

# dEFEND:

# Explainable Fake News Detection

K. Shu, L. Cui, S. Wang, D. Lee and H. Liu

Arizona State University, Penn State University

KDD 2019

# Keywords

*Fake News, Explainable Machine Learning, Social Network*

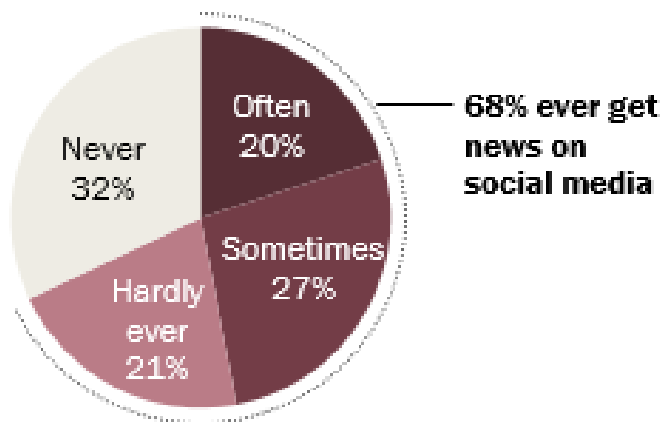


# Introduction

# News on Social Media

## About two-thirds of Americans get news on social media

*% of U.S. adults who get news on social media ...*



## But most social media news consumers expect news there to be inaccurate

*% of social media news consumers who say they expect the news they see on social media to be ...*



Note: No answer responses not shown.

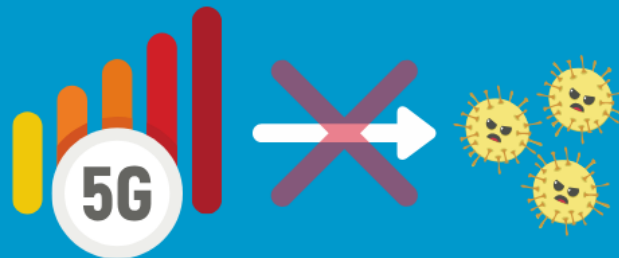
Source: Survey conducted July 30-Aug. 12, 2018.

"News Use Across Social Media Platforms 2018"

# COVID-19 Fake News

Viruses cannot travel on radio waves/mobile networks.  
COVID-19 is spreading in many countries that do not have 5G mobile networks.  
COVID-19 is spread through respiratory droplets when an infected person coughs, sneezes or speaks.  
People can also be infected by touching a contaminated surface and then their eyes, mouth or nose.

**FACT:**  
5G mobile networks  
**DO NOT** spread COVID-19



World Health  
Organization

#Coronavirus #COVID19

8 April 2020



서울시립대학교  
UNIVERSITY OF SEOUL

# Detrimental Societal Effects

## Can 'Fake News' Impact The Stock Market?



**Kenneth Rapoza** Senior Contributor

Markets

*I write about business and investing in emerging markets.*

 This article is more than 3 years old.



1. Fake news significantly weakens the public trust in governments and journalism.
2. Fake news may change the way people respond to legitimate news.
3. Rampant "online" fake news can lead to "offline" societal events.



서울시립대학교  
UNIVERSITY OF SEOUL

# Challenges on Detecting Fake News

1.

It is non-trivial to detect fake news simply based on its content.

2.

Social media data is large-scale, multi-modal, mostly user-generated, sometimes anonymous and noisy.

*Despite the success of existing deep learning based fake news detection method, cannot explain "**why**" a piece of news was detected as fake news.*

# Proposed Method

Explanation from the perspectives of **news contents**  
and **user comments**.

- ▶ News contents may contain information that is verifiably false.
- ▶ User comments have rich information from the crowd on social media, including opinions, stances, and sentiment, that are useful to detect fake news.



# Example of Fake News and Comments

## Sarah Palin Calls To Boycott Mall Of America Because “Santa Was Always White In The Bible”

...  
that comes to mind, many were highly offended. Three years after Fox’s Megyn Kelly definitively explained to America that both Jesus Christ and Santa Claus were white men, Mall of America dismissed her advice and hired Larry Jefferson, a retired

...  
need to run it into the ground, so that they never ever come up with such an offensive and sacrilegious idea again,” Palin added. “The Holy Book always said Santa Claus was white and any kind of deviation from that, regardless of its magnitude, is a sin. And we need to make an example out of Mall of America. If we

User 1

sher · 5 Dec 2016

Gee, did Santa and Jesus hang out and pound down a few beers together?

User 2

Zickler · 6 Dec 2016

St. Nicholas was white? Really?? Lol

User 3

Dean · 6 Dec 2016

FYI, this is false.

User 4

rumps · 7 Dec 2016

OMG, Santa is in the Bible

User 5

llies\_robert · 6 Dec 2016

I wanted that one. I really did.

# Related Work

# Fake News Detection

## ☐ News Contents

**Textual features** capture specific writing styles and sensational emotions.

**Visual features** capture the different characteristics for fake news.

## ☐ Social Contexts

**User-based features** from user profiles to measure their characteristics.

**Post-based features** represent users' social response.

**Network-based features** are extracted by constructing specific networks.

# Explainable Machine Learning

## ☐ Intrinsic Explainability

Constructing **self-explanatory models** which incorporate explainability directly into their structures.

## ☐ Post-hoc Explainability

Creating a **second model** to provide explanation for an existing model.

## ☐ Explanation for DNNs

Focusing on **understanding the representations** captured by neurons at intermediate layers of DNNs.

# Problem Statement

# Challenges

1.

How to perform explainable fake news detection that can improve detection performance and explainability simultaneously.

2.

How to extract explainable comments without the ground truth during training.

3.

How to model the correlation between news contents and user comments jointly for explainable fake news detection.

# Problem Statement

Let  $A$  be a news article, consisting of  $N$  sentences  $\{s_i\}_{i=1}^N$ . Each sentence  $s_i = \{w_1^i, \dots, w_{M_i}^i\}$  contains  $M_i$  words. Let  $C = \{c_1, c_2, \dots, c_T\}$  be a set of  $T$  comments related to the news  $A$ , where each comment  $c_j = \{w_1^j, \dots, w_{Q_j}^j\}$  contains  $Q_j$  words. Similar to previous research [18, 40], we treat fake news detection problem as the binary classification problem, i.e., each news article can be true ( $y = 1$ ) or fake ( $y = 0$ ). At the same time, we aim to learn a rank list  $RS$  from all sentences in  $\{s_i\}_{i=1}^N$ , and a rank list  $RC$  from all comments in  $\{c_j\}_{j=1}^T$ , according to the degree of explainability, where  $RS_k$  ( $RC_k$ ) denotes the  $k_{th}$  most explainable sentence (comment).

