Entity-Aware Dual Co-Attention Network for Fake News Detection

Sin-Han Yang [1], **Chung-Chi Chen** [2], **Hen-Hsen Huang** [3], **Hsin-Hsi Chen** [1]

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] AIST, Japan
[3] Institute of Information Science, Academia Sinica, Taiwan
b08202029@ntu.edu.tw, c.c.chen@acm.org
hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

Fake news and misinformation spread rapidly on the Internet. How to identify it and how to interpret the identification results have become important issues. In this paper, we propose a Dual Co-Attention Network (Dual-CAN) for fake news detection, which takes news content, social media replies, and external knowledge into consideration. Our experimental results support that the proposed Dual-CAN outperforms current representative models in two benchmark datasets. We further make in-depth discussions by comparing how models work in both datasets with empirical analysis of attention weights.[1]

1 Introduction

The development of the Web and social media platforms helps us obtain news quickly, but also provides a gateway for spreading false information. The impact of false information is wide, and the spread speed might be even faster than the actual one (Vosoughi et al., 2018). For example, fake news is proven empirically to influence the 2016 U.S. presidential election (Bovet and Makse, 2019; Grinberg et al., 2019; Budak, 2019). Given the impact of false information, previous studies paid a lot of effort to detect it from different aspects, including (1) news content only (Santos et al., 2020; Kim and Ko, 2021), (2) the combination of news articles and social media replies (Li et al., 2020; Lu and Li, 2020), and (3) additional publisher/user information (Long et al., 2017; Yuan et al., 2020; Del Tredici and Fernández, 2020). In this work, we focus on using both news contents and social media replies, and further add external knowledge to enhance the model's ability to capture critical entities.

Named entities play an important role in document understanding and influence text generation

performances (Narayan et al., 2021, 2022). Inspired by this notion, we design a novel model, named Dual Co-Attention Network (Dual-CAN), which takes entities' descriptions into consideration to enhance the background knowledge of the model. The proposed Dual-CAN is modified based on one of the representative fake news detection models, dEFEND (Shu et al., 2019a). There are three major improvements in the proposed Dual-CAN: (1) Inspired by Hu et al. (2021), we add entities' descriptions for enhancing the performance. (2) Instead of using LSTM-based architectures (Shu et al., 2019a; Lu and Li, 2020), we adopt attention architecture (Vaswani et al., 2017) as the backbone. (3) We further tailor-made a co-attention layer for comparing the given news article with entity descriptions. In sum, in addition to adopting entity descriptions from Wikipedia, we design a new architecture to fusion all information. Our main contribution is providing a novel model for fake news detection and pointing out a new direction for enhancing performance.

2 Related Works

Previous works in fake news detection mainly focused on two aspects: news content based and social context based. Rashkin et al. (2017) focus on the linguistic characteristics of the news content to detect fake news, and find that fake news often contain specific kinds of words. Ma et al. (2016) use recurrent neural networks (RNN) to learn the hidden representations from the contextual information of relevant posts over time. Monti et al. (2019) analyze social graph and user profile to predict fake news. Shu et al. (2019b) find that user profile features are useful in fake news detection. Shu et al. (2019a) and Lu and Li (2020) use co-attention model to leverage news content and social context. Their models not only have better performance but also provide interpretability to their models. Several works also use external

---

[1] Code repository: https://github.com/SinHanYang/Dual-CAN

knowledge to improve model's predictions. Wang et al. (2020) and Hu et al. (2021) use entity linking method to capture entity descriptions and leverage them in their models. Inspired by these works, we use external knowledge for entities to enhance performance, and use both news content and social media context in the proposed model.

3 Method

Figure 1 shows the architecture of the proposed Dual-CAN. This section describes the details of the proposed Dual-CAN model, which is composed of five components.[2] The first one is *news content encoder*, which employs word-level attention network and sentence-level encoder to generate features for the corresponding news contents. The second is *entity description encoder*. For each entity in news content, entity description encoder grabs its descriptions from the external knowledge base and creates features to represent them. The third is *user engagement encoder*, which employs the same method as *news content encoder* to create features to represent user comments. The fourth is *dual co-attention component*, which captures the relation between (news content, entity description) and (news content, user engagement) pairs. The last is *prediction component*, which combines all information from the previous components to make the final predictions.

