dEFEND: A System for Explainable Fake News Detection

Limeng Cui
Penn State University
University Park, PA
lzc334@psu.edu

Kai Shu
Arizona State University
Tempe, AZ
kai.shu@asu.edu

Suhang Wang
Penn State University
University Park, PA
szw494@psu.edu

Dongwon Lee
Penn State University
University Park, PA
dongwon@psu.edu

Huan Liu
Arizona State University
Tempe, AZ
huanliu@asu.edu

ABSTRACT

Despite recent advancements in computationally detecting fake news, we argue that a critical missing piece be the explainability of such detection–i.e., *why* a particular piece of news is *detected* as fake–and propose to exploit rich information in users' comments on social media to infer the authenticity of news. In this demo paper, we present our system for an explainable fake news detection called dEFEND, which can detect the authenticity of a piece of news while identifying user comments that can explain why the news is fake or real. Our solution develops a sentence-comment co-attention sub-network to exploit both news contents and user comments to jointly capture explainable top-*k* check-worthy sentences and user comments for fake news detection. The system is publicly accessible[1].

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Fake news; explainable machine Learning; deep learning

ACM Reference Format:
Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: A System for Explainable Fake News Detection. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3357384.3357862

1 INTRODUCTION

In recent years, fake news with various political and commercial purposes have emerged on social media. The wide and fast dissemination of fake news has made a huge negative impact on society. For example, fake news stories appeared on social media in hours after the mass shooting in Las Vegas, which have troubled the lives of many innocent people. A study showed that people of all ages and political parties have reduced their credibility with mass media [10]. Therefore, it is very important to detect fake news and prevent the spread of fake news.

Fake news detection systems are in great demand and several systems have been developed successively to detect rumors and fake news. Some systems can show users the propagation of a rumor (Hoaxy [7]). Hoaxy integrates the online information and its related fact-checking results, and provides a platform for data analysis and visualization. Others can automatically fact check rumors (ClaimBuster [3]). ClaimBuster is a platform which detects important factual claims in political discourses. However, the majority of existing systems focus on improving the detection accuracy of fake news, but rarely consider providing explanations on "why" a piece of news is detected as fake news.

In practice, explanations of why a piece of news is detected as fake news are of great importance for fake news detection systems because: (**1**) by understanding the reasons behind predictions, people are more likely to *trust* and use the system [5], which helps to prevent the spread of fake news; and (**2**) the derived explanations can provide new insights and knowledge, which can help improve fake news detection performance and ease manual fact-checking by experts if needed. Given the lack of an explainable fake news detection system and the importance of such system, in this demo, we develop a fake news detection system with high *accuracy* for detecting fake news and with self-*explanaibility* for explaining why a piece of news is fake.

The explanations provided by our system should be *user-friendly*, i.e., easily understandable and makes sense to the public. The majority of existing work on explainable machine learning approaches mainly focus on approximating the decision boundary of a complex or black-box classifier [5], which are not suitable for our system. Fortunately, we find that user comments on news can be used for explaining why a piece of news is fake. For example, for the fake claim "Pence: Michelle Obama Is The Most Vulgar First Lady We've Ever Had", a comment "Where did Pence say this? I saw him on CBS this morning and he didn't say these things." can explain why the claim is fake. In addition, such users' social engagements can facilitate fake news detection [2, 6]. For example, Ruchansky et al. [6] proposed a hybrid deep learning framework to embed the news content, the user response, and the source users promoting it together for fake news detection.

[1] http://defend.ist.psu.edu

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357862

Inspired by the rich information available in user comments on social media, we propose to derive explanations from the perspectives of news contents and user comments. First, news contents may contain linguistic clues which can verify false information. As fake news is written to deliberately mislead public opinions, the words used are prone to exaggeration and sensationalization compared with real news. Second, user comments, including opinions, stances, and sentiment, can provide complementary information. Third, users' comments are inherently related to news content and may provide accounts to explain why a given news article is fake or not, like in the above example.