# Problem Definition

**Explainable Fake News Detection.** Given a news article  $A$  and a set of related comments  $C$ , learn a fake news detection function  $f: f(A, C) \rightarrow (\hat{y}, RS, RC)$  such that it maximizes prediction accuracy with explainable sentences and comments ranked highest in RS and RC respectively.



# Proposed Framework

# **dEFEND** : Explainable Fake News Detection Framework

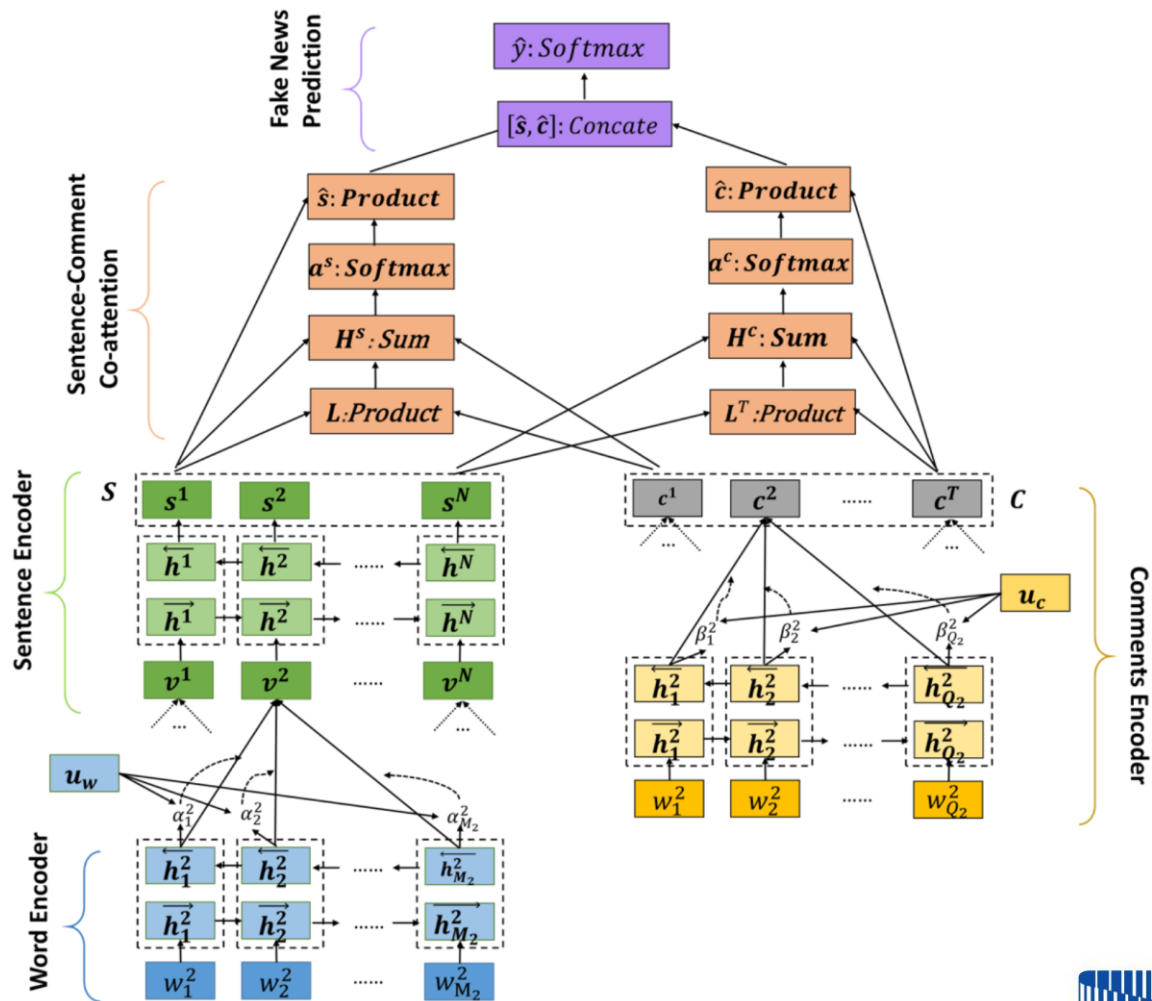
**News content encoder**

**User comment encoder**

**Sentence-comment co-attention**

**Fake news prediction**

# dEFEND



# News Contents Encoding

Fake news pieces often have opinionated and sensational language style, which have the potential to help detect fake news.

With different level(word-level and sentence-level), to provide different degrees of importance for the explainability of wht the news is fake.

*"Pence: Michelle Obama is the most **vulgar** first lady we've ever had."*

(형용사) 저속한, 천박한

# News Contents Encoding

## □ Word Encoder

→ Bidirectional Gated Recurrent Units(GRU)

$$\overrightarrow{\mathbf{h}}_t^i = \overrightarrow{GRU}(\mathbf{w}_t^i), t \in \{1, \dots, M_i\}$$

$$\overleftarrow{\mathbf{h}}_t^i = \overleftarrow{GRU}(\mathbf{w}_t^i), t \in \{M_i, \dots, 1\}$$

$$\Rightarrow \mathbf{h}_t^i = [\overrightarrow{\mathbf{h}}_t^i, \overleftarrow{\mathbf{h}}_t^i]$$

