Category-controlled Encoder-Decoder for Fake News Detection

Lianwei Wu, Yuan Rao, Cong Zhang, Yongqiang Zhao, and Ambreen Nazir

Abstract—The existing data-driven approaches typically capture credibility-indicative representations from relevant articles for fake news detection, such as skeptical and conflicting opinions. However, these methods still have several drawbacks: 1) Due to the difficulty of collecting fake news, the capacity of the existing datasets is relatively small; and 2) there is considerable unverified news that lacks conflicting voices in relevant articles, which makes it difficult for the existing methods to identify their credibility. Especially, the differences between true and fake news are not limited to whether there are conflict features in their relevant articles, but also include more extensive hidden differences at the linguistic level, such as the perspectives of emotional expression (like extreme emotion in fake news), writing style (like the shocking title in clickbait), etc., the existing methods are difficult to fully capture these differences. To capture more general and wide-ranging differences between true and fake news, in this paper, directly from the different categories of news itself, we propose a Category-controlled Encoder-Decoder model (CED) to generate examples with category-differentiated features and extend the dataset capacity to achieve data enhancement effect, thus enhancing fake news detection. Specifically, to make the generated examples enrich more news features, we develop news-guided encoder to guide relevant articles to generate news-semantic context representations. To drive the generated examples to contain more category-differentiated features, we devise category-controlled decoder which relies on pattern-shared unit to respectively capture intra-category shared features within true or fake news, and employs restriction unit to force the two types of shared features to be more different for highlighting inter-category differentiated features. The experimental results on three datasets demonstrate the superiority of CED.

Index Terms—Encoder Decoder, Fake News Detection, Social Media Analysis, Natural Language Processing

1 Introduction

As the megaphone in new era, social media with its unique equality and concealment allows everyone to become a participant and commentator for news, to convey information freely. Nevertheless, not only does the information contain true and reliable information but also plenty of fake news. Investigations illustrate that, during the US presidential election (2016), a tweeting rate for users tweeting links to websites contains news classified as fake over four times larger than that of traditional media, sometimes even adding to 25% of tweets in Twitter spread fake news [1]. More than that, 1% of users are exposed to 80% of fake news [2]. The widespread of fake news has exerted an unprecedented negative impact on personal life, social stability, and political pattern. Therefore, how to detect fake news has become one of the significant challenges of social media.

Currently, most existing studies mainly focus on capturing text features [3], [4], [5] and meta-data features [6], [7] relying on deep neural networks for fake news detection, which have achieved great success. Especially, due to most posts in social media are based on short text and lack of sufficient semantics, their relevant comments or relevant articles are widely exploited and confirmed to be powerful credibility indicators, where relevant articles represent a series of articles or comments that discuss a specific news story. In detail, Ma *et al.* [8] considered structures and semantics of both posts and comments and proposed tree-structured models based on recursive neural networks to learn representations for rumor detection. Shu *et al.* [9] developed co-attention networks to exploit both posts and comments to jointly capture top-k check-worthy sentences for fake news detection. More recently, several methods concentrate on discovering conflicting features between news and relevant articles, including differential [10], doubtful [11], disapproving [12], and refutatory [13] features for fake news detection. As concrete examples, Popat *et al.* [10] attempted to debunk false claims by constructing attention-based interaction model to learn differential and conflicting words from relevant articles. Wu *et al.* [12] proposed adaptive interaction fusion networks that fulfill cross-interaction fusion between claims and relevant articles to capture their semantic conflicts and disagreeing features for detection. They also developed evidence-aware hierarchical interactive attention networks [14] to capture the credibility semantics discussing the questionable parts of claims from relevant articles as conflict semantic fragments for claim verification.

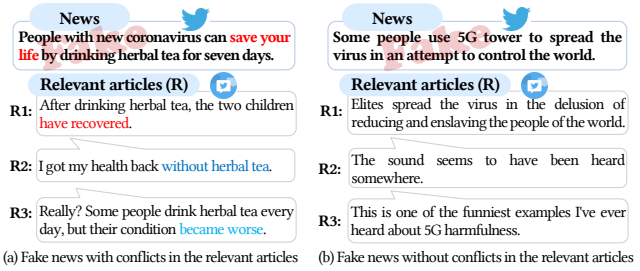


Fig. 1. The intuitive description of conflict features.

The conflicting features could be intuitively illustrated by a typical case. As shown in Figure 1(a), 'became worse' in the

- *L. Wu, Y. Rao, Y. Zhao, and A. Nazir were with Xi'an Key Lab. of Social Intelligence and Complexity Data Processing, the School of Software Engineering, Xi'an Jiaotong University; Shaanxi Joint Key Laboratory for Artifact Intelligence(Sub-Lab of Xi'an Jiaotong University), Xi'an, Shaanxi 710054, China*
  E-mail: stayhungry@stu.xjtu.edu.cn, raoyuan@mail.xjtu.edu.cn, {yongqiang1210,ambreen.nazir}@stu.xjtu.edu.cn
- *C. Zhang was with Xi'an Huanyu Satellite Control and Data Application Co. Ltd. Xi'an 710065, China. E-mail: congzhang0825@gmail.com*

relevant article 3 (R3) has conflict semantics with 'save your life' in the news and 'have recovered' in the relevant article 1 (R1). The models capturing such effective conflicting features between news and related articles or between relevant articles could achieve remarkable performance. However, these models still have several general limitations. **First**, due to the difficulty of collecting fake news, the capacity of the existing public datasets is relatively small (like PHEME dataset only contains 2,246 items, which is described in detail in Section 4.1), and the current models capturing conflict features can not expand the capacity of datasets. **Second**, when extra relevant articles are leveraged to detect fake news, the existing models lack feature filtering, which are easy to introduce noise features unrelated to news, thus interfering with the performance of the models. **Third**, there is considerable unverified news that lacks conflicting voices in relevant articles, which makes it difficult for the existing methods to identify their credibility. Particularly, the significant differences between true news and fake news are not only limited to whether there are conflict features in their relevant articles, but also include more extensive hidden differences at the linguistic level, such as the perspectives of emotional expression (like extreme emotions in fake news) [15], [16], writing styles (like shocking titles in clickbait) [17], [18], [19], [20], etc., while the existing conflict-based models are unable to explore the difference features between true and fake news from such broader perspectives. Therefore, how to design an effective model that directly touches the more wide-ranging and universal differences between true and fake news is the key to improve the performance of fake news detection.

To address the above issues, instead of elaborately capturing the conflicts features, we endeavor to explore the wide-ranging differences between the categories of news that are true and fake directly. Based on this idea, in this paper, we propose **C**ategory-controlled **E**ncoder-**D**ecoder model (henceforth, CED) to generate examples with differential features between categories and expand the capacity of the dataset to achieve data enhancement, thus improving fake news detection. The examples are generated by collecting inter-category differentiated features from the two types of intra-category shared features, where intra-category shared features denote the common features within each category of news (i.e., true or fake news), and inter-category differentiated features refer to the distinguishing features between true and fake news. Specifically, in order to capture valuable news-related features from related articles, CED devises news-guided encoder module, which relies on true and fake news to guide their relevant articles to generate true and fake examples with news characteristics, respectively. To explore more extensive hidden differences between true and fake news, CED model constructs category-controlled decoder module, which aims to consolidate the differentiation between true and fake news, which designs two units, i.e., pattern-shared unit and restriction unit. The former unit strengthens the capture of intra-category shared features within each category of news (true news with its relevant articles or fake news with its relevant articles), respectively, and the restriction unit forces both types of shared features to be more different, to highlight inter-category differentiated features. Finally, we combine the generated true and fake examples with the original news data to form an enhanced dataset for improving detection capability. Experiments on three public datasets, i.e., Snopes, PolitiFact, and PHEME, show the effectiveness of CED. The main contributions of the paper are as follows:

- A new perspective on fake news detection is explored, which considers the screening of shared features within one category (true or fake news) and the capture of individuation features between categories (true and fake news) to generate examples with differential features for strengthening fake news detection.
- The news-guided encoder module is developed, which relies on semantic matching and fusion between news and relevant articles to make the generated examples possess more valuable news semantics (Section 4.4.1 and 4.5.1).
- The pattern-shared unit with a gate filtering mechanism is able to capture intra-category shared features within true or fake news that contain multi-perspective credibility-indicative features (Section 4.4.2 and 4.5.3), and the proposed restriction unit is capable of constraining the independence and difference of the two types of shared features to underline inter-category differentiated features (Section 4.4.3).
- Our approach on three real-world datasets achieves superior improvements over state-of-the-art baselines (Section 4.3). Additionally, the combination of the generated examples by our model with the original data can significantly improve the performance of baseline methods (Section 4.4.6).

The remaining parts of this paper are organized as follows. We review some related works in Section 2. Section 3 presents the architecture of CED and explains the design of each module in detail. Experimental results and discussion are described in Section 4, and finally, Section 5 concludes our work and outlines directions for future work.

2 Related Work

Fake news detection is a long standing research topic. Readers could refer to [15] for a recent survey. Instead of the previous fake news detection methods by manually constructing features, we focus on automatic fake news detection with the help of the text generation model. Therefore, our work is related to two groups of tasks: automatic fake news detection and natural language generation.

2.1 Automatic Fake News Detection

The methods for automatic fake news detection are divided into content-based and metadata-based. Content-based methods are mainly extracted from the following aspects: grammar [21], semantics [22], emotions [23], styles [3], [5], and stances [24], [25]. For instance, Chen *et al.* [26] combined attention mechanism with Recurrent Neural Networks (RNN) to focus on text features with different attentions for capturing the hidden representation. Wu *et al.* [25] designed a sifted multi-task learning method with a selected sharing layer to filter and select feature flows between the tasks of rumour detection and stance detection. Although this kind of approach is effective, the capture of feature diversity is limited.