3.1 News Content Encoder

A news story is composed of a sequence of sentences $\mathbf{S} = [\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_N}]$, and a sentence is composed of up to M words $\mathbf{s_i} = [\mathbf{w_{i1}}, \mathbf{w_{2i}}, ..., \mathbf{w_{iM}}]$. Here, N is the maximum number of sentences in a piece of news, and M is the maximum number of words in a sentence. We perform padding to control the maximum number of sentences and words in news content. To create features to represent a news story, we use *word-level attention network* to encode each sentence, and use *sentence-level encoder* to encode all sentences in news content.

3.1.1 Word-Level Attention Network

We use Glove (Pennington et al., 2014) to create word embedding of d dimensions during the pre-processing stage for each word in sentences. For a sentence $\mathbf{s} \in \mathbb{R}^{d \times M}$, we utilize bi-directional Gating Recurrent Units (GRU) (Chung et al., 2014) to learn the word-level representation. The output
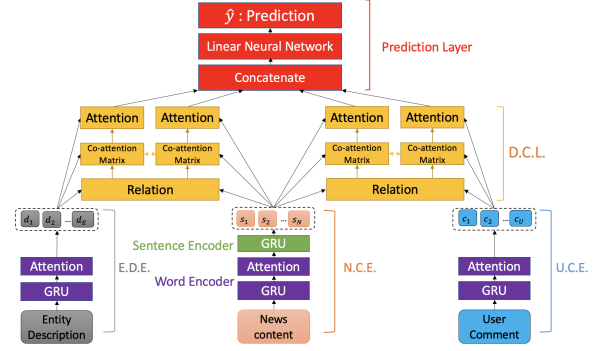
Figure 1: Architecture of Dual-CAN. D.C.L., E.D.E., N.C.E., and U.C.E. stand for dual co-attention layer, entity description encoder, news content encoder, and user comment encoder, respectively.

of the BiGRU is $\mathbf{v_i} = BiGRU(w_i) \in \mathbb{R}^{2h}, i \in \{1, 2, ...M\}$, where h is the dimension of the GRU. Next, we perform the basic attention mechanism to increase performance and interpretability (Lu and Li, 2020) of the word encoder. Attention weight α_i shows the importance of the ith word. The word-level attention network generates the representation of a sentence vector $\mathbf{v}' \in R^{2h \times 1}$ calculated as follows:

$$\mathbf{v}' = \sum_{i=1}^{M} \alpha_i \mathbf{v}_i \qquad (1)$$

where a_i is:

$$\mathbf{k_i} = tanh(\mathbf{P_n v_i} + \mathbf{b_n})$$
$$\alpha_i = \frac{exp(\mathbf{u_n k_i})}{\sum_{j=1}^{M} exp(\mathbf{u_n k_j})} \qquad (2)$$

$\mathbf{P_n} \in \mathbb{R}^{2h \times h}$, $\mathbf{u_n} \in \mathbb{R}^{h \times 1}$ are learnable parameter. We preform a linear layer on $\mathbf{v}_i$, and use a parameter $\mathbf{k}_j$ to calculate the attention weight.

3.1.2 Sentence-Level Encoder

We use BiGRU again to encode sentences in a news story. A sentence vector $\mathbf{s_i} \in R^{2h \times 1}$ is calculated from the output of *word-level attention network*:

$$\mathbf{s_i} = BiGRU(\mathbf{v'_i}), i \in \{1, 2, ..., N\} \qquad (3)$$

Finally, single news content is represented by a list of sentence vectors $\mathbf{S} = [\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_N}] \in \mathbb{R}^{2h \times N}$.

3.2 Entity Description Encoder

For each news content, we identify entities in it and grab their descriptions from Wikipedia using tools TAGME (Ferragina and Scaiella, 2010). For each entity description, we only use the first E

sentences. With the word-level attention network in Section 3.1.1, we create features that describe entity descriptions $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_E}]$. Finally, entity descriptions for a piece of news is represented by a list of sentence vectors $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_E}] \in \mathbb{R}^{2h \times E}$.