# News Contents Encoding

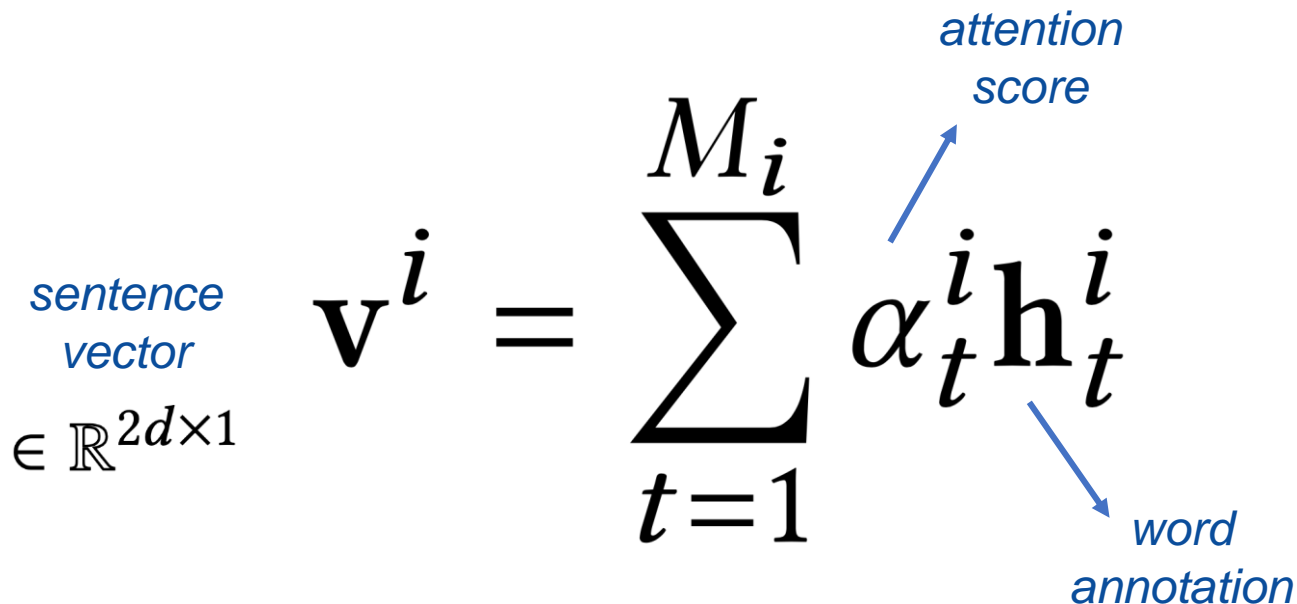
## □ Word Encoder

*sentence vector*  
 $\in \mathbb{R}^{2d \times 1}$

$$\mathbf{v}^i = \sum_{t=1}^{M_i} \alpha_t^i \mathbf{h}_t^i$$

*attention score*

*word annotation*



# News Contents Encoding

## □ Word Encoder

*hidden representation*

$$\mathbf{u}_t^i = \tanh(\mathbf{W}_w \mathbf{h}_t^i + \mathbf{b}_w)$$

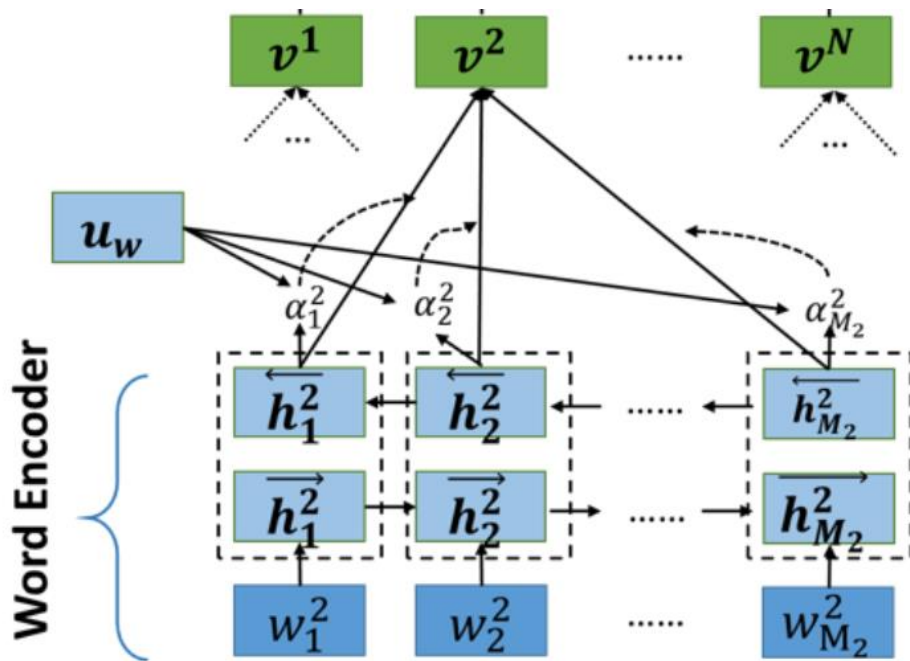
*attention score (importance)*

$$\alpha_t^i = \frac{\exp(\mathbf{u}_t^i \mathbf{u}_w^T)}{\sum_{k=1}^{M_i} \exp(\mathbf{u}_k^i \mathbf{u}_w^T)}$$

*weight parameter*

# News Contents Encoding

## □ Word Encoder





# News Contents Encoding

## □ Sentence Encoder

→ Bidirectional Gated Recurrent Units(GRU)

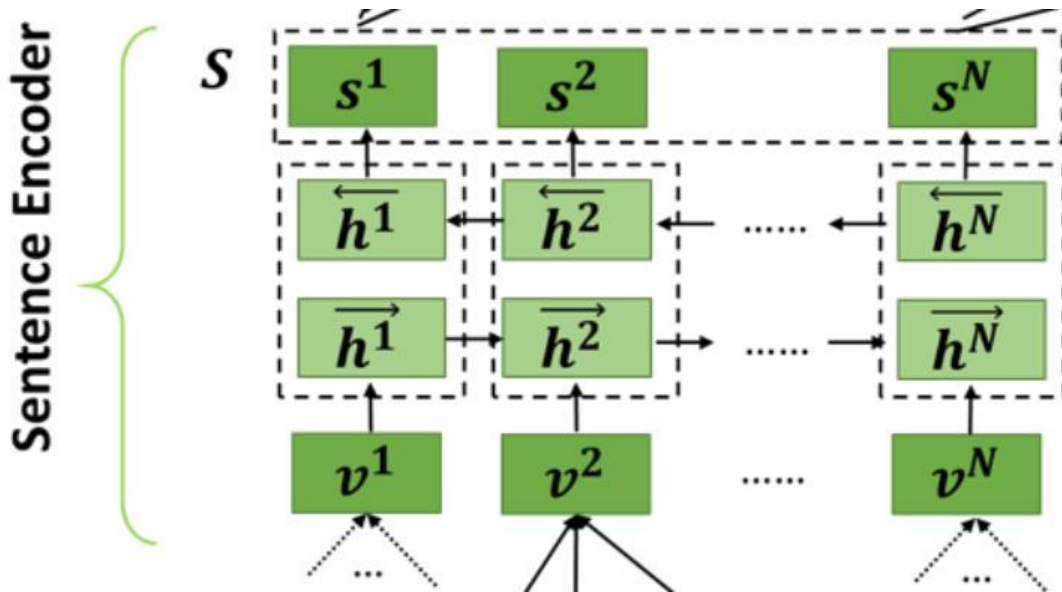
$$\vec{\mathbf{h}}^i = \overrightarrow{GRU}(\mathbf{v}^i), i \in \{1, \dots, N\}$$

$$\overleftarrow{\mathbf{h}}^i = \overleftarrow{GRU}(\mathbf{v}^i), i \in \{N, \dots, 1\}$$

$$\Rightarrow \mathbf{s}^i = [\vec{\mathbf{h}}^i, \overleftarrow{\mathbf{h}}^i]$$

