SAFE: Similarity-Aware Multi-modal Fake News Detection

Xinyi Zhou$^{(\boxtimes)}$, Jindi Wu, and Reza Zafarani

Data Lab, EECS Department, Syracuse University, Syracuse, NY 13244, USA
{zhouxinyi,reza}@data.syr.edu, jwu172@syr.edu

Abstract. Effective detection of fake news has recently attracted significant attention. Current studies have made significant contributions to predicting fake news with less focus on exploiting the relationship (similarity) between the textual and visual information in news articles. Attaching importance to such similarity helps identify fake news stories that, for example, attempt to use irrelevant images to attract readers' attention. In this work, we propose a Similarity-Aware FakE news detection method (SAFE) which investigates multi-modal (textual and visual) information of news articles. First, neural networks are adopted to separately extract textual and visual features for news representation. We further investigate the relationship between the extracted features across modalities. Such representations of news textual and visual information along with their relationship are jointly learned and used to predict fake news. The proposed method facilitates recognizing the falsity of news articles based on their text, images, or their "mismatches." We conduct extensive experiments on large-scale real-world data, which demonstrate the effectiveness of the proposed method.

Keywords: Fake news · Multi-modal analysis · Neural networks · Representation learning

1 Introduction

Following the 2016 U.S. presidential election, the impact of "fake news" has become a major concern. Based on a broad investigation of ∼126,000 verified true and fake news stories on Twitter from 2006 to 2017, Vosoughi and colleagues revealed that fake news stories spread more frequently and faster compared to true news stories [20]. As indicated by the fundamental theories on fake news in psychology and social sciences (see a comprehensive survey in Ref. [27]), the more a fake news article spreads, the higher the possibility of social media users spreading and trusting it due to repeated exposure and/or peer pressure. Such levels of trust and beliefs can easily be amplified and reinforced within social media due to its *echo chamber effect* [3]. Hence, extensive research has been conducted on effective detection of fake news to block its dissemination on social

---

X. Zhou and J. Wu—-Authors are equally contributed.

media. Fake news detection methods can be generally grouped into (1) content-based and (2) social-context-based methods. The main difference between the two types of methods is whether or not they rely on *social context* information: the information on how the news has propagated on social media, where abundant auxiliary information of social media users involved and their connections/networks can be utilized. Many innovative and significant solutions (e.g., [1,13,15]) have been proposed to exploit social context information. With more social context information available, one can often better detect fake news; however, detection becomes more challenging depending on the stage the news is currently at. It is difficult to detect fake news using social-context-based methods when it has been just published and has not been propagated (i.e., no social context information), which motivates us to further explore the role that news content can play in fake news detection.

As "a news article that is intentionally and verifiably false" [25], fake news content often contains textual and visual information. Existing content-based fake news detection methods either solely consider textual information [26], or combine both types of data ignoring the relationship (similarity) between them [4,5,23,24]. The values in understanding such relationship (similarity) for predicting fake news are two-fold. To attract public attention, some fake news stories (or news stories with low-credibility) prefer to use dramatic, humorous (facetious), and tempting images whose content is far from the actual content within the news text. Furthermore, when a fake news article tells a story with fictional scenarios or statements, it is difficult to find both pertinent and non-manipulated images to match these fictions; hence a "gap" exists between the textual and visual information of fake news when creators use non-manipulated images to support non-factual scenarios or statements.[1]

With such considerations, we propose a Similarity-Aware FakE news detection method (SAFE). The method consists of three modules, performing (1) multi-modal (textual and visual) feature extraction; (2) within-modal (or say, modal-independent) fake news prediction; (3) cross-modal similarity extraction, respectively. For each news article, we first adopt neural networks to automatically obtain the latent representation of both its textual and visual information, based on which a similarity measure is defined between them. Then, such representations of news textual and visual information with their similarity are jointly learned and used to predict fake news. The proposed method aims to recognize the falsity of a news article on either its text or images, or the "mismatch" between the text and images.

The main contributions of our work are summarized as below.

1. To our best knowledge, we present the first approach that investigates the role of the relationship (similarity) between news textual and visual information in predicting fake news;
2. We propose a new method to jointly exploit multi-modal (textual and visual) and relational information to learn the representation of news articles and predict fake news; and

---

[1] Examples at https://www.snopes.com/fact-check/rating/miscaptioned/.

3. We conduct extensive experiments on large-scale real-world data to demonstrate the effectiveness of the proposed method.

Next, we will first review the related work in Sect. 2. The proposed method will be detailed in Sect. 3, along with its iterative learning process in Sect. 4. We will detail the experiments and the results in Sect. 5. We will conclude in Sect. 6.