Metadata-based methods focus on extracting features surrounding sources [27], posts [28], [29], comments [9], users [30], [31], and propagation networks [8] for fake news detection. Concretely, the rapid development of social media has given every anonymous user the opportunity to become a publisher of information, which has greatly increased the measurement complexity of the credibility of information sources, making it

difficult to obtain better performance for the credibility of source-based methods. User-based and network-based methods generally need to construct user profiles and propagation networks, respectively, which are labor-intensive because it is as complex as the methods for constructing features manually. Due to the rich credibility features (such as questionable, argumentation, or support voices) in comments (or articles) related to news, the comment-based methods concentrate on capturing conflicting and questionable semantics from relevant comments or articles to detect fake news. As concrete examples, Ma *et al.* [8] built a tree-structured Recursive Neural Networks (RvNN) to catch the hidden representation from both propagation structures (from comments) and text contents. They also proposed a novel end-to-end hierarchical attention network [13] focusing on learning to represent coherent semantics as well as their semantic relatedness with the claim from relevant articles for fake news detection. This type of methods effectively acquires indicative credibility features, and has become one of the focuses of fake news detection currently, but it has trouble combating some news that lacks questionable voices in their relevant comments or articles. Based on this, instead of intentionally finding doubtful voices from specific news, from the perspective of the categories of news, we make great efforts to explore the differences between true and fake news from their relevant articles for enhancing fake news detection.

2.2 Natural Language Generation

The principal purpose for the task of natural language generation is to automatically generate a piece of high-quality natural language text with the aid of existing knowledge, which has been widely applied in many fields, such as neural machine translation [32], abstractive summarization [33], creative writing [34], and even fake news detection [35]. The methods [36] for the task include template-based, structural-based [37], encoder-decoder based, etc., where encoder-decoder methods [38], [39] have achieved promising performance on this task. Specifically, CopyRNN [40] first regards the generation process as a sequence-to-sequence learning task and applies an extensively available encoder-decoder framework [41] with attention [42] and copy mechanisms [43]. Based on CopyRNN, various extensions [44] are recently proposed. In addition, many studies introduce encoder-decoder generation models [35], [45] to the task of fake news detection. As a concrete example, Ma *et al.* [35] proposed a GAN-style encoder-decoder model to generate uncertain or conflicting voices to pressurize the discriminator to learn stronger rumor indicative representations for fake news detection. Nevertheless, these methods only focus on one perspective of difference features between true and fake news, i.e., conflict features, they ignore the differences based on broader perspectives, like the perspectives of emotion expression and writing styles, etc. Therefore, how to capture the wide-ranging and general difference features between true and fake news is a critical issue for fake news detection. Base on this, in this work, we develop a category-controlled generative model to collect shared features within the category, and the differential features between the categories for fake news detection.

3 Category-controlled Encoder-Decoder Model

In this section, we propose a category-controlled encoder-decoder model (CED) that generates examples with the differentiated features of news category for fake news detection. Our framework is shown in Figure 2, which consists of news-guided encoder module and category-controlled decoder module. The former relies on true and fake news to guide respectively their relevant articles for producing context representation rich in true and fake news features, and the latter aims to control decoder to generate examples with category-differentiated features (i.e., the differentiated features between true and fake news). Particularly, category-controlled decoder module designs pattern-shared unit to respectively capture the shared features within the category (i.e., true news with its relevant articles or fake news with its relevant articles) and develops restriction unit to force the two types of shared features to be more different, to obtain inter-category differentiated features.

3.1 News-guided Encoder Module

As shown in Figure 3, news-guided encoder module is composed of a sequence encoding layer, a guided matching layer, and a fusion merging layer. First, the sequence encoding layer reads the content of the relevant article and news, and learns their contextual representations separately. Then, the guided matching layer gathers the news information for each word in relevant articles reflecting the important parts of the context of this word. Finally, the fusion merging layer merges the aggregated news semantics into each word of the relevant article to produce the final news-guided context representation. Especially, true news-guided encoder and fake news-guided encoder generate representations in the same way. The following detailed process applies to both true and fake news.

Sequence Encoding Layer We employ a bidirectional gated recurrent unit (Bi-GRU) to encode relevant article and news sequences. For any sequence X, each word $w_i \in X$ is first embedded into a D-dimensional embedding vector $v_i \in \mathbb{R}^D$. Particularly, we adopt X_a, X_c to denote the embeddings of the relevant article and the news, respectively. Then each word is mapped into forward and backward hidden states (denoted as $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$) with the following defined operations:

$$\overrightarrow{\mathbf{h}}_i = \mathrm{GRU}(v_i, \mathbf{h}_{i-1}) \tag{1}$$

$$\overleftarrow{\mathbf{h}}_i = \mathrm{GRU}(v_i, \mathbf{h}_{i+1}) \tag{2}$$

The concatenation of $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$, $[\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$, serves as w_i's hidden state in encoder, denoted as $\mathbf{h}_i \in \mathbb{R}^d$. $\mathbf{h}_i^a$ and $\mathbf{h}_j^c$ serve as the contextual vectors for the i-th word of the relevant article and the j-th word of the news, respectively.

Guided Matching Layer In the matching layer, considering that there may be information in the relevant article that reveals the truth of fake news, which leads to conflicting semantics between the relevant article and the news. To remain the different relationships between different types of news and comments, we design gated mechanism aiming different types of news, i.e.,screening the conflict semantics between fake news and relevant articles, and filtering noisy semantics between true news and relevant articles. Additionally, to reflect the importance of information in the relevant article based on the fact that the highly news-related information in the relevant article should contain more core information, we engage news-guided attention mechanism to aggregate the relevant screened semantics from the news for

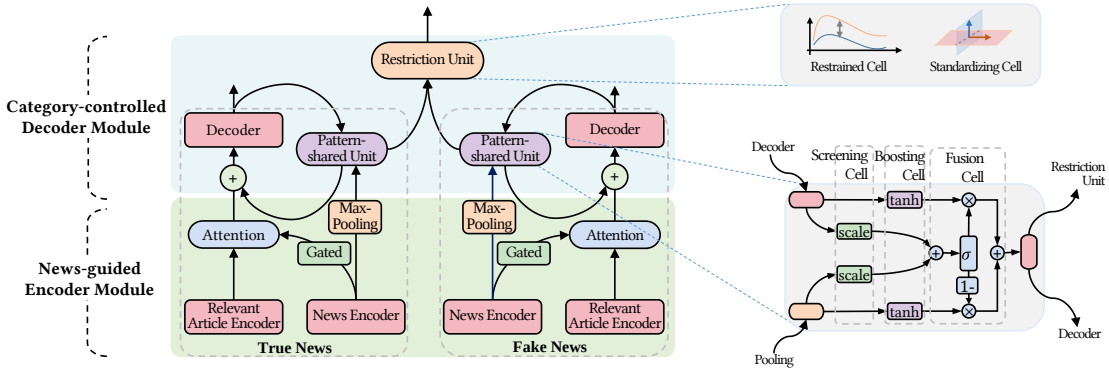


Fig. 2. Overall architecture of CED. The model consists of two modules: news-guided encoder module (the details are shown in Figure 3) and category-controlled decoder module including a pattern-shared unit and a restriction unit except decoder.

each word of the relevant article. Formally, the aggregation process of the i-th word (at step i) in the relevant article $\mathrm{att}_i = \mathrm{attn}(\mathbf{h}_i^a, [\mathbf{h}_1^c, \mathbf{h}_2^c, ..., \mathbf{h}_{|c|}^c]; \mathbf{W}_{att})$ is represented as:

$$\mathrm{att}_i = \sum_{j=1}^{|c|} \alpha_{i,j} \mathbf{h}_{g_j}^c \tag{3}$$

$$\mathbf{h}_{g_j}^c = \sigma(\mathbf{W}_g \mathbf{h}_i^a + \mathbf{b}_g) \odot \mathbf{h}_j^c \tag{4}$$

$$\alpha_{i,j} = \exp(z_{i,j}) / \sum_{k=1}^{|c|} \exp(z_{i,k}) \tag{5}$$

$$z_{i,j} = (\mathbf{h}_i^a)^T \mathbf{W}_{att} \mathbf{h}_j^c \tag{6}$$

where $\mathbf{W}_g$, $\mathbf{W}_{att}$, and $\mathbf{b}_g$ are trainable parameters. $\sigma(\cdot)$ is sigmoid activation function, and $\alpha_{i,j}$ is the normalized attention score between $\mathbf{h}_i^a$ and $\mathbf{h}_i^c$.

Thus, through these two structures, the guided matching layer can not only capture the parts highly related to the news from the relevant article, but also earn the conflicting semantics between the related article and the news.

Fusion Merging Layer Finally, the original vector $\mathbf{h}_i^a$ of the relevant article and the aggregated information vector att_i act as the inputs to information fusion merging layer:

$$\overrightarrow{\mathbf{m}}_i = \mathrm{GRU}([\mathbf{h}_i^a; \mathrm{att}_i], \overrightarrow{\mathbf{m}}_{i-1}) \tag{7}$$

$$\overleftarrow{\mathbf{m}}_i = \mathrm{GRU}([\mathbf{h}_i^a; \mathrm{att}_i], \overleftarrow{\mathbf{m}}_{i+1}) \tag{8}$$

$$\tilde{\mathbf{m}}_i = \lambda \mathbf{h}_i^a + (1-\lambda)[\overrightarrow{\mathbf{m}}_i; \overleftarrow{\mathbf{m}}_i] \tag{9}$$

where $[\mathbf{h}_i^a; \mathrm{att}_i] \in \mathbb{R}^{2d}$, $\overrightarrow{\mathbf{m}}_i \in \mathbb{R}^{d/2}$, $[\overrightarrow{\mathbf{m}}_i; \overleftarrow{\mathbf{m}}_i] \in \mathbb{R}^d$, and $\tilde{\mathbf{m}} \in \mathbb{R}^d$. $\mathbf{h}_i^a$ in Eq. (9) is a residual connection, and $\lambda \in (0,1)$ is the hyperparameter. Eventually, we obtain the news-guided contextual representation of the context (i.e., $\tilde{\mathbf{m}} = [\tilde{\mathbf{m}}_1, \tilde{\mathbf{m}}_2, ..., \tilde{\mathbf{m}}_{|a|}]$), where $|a|$ denotes the sequence length of the relevant article, which is regarded as a repository for later decoding. Particularly, $\tilde{\mathbf{m}}_{true}$ and $\tilde{\mathbf{m}}_{false}$ denote true news-guided contextual representation and fake news-guided contextual representation, respectively.