3.3 User Comment Encoder

For all user comments related to a news story, we only use the first U sentences. We extract features to describe user comments $\mathbf{C} = [\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_U}]$ with the word-level attention network in Section 3.1.1. Finally, user comments for a news story are represented by a list of sentence vectors $\mathbf{C} = [\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_U}] \in \mathbb{R}^{2h \times U}$.

3.4 Dual Co-Attention Component

Because we want to know whether the entity description confirms/refutes the news content and whether user comments reflect the character of the news content, we adopt co-attention network for capturing the relationship between news content and entity descriptions, and another co-attention network for linking the relationship between news content and user comments. Given news content feature vectors $\mathbf{S} = [\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_N}] \in \mathbb{R}^{2h \times N}$, entity description feature vectors $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_E}] \in \mathbb{R}^{2h \times E}$, and user comments feature vectors $\mathbf{C} = [\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_U}] \in \mathbb{R}^{2h \times U}$, we use *dual co-attention mechanism* for interpreting model predictions.

3.4.1 Entity Description Co-attention

First, we compute a relation matrix $\mathbf{F}$

$$\mathbf{F} = tanh(\mathbf{D}\mathbf{W_r}\mathbf{S}) \in \mathbb{R}^{E \times N} \qquad (4)$$

to capture the relationship between news content and entity descriptions, where $\mathbf{W_r} \in \mathbb{R}^{2h \times 2h}$ is a learnable parameter. Second, we calculate interaction maps for news content H_s and entity description H_c,

$$\begin{aligned} \mathbf{H_s} &= tanh(\mathbf{W_s}\mathbf{S} + \mathbf{W_d}\mathbf{D}\mathbf{F}^T) \\ \mathbf{H_d} &= tanh(\mathbf{W_d}\mathbf{D} + \mathbf{W_s}\mathbf{S}\mathbf{F}) \end{aligned} \qquad (5)$$

where $\mathbf{W_s}, \mathbf{W_d} \in \mathbb{R}^{2h \times 2h}$ are learnable parameters. Third, we calculate attention weights on each sentence in news content and entity descriptions.

$$\begin{aligned} \mathbf{a_{s_1}} &= softmax(\mathbf{w_{hs}}\mathbf{H_s}) \\ \mathbf{a_d} &= softmax(\mathbf{w_{hd}}\mathbf{H_d}) \end{aligned} \qquad (6)$$

where $\mathbf{w_{hs}}$ and $\mathbf{w_{hd}} \in \mathbb{R}^{1 \times 2h}$ are learnable parameters. After we get attention weights $\mathbf{a_{s_1}} \in \mathbb{R}^{1 \times N}, \mathbf{a_d} \in \mathbb{R}^{1 \times E}$, we generate new feature vectors for news contents and entity descriptions:

$$\begin{aligned} \hat{\mathbf{s_1}} &= \mathbf{a_{s_1}}\mathbf{S}^T \\ \hat{\mathbf{d}} &= \mathbf{a_d}\mathbf{D}^T \end{aligned} \qquad (7)$$

Finally, we represent news content in a feature vector $\hat{\mathbf{s_1}} \in \mathbb{R}^{1 \times 2h}$, and entity descriptions in a feature vector $\hat{\mathbf{d}} \in \mathbb{R}^{1 \times 2h}$.

3.4.2 User Comment Co-attention

We apply co-attention model as shown in Section 3.4.1 to news content and user comments. We represent news content in a feature vector $\hat{\mathbf{s_2}} \in \mathbb{R}^{1 \times 2h}$, and user comments in a feature vector $\hat{\mathbf{c}} \in \mathbb{R}^{1 \times 2h}$. The attention weights vector for news content and user comments are $\mathbf{a_{s_2}} \in \mathbb{R}^{1 \times N}$ and $\mathbf{a_c} \in \mathbb{R}^{1 \times U}$.