2   Related Work

There has been extensive research on fake news detection. Fake news detection methods can be generally grouped into (I) content-based and (II) social-context-based methods.

I. Content-Based Fake News Detection. Content-based methods detect fake news by utilizing news content, i.e., the *textual information* and/or *visual information* within news content.

Most content-based methods have comprehensively investigated news textual information. Within a traditional statistical natural language processing framework, such investigation has crossed multiple levels of language. By assuming that fake news differs from true news in linguistic/writing styles in the content, various hand-crafted features have been extracted from news content for representation and used for classification by, e.g., SVM and random forest. For example, Pérez-Rosas et al. employed lexical features by using bag-of-words with n-gram models, semantic features relying on LIWC [10], syntactic features such as context-free grammars, and news readability [11]. Instead of extracting features based on experience, Zhou et al. [26] validated the role of fundamental theories in psychology and social science in guiding fake news feature engineering. Rhetorical structures among sentences or phrases within news content have also been investigated with either a vector space model [14] or Bi-LSTM [6]. Researchers have also explored the political bias [12] and homogeneity [2] of news publishers by mining news content that they have published, and have demonstrated how such information can help detect fake news.

In addition to textual information, greater – while still limited – attention has been recently paid to visual information within news content. Jin et al. analyzed images between true news and fake news in terms of, e.g., their clarity [5]. Along with the recent advances in deep learning, various RNNs and CNNs have been developed for multi-modal fake news detection and related tasks [4,7,18,21,23,24]. To learn the multi-modal (textual and visual) representation of news content, Jin et al. developed VGG-19 and LSTM with an attention mechanism [4], and Khattar et al. designed an encoder-decoder mechanism [7]. Yang et al. proposed TI-CNN, which detects fake news by extracting both explicit and latent multi-modal features within news content [24]. Wang et al. proposed Event Adversarial Neural Network (EANN) to learn event-invariant features representative of news content across various topics and domains [23]. While current techniques have
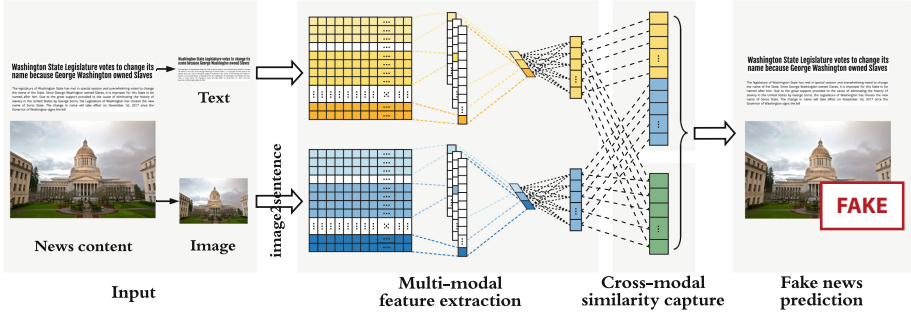
Fig. 1. Overview of the SAFE framework

facilitated the development of multi-modal fake news detection, the relationship across modalities has been barely explored and exploited. Our work bridges this gap by directly capturing the relationship (similarity) between the textual and visual information within news content, and firstly learning the representation of news articles through mining its multi-modal information and the relationship across modalities.

II. Social-Context-Based Fake News Detection. Social-context-based methods detect fake news by investigating social-context information related to news articles, i.e., how news articles spread on social media. Significant contributions have been made on identifying the differences in propagation patterns between fake news and the truth [20]. Such contributions have also focused on how user profiles [1] and opinions [13,15] can help news verification using feature engineering [1] and neural networks [13,15]. Nevertheless, verifying a news article that has been published online, e.g., on a news outlet such as BuzzFeed (https://www.buzzfeed.com/), before it has been disseminated on social media demands content-based methods as social-context information at this stage does not exist. For this purpose, we focus on mining news content in this work, where the proposed method will be detailed next.

3 Methodology

In this section, the proposed method (SAFE) is detailed in terms of its three modules performing: (I) multi-modal feature extraction (Sect. 3.1), (II) modal-independent fake news prediction (Sect. 3.2), and (III) cross-modal similarity extraction (Sect. 3.3). Then, we detail in Sect. 3.4 how various modules can work collectively to predict fake news. An overview of the SAFE framework is presented in Fig. 1. Before further specification, we formally define the problem and introduce some key notations as follows.