# News Contents Encoding

## □ Sentence Encoder



# User Comments Encoding

Comments may contain useful semantic information to help fake news detection.

$$\vec{h}_t^j = \overrightarrow{GRU}(\mathbf{w}_t^j), t \in \{1, \dots, Q_j\}$$

$$\overleftarrow{h}_t^j = \overleftarrow{GRU}(\mathbf{w}_t^j), t \in \{Q_j, \dots, 1\}$$

$$\Rightarrow \mathbf{h}_t^j = [\vec{h}_t^j, \overleftarrow{h}_t^j]$$

# User Comments Encoding

*comment  
vector*  
 $\in \mathbb{R}^{2d \times 1}$

$$\mathbf{c}^j = \sum_{t=1}^{Q_j} \beta_t^j \mathbf{h}_t^j$$

*attention  
score*

*word  
annotation*

# User Comments Encoding

*hidden representation*

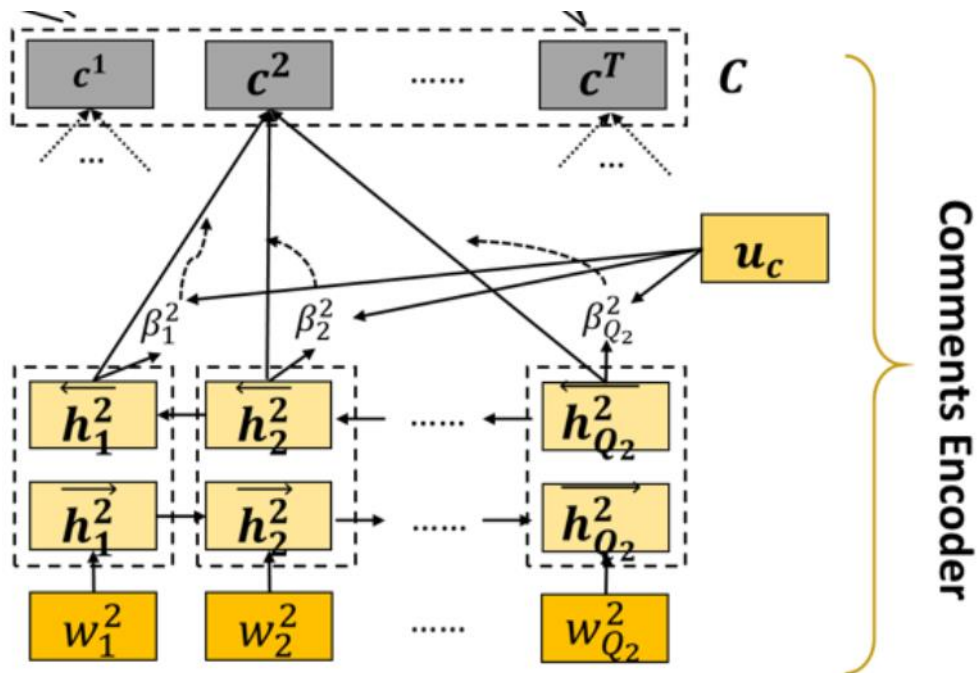
$$\mathbf{u}_t^j = \tanh(\mathbf{W}_c \mathbf{h}_t^j + \mathbf{b}_c)$$

*attention score (importance)*

$$\beta_t^j = \frac{\exp(\mathbf{u}_t^j \mathbf{u}_c^T)}{\sum_{k=1}^{Q^j} \exp(\mathbf{u}_k^j \mathbf{u}_c^T)}$$

*weight parameter*

# User Comments Encoding



# Sentence-Comment Co-attention

To design attention mechanisms to give high weights of representations of news sentences and comments that are beneficial to fake news detection.

*feature matrix  
of news sentences*  $\mathbf{S} = [\mathbf{s}^1; \cdots, \mathbf{s}^N] \in \mathbb{R}^{2d \times N}$

*feature map  
of user comments*  $\mathbf{C} = \{\mathbf{c}^1, \cdots, \mathbf{c}^T\} \in \mathbb{R}^{2d \times T}$

*affinity  
matrix*  $\mathbf{F} = \tanh(\mathbf{C}^\top \mathbf{W}_l \mathbf{S})$   
 $\in \mathbb{R}^{T \times N}$

# Sentence-Comment Co-attention

*sentence  
attention map*  $\mathbf{H}^s = \tanh(\mathbf{W}_s \mathbf{S} + (\mathbf{W}_c \mathbf{C}) \mathbf{F})$

*comment  
attention map*  $\mathbf{H}^c = \tanh(\mathbf{W}_c \mathbf{C} + (\mathbf{W}_s \mathbf{S}) \mathbf{F}^\top)$

*attention probabilities  
of each sentences*  $\mathbf{a}^s = \text{softmax}(\mathbf{w}_{hs}^\top \mathbf{H}^s)$

*attention probabilities  
of each comments*  $\mathbf{a}^c = \text{softmax}(\mathbf{w}_{hc}^\top \mathbf{H}^c)$