3.2 Category-controlled Decoder Module

To drive the decoder to generate true and fake examples with more category-differentiated features, we first design pattern-shared unit to capture the shared features within the category (true news with its relevant articles or fake news with its relevant articles). Simultaneously, we explore the restriction unit to constraint the two types of shared features to make them more different, to gain inter-category differentiated features. At this time, the controlled decoder will generate more examples with differentiated features between true and fake news.

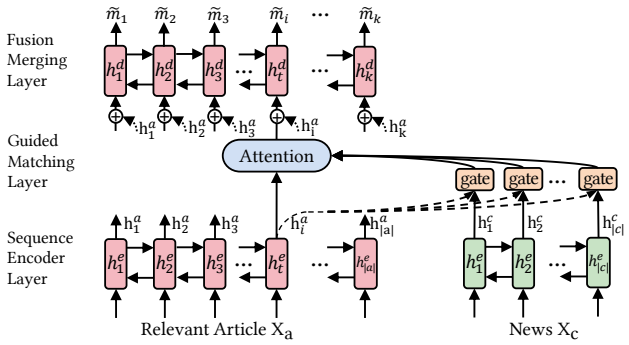


Fig. 3. The snapshot of the news-guided encoder module at step i.

3.2.1 Pattern-shared Unit

To acquire the shared features within true or fake news, respectively, we design a pattern-shared unit consisting of screening cell, filtering cell, and fusion cell, which chooses features from the generated example and the news. Particularly, the pattern-shared unit for true news and that for fake news are the same. Take the generated example at time step t as an example, the details of pattern-shared unit are described as:

Screening Cell We adopt affine transformation to screen news semantics and generated semantics for discovering eigen-invariant vectors.

$$S_e(e_{t-1}) = \mathbf{W}_{se} e_{t-1} + \mathbf{b}_{se} \tag{10}$$

$$S_c(\mathbf{E}_c^{pool}) = \mathbf{W}_{sc} \mathbf{E}_c^{pool} + \mathbf{b}_{sc} \tag{11}$$

where $\mathbf{E}_c^{pool}$ denotes the max-pooling vector of news at encoding layer. $e_{t-1} \in R^D$ is the embeddings of the $(t-1)$-th predicted word in decoder (in subsection 3.2.2). $\mathbf{W}_{se} \in \mathbb{R}^{D \times d}$, $\mathbf{b}_{se} \in \mathbb{R}^d$, $\mathbf{W}_{sc} \in \mathbb{R}^{2d \times d}$, and $\mathbf{b}_{se} \in \mathbb{R}^d$ are trainable parameters.

Boosting Cell To enhance the non-linear ability of semantic features, we exploit the activation function - $\tanh$ to map news semantics and generated semantics.

$$t_e(e_{t-1}) = \tanh(\mathbf{W}_e e_{t-1} + \mathbf{b}_e) \tag{12}$$

$$t_c(\mathbf{E}_c^{pool}) = \tanh(\mathbf{W}_c \mathbf{E}_c^{pool} + \mathbf{b}_c) \tag{13}$$

where $\mathbf{W}_e \in \mathbb{R}^{D\times d}$, $\mathbf{b}_e \in \mathbb{R}^d$, $\mathbf{W}_c \in \mathbb{R}^{2d\times d}$, and $\mathbf{b}_c \in \mathbb{R}^d$ are trainable parameters.

Fusion Cell To capture the shared features within one category (i.e., true or fake news) from the news semantics and generated semantics, we build fusion cell to integrate the eigen-invariant semantics and the non-linear semantics.

$$\alpha = \sigma(\mathbf{W}_{ce}[S_e(e_{t-1}); S_c(\mathbf{E}_c^{pool})] + \mathbf{b}_{ce}) \tag{14}$$

$$\mathbf{F}_t = \alpha t_e(e_{t-1}) + (1-\alpha)t_c(\mathbf{E}_c^{pool}) \tag{15}$$

where $\mathbf{W}_{ce} \in \mathbb{R}^{2d\times d}$, $\mathbf{b}_{ce} \in \mathbb{R}^d$ are trainable parameters. Specially, $\mathbf{F}_t^{true}$ (or $\mathbf{F}_t^{false}$) are the shared features within true news (or fake news) with its relevant articles.

3.2.2 Decoder

After learning the news-guided contextual representation, we engage an attention-based GRU decoder to generate a word sequence as the example. In order to make true and fake examples generated by the decoder rich in category features, we rely on the intra-category features learned from the pattern-shared unit to control the generation of the decoder. Concretely, when generating the t-th word in the example, the decoder emits a hidden state vector $\mathbf{h}_t \in \mathbb{R}^d$ and puts a global attention over $\tilde{\mathbf{m}}$. The attention desires to secure indicative representations from $\tilde{\mathbf{m}}$ and integrates them into a context vector $\tilde{\mathbf{h}}_t$ defined as:

$$\mathbf{h}_t = \mathrm{GRU}([e_{t-1}; \mathbf{F}_t; \tilde{\mathbf{h}}_{t-1}], \mathbf{h}_{t-1}) \tag{16}$$

$$\mathbf{c}_t = \mathrm{attn}(\mathbf{h}_t, \tilde{\mathbf{m}}, \mathbf{W}_1) \tag{17}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_2[\mathbf{c}_t; \mathbf{h}_t]) \tag{18}$$

where $t = 1, 2, ..., L_y$, and L_y is the length of the generated example. $e_{t-1} \in R^D$ is the embeddings of the $(t-1)$-th predicted word wherein e_0 is the embeddings of the start token. $\mathbf{F}_t \in \mathbb{R}^d$ is the outputs of pattern-shared unit at time step t. On the true news side, $\mathbf{F}_t = \mathbf{F}_t^{true}$, and on the fake news side, $\mathbf{F}_t = \mathbf{F}_t^{false}$. $\mathbf{c}_t \in \mathbb{R}^d$ is the attentional vector, and $\tilde{\mathbf{h}}_t \in \mathbb{R}^d$ is the attentional vector at time step t.

The predicted probability distribution over the predefined vocabulary V for current step is computed by:

$$Pr(y_t|y_{<t}, X_a, X_c, \mathbf{F}) = \mathrm{softmax}(\mathbf{W}\tilde{\mathbf{h}}_t + \mathbf{b}) \tag{19}$$

which reflects how likely a word to be the t-th word in the generated example. Here, $y_{<t}$ refers to $(y_1, y_2, ..., y_{t-1})$. $\mathbf{W} \in \mathbb{R}^{V\times d}$ and $\mathbf{b} \in \mathbb{R}^V$ are trainable weights. In particular, <SOS> and <EOS> are applied in the generated sequence to enable the decoder to understand the beginning and end of a sentence.

3.2.3 Restriction Unit

To make the generated true and fake examples by decoder contain more category-differentiated features, i.e., exploring the more differential features in true and fake news (inter-category features), we design restriction unit to restrain the two types of shared features captured from the true or fake news. In detail, restriction unit consists of two cells, i.e., restrained cell and standardizing cell.

Restrained Cell Restrained cell encourages the two types of shared features to be as different as possible.

$$\mathcal{L}_{simi} = \sum_i^d \mathbf{F}_{t_i}^{true} \log\left(\frac{\mathbf{F}_{t_i}^{true}}{\mathbf{F}_{t_i}^{false}}\right) \tag{20}$$

$$\mathcal{L}_{diff} = 1/\mathcal{L}_{simi} \tag{21}$$

where d denotes the vector length of both $\mathbf{F}_t^{true}$ and $\mathbf{F}_t^{false}$.

Standardizing Cell Standardizing cell alleviates the correlation between the two types of shared features and ensures their independence. Concretely:

$$\mathcal{L}_{rc} = ||\mathbf{F}_t^{true} - \mathbf{F}_t^{false}||_F^2 \tag{22}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm [46].

All loss of the restriction unit could be integrated as:

$$\mathcal{L}_r = \alpha_1 \mathcal{L}_{rc} + (1-\alpha_1)\mathcal{L}_{diff} \tag{23}$$

where α_1 is the hyper-parameter.

3.3 Training

During the training stage, we apply stochastic gradient descent to minimize the loss function of our model:

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{t=1}^{L_y} \log((Pr(y_t|y_{<t}, X_a, X_c, \mathbf{F}; \Theta)) + \mathcal{L}_r) \tag{24}$$

where N is the number of training instances, and Θ means the set of all learnable parameters. Besides, we use beam search with beam width 1 as the decoding algorithm during testing.

3.4 Fake News Detection

After generation and optimization, generated examples obtain available differentiated features between news categories. To enhance detection ability, we combine generated examples with original data to form a new augmented dataset, i.e., $\mathrm{A} = \{\{\mathrm{G}\} \cup \{\mathrm{P}\}\}$. In more detail, $\mathrm{A} = \{\{\mathrm{G_F}\} \cup \{\mathrm{P_F}\} \cup \{\mathrm{G_T}\} \cup \{\mathrm{P_T}\}\}$, where $\mathrm{G_F}$ and $\mathrm{G_T}$ represent generated fake and true examples, respectively, and $\mathrm{P_F}$ and $\mathrm{P_T}$ represent all original fake and true news, respectively. Especially, the number of generated examples is the same as the number of original news items, that is, we generate a new example for each original news item.

Next, we utilize the augmented dataset A to verify news by multi-head self-attention networks [47] which not only capture global dependencies of the whole sequence but also learn inner structure features of the sequence. We briefly express it:

$$\mathbf{E} = \mathrm{self\text{–}attention}([\mathrm{P_i}; \mathrm{G_i}], \theta_{att}) \tag{25}$$

$$p = \mathrm{softmax}(\mathbf{W}_\mathrm{p}\mathbf{E} + \mathbf{b}_\mathrm{p}) \tag{26}$$

where $[\mathrm{P_i}; \mathrm{G_i}]$ represents the concatenation of a piece of news $\mathrm{P_i}$ and a generated example $\mathrm{G_i}$ in one item in A. $\mathbf{E}$ is the credibility vector learned by self-attention networks. θ_{att}, $\mathbf{W}_\mathrm{p}$, and $\mathbf{b}_\mathrm{p}$ are all trainable parameters.

