–(f) shows the performance comparison between text-only model and multimodal in terms of accuracy, precision, recall, false positive, false negative, and F-1 score. From Table III and Fig. 12, it is easy to make a conclusion that the multimodal fake news detection approach outperforms single-modal approaches in almost all the metrics. BERT performs the best among all the single-modal techniques. Deep learning approaches behave like black box in online fake news detection. The results become even more difficult to understand and interpret when visual information is included in the detection pipeline. Combined the results in the section with the statistical evaluation in Section IV-D, we understand that in the training phase, the similarity between textual vectors and visual vectors in fake news is optimized to become larger than that in the real news circumstances. The results reported in this work can help researchers better understand why visual information is essential in fake news detection and how visual information is evolved in the training optimization.

TABLE III
COMPARISON BETWEEN TEXTUAL-BASED MODEL AND MULTIMODAL FOR FAKE NEWS DETECTION

	Acc.	Pr.	Re.	F-1	FP	FN
BERT [49]	0.9234	0.9354	0.9386	0.937	0.1000	0.0614
BERTweet [50]	0.8720	0.9100	0.8870	0.8990	0.1534	0.1129
DistilBERT [51]	0.8722	0.8954	0.8892	0.8923	0.1530	0.1107
RoBERTa [52]	0.8928	0.9310	0.8858	0.9078	**0.0969**	0.1289
ELECTRAL [53]	0.8989	0.9330	0.9060	0.9198	0.1142	0.1032
ViLBERT [45]	**0.9627**	**0.9523**	**0.9539**	**0.9530**	0.1000	**0.0461**

V. CONCLUSION

In this work, we conduct a practical and novel research problem of evaluating the image–text similarity for real news and fake news. Specifically, we give the definitions of four novel similarities, i.e., image–text textual similarity $\texttt{Sim}_{\texttt{text}}(t, v)$, semantic similarity $\texttt{Sim}_{\texttt{sem}}(t, v)$, contextual similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$, and post-training similarity $\texttt{Sim}_{\texttt{post}}(t, v)$. Here, $\texttt{Sim}_{\texttt{text}}(t, v)$ and $\texttt{Sim}_{\texttt{sem}}(t, v)$ are based on the cosine similarity between news text and image caption/enhanced caption; $\texttt{Sim}_{\texttt{cont}}(t, v)$ is computed by Jaccard similarity regarding contextual entities related to both the text and the image; and $\texttt{Sim}_{\texttt{post}}(t, v)$ is evaluated after a fake news detection model is fully trained. In the process of calculating image–text semantic similarity $\texttt{Sim}_{\texttt{cont}}(t, v)$, we design a novel method to generate the enhanced image caption based

on both traditional image captioning technique and Google image search engine. We evaluate the four similarities using three real-world fake news datasets in the experimental evaluation. The experimental results are interesting, since opposed to the existing research works, we find that fake news image–text similarities are higher than that for real news in most cases. In addition, the comparison of models' detection performance further validates the essentiality of visual information in this problem. Instead of proving that the image–text similarity can increase the fake news detection performance, which has already been validated in many previously published work, our experimental results are essential and useful for researchers to understand the mechanisms of fake news creation and design effective detection and mitigation solutions.

To generate more misleading information, the fake news creators tend to find images highly related to the news content. Psychologically, such news can draw more attention to the readers and cannot be easily detected as false. In the future, we plan to explore the mathematical intuition behind the post-training similarity and try to utilize the image–text similarity as an influential feature for improving the model's performance.

REFERENCES

[1] Statista. *Number of Social Network Users Worldwide From 2017 to 2025*. Accessed: Feb. 12, 2021. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114208.

[3] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 112986.

[4] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell.*, vol. 51, no. 7, pp. 4214–4235, 2021.

[5] D. Varshney and D. K. Vishwakarma, "Artimarker: A novel artificially inflated video marking and characterization method on YouTube," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (5th Int. Conf. Comput., Commun. Signal Process.*, May 2021, vol. 51, no. 7, pp. 244–249.