# Sentence-Comment Co-attention

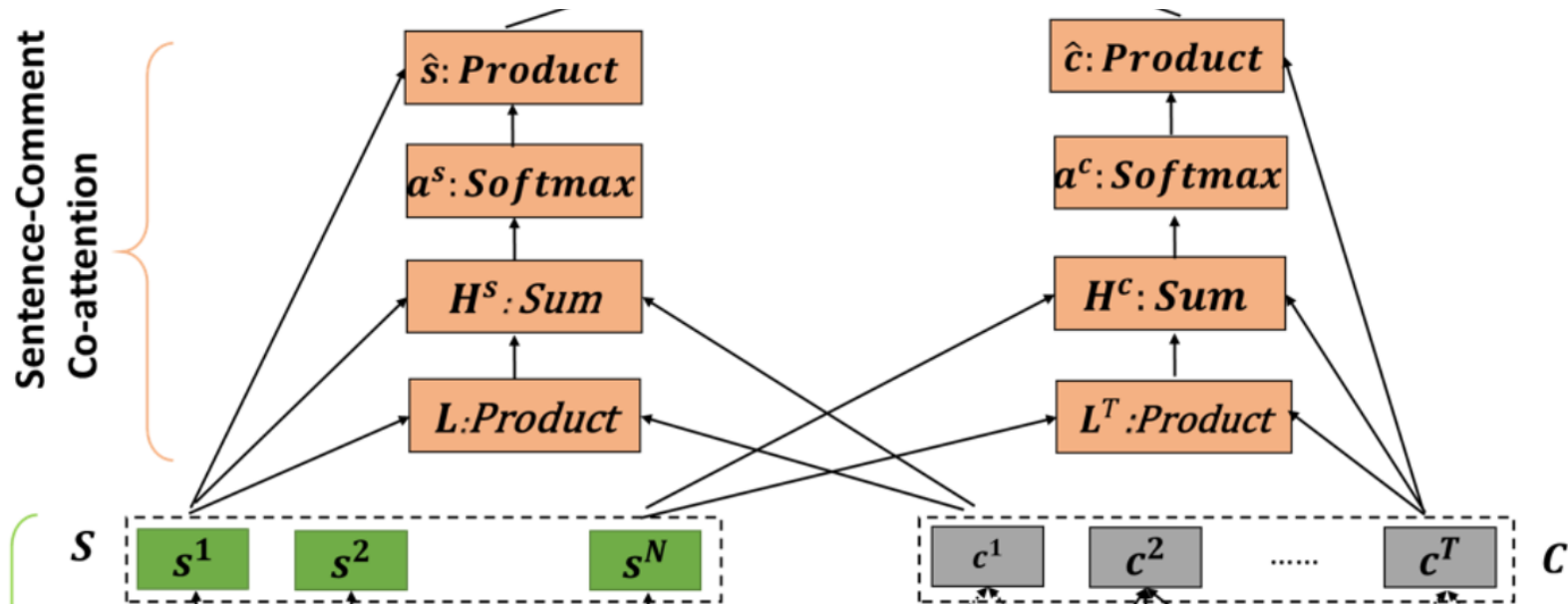
*learned features  
for news sentences*

$$\hat{\mathbf{s}} = \sum_{i=1}^N \mathbf{a}_i^s \mathbf{s}^i$$

*learned features  
for user comments*

$$\hat{\mathbf{c}} = \sum_{j=1}^T \mathbf{a}_j^c \mathbf{c}^j$$

# Sentence-Comment Co-attention



# The Proposed Framework: dEFEND

$$\hat{\mathbf{y}} = \text{softmax}([\hat{\mathbf{s}}, \hat{\mathbf{c}}]\mathbf{W}_f + \mathbf{b}_f)$$

*predicted probability vector*  $\hat{\mathbf{y}} = [\hat{y}_0, \hat{y}_1]$

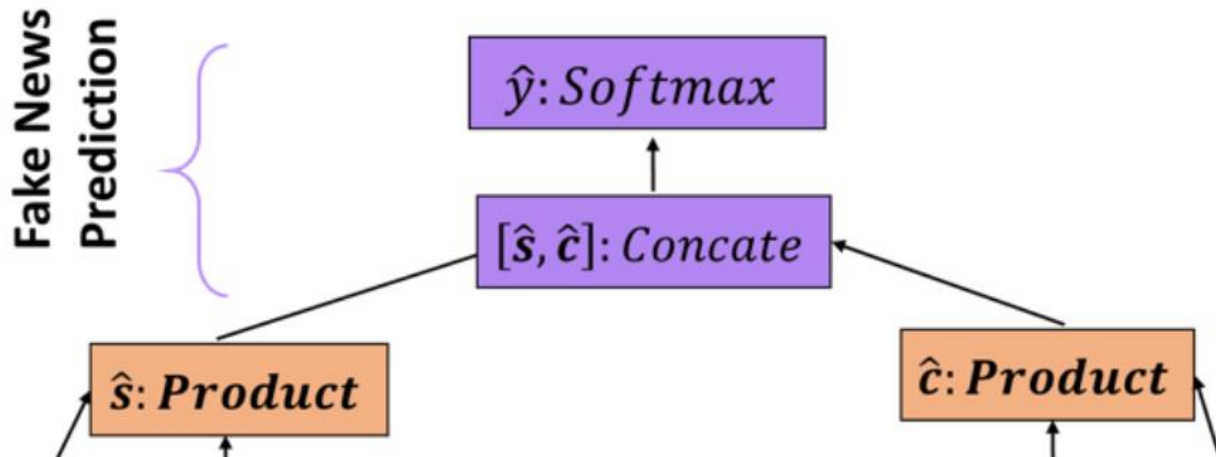
the predicted probability of real news (label 0)

the predicted probability of fake news (label 1)