Finally, we train the networks to minimize cross-entropy error for a single instance with ground-truth label y:

$$\mathcal{L}_{task} = -\sum y \log p \tag{27}$$

4 Experiments

In this section, to demonstrate the effectiveness of CED extensively, we first describe three public benchmark datasets (i.e., Snopes, PolitiFact, and PHEME) and experimental settings. We then systematically evaluate the performance of CED on these datasets around the following perspectives: the performance of the model, model ablation, assessing the utility of the generated examples on different baseline

methods, the effect of the quantity of generated examples, and the training process. Next, we visualize the features learned by the modules of CED for a more transparent understanding to end-users. Finally, the limitations of CED are analyzed according to the above experimental results.

4.1 Datasets

We evaluate CED and demonstrate its generality by performing experiments on three public competitive datasets: Snopes, PolitiFact, and PHEME. Their details are shown as follows.

Snopes and PolitiFact are provided by Popat *et al.* [10], containing 4,341 and 3,568 pieces of news, along with 29,242 and 29,556 relevant articles collected from various web sources, respectively. For the labels of the two datasets: each news in Snopes is labeled as true and false, while each news in PolitiFact is originally divided into one of the following six veracity categories: true, mostly true, half true, mostly false, false, and pants on fire. To capture the different features of true and false news, we merge true, mostly true, and half true as true, and the other categories are treated as false.

PHEME [48] is leveraged for binary classification of true and fake with respect to a tweet (news) via its relevant tweets (comments). The dataset includes Twitter conversation threads associated with nine newsworthy events containing Charlie Hebdo, Ottawa shooting, etc. A conversation thread consists of a tweet and a series of comments. We filter out claims with less than 10 tweets and balance the number of instances of the two categories, i.e., 1,123 pieces of true news and 1,123 pieces of fake news (2,246 conversation threads).

We divide the datasets into training, validation, and testing subsets with the proportion of 70%, 10%, and 20%, respectively.

4.2 Settings

We strictly tune all hyper-parameters on the validation dataset and gain the best performance via a small grid search. The details of the tuned hyper-parameters are shown as follows: 1) The pre-trained BERT-base model [49] is used to initialize the sequence embeddings of news and relevant articles; 2) The dimension D is set to 768; 3) Bi-GRU is the single-layer and the hidden size of GRU is assigned as 200; 4) Since the length of each news and relevant article are different, we respectively perform zero padding here by setting a maximum length; 5) The parameter α_1 is finally trained as 0.65, 0.60, and 0.80 on Snopes, PolitiFact, and PHEME, respectively; 6) In self-attention networks, attention heads and blocks are set to 6 and 4, respectively; 7) The dropout of multi-head attention is set to 0.6; 8) L2-regularizers with fully connected layers as well as dropout is used for training; 9) All the models are trained using Adam optimizer [50] with a learning rate of 0.002 and mini-batches size of 128 to minimize categorical cross-entropy loss; and 10) We decay the learning rate into the half when the evaluation perplexity stops dropping. Early stopping is applied when the validation perplexity stops dropping for three continuous generation evaluations.

4.3 Performance Comparison

4.3.1 Results on Snopes and PolitiFact

We compare CED and the following state-of-the-art baselines on Snopes and PolitiFact:

SVM: A linear SVM adopts a collection of linguistic features handcrafted around news content for fake news detection [51].

CNN: A CNN model captures news semantics through different convolutional window sizes serving as n-grams for fake news detection [52]. Here, we only consider news content without meta-data features.

LSTM: LSTM models word sequences of news to learn and represent semantics for fact-checking [53].

DeClarE: Popat *et al.* [10] propose an evidence-aware assessment model to aggregate signals from relevant articles, the language of the articles, and the trustworthiness of their sources.

ADV-VIR: A data argumentation-based method [54] extends adversarial and virtual adversarial training to text classification by applying perturbations to the word embeddings in recurrent neural networks.

GAN-ED: Ma *et al.* [35] develop a GAN-style encoder-decoder model to produce uncertain or conflicting voices for pressurizing the discriminator to learn stronger rumor indicative presentations for rumor detection.

dEFEND: Sentence-comment co-attention networks [9] exploit both news contents and comments to jointly capture explainable top-k check-worthy sentences for detection.

HAN: A hierarchical attention network [13] focuses on learning to represent coherent evidence as well as their semantic relatedness with claim for claim verification.

On Snopes and PolitiFact, we employ 10% of the datasets for tuning the hyperparameters and conduct 10-fold cross-validation on the rest of the datasets. We resort to micro-/macro-averaged F1, class-specific F1-score as evaluation metrics. Additionally, we implement our model with Tensorflow[1]. Due to DeClarE is not open-source, we replicate it based on its network structures with Theano[2]. Other baselines are implemented through the source codes they published. As shown in Table 1, we observe that:

- Compared with SVM, CNN, and LSTM, DeClarE obtains the best performance, presenting at least 3.5% and 3.2% boost in micF1 on Snopes and PolitiFact, respectively, which indicates that DeClarE not only learns deep semantics from news content but also captures salient words from relevant articles to enhance representation of credibility features.
- The performance of ADV-VIR is obviously better than that of DeClarE, which expresses that data enhancement based on constraints could elevate the detection ability of models. GAN-ED wins more remarkable performance than ADV-VIR, which explains that relying on GAN to guide encoder-decoder model to generate conflicting features for heightening indicative representations could effectively improve performance. dEFEND achieves more eminent performance than GAN-ED, which reflects the utility of interaction between news and comments. Moreover, HAN is superior to dEFEND, which illustrates that it is effective for detection to consider coherent evidence and common features from relevant articles.
- CED consistently outperforms the other baseline methods, presenting 82.8% and 80.5% in micF1 on Snopes and PolitiFact, respectively, and its advantages could be expressed in the following two perspectives: 1) Our model is able to strengthen representations of credibility-indicative features through utilizing the news content to guide encoder-decoder model to capture salient words from relevant articles; and 2) CED

1. https://www.tensorflow.org
2. http://deeplearning.net/software/theano

TABLE 1
Performance Comparison on Snopes and PolitiFact

Methods	Snopes				PolitiFact			
	micF1	macF1	F1-score True	F1-score False	micF1	macF1	F1-score True	F1-score False
SVM	0.704	0.649	0.511	0.786	0.658	0.623	0.516	0.710
CNN	0.721	0.636	0.460	0.812	0.654	0.610	0.505	0.745
LSTM	0.689	0.642	0.517	0.771	0.673	0.629	0.527	0.753
DeClarE	0.756	0.690	0.550	0.831	0.705	0.686	0.542	0.775
ADV-VIR	0.783	0.710	0.556	0.845	0.733	0.714	0.571	0.796
GAN-ED	0.792	0.746	0.567	0.851	0.771	0.743	0.602	0.809
dEFEND	0.801	0.750	0.581	0.860	0.778	0.748	0.611	0.813
HAN	0.803	0.753	0.646	0.864	0.783	0.751	0.622	0.821
Ours	**0.828**	**0.778**	**0.667**	**0.887**	**0.805**	**0.774**	**0.656**	**0.842**

TABLE 2
Performance Comparison of CED Against the Baselines on PHEME

Methods	Accuracy	Precision		Recall		F1-score	
		False	True	False	True	False	True
DT-Rank	0.562	0.588	0.549	0.421	0.704	0.491	0.617
DTC	0.581	0.582	0.579	0.473	0.788	0.478	0.584
SVM-TS	0.651	0.663	0.642	0.617	0.786	0.639	0.663
BOW	0.704	0.724	0.687	0.675	0.734	0.699	0.710
CNN	0.665	0.671	0.661	0.652	0.679	0.661	0.669
GRU	0.742	0.737	0.754	0.753	0.730	0.745	0.739
CVM	0.767	0.760	0.782	0.787	0.752	0.768	0.764
SHG	0.774	0.765	0.782	0.787	0.752	0.776	0.767
GAN-ED	0.781	0.773	0.791	0.796	0.766	0.784	0.778
dEFEND	0.790	0.782	0.804	0.806	0.772	0.794	0.788
Ours	**0.803**	**0.795**	**0.814**	**0.819**	**0.788**	**0.807**	**0.801**

leverages the category-based features to guide encoder-decoder to generate category-indicative representations, highlighting the differences between true and fake news.

4.3.2 Results on PHEME

To further evaluate the effectiveness of CED, we conduct experiments on PHEME dataset. Particularly, due to our model in this paper does not take into account association relationships between relevant articles, so instead of utilizing the FEVER dataset that possesses associations between related articles, we exploit three datasets, i.e., Snopes, PolitiFact, and PHEME, which contain relatively independent comments/relevant articles, to evaluate our CED. We compare our CED with the following baselines on PHEME:

DT-Rank: Decision-Tree-based Ranking method [55] identifies rumors trending through finding entire clusters of posts whose topic is a disputed factual claim.

DTC: A Decision Tree Classifier model [56] utilizes a series of handcrafted features including message-based, user-based, topic-based, and propagation-based features from tweets for information credibility evaluation.

SVM-TS: A linear SVM classification model [57] captures the temporal characteristics of social context information based on the time series of the rumor's lifecycle for detection.

BOW: A naive baseline for fake news detection relies on Bag-Of-Words to obtain news text representation and construct detection classifier with linear SVM.

CNN: A CNN-based model [58] gains rumor representations by framing relevant tweets as fixed-length sequences.

GRU: An RNN-based model with GRU [59] learns representations of relevant tweets over time for fake news detection.

CVM: Conflicting Viewpoint Method [60] utilizes topic models to discover conflicting semantics and constructs a credibility propagation network of tweets linked with supporting or opposing relations for generating evaluation results.

SHG: Stylized Headline Generation [45] generates readable and realistic headlines to enlarge original training data for improving the classification capacity of clickbait detection.

GAN-ED and **dEFEND:** They have been introduced in subsection 4.3.1.