[6] C. M. Greene and G. Murphy, "Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation," *J. Exp. Psychol., Appl.*, viol. 27, no. 4, p. 773, 2021.

[7] D. Varshney and D. K. Vishwakarma, "Hoax news-inspector: A real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 9, pp. 8961–8974, Sep. 2021.

[8] D. Varshney and D. K. Vishwakarma, "Analysing and identifying crucial evidences for the prediction of false information proliferated during COVID-19 outbreak: A case study," in *Proc. 8th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jul. 2021, pp. 47–51.

[9] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 115002.

[10] D. K. Vishwakarma and C. Jain, "Recent state-of-the-art of fake news detection: A review," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 1–6.

[11] B. Malhotra and D. K. Vishwakarma, "Classification of propagation path and tweets for rumor detection using graphical convolutional networks and transformer based encodings," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2020, pp. 183–190.

[12] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[13] S. Dadkhah, F. Shoeleh, M. M. Yadollahi, X. Zhang, and A. A. Ghorbani, "A real-time hostile activities analyses and detection system," *Appl. Soft Comput.*, vol. 104, Jun. 2021, Art. no. 107175.

[14] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5644–5651.

[15] E. Okoro, B. Abara, A. Umagba, A. Ajonye, and Z. Isa, "A hybrid approach to fake news detection on social media," *Nigerian J. Technol.*, vol. 37, no. 2, pp. 454–462, 2018.

[16] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, "Fake news detection in social networks via crowd signals," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 517–524.

[17] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, 2019, pp. 312–320.

[18] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal fake news detection with textual, visual and semantic information," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, 2020, pp. 30–38.

[19] P. Meel and D. K. Vishwakarma, "Machine learned classifiers for trustworthiness assessment of web information contents," in *Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS)*, Feb. 2021, pp. 29–35.

[20] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 849–857.

[21] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, 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.

[22] 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.

[23] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Advances in Knowledge Discovery and Data Mining*, vol. 12085. Cham, Switzerland: Springer, 2020, pp. 354–367.

[24] V. K. Singh, I. Ghosh, and D. Sonagara, "Detecting fake news stories via multimodal analysis," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 1, pp. 3–17, Jan. 2021.

[25] P. Meel and D. K. Vishwakarma, "Deep neural architecture for veracity analysis of multimodal online information," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2021, pp. 7–12.

[26] D. Jakus, "Visual communication in public relations campaigns," *Marketing Sci. Res. Organizations*, vol. 27, no. 1, pp. 25–36, 2018.

[27] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manage.*, vol. 58, no. 5, Sep. 2021, Art. no. 102610.

[28] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 795–816.

[29] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 153–162.

[30] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," *Inf. Sci.*, vol. 567, pp. 23–41, Aug. 2021.

[31] D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cognit. Syst. Res.*, vol. 58, pp. 217–229, Dec. 2019.

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

[33] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.

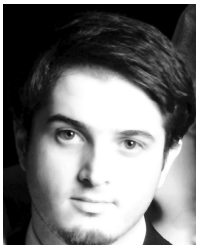
[36] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, "Discriminability objective for training descriptive captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6964–6974.

[37] A. Mousa and B. Schuller, "Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics:*, vol. 1, 2017, pp. 1023–1032.

[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
[40] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2422–2431.
[41] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2407–2415.
[42] S. Ding, S. Qu, Y. Xi, A. K. Sangaia, and S. Wan, "Image caption generation with high-level image features," *Pattern Recognit. Lett.*, vol. 123, pp. 89–95, May 2019.
[43] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102178.
[44] X. Zhang, S. Dadkhah, S. Mahdavifar, R. Lu, and A. A. Ghorbani, "An entity matching-based image topic verification framework for online fact-checking," *Int. J. Multimedia Intell. Secur.*, vol. 4, no. 1, pp. 65–85, 2022.
[45] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[47] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[48] M. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. F. Jones, "The benchmarking initiative for multimedia evaluation: MediaEval 2016," *IEEE MultimediaMag.*, vol. 24, no. 1, pp. 93–96, Jan. 2017.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[50] D. Quoc Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," 2020, *arXiv:2005.10200*.
[51] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[52] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.
[53] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.