Problem Definition and Key Notation. Given a news article $A = \{T, V\}$ consisting of textual information T and visual information V, we denote $\mathbf{t} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$ as the corresponding representations, where $\mathbf{t} = \mathcal{M}_t(T, \theta_t)$ and $\mathbf{v} = \mathcal{M}_v(V, \theta_v)$. Let $s = \mathcal{M}_s(\mathbf{t}, \mathbf{v})$ denote the similarity between $\mathbf{t}$ and $\mathbf{v}$, where $s \in [0, 1]$. Our goal is to predict whether A is a fake news article ($\hat{y} = 1$) or a true one ($\hat{y} = 0$) by investigating its textual information, visual information, and their relationship, i.e., to determine $\mathcal{M}_p : (\mathcal{M}_t, \mathcal{M}_v, \mathcal{M}_s) \xrightarrow{(\theta_t, \theta_v, \theta_p)} \hat{y} \in \{0, 1\}$, where θ_* are parameters to be learned.

3.1   Multi-modal Feature Extraction

The multi-modal feature extraction module of SAFE aims to represent the (I) textual information and (II) visual information of a given news article in d-dimensional space, respectively.

Text. We extend Text-CNN [8] by introducing an additional fully connected layer to automatically extract textual features for each news article. The architecture of Text-CNN is provided in Fig. 2, which contains a convolutional layer and max pooling. Given a piece of content with n words, each word is first embedded as $\mathbf{x}_t^l \in \mathbb{R}^k, l = 1, 2, \cdots, n$ [9]. The convolutional layer is used to produce a *feature map*, denoted as $C_t = \{c_t^i\}_{i=1}^{n-h+1}$, from a sequence of local inputs $\{\mathbf{x}_t^{i:(i+h-1)}\}_{i=1}^{n-h+1}$, via a *filter* $\mathbf{w}_t$. As shown in Fig. 2, each local input is a group of h continuous words. Mathematically,

$$c_t^i = \sigma(\mathbf{w}_t \cdot \mathbf{x}_t^{i:(i+h-1)} + b_t), \tag{1}$$

$$\mathbf{x}_{i:(i+h-1)} = \mathbf{x}_i \oplus \mathbf{x}_{i+1} \oplus \cdots \oplus \mathbf{x}_{i+h-1}, \tag{2}$$

where $\mathbf{w}_t, \mathbf{x}_t^{i:(i+h-1)} \in \mathbb{R}^{hk}$, $b_t \in \mathbb{R}$ is a *bias*, $\oplus$ is the concatenation operator, and σ is ReLU function. Note that $\mathbf{w}_t$ and b_t are all parameters within Text-CNN to be learned. Then, a max-over-time pooling operation is applied on the obtained feature map for dimension reduction, i.e., $\hat{c}_t = \max\{c_t^i\}_{i=1}^{n-h+1}$. Finally, the representation of the news text can be obtained by $\mathbf{t} = \mathbf{W}_t \hat{\mathbf{c}}_t + \mathbf{b}_t$, where $\hat{\mathbf{c}}_t \in \mathbb{R}^g$, g is the different number of window sizes chosen; $\mathbf{W}_t \in \mathbb{R}^{d \times g}$ and $\mathbf{b}_t \in \mathbb{R}^d$ are parameters to be learned.

Image. For representing news images, we also use Text-CNN with an additional fully connected layer while we first process visual information within news content using a pre-trained `image2sentence` model[2] [19]. Compared to existing multi-modal fake news detection studies that often directly apply a pre-trained CNN (e.g., VGG) model to obtain the representation of news images [4,23], we adopt the aforementioned processing strategy for consistency and to increase

---

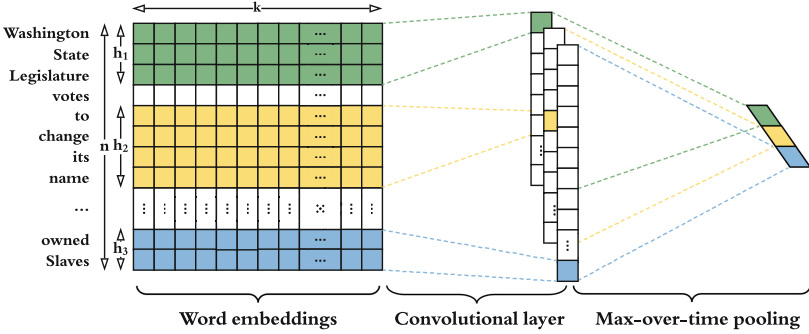[2] https://github.com/nikhilmaram/Show_and_Tell.