3.5 Prediction Component

Our task is a binary classification task with real/fake labels. First, we concatenate all feature vectors $\mathbf{f} = [\hat{\mathbf{s_1}}, \hat{\mathbf{d}}, \hat{\mathbf{s_2}}, \hat{\mathbf{c}}]$, and feed the result into a 2-layer linear neural network. It is calculated by:

$$\hat{\mathbf{y}} = \mathbf{W_2}(\mathbf{W_1}\mathbf{f} + \mathbf{b_1}) + \mathbf{b_2} \qquad (8)$$

where $\mathbf{W_1}$ and $\mathbf{W_2}$ are learnable parameters and $\mathbf{b_1}, \mathbf{b_2}$ are bias terms. The prediction result $\hat{\mathbf{y}} = [y_0, y_1]$ indicates the probabilities of label 0 is y_0, and label 1 is y_1. We choose cross entropy as our loss function:

$$\mathcal{L}(\theta) = -ylog(\hat{y_1}) - (1-y)log(1-\hat{y_0}) \qquad (9)$$

where θ is all parameters in our model. We choose Adam optimizer (Kingma and Ba, 2014) to optimize all parameters θ.

4 Experiments

4.1 Datasets

We adopt two datasets in our experiment. The first dataset is GossipCop (Shu et al., 2018), which collects both news content and social context from fact-checking website. The second dataset is CoAID (Cui and Lee, 2020), which is a benchmark dataset for COVID-19 misinformation. Please refer to Appendix A for the statistics of the datasets. We follow the evaluation settings as previous studies (Shu et al., 2018; Cui and Lee, 2020) to use (Accuracy, F1, Precision, Recall) for *GossipCop* and use (PR-AUC, F1, Precision, Recall) for *CoAID*.

Model (Input) (# of Parameters)	GossipCop					CoAID		
	Accuracy	F1	Precision	Recall	PR-AUC	F1	Precision	Recall
BiGRU (N+C+E) (28M)	0.580	0.367	0.290	0.500	0.876	0.782	0.769	0.804
BERT (N+C+E) (339M / 110M)	0.787	0.776	0.787	0.771	0.940	0.877	0.901	0.859
RoBERTa (N+C+E) (384M / 125M)	0.894	0.890	0.896	0.887	0.918	0.877	0.901	0.859
LinkBERT (N+C+E) (330M / 110M)	0.824	0.811	0.841	0.802	0.927	0.880	0.903	0.863
dEFEND (N+C) (5M)	0.771	0.758	0.771	0.754	0.749	0.799	0.792	0.808
Dual-CAN (N+E) (33M)	0.895	0.891	0.901	0.885	0.853	0.884	0.905	0.868
Dual-CAN (N+C) (33M)	0.914	0.912	0.913	0.911	0.937	**0.887**	**0.907**	**0.872**
Dual-CAN (N+C+E) (33M)	**0.949**	**0.947**	**0.946**	**0.949**	**0.954**	0.884	0.905	0.868

Table 1: Experimental results. N, C, and E denote news content, user comments, and entity description, respectively. BERT-based models are implemented in two methods (details in Appendix B) with different number of parameters.

4.2 Results

We compare the results with the following representative models: BiGRU (Chung et al., 2014), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), LinkBERT (Yasunaga et al., 2022)), and dEFEND (Shu et al., 2019a).[3] Table 1 shows our experimental results. Our Dual-CAN outperforms all baselines in both datasets. In addition, our Dual-CAN uses fewer parameters than BERT-based models. Our approach also performs better than dEFEND (Shu et al., 2019a) when no entity descriptions are provided. This is because we use different preprocessing methods, and the differences between two model architectures. The bottom half of Table 1 shows ablation analysis of the proposed model. The results indicate the importance of adding entity information to the proposed model, especially in *GossipCop*. However, only a few improvements in PR-AUC when experimenting with *CoAID*. *CoAID* usually are short posts that contain few entities, which results in the limitation of the proposed entity-aware concept. The main source to predict whether a piece of news is fake is the news content itself. Therefore, N+E, N+C, and N+C+E results only have small differences because they both contain N. The roles of C and E are to improve the predictions.

5 Interpretability

We examine attention weights $[\mathbf{a_{s_1}}, \mathbf{a_d}, \mathbf{a_{s_2}}, \mathbf{a_{s_c}}]$ to find those sentences that the proposed model is focusing on when making predictions. Figure 2 illustrates the results. We find that our model pays a certain degree of attention to the first sentence in the entity descriptions of both datasets (Figure 2a,2c).
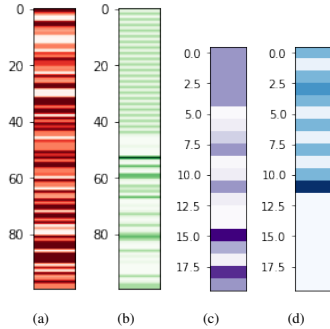

Figure 2: Attention weights of: (a) *GossipCop* entity description, (b) *GossipCop* user comments, (c) *CoAID* entity description, and (d) *CoAID* user comments. Dark colors means higher attention weights. The vertical axis means the index of the sentence.