Therefore, our system is built upon a novel deep learning framework for explainable fake news detection, which can jointly learn explainable information from news contents and users' comments. Our system provides a web-based interface enabling users to check the authenticity of a new. The backend detection algorithm serves users with not only the detection result but also all the arguments that support the detection result including crucial sentences in the article and explainable comments from social media platforms. The main **contributions** of the demonstration paper are:

- We develop a novel fake news detection system dEFEND, which considers both accuracy and explainability.
- The backend of dEFEND is built upon on a novel deep learning framework that exploits both user comments and news contents for improving the detection accuracy and proving explainability. We conduct extensive experiments on real-world datasets and demonstrate the framework significantly outperforms 7 state-of-the-art fake news detection methods by at least 5.33% in F1-score [8].
- To intuitively show the news and enhance the interactivity of the system, dEFEND integrates exploratory functions such as news propagation network visualization and related and trending news.

2 SYSTEM OVERVIEW

The dEFEND system consists of two major components as shown in Figure 1: a web-based user interface and a backend which integrates our fake news detection model. The backend consists of multiple components: (1) a database to store the pre-trained results as well as a crawler to extract unseen news and its comments, (2) the dEFEND algorithm module based on explainable deep learning fake news detection, which gives the detection result and explanations simultaneously and (3) an exploratory component that shows the propagation network of the news, trending and related news.

2.1 User Interface

We design a web-based interface which provides users with explainable fact-checking of news. A user can enter either the tweet URL or the title of the news. A screenshot was shown in Figure 3. On typical fact-checking websites, a user just sees the check-worthy score of news (like Gossip Cop[2]) or each sentence (like ClaimBuster[3]). In our approach, the user can not only see the detection result (in the right of Figure 3(a)), but also can find all the arguments that support the detection result, including crucial sentences in the article (in the middle of Figure 3(b)) and explainable comments from social media platforms (in the right of Figure 3(b)). At last, the user can also review the results and find related news/claims.

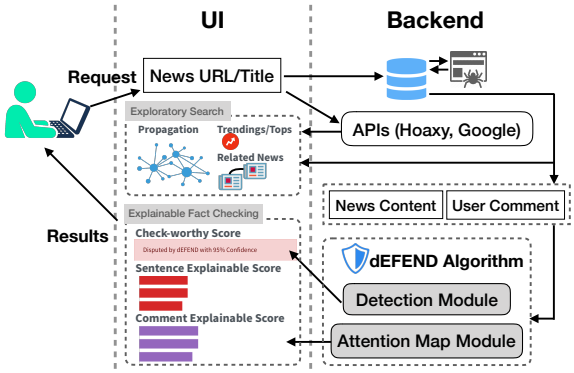


Figure 1: dEFEND System Overview

The system also provides exploratory search functions including news propagation network, trending news, top claims and related news. The news propagation network (in the left of Figure 3(b)) is to help readers understand the dynamics of real and fake news sharing, as fake news are normally dominated by very active users, while real news/fact checking is a more grassroots activity [7]. Trending news, top claims and related news (in the lower left of Figure 3(a)) can give some query suggestions to users.

2.2 Interactive Backend

The explainable fact-checking is the main task of the backend, which is to compute the check-worthy results of a given news and related comments. After getting the tweet URL/title as input, the backend searches in the database or crawls corresponding news content and users' comments online, and provides the data to the dEFEND algorithm component. It is pre-trained on the FakeNewsNet database [9]. It contains two modules: the detection module and the attention map module. The detection module generates a check-worthy score between 0 and 1 by using the news content and users' comments, which is displayed on the frontend through a warning sign. The attention map module outputs the attentions weights of each sentence and comment, which show how much they are related to the major claim of the news, in other words, the explainability.

To provide exploratory search functions, the backend calls the Hoaxy API[4] to get the diffusion paths (retweets, quotes and mentions), and visualizes the spread of news and related fact-checking online on the frontend. It uses Google News API[5] to get the trending news and top claims.

3 EXPLAINABLE DETECTION

In this section, we present details of the explainable fake news detection algorithm of dEFEND. It consists of four parts as shown in Figure 2: (1) a news content encoder, (2) a user comment encoder, (3) a sentence-comment co-attention component, and (4) a fake news prediction component.