We make use of 10% of the tweets in PHEME for tuning the hyper-parameters, and the rest of the claims are conducted for 5-fold cross-validation. Additionally, we utilize accuracy, precision, recall, and F1-score as evaluation metrics.

As shown in Table 2, we gain the following observations:

- In all baselines, the first three baselines (i.e., DT-Rank, DTC, and SVM-TS) based on hand-crafted features perform clearly worse than the other baselines, reflecting ranging from 1.4% to 22.8% degradation in accuracy, which explains that the automatically extracted methods are able to discover more hidden credibility-indicative features than the hand-crafted methods.
- In automatically extracted methods, GRU achieves more superior performance than BOW and CNN on PHEME, which illustrates that GRU is capable of capturing complex hidden credibility-indicative features beyond explicit and shallow patterns. CVM outperforms GRU on PHEME, which demonstrates excavating the differentiated features between news and comments is conducive to improving the performance of fake news. GAN-ED and SHG respectively boost 0.7% and 1.4% performance in accuracy on PHEME compared with CVM, which indicates that the automatically extracted methods relying on generative model produce more credibility features for fake news detection.
- Our model consistently outperforms other baselines on PHEME, presenting at least 15.2% boost than hand-crafted methods and at least 1.3% boost in accuracy than automatically extracted methods, which confirms the effectiveness of CED relying on encoder-decoder generative model.
- According to Table 1, and 2, the performance of fake news is more remarkable than that of true news, while the performance of fake news on PHEME has no similar phenomenon. The reason may be that the relevant articles of fake news on Snopes and PolitiFact contain more credibility features of fake news, while the comments in PHEME include more irrelevant noise, like irrelevant banter and AD, which may interfere with the acquisition of credibility features of models.
- Compared with Section 4.3.1 and 4.3.2, our model employs different baseline models for different datasets. The reason is that some baseline models are only appropriate for a particular dataset, that is, some baseline models perform well on certain datasets and perform poorly on other datasets.

4.3.3 Quality Analysis of Model Generation on Three Datasets

To analyze the generated quality of CED, we adopt three evaluation metrics, including Perplexity, BLEU-1, and BLEU-3, to evaluate the generation performance of our model as well as the competitive generation models, i.e., **ADV-VIR**,

TABLE 3
Quality Analysis of CED Against Several Generated Baselines

Methods	Snopes			PolitiFact			PHEME		
	Perp.	B-1	B-3	Perp.	B-1	B-3	Perp.	B-1	B-3
ADV-VIR	35.24	19.27	2.34	38.23	18.78	1.97	38.50	13.04	1.64
SHG	30.49	21.43	2.89	33.84	20.39	2.42	35.98	14.90	2.02
GAN-ED	25.72	24.56	3.26	27.61	22.16	2.76	32.20	16.85	2.45
Ours	**22.45**	**26.12**	**3.68**	**25.26**	**23.44**	**3.12**	**30.42**	**18.57**	**2.78**

TABLE 4
Results of Ablation Test of Our CED on the Three Datasets

Methods	Snopes		PolitiFact		PHEME
	micF1	macF1	micF1	macF1	Accuracy
-News Enc.	0.804	0.760	0.783	0.758	0.786
-Gated	0.812	0.769	0.792	0.761	0.793
-Pattern	0.786	0.741	0.776	0.737	0.779
-Restriction	0.795	0.753	0.783	0.755	0.787
-BERT Emb.	0.811	0.765	0.793	0.766	0.794
-SelfAtt.	0.813	0.769	0.798	0.769	0.798
CED	0.828	0.778	0.805	0.774	0.803

SHG, and **GAN-ED**. Perplexity is the standard measure for evaluating language models, and BLEU [61] measures the ratios of the co-occurrences of n-grams between the generated example and the original relevant article.

Table 3 shows the performance of all the compared methods on the three datasets. We could learn that: First, among the three generated methods, SHG and GAN-ED perform better than ADV-VIR. The two models have similar network architectures, i.e., two encoder-decoder architectures, which confirms the effectiveness of the architecture. Simultaneously, we also adopt it to our model as the basic framework. Second, compared with SHG, GAN-ED utilizes GAN-style strategy to capture conflicting features from relevant articles and generate higher-quality text sequences, which indicates that there are abundant conflict or controversial semantics in news-related comments (or relevant articles). Third, our model outperforms all the baselines with a large margin, reflecting at least 0.42%, 0.36%, and 0.33% boost in BLEU-3 on the three datasets, respectively. A major reason is that the baselines are lack of constraints on the discriminator or decoding layer, while our model filters noise information and enhances valuable semantics with the help of pattern-shared unit in the decoder.

4.4 Discussion

4.4.1 Ablation Analysis

To evaluate the effectiveness of different components of CED, we ablate CED into the following simplified models: 1) **-News Enc.** denotes CED removes news encoder components from true and fake news modules, respectively; 2) **-Gated** means CED replaces gated mechanism with concatenation; 3) **-Pattern** means CED removes pattern shared units from true and fake news modules and adopt concatenation operation as the connection way; 4) **-Restriction** represents that CED removes restriction unit for fake news detection, 5) **-BERT Emb.** is that CED replaces BERT embeddings to word2vec embeddings as input embeddings, and 6) **-SelfAtt.** is that CED replaces self-attention networks to BiLSTM for fake news detection.

As shown in Table 4, we have the following observations:

- **Effectiveness of news encoder.** CED boosts at least 1.7% performance than -News Enc., which indicates that the effectiveness of news guiding encoder-decoder model to generate examples for fake news detection.
- **Effectiveness of gated unit.** Analysis of the results of -Gated and CED, CED obtains the superior performance relying on the gated unit, which illustrates the effectiveness of CED utilizing gate mechanism to filter news features.
- **Effectiveness of pattern-shared unit.** When compared with -Pattern, CED significantly improves performance with the help of pattern-shared unit, which explains the effectiveness of pattern-shared unit capturing the shared features within the same category.
- **Effectiveness of restriction unit.** By introducing restriction unit, CED boosts the performance as compared with -Restriction, which confirms that the effectiveness of CED designing restriction unit to acquire inter-category differentiated features between true and fake news.
- **Effectiveness of BERT embeddings.** The input embeddings of CED are replaced with word2vec embeddings by BERT embeddings, and the model achieves weaker performance, which proves the effectiveness of the model using BERT embeddings as input embeddings.
- **Effectiveness of the framework of CED.** By replacing self-attention networks with BiLSTM for fake news detection, CED reflects 1.5%(micF1), 0.7%(micF1), and 0.5%(accuracy) performance reduction on the three datasets, which illustrates the effectiveness of self-attention networks. Furthermore, compared with Table 1 and 2, CED with the aid of BiLSTM is superior to the latest baseline models (such as HAN and dEFEND), which conveys the effectiveness of our CED framework.

TABLE 5
Ablation Results of Pattern-shared Unit

Methods	Snopes		PolitiFact		PHEME
	micF1	macF1	micF1	macF1	Accuracy
-boosting	0.809	0.766	0.797	0.763	0.797
-screening	0.804	0.762	0.792	0.756	0.794
-fusion	0.798	0.758	0.785	0.745	0.787
-Pattern	0.786	0.741	0.776	0.737	0.779
CED	0.828	0.778	0.805	0.774	0.803

4.4.2 Pattern-shared Unit Evaluation

Table 5 provides the performance of each part of pattern-shared unit by the following simplified models: 1) **-boosting** means pattern-shared unit removing the boosting cell; 2) **-screening** implies that pattern-shared unit removing the screening cell; and 3) **-fusion** denotes the fusion cell is replaced with concatenation operation. Additionally, **-Pattern** means pattern shared unit is removed from CED, which has introduced in Section 4.4.1. From Table 5, we win the following observations:

- Ablating any parts of pattern-shared unit could reduce the performance of pattern-shared unit, weakening the performance by from 0.8% to 3.0% in micF1 on Snopes and PolitiFact and from 0.6% to 1.6% in accuracy on PHEME, which demonstrates the effectiveness of each part of pattern-shared unit.

- Compared with -boosting, -screening fulfils both 0.5% reduction in micF1 on Snopes and PolitiFact, respectively, and 0.3% reduction in accuracy on PHEME, which presents that for the pattern-shared unit, capturing eigen-invariant features is more effective than strengthening specific features for improving the performance of fake news detection.
- -fusion reflects the worst performance compared to the models that only ablate one part, which elaborates that the organic integration of news semantics and generated semantics is able to capture more effectively shared features for improving the performance of detection.

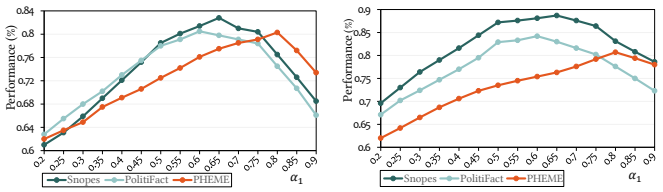
4.4.3 Restriction Unit Evaluation

To evaluate the contribution of each cell of restriction unit to the inter-category differentiated features, we test the performance of CED with the parameter α_1. Specifically, we take α_1 from 0.2 to 0.9 in units of 0.05 on the three datasets and test the overall performance (micF1 on Snopes and PolitiFact, and accuracy on PHEME) of the model and the specific performance on false news (i.e., the term of False F1-score). As shown in Figure 4, we observe that:

- On the whole, when α_1 is less than 0.5, the performance of the model improves with the increase of α_1, but the performance of the model on Snopes and PolitiFact is significantly faster than that on PHEME. At the same time, when the model obtains the optimal performance on three datasets, α_1 on Snopes is more similar to that on PolitiFact, but it is quite different from that on PHEME, the specific values of α_1 on the three datasets are 0.65, 0.60, and 0.80, respectively. These reflect from a certain aspect that the similarity of data structure between Snopes and PolitiFact datasets as well as the differences from those on PHEME.
- When the model achieves the most excellent performance on the three datasets, their parameters α_1 are all greater than 0.5, which indicates that the standardizing cell in restriction unit contributes more to the capture of inter-category differentiated features than the restrained cell. Furthermore, when α_1 exceeds 0.75, the performance of the model drops sharply on Snopes and PolitiFact, which expresses that weakening the role of restrained cell is not conducive to the performance of the model. Taken together, these show that the two cells complement each other, and their organic combination could exert the maximum potential.
- In the term of False F1-score, when α_1 is less than 0.5, the performance of the model is improved quickly, and when α_1 is greater than 0.5, its performance is boosted slowly on Snopes and PolitiFact. However, when α_1 goes from 0.2 to 0.8, it exhibits a stable trend of ascending on PHEME. This may be that fake news in Snopes and PolitiFact has more obvious category-based credibility-indicative features, compared with that in PHEME.