Our intuition about this phenomenon is that the first sentence always provides a brief definition of the entity, and it would be helpful for models to understand the given entity. On the other hand, model's attention weights on user comments of both datasets are in the middle replies, as shown in Figure 2b and Figure 2d. It follows our intuition because the sentences like "FYI. It's a fake news." for clarifying the given news/post is fake news always appears later than some discussions. Based on Figure 2d, we also find that models give little attention weight to the twelfth or later sentences. Besides weight distributions studies, we also did some case studies in Appendix C. The results show that attention weights do reflect the important parts of the input, which help us interpret the model better. For example, we understood the importance and usage of entity descriptions from attention weights.

6 Conclusion

We propose a dual co-attention network for fake news detection, which improves the previous representative model, dEFEND, by (1) adding entity

---

[3]Because Shu et al. (2019a) did not release the information for dataset separation, we use the same hyperparameter reported in their work to reproduce the results. All implemental details are provide in Appendix B

description as external knowledge and (2) redesigning co-attention architecture for using all input information. Our results support the usefulness of the proposed Dual-CAN model. The interpretability based on the attention weight is also discussed.

Limitations

The major limitation of the proposed model is that when the given text (news article or social media post) is short, and the performance of adding entity description may not be significantly improved. It is because such text provides few entities in the narrative, and it will limit the proposed entity-aware concept.

Ethical Statement

We will follow the licenses of GossipCop (Shu et al., 2018) and CoAID (Cui and Lee, 2020) to share the training, development, and test datsets in our experiments.

Acknowledgments

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Ceren Budak. 2019. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*, pages 139–150.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Limeng Cui and Dongwon Lee. 2020. CoAID: Covid-19 healthcare misinformation dataset.

Marco Del Tredici and Raquel Fernández. 2020. Words are the window to the soul: Language-based user representations for fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5467–5479, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1625–1628, New York, NY, USA. Association for Computing Machinery.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Gihwan Kim and Youngjoong Ko. 2021. Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3276–3280, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 505–514, Online. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Measuring the impact of readability features in fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1404–1413, Marseille, France. European Language Resources Association.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

Datasets	GossipCop		CoAID	
Word Embedding	Glove 100d	Glove 300d	Glove 100d	Glove 300d
Max sentence length	120		120	
Max sentence number per news N	40		4	
Max sentence number per entity description	4		4	
Max sentence number per user comment	2		2	
Max sentence number of total entity description E	100		20	
Max sentence number of total user comment U	100		20	
Embedding dimension	100	300	100	300
h	100	300	100	300
Batch size	16		32	
learning rate	0.001		0.001	

Table 2: Model parameters.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019b. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. volume 30.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2020. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Dataset

Table 3 reports the statistics of the datasets. We will release the datasets for reproduction, and follow the same license of GossipCop

Datasets	GossipCop	CoAID
Total news	4,273	2,162
True news	2,562	1,590
Fake news	1,711	572
User Comments	309,059	37,187
Entity Descriptions	95,150	5,666

Table 3: Dataset statistics.

and CoAID. Due to the size limits, we cannot upload the dataset via the submission system. Please download it via the following anonymous link: https://drive.google.com/file/d/1QuZeINFHqy8OF1Av5627zTyyVVg7g2HD/view?usp=sharing.

B Implementation Detail

Below are the implementation details of the baseline models:

- **BiGRU** (Chung et al., 2014): We use Glove 300d for word embedding of news content, entity descriptions and user comments. The word embedding of three resources are feed into BiGRU and concatenate their results $\mathbf{T} = [v_n, v_d, v_c]$. Second, we feed $\mathbf{T}$ into linear neural network described in Section 3.5 to get final result.

- Pretrained language models (**BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), **LinkBERT** (Yasunaga et al., 2022)): We adopt three representative pretrained language models for comparison, and implemented in two different ways.