[2]https://www.gossipcop.com/
[3]https://idir-server2.uta.edu/claimbuster/
[4]https://rapidapi.com/truthy/api/hoaxy
[5]https://newsapi.org/s/google-news-api

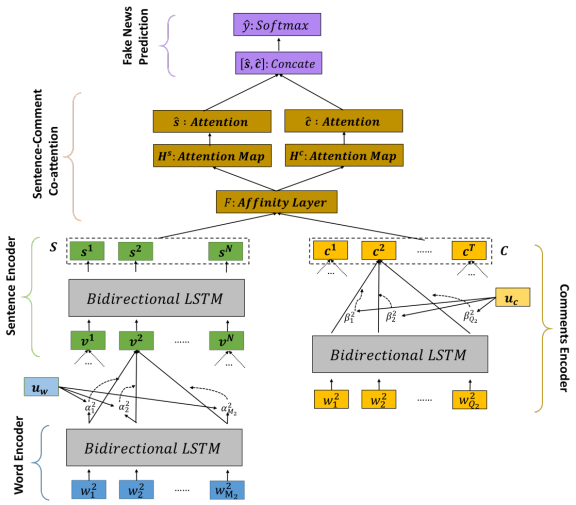


Figure 2: dEFEND Algorithm

3.1 News Contents Encoding

A news document contains linguistic cues with different levels such as word-level and sentence-level, which provide different degrees of explainability of why the news is fake. For example, in a fake news claim "Pence: Michelle Obama is the most vulgar first lady we've ever had", the word "vulgar" contributes more signals to decide whether the news claim is fake rather than other words. Hence we propose to learn the news content representations through a hierarchical structure. Specifically, we first learn the sentence vectors by using the word encoder with attention and then learn the sentence representations through sentence encoder component.

Word Encoder: We learn the sentence representation via a bidirectional Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU). The bidirectional GRU contains the forward GRU $\overrightarrow{f}$ which reads sentence s_i from word w_1^i to $w_{M_i}^i$ and a backward GRU $\overleftarrow{f}$ which reads sentence s_i from word $w_{M_i}^i$ to w_1^i:

$$\begin{aligned}\overrightarrow{\mathbf{h}_t^i} &= \overrightarrow{GRU}(\mathbf{w}_t^i), t \in \{1, \ldots, M_i\} \\ \overleftarrow{\mathbf{h}_t^i} &= \overleftarrow{GRU}(\mathbf{w}_t^i), t \in \{M_i, \ldots, 1\}\end{aligned} \tag{1}$$

We then obtain an annotation of word w_t^i by concatenating the forward hidden state $\overrightarrow{\mathbf{h}_t^i}$ and backward hidden state $\overleftarrow{\mathbf{h}_t^i}$, which contains the information of the whole sentence centered around w_t^i. As not all words contribute equally to the representation of the sentence, we adopt an attention mechanism to learn the weights to measure the importance of each word. The sentence vector $\mathbf{v}^i \in \mathbb{R}^{2d\times 1}$ is computed as $\mathbf{v}^i = \sum_{t=1}^{M_i} \alpha_t^i \mathbf{h}_t^i$, where α_t^i is the attention weight, which measures the importance of t-th word in sentence s_i.

Sentence Encoder: Similar to word encoder, we use RNN with GRU to encode each sentence in news articles. Through the sentence encoder, we can learn the sentence representations. The annotation of sentence $\mathbf{s}_i \in \mathbb{R}^{2d\times 1}$ is obtained by concatenating the forward hidden state $\overrightarrow{\mathbf{h}^i}$ and backward hidden state $\overleftarrow{\mathbf{h}^i}$, which captures the context from neighbor sentences.

3.2 User Comments Encoding

People express their opinions towards fake news through social media posts such as comments, which may contain useful semantic information that has the potential to help fake news detection. The comments extracted from social media are usually short text, so we adopt bidirectional GRU to model the word sequences in comments. We further obtain the annotation of word w_t^j by concatenating $\overrightarrow{\mathbf{h}_t^j}$ and $\overleftarrow{\mathbf{h}_t^j}$. Similarly, the attention mechanism is introduced to learn the weights to measure the importance of each word.