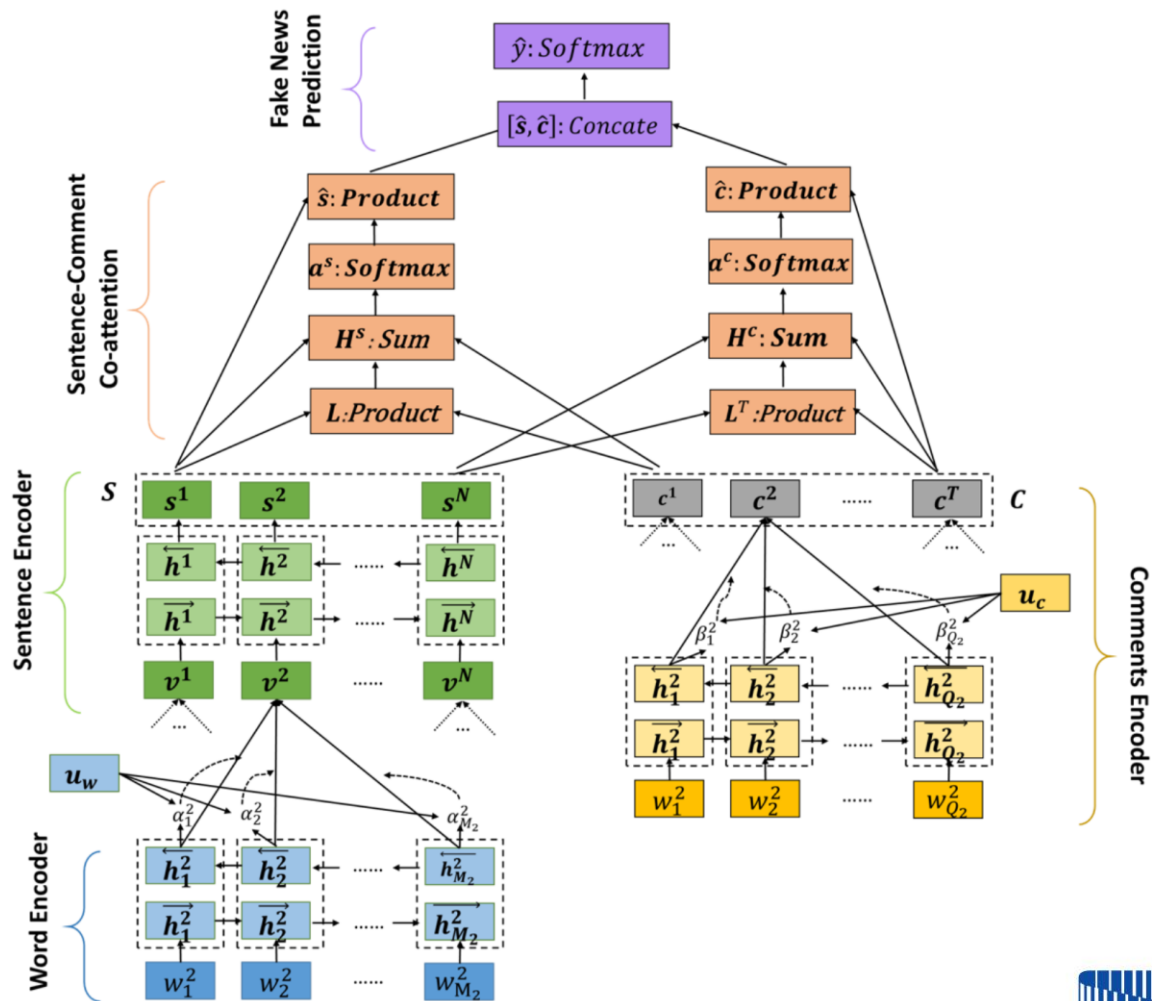
$$\mathcal{L}(\theta) = -y \log(\hat{y}_1) - (1 - y) \log(1 - \hat{y}_0)$$

where  $\theta$  denotes the parameters of the network.

# The Proposed Framework: dEFEND



# dEFEND



# Experiments

# Experimental Questions

EQ1.

Can dEFEND improve fake news classification performance by modeling news contents and user comments simultaneously?

EQ2.

How effective are news contents and user comments, respectively, in improving the detection performance of dEFEND?

EQ3.

Can dEFEND capture the news sentences and user comments that can explain why a piece of news is fake?

# Dataset

**Table 1: The statistics of FakeNewsNet dataset**

Platform	PolitiFact	GossipCop
# Users	68,523	156,467
# Comments	89,999	231,269
# Candidate news	415	5,816
# True news	145	3,586
# Fake news	270	2,230



# Compared Fake News Detection Methods

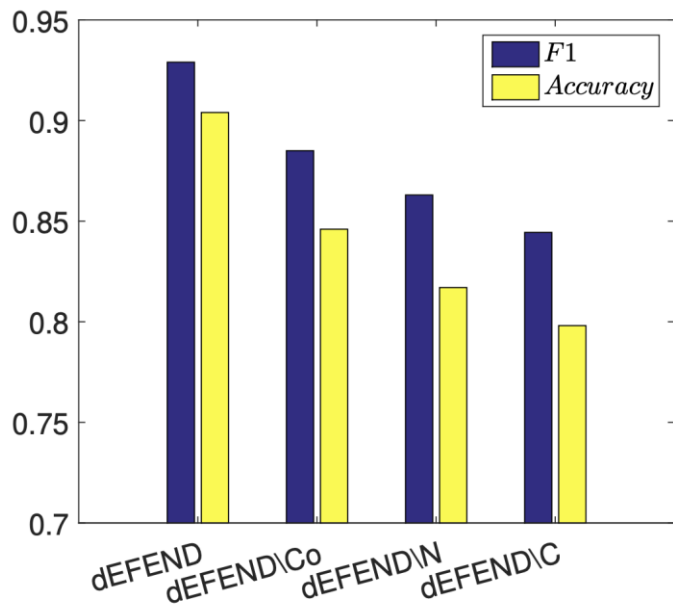
- **RST** [36]: RST stands for Rhetorical Structure Theory, which builds a tree structure to represent rhetorical relations among the words in the text. RST can extract news style features by mapping the frequencies of rhetorical relations to a vector space<sup>7</sup>.
- **LIWC** [32]: LIWC stands for Linguistic Inquiry and Word Count, which is widely used to extract the lexicons falling into psycholinguistic categories. It learns a feature vector from psychology and deception perspective<sup>8</sup>.
- **HAN** [50]: HAN utilizes a hierarchical attention neural network framework on news contents for fake news detection. It encodes news contents with word-level attentions on each sentence and sentence-level attentions on each document.
- **text-CNN** [23]: text-CNN utilizes convolutional neural networks to model news contents, which can capture different granularity of text features with multiple convolution filters.
- **TCNN-URG** [35]: TCNN-URG consists of two major components: a two-level convolutional neural network to learn representations from news content, and a conditional variational auto-encoder to capture features from user comments.
- **HPA-BLSTM** [13]: HPA-BLSTM is a neural network model that learns news representation through a hierarchical attention network on word-level, post-level, and sub-event level of user engagements on social media. In addition, post features are extracted to learn the attention weights during post-level.
- **CSI** [37]: CSI is a hybrid deep learning model that utilizes information from text, response, and source. The news representation is modeled via an LSTM neural network with the Doc2Vec [25] embedding on the news contents and user comments as input, and for a fair comparison, the user features are ignored.