Xichen Zhang received the B.E. degree from the Changsha University of Science and Technology, Changsha, China, in 2010, and the M.S. degree in computer science from the Canadian Institute for Cybersecurity (CIC), Faculty of Computer Science (FCS), University of New Brunswick (UNB), Fredericton, NB, Canada, in 2018. He is currently pursuing the Ph.D. degree with FCS, UNB.

After his studies, he worked as a Research Assistant at CIC, UNB. His research interests are data mining in cybersecurity, privacy enhancing technologies, and the Internet of Things (IoT)-big data security and privacy.



Sajjad Dadkhah is currently a Research Associate, a Cybersecurity Team Leader, and a Faculty Member at the Canadian Institute of Cybersecurity (CIC), University of New Brunswick (UNB), Fredericton, NB, Canada. He has over ten years of experience in digital multimedia security, computer security, and machine learning-based detection systems. He has been involved in several security projects as a team leader, a researcher, and a security consultant in different organizations, such as Kyushu University, Fukuoka, Japan; Universiti Malaya (UM), Kuala Lumpur, Malaysia; IRIS Smart Technology Complex, Kuala Lumpur, and the Kyushu Institute of Technology, Kitakyushu, Japan. In September 2016, he was awarded a fellowship by the Kyushu Institute of Technology to continue his research for two years.

Mr. Dadkhah has been the Managing Editor and a Board Member of *Applied Soft Computing* (ASOC) Elsevier journal since 2016.

Alexander Gerald Weismann is currently pursuing the bachelor's degree in computer science and media arts and culture (concurrent) with the University of New Brunswick, Fredericton, NB, Canada.

He worked as a Security Software Developer at the Canadian Institute for Cybersecurity, University of New Brunswick. His research areas focus on neural network development, Twitter dataset creation, and fake news detection with multimodal analyses.

Mohammad Amin Kanaani received the bachelor's degree from the Department of Computer Engineering, University of Guilan, Rasht, Iran, in September 2015. He is currently pursuing the M.Sc. degree with the University of New Brunswick, Fredericton, NB, Canada.

He is currently an Natural Language Processing (NPL) Researcher at the Canadian Institute for Cybersecurity, University of New Brunswick. His research interests include NPL, fake news detection, and explainable artificial intelligence.

Ali A. Ghorbani (Senior Member, IEEE) has held a variety of academic positions for the past 39 years. He is currently a Professor of Computer Science, the Tier 1 Canada Research Chair in Cybersecurity, and the Director of the Canadian Institute for Cybersecurity, University of New Brunswick, Fredericton, NB, Canada, which he established in 2016. He was the Dean of the Faculty of Computer Science, University of New Brunswick, from 2008 to 2017. He is also the Founding Director of the Laboratory for Intelligence and Adaptive Systems Research. He has spent over 29 years of his 39-year academic career, carrying out fundamental and applied research in machine learning, cybersecurity, and critical infrastructure protection. He is the co-inventor on three awarded and one filed patent in the fields of cybersecurity and web intelligence and has published over 280 peer-reviewed articles during his career. He has supervised over 190 research associates, post-doctoral fellows, and students during his career. He has authored the book, *Intrusion Detection and Prevention Systems: Concepts and Techniques* (Springer, October 2010). He developed several technologies adopted by high-tech companies and co-founded three startups, Sentrant Security, EyesOver Technologies, and Cydarien Security, Fredericton, Canada, in 2013, 2015, and 2019, respectively.

Dr. Ghorbani was a recipient of the 2017 Startup Canada Senior Entrepreneur Award and the Canadian Immigrant Magazine's RBC top 25 Canadian immigrants of 2019. He is the Co-Founder of the Privacy, Security, Trust (PST) Network in Canada and its annual international conference and served as the Co-Editor-in-Chief for *Computational Intelligence: An International Journal* from 2007 to 2017.