3.3 Sentence-comment Co-attention

We observe that not all sentences in news contents are fake, and in fact, many sentences are true but only for supporting wrong claim sentences [1]. Thus, news sentences may not be equally important in determining and explaining whether a piece of news is fake or not. For example, the sentence "Michelle Obama is so vulgar she's not only being vocal.." is strongly related to the major fake claim "Pence: Michelle Obama Is The Most Vulgar First Lady We've Ever Had". Similarly, user comments may contain relevant information about the important aspects that explain why a piece of news is fake. For example, a comment "Where did Pence say this? I saw him on CBS this morning and he didn't say these things.." is more explainable and useful to detect the fake news. Thus, we aim to select such news sentences and user comments. This suggests us to design attention mechanisms to give high weights of representations of news sentences and comments that are beneficial to fake news detection. Specifically, we use sentence-comment co-attention because it can capture the semantic affinity of sentences and comments and further help learn the attention weights of them.

We can construct the feature matrix of news sentences $\mathbf{S} = [\mathbf{s}^1; \cdots, \mathbf{s}^N] \in \mathbb{R}^{2d\times N}$ and the feature map of user comments $\mathbf{C} = \{\mathbf{c}^1, \cdots, \mathbf{c}^T\} \in \mathbb{R}^{2d\times T}$, the co-attention attends to the sentences and comments simultaneously. Similar to [4], we first compute the affinity matrix $\mathbf{F} \in \mathbb{R}^{T\times N}$ as follows,

$$\mathbf{F} = \tanh(\mathbf{C}^\mathsf{T}\mathbf{W}_l\mathbf{S}) \tag{2}$$

where $\mathbf{W}_l \in \mathbb{R}^{2d\times 2d}$ is a weight matrix to be learned through the networks. Following the optimization strategy in [4], we can consider the affinity matrix as a feature and learn to predict sentence and comment attention maps as follows,

$$\mathbf{H}^s = \tanh(\mathbf{W}_s\mathbf{S} + (\mathbf{W}_c\mathbf{C})\mathbf{F}), \quad \mathbf{H}^c = \tanh(\mathbf{W}_c\mathbf{C} + (\mathbf{W}_s\mathbf{S})\mathbf{F}^\mathsf{T}) \tag{3}$$

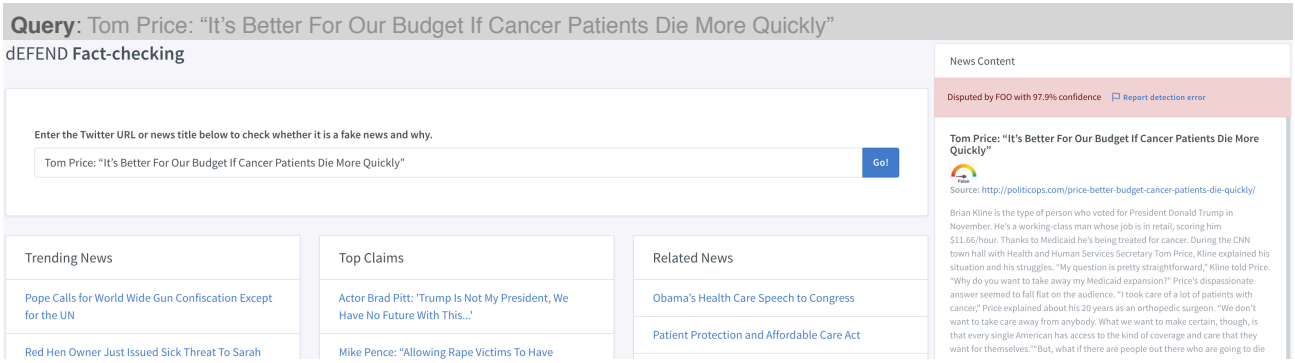
where $\mathbf{W}_s, \mathbf{W}_c \in \mathbb{R}^{k\times 2d}$ are the weight parameters. The attention weights of sentences and comments are given as,

$$\mathbf{a}^s = \text{softmax}(\mathbf{w}_{hs}^\mathsf{T}\mathbf{H}^s), \quad \mathbf{a}^c = \text{softmax}(\mathbf{w}_{hc}^\mathsf{T}\mathbf{H}^c) \tag{4}$$