# Fake News Detection Performance

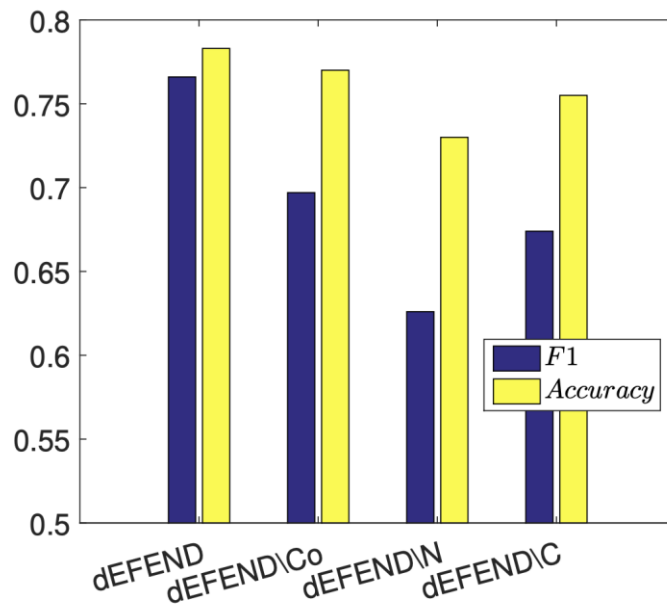
Datasets	Metric	RST	LIWC	text-CNN	HAN	TCNN-URG	HPA-BLSTM	CSI	dEFEND
<b>PolitiFact</b>	Accuracy	0.607	0.769	0.653	0.837	0.712	0.846	0.827	<b>0.904</b>
	Precision	0.625	0.843	0.678	0.824	0.711	0.894	0.847	<b>0.902</b>
	Recall	0.523	0.794	0.863	0.896	0.941	0.868	0.897	<b>0.956</b>
	F1	0.569	0.818	0.760	0.860	0.810	0.881	0.871	<b>0.928</b>
<b>GossipCop</b>	Accuracy	0.531	0.736	0.739	0.742	0.736	0.753	0.772	<b>0.808</b>
	Precision	0.534	<b>0.756</b>	0.707	0.655	0.715	0.684	0.732	0.729
	Recall	0.492	0.461	0.477	0.689	0.521	0.662	0.638	<b>0.782</b>
	F1	0.512	0.572	0.569	0.672	0.603	0.673	0.682	<b>0.755</b>

\* train 75% / test 25%, average performance of 5 times experiments.