Fig. 2. Text-CNN architecture

insights when computing the similarity across modalities. As we will demonstrate later in our experiments, it also leads to performance improvements. Let $\hat{\mathbf{c}}_v$ denote the output of the neural network with parameters $\mathbf{w}_v$ (filter) and b_v (bias). Similarly, the final representation of news visual information is then computed by $\mathbf{v} = \mathbf{W}_v \hat{\mathbf{c}}_v + \mathbf{b}_v$, where $\mathbf{W}_v$ and $\mathbf{b}_v$ are parameters to be learned.

3.2 Modal-Independent Fake News Prediction

To properly represent news textual and visual information in predicting fake news, we aim to correctly map the extracted textual and visual features of news content to their possibilities of being fake, and further to their actual labels. Mathematically, such possibilities can be computed by

$$\mathcal{M}_p(\mathbf{t}, \mathbf{v}) = \mathbf{1} \cdot \text{softmax}(\mathbf{W}_p(\mathbf{t} \oplus \mathbf{v}) + \mathbf{b}_p), \tag{3}$$

where $\mathbf{1} = [1,0]^\top$, $\oplus$ is the concatenation operator, $\mathbf{W}_p \in \mathbb{R}^{2 \times 2d}$ and $\mathbf{b}_p \in \mathbb{R}^2$ are parameters. To let the computed possibilities of news articles being fake approach their actual labels, a cross-entropy-based loss function is defined:

$$\mathcal{L}_p(\theta_t, \theta_v, \theta_p) = -\mathbb{E}_{(a,y) \sim (A,Y)}(y \log \mathcal{M}_p(\mathbf{t}, \mathbf{v}) + (1-y) \log(1 - \mathcal{M}_p(\mathbf{t}, \mathbf{v}))), \tag{4}$$

where $\theta_p = \{\mathbf{W}_p, \mathbf{b}_p\}$, $\theta_t = \{\mathbf{W}_t, \mathbf{b}_t, \mathbf{w}_t, b_t\}$, $\theta_v = \{\mathbf{W}_v, \mathbf{b}_v, \mathbf{w}_v, b_v\}$, and

$$(\hat{\theta}_t, \hat{\theta}_v, \hat{\theta}_p) = \arg \min_{\theta_t, \theta_v, \theta_p} \mathcal{L}_p(\theta_t, \theta_v, \theta_p). \tag{5}$$

3.3 Cross-Modal Similarity Extraction

When attempting to correctly map the multi-modal features of news articles to their labels, features belonging to two different modals are considered separately – concatenating them with no relation between them explored (see Sect. 3.2). However, besides that, the falsity of a news article can be also detected by assessing how (ir)relevant the textual information is compared to its visual

information; fake news creators sometimes actively use irrelevant images for false statements to attract readers' attention, or passively use them due to the difficulty in finding a supportive non-manipulated image (see case studies in Sect. 5 for examples). Compared to news articles delivering relevant textual and visual information, those with disparate statements and images are more likely to be fake. We define the relevance between news textual and visual information as follows by slightly modifying cosine similarity:

$$\mathcal{M}_s(\mathbf{t}, \mathbf{v}) = \frac{\mathbf{t} \cdot \mathbf{v} + \|\mathbf{t}\| \|\mathbf{v}\|}{2 \|\mathbf{t}\| \|\mathbf{v}\|} \tag{6}$$

In such a way, it is guaranteed that $\mathcal{M}_s(\mathbf{t}, \mathbf{v})$ is positive and $\in [0,1]$ (to be utilized in Eq. (7)); 0 indicates that $\mathbf{t}$ and $\mathbf{v}$ are far from being similar, while 1 indicates that $\mathbf{t}$ and $\mathbf{v}$ are exactly the same.

Then, we can define the loss function based on cross-entropy as below, which assumes that news articles formed with mismatched textual and visual information are more likely to be fake compared to those with matching textual statements and images, when analyzing from a pure similarity perspective:

$$\mathcal{L}_s(\theta_t, \theta_v) = -\mathbb{E}_{(a,y) \sim (A,Y)}(y \log(1 - \mathcal{M}_s(\mathbf{t}, \mathbf{v})) + (1 - y) \log \mathcal{M}_s(\mathbf{t}, \mathbf{v})), \tag{7}$$

$$(\hat{\theta}_t, \hat{\theta}_v) = \arg \min_{\theta_t, \theta_v} \mathcal{L}_s(\theta_t, \theta_v). \tag{8}$$