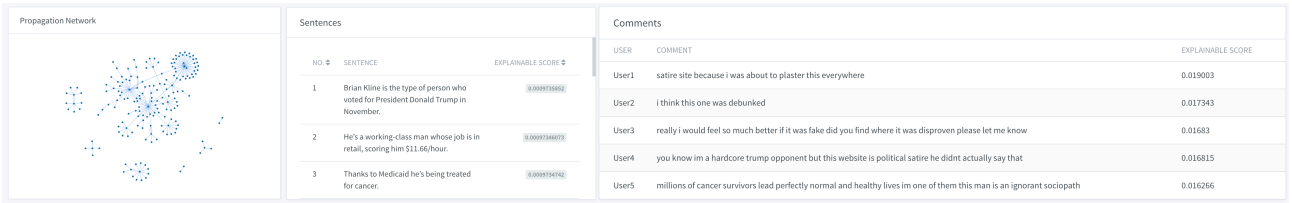
where $\mathbf{a}^s \in \mathbb{R}^{1\times N}$ and $\mathbf{a}^c \in \mathbb{R}^{1\times T}$ are the attention probabilities of each sentence $\mathbf{s}^i$ and comment $\mathbf{c}^j$, respectively. $\mathbf{w}_{hs}, \mathbf{w}_{hc} \in \mathbb{R}^{1\times k}$ are the weight parameters.

3.4 Explainable Fake News Detection

We further concatenate the above outputs together and then add a softmax layer on the top, the output is the probabilities that the news is real and fake respectively. Thus, the goal is to minimize the cross-entropy loss. The sentences and comments with top-ranked



(a) User Interface of Search: the input box (upper left), query suggestions (lower left) and detection result (right).



(b) Explainable Fact Checking: intuitive propagation network (left), explainable sentences (middle) and comments (right).

Figure 3: Demonstration of dEFEND

attention weights are related to the major claims in fake news, which are likely to be check-worthy.

4 DEMONSTRATION

In this section, we show two scenarios of how dEFEND can be used for fact-checking.

Exploratory Search: The system provides users with browsing functions. Consider a user who doesn’t know what to check specifically. By browsing the trending news, top claims and news related to the previous search right below the input box, the user can get some ideas about what he could do. News can be the coverage of an event, such as “Seattle Police Begin Gun Confiscations: No Laws Broken, No Warrant, No Charges” and claim is the coverage around what a celebrity said, such as “Actor Brad Pitt: ’Trump Is Not My President, We Have No Future With This...”’. Users can search these titles by clicking on them. The news related to the user’s previous search is recommended.

Explainable Fact Checking: Consider a user who wants to check whether Tom Price has said “It’s Better For Our Budget If Cancer Patients Die More Quickly”. The user first enters the tweet URL or the title of a news in the input box in Figure 3(a). The system would return the check-worthy score, the propagation network, sentences with explainable scores, and comments with explainable scores to the user in Figure 3(b). The user can zoom in the network to check the details of the diffusion path. Each sentence is shown in the table along with its score. The higher the score, the more likely the sentence contains check-worthy factual claims. The lower the score, the more non-factual and subjective the sentence is. The user can sort the sentences either by the order of appearance or by the score. Comments’ explainable scores are similar to sentences’. The top-5 comments are shown in the descending order of their explainable score.

5 CONCLUSION

In this demo paper, we present a system for explainable fake news detection which provides the check-worthy scores and explainable results at the same time. The algorithm relies on the rich information available in user comments on social media.

6 ACKNOWLEDGMENT

This material is in part supported by the NSF awards #1614576, #1742702, #1820609, and #1915801, ONR grant N00014-17-1-2605 and N000141812108, and ORAU-directed R&D program in 2018.

REFERENCES

[1] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *ACL*. 171–175.
[2] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor Detection with Hierarchical Social Attention Network. In *CIKM*. ACM, 943–951.
[3] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *VLDB* 10, 12 (2017), 1945–1948.
[4] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*. 289–297.
[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. 1135–1144.
[6] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM*. 797–806.
[7] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *WWW*. 745–750.
[8] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *KDD*. 395–405.
[9] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. 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).
[10] Art Swift. 2016. Americans’ trust in mass media sinks to new low. *Gallup News* 14 (2016), 6.