4.4.4 The Limitation of the Method Capturing Conflict Features

We randomly select 200 items from three datasets, i.e., Snopes, PolitiFact, and PHEME, and we make a manual comparison and find that 32%, 36%, and 26% of the items on the three datasets have no conflicts in relevant articles. To confirm the limitation of fake news detection using conflicting features, we conduct experiments to compare CED model and a typical method relying on the conflict features between news and comments to detect fake news, AIFN [12]. The experimental results are shown in Figure 5, we find that our model is obviously better than AIFN model, showing at least 2.8%(micF1), 2.6%(micF1), and 1.4%(accuracy) performance on the three datasets, respectively, which confirms that the advantages of our CED in capturing category differences and the limitations of capturing conflict semantics for fake news detection.



(a) Overall performance changes of CED with α_1. (b) Performance changes of CED on false news with α_1.

Fig. 4. Performance changes of CED with α_1 of restriction unit on Snopes, PolitiFact, and PHEME.

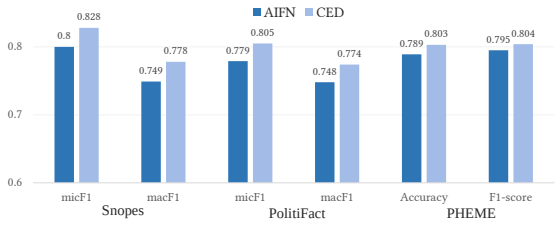
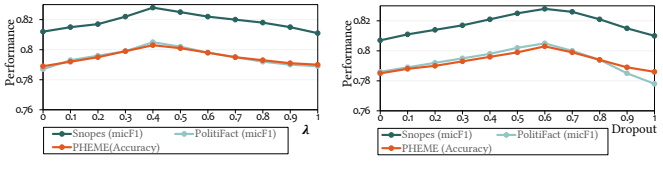


Fig. 5. The performance comparison between our CED and AIFN.

4.4.5 The Impact of Several Parameters on Model Performance

In this section, we examine the effect of λ in news-guided encoder and dropout on model performance. First, we evaluate the influence of λ in news-guided encoder module on the performance of CED, we adjust the value of λ to measure the performance change of the model from 0 to 1. The experimental results are shown in Figure 6(a). We observe that when λ is 0.4, the model achieves the best performance, and when λ is less than 0.3 and higher than 0.7, the performance of CED becomes poor, which shows the effectiveness of our model in fusing the original encoding features and the aggregated information features to generate examples. Furthermore, in overall, with the change of λ, the performance of the model does not change significantly, floating around 2%, which also shows that the integration of sequence features has limited effect on news-guided encoder.

Second, we measure the impact of dropout of multi-head self-attention networks on model performance, we change the dropout value from 0 to 1 to obtain the experimental results of the model, as shown in Figure 6(b). We observe that,



(a) The impact of λ on model performance (b) The impact of dropout of self-attention networks on model

Fig. 6. The impact of λ and dropout of self-attention networks on model performance

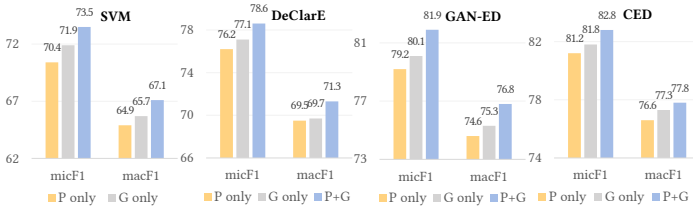


Fig. 7. Performance comparison of the typical models based on original news and the generated examples on Snopes.

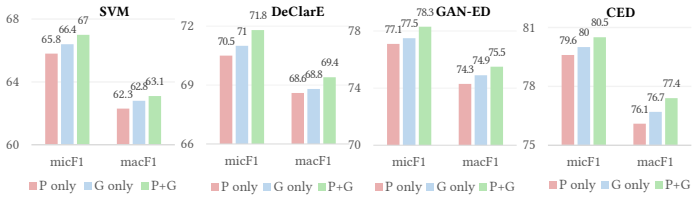


Fig. 8. Performance comparison of the typical models based on original news and the generated examples on PolitiFact.

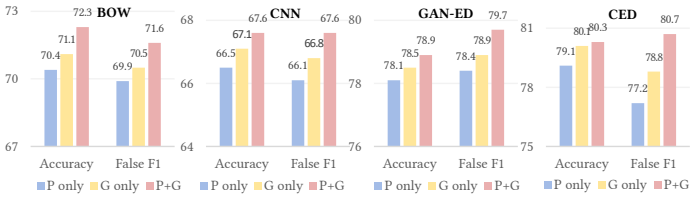


Fig. 9. Performance comparison of the typical models based on original news and the generated examples on PHEME.

at the beginning, the performance of the model improves with the increase of dropout, which confirms that the utilization of dropout to prevent overfitting of the model could effectively improve the performance of CED. In particular, when the dropout is 0.6, the performance of the model is optimal. After that, the performance of the model decreases with the increase of dropout. Therefore, we choose dropout of 0.6 to obtain the most excellent performance of CED.

4.4.6 Assessing Utility of Generated Examples

To evaluate the utility of the generated examples with rich inter-category differentiated features, we perform experiments on multiple models on three types of data, i.e., only the original dataset (P only), only the generated examples (G only), and the integration of the original data and the generated examples (G+P). The experimental results are shown in Figures 7, 8, and 9. We observe that:

- Models achieve more satisfactory performance on the three datasets using only the generate examples than only using the original dataset. Specifically, the four models show at least 0.4% boosts in micF1. This illustrates that the examples generated by CED include more inter-category differentiated features than the original dataset.
- On the whole, all models with generated examples achieve varying degrees of improvements than these without generated examples. SVM obtains the most obvious improvement on Snopes and PHEME, and DeClarE wins 1.3% boost in micF1 on PolitiFact. All these elaborate the usefulness of generated examples and also convey that the improvements of the model mainly comes from the generated examples.
- The performance of most models is more prominent on Snopes than on PolitiFact. For instance, SVM(P+G) and DeClarE(P+G) achieve 3.1% and 2.4% boost in micF1 on Snopes than SVM(P only) and DeClarE(P only), respectively, while reflecting 1.2% and 1.3% boost on PolitiFact, respectively. The main reason can be attributed to the differences in classification. In detail, PolitiFact originally divides news into six categories, and we divide them into two categories, i.e., true and fake news, resulting in some semi-true and semi-false features being blend into true or false news to affect the capture of shared features within the same category.
- Models achieve a relatively low-performance boost on PHEME. Specifically, models obtain an improvement of more than 2.5% (in micF1) on the first two datasets, while only reflecting an average performance improvement of about 1.2% (in accuracy) on PHEME. The reason may be that we utilize comments of news in PHEME as inputs of CED, and their text quality is poorer than that of the relevant articles in Snopes and PolitiFact.

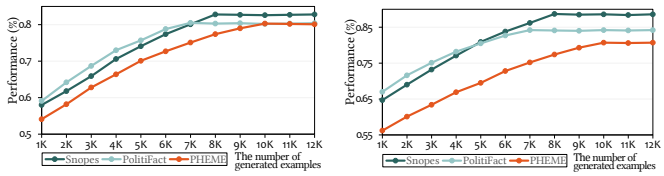
4.4.7 Effect of the Quantity of Generated Examples on CED

From subsection 4.4.6, we have learned that the generated examples play a decisive role in improving the performance of CED. To further understand the impact of generated examples on the model, we conduct two groups of experiments to analyze the changes of the performance of CED with the number of generated examples captured from the three datasets, as shown in Figure 10. One group (Figure 10(a)) examines the overall performance of CED using the metric of micF1 on Snopes and PolitiFact, and the metric of accuracy on PHEME, and the other group (Figure 10(b)) investigates the performance of the specific category (fake news) utilizing the metric of false F1-score on the three datasets. From the figure, we gain the following observations:

- In the early stage, i.e., the number of generated examples is less than 6K, the performance of CED continues to improve with the increase of the number of generated examples, and the overall performance change is generally consistent with that of a single category (fake news), which shows that the increase of generated examples continuously provides the model with more inter-category differentiated features for distinguishing the credibility of news.
- As the number of generated examples increases, the model finally achieves stable and optimal performance. When the generated examples keep on increment, it is difficult for the model to continue to improve its performance, which indicates that CED has secured the maximum credibility-indicative features from the examples, providing constructive guidance for model training and optimization.
- The number of generated examples required for the model to reach stable performance is different, i.e., 8K, 7K, and 10K on Snopes, PolitiFact, and PHEME, respectively. This explains from one side that the model training on PHEME is more laborious than on Snopes and PolitiFact.

4.4.8 Analysis of Training Process

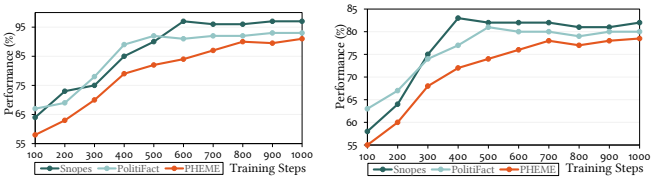
In addition to the influence of the above aspects on CED, we also conduct a study on the training process of CED on the three datasets. During the training, a batch data is



(a) The overall performance changes of CED with the number of generated examples.

(b) The changes of the specific performance of CED on false news with the number of generated examples.

Fig. 10. Performance changes of CED with the number of generated examples on Snopes, PolitiFact, and PHEME.