3.4 Model Integration and Joint Learning

When detecting fake news, we aim to correctly recognize fake news stories whose falsity is in their (1) textual and/or visual information, or (2) their relationship, as specified in Sect. 3.2 and Sect. 3.3, respectively. To involve both cases, we specify our final loss function as

$$\mathcal{L}(\theta_t, \theta_v, \theta_p) = \alpha \mathcal{L}_p(\theta_t, \theta_v, \theta_p) + \beta \mathcal{L}_s(\theta_t, \theta_v), \tag{9}$$

where parameters can be jointly learned by

$$(\hat{\theta}_t, \hat{\theta}_v, \hat{\theta}_p) = \arg \min_{\theta_t, \theta_v, \theta_p} \mathcal{L}(\theta_t, \theta_v, \theta_p). \tag{10}$$

4   Optimization

We outline the optimization process to learn the model parameters, i.e., iteratively solving Eq. (10). The process is summarized in Algorithm 1. The updating rule for each parameter is as follows:

Update θ_p. Let γ be the learning rate, the partial derivative of $\mathcal{L}$ w.r.t. θ_p is:

$$\theta_p \leftarrow \theta_p - \gamma \cdot \alpha \frac{\partial \mathcal{L}_p}{\partial \theta_p}. \tag{11}$$

Algorithm 1: SAFE

Input: $A = \{(T_j, V_j)\}_{j=1}^m$, $Y = \{y_j\}_{j=1}^m$, $H = \{h_k\}_{k=1}^g$, γ

Output: $\theta_p = \{\mathbf{W}_p, \mathbf{b}_p\}$, $\theta_t = \{\mathbf{W}_t, \mathbf{b}_t, \mathbf{w}_t, b_t\}$, $\theta_v = \{\mathbf{W}_v, \mathbf{b}_v, \mathbf{w}_v, b_v\}$

1 Randomly initialize $\mathbf{W}_p, \mathbf{b}_p, \mathbf{W}_t, \mathbf{b}_t, \mathbf{w}_t, b_t, \mathbf{W}_v, \mathbf{b}_v, \mathbf{w}_v, b_v$;

2 while *not convergence* **do**

3   **foreach** (T_j, V_j) **do**

4   Update θ_p: $\{\mathbf{W}_p, \mathbf{b}_p\} \leftarrow$ Eq. (12);

5   **foreach** h_k **do**

6   Update θ_t: $\{\mathbf{W}_t, \mathbf{b}_t, \mathbf{w}_t, b_t\} \leftarrow$ Eqs. (14-18);

7   Update θ_v: similar to updating θ_t;

8   **end**

9   **end**

10 end

11 return $\mathbf{W}_p, \mathbf{b}_p, \mathbf{W}_t, \mathbf{b}_t, \mathbf{w}_t, b_t, \mathbf{W}_v, \mathbf{b}_v, \mathbf{w}_v, b_v$

As $\theta_p = \{\mathbf{W}_p, \mathbf{b}_p\}$, updating θ_p is equivalent to updating both $\mathbf{W}_p$ and $\mathbf{b}_p$ in each iteration, which respectively follow the following rules:

$$\mathbf{W}_p \leftarrow \mathbf{W}_p - \gamma \cdot \alpha \Delta \mathbf{y} (\mathbf{t} \oplus \mathbf{v})^\top, \quad \mathbf{b}_p \leftarrow \mathbf{b}_p - \gamma \cdot \alpha \Delta \mathbf{y}, \tag{12}$$

where $\Delta \mathbf{y} = [\hat{y} - y, y - \hat{y}]^\top$.

Update θ_t. The partial derivative of $\mathcal{L}$ w.r.t. θ_t is generally computed by

$$\theta_t \leftarrow \theta_t - \gamma (\alpha \frac{\partial \mathcal{L}_p}{\partial \mathcal{M}_t} \frac{\partial \mathcal{M}_t}{\partial \theta_t} + \beta \frac{\partial \mathcal{L}_s}{\partial \mathcal{M}_t} \frac{\partial \mathcal{M}_t}{\partial \theta_t}). \tag{13}$$

Let $\nabla \mathcal{L}_*(\mathbf{t}) = \frac{\partial \mathcal{L}_*}{\partial \mathcal{M}_t}$, $\mathbf{t}_0 = \frac{\mathbf{t}}{||\mathbf{t}||}$, $\mathbf{v}_0 = \frac{\mathbf{v}}{||\mathbf{v}||}$, and $\mathbf{W}_{p,L}$ denote the first d columns of $\mathbf{W}_p$, we can have

$$\nabla \mathcal{L}_p(\mathbf{t}) = \mathbf{W}_{p,L}^\top \Delta \mathbf{y}, \tag{14}$$