# Impacts of News Contents and User Comments



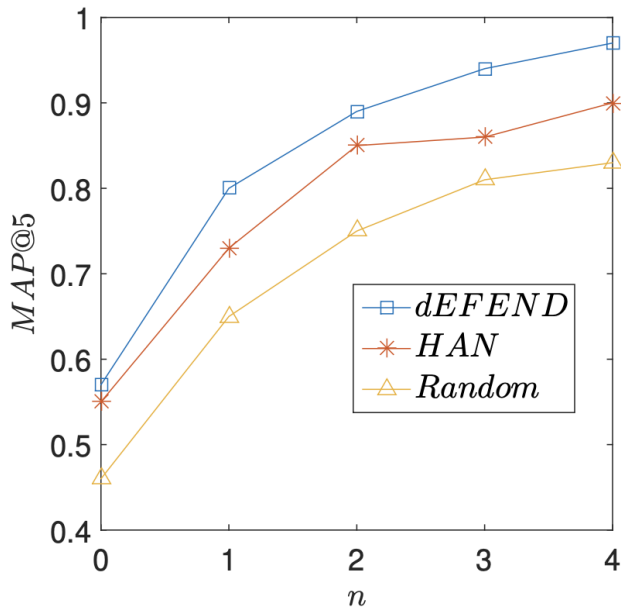
(a) PolitiFact



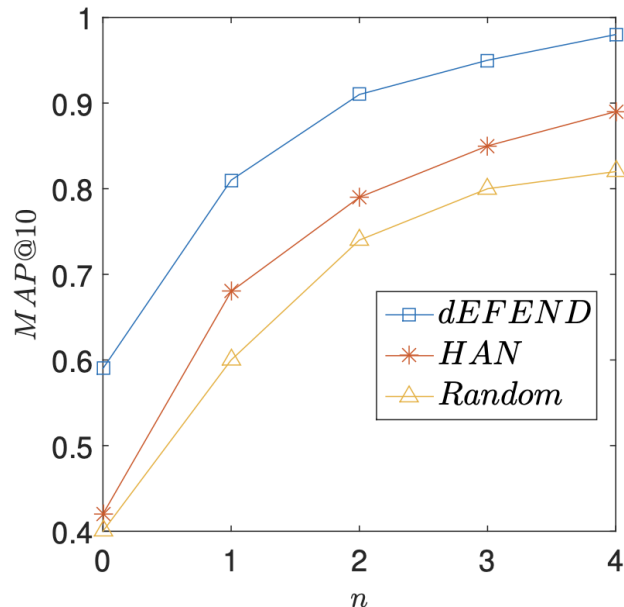
(b) GossipCop

# Explainability Evaluation

## □ News Sentence Explainability



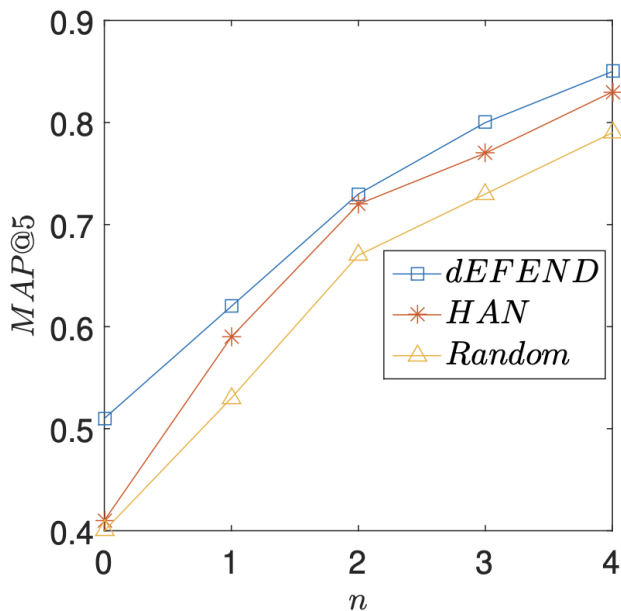
(a) MAP@5 on PolitiFact



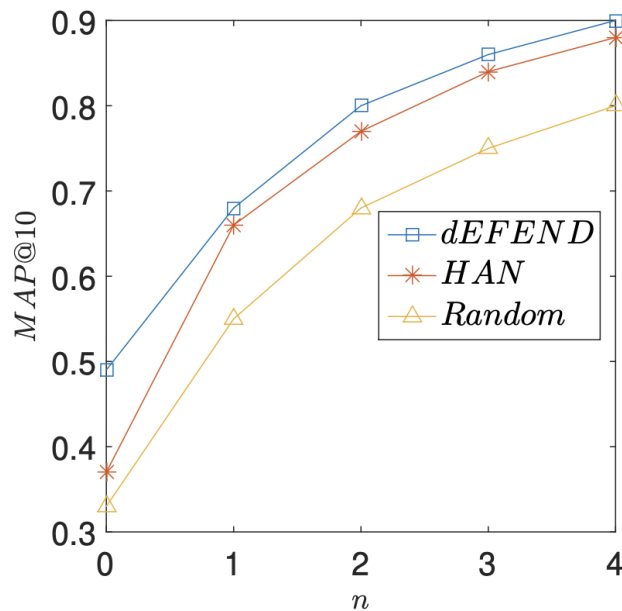
(b) MAP@10 on PolitiFact

# Explainability Evaluation

## □ News Sentence Explainability



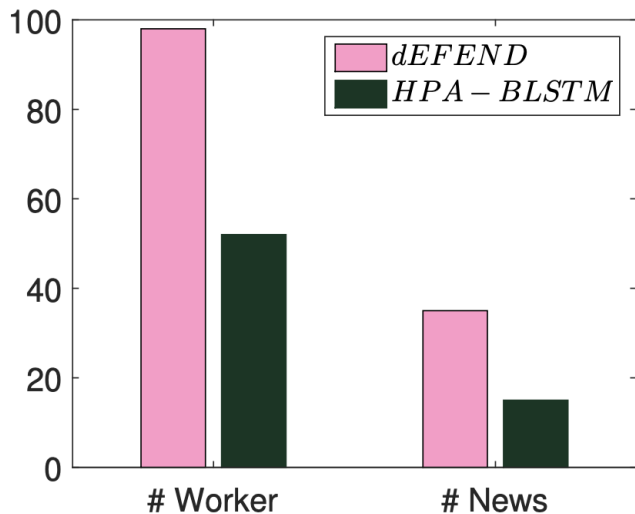
(c) MAP@5 on GossipCop



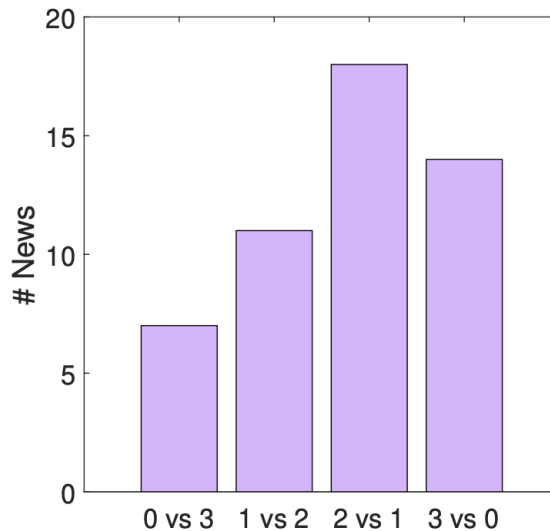
(d) MAP@10 on GossipCop

# Explainability Evaluation

## □ User Comments Explainability



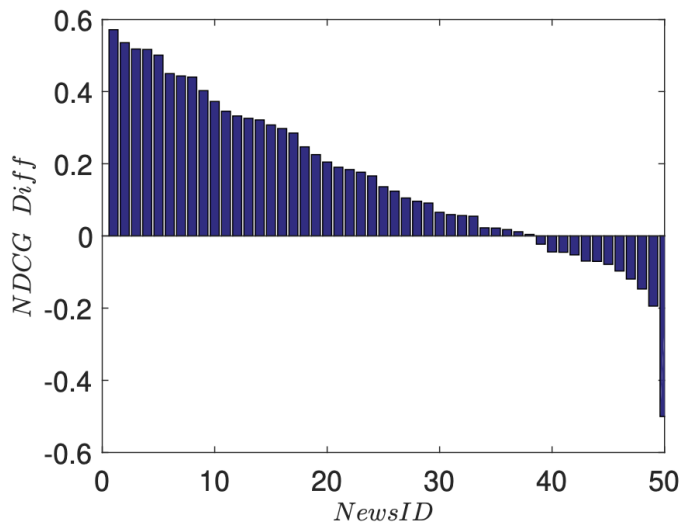
(a) Winning Count



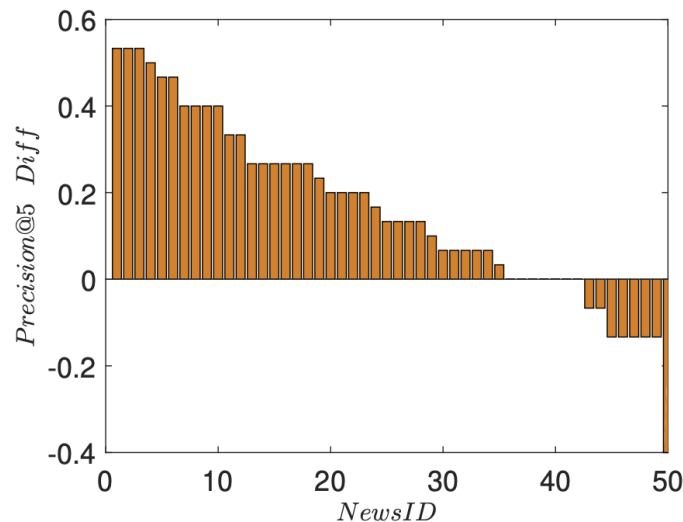
(b) Worker Voting Ratio (*dEFEND* vs. *HPA-BLSTM*)

# Explainability Evaluation

## ❑ User Comments Explainability



(a) NDCG



(b) precision@5

# Case Study

## Fake News

### Iranian Official Drops Bombshell: Obama Secretly Gave Citizenship to 2500 Iranians as Part of Nuke Deal

By [redacted] - July 2, 2018

148 Comments

A senior Iranian cleric and member of parliament has just dropped a bombshell.

He is claiming that the Obama administration, as part of negotiating during the Iran Deal, granted U.S. citizenship to 2500 Iranians including family members of government officials.

...

There have been so many things hidden from the public about the Iran Deal if this was one more thing given up in bribe, it wouldn't be hard to believe.

## Comments

If you had done your research, you would know that the president does not have the power to give citizenship. This would have to be done as an act of congress... (0.0160)

Isn't graft and payoffs normally an offense even for an ex-president? (0.0086)

Wow! What's frightening is where will it end? We could be seeing some serious issues here. (0.0051)

Walk away from them (0.0080)