(a) Performance changes of CED in training set on three datasets.

(b) Performance changes of CED in validation set on three datasets.

Fig. 11. Performance changes of CED in training set and validation set on three datasets during training process.

provided at a time, the weights of our model are updated in the direction that minimizes the loss, which is defined as a training step. We analyze the training process of the model through observing the performance (in micF1 on Snopes and PolitiFact, in accuracy on PHEME) of CED in training set and validation set under incremental training steps. The experimental results are shown in Figure 11. We could clearly obtain the following observations:

- The accuracy of our model gradually increases and tends to be stable as the number of training steps increases. Specifically, our model obtains the best performance in accuracy at about 600, 500, and 800 training steps in the training set, and at about 500, 400, and 700 training steps in the validation set on the three datasets, respectively.
- CED is difficult to converge in the early training phase. In detail, at the training step of 300 in the validation set, our model only secures 75.69% (in micF1), 74.34% (in micF1), and 68.21% (in accuracy) performance on Snopes, PolitiFact, and PHEME, respectively. We speculate that the pattern-shared unit of CED has not learned enough intra-category common features, which leads to the non-obvious inter-category differentiated features.
- Our model obtains the faster convergence on Snopes and PolitiFact (at about 600 and 500 training steps in the training set, respectively), while achieving relatively slower convergence on PHEME (at about 800 training steps in the training set). The reason might be that on Snopes and PolitiFact, the high-quality articles with long text, as the inputs of the model, could capture more credibility features in the process of generation, compared with the highly sparse comments with short text are used as the inputs of the model on PHEME.

4.5 Case Study

To gain a more transparent understanding of the features captured by our model, we visualize the outputs of news-guided encoder and that of category-controlled decoder, respectively. The details are as follows:

4.5.1 The Visualization of News-guided Encoder

To express intuitively what features CED has learned from news and relevant articles, we carry out experiments to visualize the outputs of news-guided encoder. Specifically, we first employ max-pooling operation to pool the outputs of fusion merging layer $\tilde{m}_i$ and then map them into the corresponding elements in the input layer, respectively. Finally the interesting patterns are obtained and visualized in Figure 12. We observe that:

- Encoder without news guidance only obtains the key semantics of the relevant articles (like 'no crime wave' and 'no menace lurking'), while encoder using news guidance not only acquires the key semantics of the relevant articles but also wins the key semantics of the news, like 'ban black cars' and 'curb global warming' (relevant article 1).
- Encoder using news guidance is able to focus on the most relevant parts of the news through the relevant articles, such as 'hotter climates' and 'lots of questions' in relevant article 2 and 'curb global warming' in news semantics.
- Our model is also capable of capturing some of the rich and diverse core semantic features from relevant articles, which helps to reveal the wrong parts of the claim, e.g., 'california agencies' and 'overreaching regulations', while encoder without news guidance only secures non-key words that are not strongly related to news semantics, e.g., 'pitched competition' and 'california agencies' (relevant article 3).

4.5.2 The Visualization of Category-controlled Decoder

We conduct experiments to visualize how category-controlled decoder enriches the semantics of generated examples to possess more category-differential features. We sample one piece of true news and one piece of fake news from Snopes, and present some generated contents in Figures 13 and 14, respectively. We observe that:

- The generated examples present weak grammaticality, but are relevant to the input news, like the generated content 'giulian's misquotation of clinton' and the news content 'candidate said' in Figure 13.
- Decoder with category control indicates rich credibility semantics, e.g., 'absurd' and 'unconvincing' (in red words) while decoder without category control only focuses on the key semantics, e.g., 'muslim' and 'not be allowed' (in grey words) in Figure 14.
- The generated examples produce different credibility-indicative semantics aiming at different categories of information, like 'will be a joke' (red words) in fake news and blue words 'no rumor just fact' (blue words) in true news.

4.5.3 The Visualization of Pattern-shared Unit

We have observed from the latest subsection that the generated true and fake examples are able to generate different credibility semantics. To eloquently explore the differences in the generation of true and fake examples, we visualize the features captured in pattern-shared units that are corresponding to true news and fake news, respectively. Specifically, based on the three datasets, we first look up these elements with the largest values from F_t in fusion cell of pattern-shared units, then these elements are mapped into the corresponding values in E_c^{pool} of news and the

News: California planning ban black cars order curb global warming	***News:*** California planning ban black cars order curb global warming
Relevant Article 1: Described that we have yet to hear about but even if that proves to be the case there is clearly no crime wave no ever present danger to motorists everywhere no menace lurking in the nation's parking lots the urban the urban legend california is planning to ban black cars in order to curb global warming the Snopes response	***Relevant Article 1:*** Described that we have yet to hear about but even if that proves to be the case there is clearly no crime wave no ever present danger to motorists everywhere no menace lurking in the nation's parking lots the urban the urban legend california is planning to ban black cars in order to curb global warming the Snopes response.
Relevant Article 2: are white cars cooler than dark cars people who are relocating to hotter climates have lots of questions about their cars should you sell your dark blue car because it will be like an oven in phoenix	***Relevant Article 2:*** are white cars cooler than dark cars people who are relocating to hotter climates have lots of questions about their cars should you sell your dark blue car because it will be like an oven in phoenix
Relevant Article 3: in any disputes there is a pitched competition between california agencies for which is the most nonsensical in its implementation of overreaching regulations	***Relevant Article 3:*** in any disputes there is a pitched competition between california agencies for which is the most nonsensical in its implementation of overreaching regulations
(a) Encoder without news guidance	(b) Encoder using news guidance

Fig. 12. The visualization of the encoder with/without news guidance on a piece of unverified news with three relevant articles.

News [Fake]: Leading democratic candidate said unfettered free market destructive force modern America	
Generated Example: The details on these issues regarding giulian's misquotation of Clinton, these words are absurd and unconvincing, will be a joke…	***Generated Example:*** The free market is one of our greatest assets and Hillary once said that the unfettered free market is the most destructive force in modern America…
(a) The decoder with category control	(b) The decoder without category control

Fig. 13. The visualization of the decoder with/without category control on a piece of fake news.

News [True]: says roy moore said practicing muslim not member congress faith	
Generated Example: No rumor just fact roy moore say muslims shouldn't be allowed to serve in congress …	***Generated Example:*** Alabama said that a practicing muslim should not be allowed to serve in congress by reason of his faith moore who is currently facing allegations of sexual misconduct…
(a) The decoder with category control	(b) The decoder without category control

Fig. 14. The visualization of the decoder with/without category control on a piece of true news.

generated words in decoder and then find out the corresponding specific words in the news sequence and the decoding sequence. The visualized results are shown in Figure 15. We have the following observations:

- In pattern-shared unit for true news, many words possess close semantics, like 'agree' and 'sure'. Equally, in pattern-shared unit for false news, several words have similar meanings, such as 'mess'. These illustrate that our model could learn the intra-category shared credibility features within the generated true and fake examples.
- Both pattern-shared units capture words that are mostly commonly-used words, like 'get' and 'fact' for true news, and 'obviously' and 'but' for fake news, which rarely involves domain words or specific words that reflect one specific news or one event. It presents that pattern-shared unit could effectively obtain the eigen-invariant features.
- Both pattern-shared units acquire questionable and skeptical words, such as 'probably', 'correct', and 'guess' for true news and 'suspect', 'incredible', and 'wrong' for fake news, which indicates that our model is capable of learning credibility-indicative features, and it also confirms to our consensus that both true and fake news may be questioned.
- More emotional smooth words (like 'delightful', 'joyful', and 'sad') are captured in the pattern-shared unit for true news while the pattern-shared unit for fake news learns more negative words or extreme words, like 'extremely', 'rage', and 'manic', which demonstrates that pattern-shared unit effectively enhances the category-differentiated features at emotion level.

Multi-perspectives	Pattern-shared Unit for True News	Pattern-shared Unit for Fake News
Credibility Perspective	No doubt; Agree; Clear; Sure; Get Correct; Believe; Yes; Fact	But; Obviously; Seriously; Really Mess; Wrong; Can not; Absolutely
Conflict Perspective	Probably; Correct; Guess	Suspect; Incredible; Wrong; Confuse; Probably
Emotion Perspective	Delightful; Appreciate; Joyful; Sad; Boring	Extremely; Rage; Manic; Despair
Style Perspective	Should; Consider; Presumably; Reported	Shocking; Unexpectedly; Completely; Absolutely

Fig. 15. The visualization of pattern-shared units for true news and that for fake news on three datasets.

- Some words with different styles are captured by both pattern-shared units. Specifically, the unit for fake news could capture some shock style-related words, like 'shocking', 'unexpected', and 'absolutely', while the unit for true news focus more on objective style-related words, like 'should', 'consider', and 'reported'.
- In general, there are differences between the words captured based on true and fake news from multiple perspectives, which not only learn the conflict features often captured by the existing models, but also obtain the differences from the perspectives of extreme emotion and writing styles.

4.6 Error Analysis

Although CED outperforms previous state-of-the-art methods on three datasets (From subsection 4.3), we also discover the following defects based on the above experiments:

- From the subsection 4.5.3, it is found that CED can capture shared skepticism-indicative features from both true and fake news, but cannot accurately learn the false parts of the unverified news content, which indicates that our model is difficult to apply to the interpretability of fake news.
- It is found from subsection 4.4.8 that our model not only achieves weaker performance but also takes longer training time on PHEME compared with on Snopes and PolitiFact. The reason is that PHEME is structured as 'news–comments', while Snopes and PolitiFact are structured as 'news–relevant articles', in which comments are inferior to relevant articles in terms of length, text quality, and relevance with news. This elaborates that our model performs relatively poorly in terms of performance and computational efficiency in dealing with low-quality comments compared to dealing with high-quality relevant articles.
- In addition, our model is designed for the binary classification of true and fake news, and our model cannot be directly applied to fine-grained fake news detection(such

as the six types of news on PolitiFact). For fine-grained fake news detection, we propose to extend our CED from two types of generated examples to multiple types of examples, and then strengthen the differences between these examples by restriction unit.