$$\nabla \mathcal{L}_s(\mathbf{t}) = \frac{1 - y}{2s \, ||\mathbf{t}||} ((2s - 1)\mathbf{t}_0 - \mathbf{v}_0), \tag{15}$$

based on which the parameters in θ_t are respectively updated as follows:

$$\mathbf{W}_t \leftarrow \mathbf{W}_t - \gamma \cdot \mathbf{D}_t \mathbf{B}_t, \quad \mathbf{b}_t \leftarrow \mathbf{b}_t - \gamma \cdot \mathbf{B}_t, \tag{16}$$

$$\mathbf{w}_t \leftarrow \mathbf{w}_t - \gamma \cdot \mathbf{x}_t^{\hat{i}:(\hat{i}+h-1)} \mathbf{W}_t^\top \mathbf{B}_t, \quad b_t \leftarrow b_t - \gamma \cdot \mathbf{W}_t^\top \mathbf{B}_t, \tag{17}$$

where $\hat{i} = \arg \max_i \{c_t^i\}_{i=1}^{n-h+1}$, $\mathbf{D}_t \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entry value $c_t^{\hat{i}}$, and

$$\mathbf{B}_t = \alpha \nabla \mathcal{L}_p(\mathbf{t}) + \beta \nabla \mathcal{L}_s(\mathbf{t}). \tag{18}$$

Update θ_v. It is similar to updating θ_t; we omit details due to space constraints.

5   Experiments

We detail experimental setup in Sect. 5.1, followed by evaluating SAFE in Sect. 5.2.

5.1   Experimental Setup

We detail (I) the data used in our experiments, (II) the baselines SAFE is compared to, and (III) implementation details such as how data was pre-processed and SAFE hyper-parameters were set.

Datasets. Our experiments are conducted on two well-established public benchmark datasets of fake news detection[3] [16]. News articles in datasets are respectively collected from PolitiFact and GossipCop. PolitiFact (https://www.politifact.com/) is a well-known non-profit fact-checking website of political statements and reports in the U.S. [22]. GossipCop (https://www.gossipcop.com/) is a website that fact-checks celebrity reports and entertainment stories published in magazines and newspapers. News articles in PolitiFact dataset were published from May 2002 to July 2018 and those in GossipCop dataset were published from July 2000 to December 2018. Ground truth labels (*fake* or *true*) of news articles in both datasets were provided by domain experts, which guarantees the quality of news labels. Statistics of the two datasets are provided in Table 1.

Baselines. We compare to the following baselines, which detect fake news using (i) textual (LIWC [10]), (ii) visual (VGG-19 [17]), or (iii) multi-modal information (att-RNN [4]).

- **LIWC** [10]: LIWC is a widely-accepted psycho-linguistics lexicon. Given a news story, LIWC can count the words in the text falling into one or more of over 80 linguistic, psychological, and topical categories. These numbers act as hand-crafted features used by, e.g., random forest, to predict fake news;
- **VGG-19**[4] [17]: VGG-19 is a widely-used CNN with 19 layers for image classification. We use a fine-tuned VGG-19 as one of the baselines; and
- **att-RNN** [4]: att-RNN is a deep neural network model applicable for multimodal fake news detection. It employs LSTM and VGG-19 with attention mechanism to fuse textual, visual and social-context features of news articles. We set the hyper-parameters the same as that in [4] and exclude the social-context features for a fair comparison.

Table 1. Data statistics

	PolitiFact			GossipCop		
	Fake	True	Overall	Fake	True	Overall
# News articles	432	624	1,056	5,323	16,817	22,140
– with textual information	420	528	948	4,947	16,694	21,641
– with visual information	336	447	783	1,650	16,767	18,417

We also include the following variants of the proposed SAFE method:

– **SAFE\T**: The proposed SAFE method without using textual information;
– **SAFE\V**: The proposed SAFE method without using visual information;
– **SAFE\S**: SAFE without capturing the relationship (similarity) between news textual and visual information. In this case, the extracted multi-modal features of each news article are fused by concatenating them; and
– **SAFE\W**: The proposed method when only the relationship between textual and visual information is assessed. In this case, the classifier is directly connected with the output of the cross-modal similarity extraction module, i.e., $\hat{y} \leftarrow \mathrm{softmax}(\mathbf{W}[\mathcal{M}_s, 1 - \mathcal{M}_s]^\top + \mathbf{b})$, where $\mathbf{W}$ and $\mathbf{b}$ are parameters.