  - **Method 1**
    For each experiment, we feed news content, entity descriptions, and user comments into three tokenizers respectively. Afterwards, we feed input id and attention masks of each resource into three pretrained language models respectively. Each pretrained model handles one resource. Finally, we concatenate the outputs of three pretrained models, and pass a linear layer to output probability of two labels $\hat{y}$. This method is used for *GossipCop* dataset in Table 1.

  - **Method 2**
    The number of parameters in **Method 1** is huge, but it's necessary for *GossipCop* dataset. We tried another method

to reduce the number of parameters. We concatenate all three resources and feed into one tokenizer. Second, we feed input id and attention masks into one pretrained models. The final procedure is same as the previous method. The experiment results for *CoAID* dataset are in Table 1, and the results are better than **Method 1**'s. The experiment results for *GossipCop* dataset is in Table 4. The performances are worse than **Method 1**'s. We believe it's because *GossipCop* dataset's data are too long for a single pretrained model. Therefore, we tried **Longformer** (Beltagy et al., 2020) which accept longer input. The performance becomes better, but this methods uses more parameters.

- **dEFEND** (Shu et al., 2019a): dEFEND is one of the representative fake news detection methods. It is based on co-attention model to increase explanability.[4]

Table 2 reports the hyperparameters used in the proposed Dual-CAN. In the ablation study, we remove the original data of entity description E or user comments C, and replace them with padding token <PAD>. Therefore, the model architecture remains the same as Section 3 stated. We have submitted the code for review, and it will be released on GitHub.

C Case Study of Interpretability

We analyzed individual sentences and words which have higher attention weight, in order to figure out the explanability of the attention weight.

For sentence-level analysis, entity descriptions that define an entity would have higher attention weights. Here are two example entity descriptions that have higher attention weights:

1. **{Dataset: GossipCop, id: 587, attention weight: 0.036 >average 0.01}**: *"IMDb (an abbreviation of Internet Movie Database) is an online database of information related to films, television series, home videos,. . . "*

---

[4]Because Shu et al. (2019a) did not release the information for dataset separation, we use the same hyperparameter reported in their work to reproduce the results. We will release the datasets for reproduction.

Model (Input) (# of Parameters)	GossipCop			
	Accuracy	F1	Precision	Recall
BERT (N+C+E) (110M)	0.643	0.587	0.645	0.599
RoBERTa (N+C+E) (125M)	0.698	0.631	0.774	0.646
LinkBERT (N+C+E) (110M)	0.702	0.694	0.694	0.693
Longformer (N+C+E) (148M)	0.752	0.742	0.758	0.723

Table 4: **Method 2** experiment results of GossipCop dataset. N, C, and E denote news content, user comments, and entity description, respectively.

2. **{Dataset: CoAID, id: 48, attention weight: 0.182 >average 0.05}**: *"The Centers for Disease Control and Prevention (CDC) is the national public health agency of the United States."*

3. **{Dataset: CoAID, id: 1304, attention weight: 0.119 > average 0.05}**: *"Getty Images, Inc. is a British-American visual media company and is a supplier of stock images, editorial photography, video and music for business and consumers, with a library of over 477 million assets."*

Moreover, we can see some correlation between highlighted entity descriptions and news content that contain them. For example, the news sentence which contains entity (2,3), both have higher attention weight than average.

1. **{Dataset: CoAID, id: 48, attention weight: 0.33 >average 0.25}**: *"enters for disease control and prevention, cdc twenty four seven, saving lives protecting people centers for disease control and prevention"*

2. **{Dataset: CoAID, id: 1304, attention weight: 0.33 > average 0.25}**: *"getty images the antimalarial drug hydroxychloroquine is being widely promoted as a cure for covid-19 but we still lack good data on its true benefits."*

Case studies indicate that our model performs like it is doing "fact-checking", which is an useful and important strategy for fake news detection. Meanwhile, entity descriptions are essential for fact-checking. Therefore, with the good usage of entity descriptions, fake news detection can achieve better performance, same as the ablation studies in Section 4.2 shown.

For word-level analysis, we discovered similar results as (Lu and Li, 2020) did. Some fake news contains emotional words or words that catch people's attention like "Breaking"or "warn".