5 Conclusion

In this paper, we proposed a novel category-controlled encoder-decoder model (CED) to generate examples with differentiated features between categories (i.e., true and fake news) for fake news detection, which learned inter-category differential features from intra-category shared features. Our model designed gated pattern-shared unit to strengthen the representation learning of intra-category shared features within true or fake news and developed restriction unit to force the two types of intra-category shared features to be more different for gaining inter-category differentiated features. We verified the performance of CED on three real-world benchmark datasets, i.e., Snopes, PolitiFact, and PHEME. From these experiments, we have observed that CED outperformed existing hand-crafted, shallow, and automatic deep extracted methods. We noted that the generated examples were able to effectively improve the performance of different types of methods. We also gained a transparent understanding to end-users about the features captured by our model through visualizing the different modules of CED.

This work has tackled fake news detection on social media through learning the content semantics of news and relevant articles to capture the shared features within the category for exploring inter-category differentiated features. While social media is rich in meta-data information, we will study the following directions in future. First, we will enhance our CED framework by fusing multiple meta-data features, e.g., user profile, to strengthen the capture of the features within one category. Second, we will develop (semi-)supervised variants for pattern-shared unit, to learn fine-grained and category-oriented shared features for different types of information, rather than just limited to the binary classification between true and fake news. Third, we will consider improving the efficiency of CED by driving the interaction between news and relevant articles to acquire more high-quality relevant articles.

Acknowledgments

The research work is supported by 'National key research and development program in China' (2019YFB2102300); 'the World-Class Universities (Disciplines) and the Characteristic Development Guidance Funds for the Central Universities of China' (PY3A022); Ministry of Education Fund Projects (No. 18JZD022 and 2017B00030); Shenzhen Science and Technology Project (JCYJ20180306170836595); Basic Scientific Research Operating Expenses of Central Universities (No.ZDYF2017006); Xi'an Navinfo Corp. & Engineering Center of Xi'an Intelligence Spatial-temporal Data Analysis Project (C2020103); Beilin District of Xi'an Science & Technology Project (GX1803).

References

[1] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nat. Commun.*, vol. 10, no. 1, p. 7, 2019.

[2] L. F. B. S.-T. N. Grinberg, K. Joseph and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[3] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[4] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *Proc. The Web Conference*. ACM, 2019, pp. 2915–2921.

[5] T. Gröndahl and N. Asokan, "Text analysis in adversarial settings: Does deception leave a stylistic trace?" *ACM Comput. Surv.*, vol. 52, no. 3, p. 45, 2019.

[6] 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*, 2019.

[7] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. B. Hu, "Xfake: Explainable fake news detector with visualizations," in *Proc. The Web Conference*. ACM, 2019, pp. 3600–3604.

[8] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proc. ACL*, 2018, pp. 1980–1989.

[9] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proc. SIGKDD*, 2019, pp. 395–405.

[10] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," in *Proc. EMNLP*, 2018, pp. 22–32.

[11] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Proc. AAAI*, vol. 33, 2019, pp. 6859–6866.

[12] L. Wu and Y. Rao, "Adaptive interaction fusion networks for fake news detection," in *Proc. ECAI*, 2020.

[13] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *Proc. ACL*, 2019, pp. 2561–2571.

[14] L. Wu, Y. Rao, X. Yang, W. Wang, and A. Nazir, "Evidence-aware hierarchical interactive attention networks for explainable claim verification," in *Proc. IJCAI*, 2020.

[15] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *Proc. SIGKDD*, vol. 19, no. 1, pp. 22–36, 2017.

[16] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," *TOIT*, vol. 20, no. 2, pp. 1–18, 2020.

[17] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: recognizing clickbait as" false news"," in *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, 2015, pp. 15–19.

[18] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, 2016, pp. 7–17.

[19] W. Zhong, D. Tang, Z. Xu, R. Wang, N. Duan, M. Zhou, J. Wang, and J. Yin, "Neural deepfake detection with factual structure of text," in *Proc. EMNLP*, 2020, pp. 2461–2470.

[20] J. P. Baptista and A. Gradim, "Understanding fake news consumption: A review," *Social Sciences*, vol. 9, no. 10, p. 185, 2020.

[21] S. De Sarkar, F. Yang, and A. Mukherjee, "Attending sentences to detect satirical fake news," in *Proc. COLING*, 2018, pp. 3371–3380.

[22] N. T. Tam, M. Weidlich, B. Zheng, H. Yin, N. Q. V. Hung, and B. Stantic, "From anomaly detection to rumour detection using data streams of social platforms," *Proc. VLDB Endowment*, vol. 12, no. 9, pp. 1016–1029, 2019.

[23] A. Giachanou, P. Rosso, and F. Crestani, "Leveraging emotional signals for credibility detection," in *Proc. SIGIR*. ACM, 2019, pp. 877–880.

[24] J. Ma, W. Gao, and K.-F. Wong, "Detect rumor and stance jointly by neural multi-task learning," in *Proc. The Web Conference*, 2018, pp. 585–593.

[25] L. Wu, Y. Rao, H. Jin, A. Nazir, and L. Sun, "Different absorption from the same sharing: Sifted multi-task learning for fake news detection," *arXiv preprint arXiv:1909.01720*, 2019.

[26] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Proc. PAKDD*. Springer, 2018, pp. 40–52.

[27] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proc. CIKM*. ACM, 2017, pp. 797–806.

[28] Q. Zhang, A. Lipani, S. Liang, and E. Yilmaz, "Reply-aided detection of misinformation via bayesian deep learning," in *Proc. The Web Conference*. ACM, 2019, pp. 2333–2343.

[29] K. Zhou, C. Shu, B. Li, and J. H. Lau, "Early rumour detection," in *Proc. NAACL*, 2019, pp. 1614–1623.
[30] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The Role of User Profile for Fake News Detection," *arXiv:1904.13355*, 2019.
[31] Q. Li, Q. Zhang, and L. Si, "Rumor detection by exploiting user credibility information, attention and multi-task learning," in *Proc. ACL*, 2019, pp. 1173–1179.
[32] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
[33] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv preprint arXiv:1705.04304*, 2017.
[34] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proc. ACL*, 2018, pp. 889–898.
[35] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *Proc. The Web Conference*. ACM, 2019, pp. 3049–3055.
[36] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, 2018.
[37] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. NeurIPS*, 2015, pp. 3483–3491.
[38] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.
[39] J. Juraska, P. Karagiannis, K. K. Bowden, and M. A. Walker, "A deep ensemble model with slot alignment for sequence-to-sequence natural language generation," in *Proc. NAACL*, 2018, pp. 152–162.
[40] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, "Deep keyphrase generation," in *Proc. ACL*, 2017, pp. 582–592.
[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NeurIPS*, 2014, pp. 3104–3112.
[42] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[43] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," in *Proc. ACL*, 2016, pp. 1631–1640.
[44] W. Chen, Y. Gao, J. Zhang, I. King, and M. R. Lyu, "Title-guided encoding for keyphrase generation," in *Proc. AAAI*, vol. 33, 2019, pp. 6268–6275.
[45] K. Shu, S. Wang, T. Le, D. Lee, and H. Liu, "Deep headline generation for clickbait detection," in *Proc. ICDM*. IEEE, 2018, pp. 467–476.
[46] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. NeurIPS*, 2016, pp. 343–351.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[48] E. Kochkina, M. Liakata, and A. Zubiaga, "All-in-one: multi-task learning for rumour verification," in *Proc. COLING*, 2018, pp. 3402–3413.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[51] J. Thorne and A. Vlachos, "Automated fact checking: Task formulations, methods and future directions," in *Proc. COLING*, 2018, pp. 3346–3359.
[52] W. Y. Wang, "liar, liar pants on fire: A new benchmark dataset for fake news detection," in *Proc. ACL*, 2017, pp. 422–426.
[53] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proc. EMNLP*, 2017, pp. 2931–2937.
[54] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial Training Methods for Semi-Supervised Text Classification," in *Proc. ICLR*, 2017.
[55] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proc. The Web Conference*, 2015, pp. 1395–1405.
[56] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. The Web Conference*, 2011, pp. 675–684.
[57] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. CIKM*, 2015, pp. 1751–1754.
[58] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proc. IJCAI*, 2017, pp. 3901–3907.
[59] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proc. IJCAI*, 2016, pp. 3818–3824.
[60] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. AAAI*, 2016.
[61] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

Lianwei Wu is currently working towards the PhD degree with the School of Software Engineering, Xi'an Jiaotong University, China. He has published over 10 research papers in major international conferences and journals including: AAAI, IJCAI, ACL, EMNLP, ECAI, IEEE Transactions on Affective Computing, and Information Sciences, etc. His research interests include information credibility evaluation, social media analysis, and natural language processing.

Yuan Rao received the PhD degree from Xi'an Jiaotong University, China, in 2005. He was a post-doctoral researcher at Tsinghua University from 2005 to 2007. He was a Visiting Scholar at the LTI Lab. of Carnegie Mellon University from 2015 to 2016. He is currently an associate professor and deputy director of the department of business analysis and technology, in the School of Software Engineering with Xi'an Jiaotong University. He is also the Director of the Laboratory of Social Intelligence and Complexity Data Processing (SICDP), Xi'an Jiaotong University. His main research interests include natural language processing, social media analysis, and machine learning.

Cong Zhang received master's degrees from INSEEC, France, in 2015. She is currently a HR assistant in with Xian Huanyu Satellite Control and Data Application Co. Ltd. Her research interests include social media analysis and data mining.

Yongqiang Zhao received the BS degree in mechanical engineering from Taiyuan University of Science and Technology, Taiyuan, China, in 2017. He is currently working towards the master degree in the School of Software Engineering, Xi'an Jiaotong University. His research interests include computer vision, object detection, and image captioning.

Ambreen Nazir received the BSc and MSc degrees in software engineering from the University of Science and Technology, Taxila, Pakistan in 2012 and 2014, respectively. She worked as a lecturer at Comsats Institute of Information and Technology, Pakistan for 2 years. She is currently working towards the PhD degree at the Xi'an Jiaotong University, focusing on aspect-level sentiment analysis and its applications, and how to move towards a more contextual semantic-oriented form of sentiment analysis.