Implementation Details. In our experiments, each dataset was separated into 80% for training and 20% for testing based on the publication dates of news articles, where newly published articles were treated as test data. five-fold cross-validation was used for model training. We set the learning rate as 10^{-4}, the number of iterations as 100, and the strides (H) as $\{3, 4\}$.
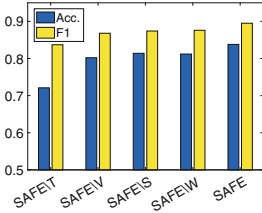
5.2 Performance Analysis

We evaluate the general performance of SAFE by comparing it with (I) state-of-the-art fake news detection methods and (II) its variants. Next, (III) parameters within SAFE are analyzed and (IV) case studies are presented to validate its effectiveness. We use accuracy, precision, recall, and F_1 score to evaluate how well the representation and prediction perform.

General Performance Analysis. The general performance of SAFE and baselines are provided in Table 2. Results indicate when predicting fake news, SAFE can outperform all baselines based on the accuracy values and F_1 scores for both datasets. Based on PolitiFact data, the general performance of methods is SAFE > att-RNN ≈ LIWC > VGG-19; while for GossipCop data, such performance is SAFE > VGG-19 > att-RNN > LIWC. Note that multiple supervised learners (such as SVM, decision tree, logistic regression, and k-NN) have been used with LIWC in our experiments, where we present the best performance (obtained from random forest) in Table 2.

Table 2. Performance of methods in detecting fake news

|		LIWC[a]	VGG-19[b]	att-RNN[c]	SAFE\T[b]	SAFE\V[a]	SAFE\S[c]	SAFE\W[c]	SAFE[c]
PolitiFact	Acc.	0.822	0.649	0.769	0.674	0.721	0.796	0.738	**0.874**
	Pre.	0.785	0.668	0.735	0.680	0.740	0.826	0.752	**0.889**
	Rec.	0.846	0.787	**0.942**	0.873	0.831	0.801	0.844	0.903
	F_1	0.815	0.720	0.826	0.761	0.782	0.813	0.795	**0.896**
GossipCop	Acc.	0.836	0.775	0.743	0.721	0.802	0.814	0.812	**0.838**
	Pre.	**0.878**	0.775	0.788	0.734	0.853	0.875	0.853	0.857
	Rec.	0.317	0.970	0.913	**0.974**	0.883	0.872	0.901	0.937
	F_1	0.466	0.862	0.846	0.837	0.868	0.874	0.876	**0.895**

a: Text-based methods
b: Image-based methods
c: Multi-modal methods
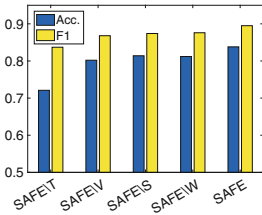
(a) PolitiFact
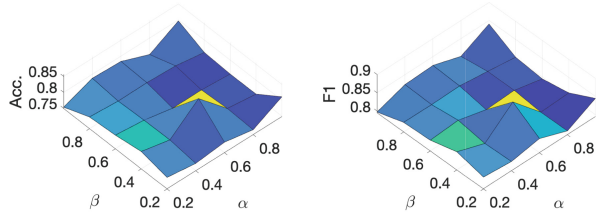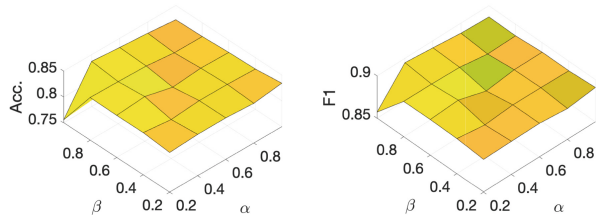
(b) GossipCop

Fig. 3. Module analysis

(a) PolitiFact

(b) GossipCop

Fig. 4. Parameter analysis

Module Analysis. The performance of SAFE and its variants are presented in Table 2 and Fig. 3. Results indicate when predicting fake news, (1) integrating news textual information, visual information, and their relationship (SAFE) performs best among all variants, (2) using multi-modal information (SAFE\S or SAFE\W) performs better compared to using single-modal information (SAFE\T or SAFE\V); (3) it is comparable to detect fake news by either independently using multi-modal information (SAFE\S) or mining their relationship (SAFE\W); and (4) textual information (SAFE\V) is more important compared to visual information (SAFE\T).

Parameter Analysis. In Eq. (9), α and β are used to allocate the relative importance between the extracted multi-modal features (α) and the similarity across modalities (β). To assess their influence in method performance, we

(a) $s = 0.024$ (b) $s = 0.044$ (c) $s = 0.001$

Fig. 5. Fake news

(a) $s = 0.966$ (b) $s = 0.975$ (c) $s = 0.983$

Fig. 6. True news

changed the value of α and β respectively from 0 to 1 with a step size of 0.2. Results in Fig. 4 show that various parameter values lead to the accuracy (or F_1 score) of SAFE ranging from 0.75 to 0.85 (or from 0.8 to 0.9) for both datasets. The proposed method performs best when $\alpha : \beta = 0.4 : 0.6$ in PolitiFact and $\alpha : \beta = 0.6 : 0.4$ in GossipCop, which again validates the importance of both multi-modal information and cross-modal relationship in predicting fake news.

Case Study. In our case studies, we aim to answer the following questions: is there any real-world fake news story whose textual and visual information are not closely related to each other? If there is, can SAFE correctly recognize such irrelevance and further recognize its falsity? For this purpose, we went through the news articles in the two datasets, and compared their ground truth labels with their similarity scores computed by SAFE. Several examples are presented in Figs. 5–6. It can be observed that (**I**) the gap between textual and visual information exist for some fictitious stories for (but not limited to) two reasons. First, such stories are difficult to be supported by non-manipulated images. An example is in Fig. 5a, where no voting- and bill-related image is actually available. Compared to the couples having a real intimate relationship (see Fig. 5c), the fake ones often have rare group photos or use collages (see Fig. 5c). Second, using "attractive" though not closely relevant images can help increase the news traffic. For example, the fake news in Fig. 5b includes an image with a smiling individual that conflicts with the death story. (**II**) SAFE helps correctly assess

the relationship (similarity) between news textual and visual information. For fake news stories in Fig. 5, their corresponding similarity scores are all low and SAFE correctly labels them as fake news. Similarly, SAFE assigns all true news stories in Fig. 6 a high similarity score, and predicts them as true news.

6   Conclusion

In this work, a similarity-aware multi-modal method, named SAFE, is proposed to predict fake news. The method extracts both textual and visual features of news content, and investigates their relationship. Experimental results indicate multi-modal features and the cross-modal relationship (similarity) are valuable with a comparable importance in fake news detection. Case studies conducted further validate the effectiveness of the proposed method in assessing such similarity and predicting fake news. Nevertheless, we should point out the proposed method investigates textual and visual information without considering, e.g., network and video information. Additionally, relationships within modalities are valuable as well such as the textual (or visual) similarity among or between pairwise news articles, which both will be part of our future work.

References

1. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: The World Wide Web Conference, pp. 675–684. ACM (2011)
2. Horne, B.D., Nørregaard, J., Adalı, S.: Different spirals of sameness: a study of content sharing in mainstream and alternative media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 257–266 (2019)
3. Jamieson, K.H., Cappella, J.N.: Echo Chamber: Rush Limbaugh and the Conservative Media Establishment. Oxford University Press, Oxford (2008)
4. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 795–816. ACM (2017)
5. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE Trans. Multimedia **19**(3), 598–608 (2017)
6. Karimi, H., Tang, J.: Learning hierarchical discourse-level structure for fake news detection. arXiv preprint arXiv:1903.07389 (2019)
7. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: MVAE: multimodal variational autoencoder for fake news detection. In: WWW, pp. 2915–2921. ACM (2019)
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Technical report (2015)
11. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. arXiv preprint arXiv:1708.07104 (2017)

12. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)
13. Qian, F., Gong, C., Sharma, K., Liu, Y.: Neural user response generator: fake news detection with collective user intelligence. In: IJCAI, pp. 3834–3840 (2018)
14. Rubin, V.L., Lukoianova, T.: Truth and deception at the rhetorical structure level. J. Assoc. Inf. Sci. Technol. **66**(5), 905–917 (2015)
15. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of the ACM on Conference on Information and Knowledge Management, pp. 797–806. ACM (2017)
16. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: a data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Truong, Q.T., Lauw, H.: Multimodal review generation for recommender systems. In: The World Wide Web Conference, pp. 1864–1874. ACM (2019)
19. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 652–663 (2016)
20. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
21. Wang, H., Sahoo, D., Liu, C., Lim, E.-P., Hoi, S.C.: Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In: CVPR, pp. 11572–11581 (2019)
22. Wang, W.Y.: "Liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
23. Wang, Y., et al.: EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849–857. ACM (2018)
24. Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: TI-CNN: convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749 (2018)
25. Zafarani, R., Zhou, X., Shu, K., Liu, H.: Fake news research: theories, detection strategies, and open problems. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3207–3208. ACM (2019)
26. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: a theory-driven model. arXiv preprint arXiv:1904.11679 (2019)
27. Zhou, X., Zafarani, R.: Fake news: a survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315 (2018)