Connecting the Dots Between Fact Verification and Fake News Detection

Qifei Li* **Wangchunshu Zhou***
Beihang University
{liqifei, zhouwangchunshu}@buaa.edu.cn

Abstract

Fact verification models have enjoyed a fast advancement in the last two years with the development of pre-trained language models like BERT and the release of large scale datasets such as FEVER. However, the challenging problem of fake news detection has not benefited from the improvement of fact verification models, which is closely related to fake news detection. In this paper, we propose a simple yet effective approach to connect the dots between fact verification and fake news detection. Our approach first employs a text summarization model pre-trained on news corpora to summarize the long news article into a short claim. Then we use a fact verification model pre-trained on the FEVER dataset to detect whether the input news article is real or fake. Our approach makes use of the recent success of fact verification models and enables zero-shot fake news detection, alleviating the need of large scale training data to train fake news detection models. Experimental results on FakenewsNet, a benchmark dataset for fake news detection, demonstrate the effectiveness of our proposed approach.

1 Introduction

Recently, fake news has been appearing in large numbers, which are easily accessed and disseminated in the online world with the booming development of online social networks. Users can be affected due to the deceptive words intentionally and verifiably false, which makes fake news detection urgent and important for maintaining social order (Shu et al., 2017).

Most existing methods for detecting fake news rely heavily on supervised learning on a large scale dataset with news articles labeled as fake or real by human experts. However, such a labeled dataset is difficult and time-consuming to obtain while few large scale fake news detection datasets are publicly available (Oshikawa et al., 2018). Meanwhile, with the fast development of pre-trained language models (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018) and the release of large scale datasets (Thorne et al., 2018), research on the fact verification task have enjoyed fast advancement (Nie et al., 2019; Hanselowski et al., 2018; Zhou et al., 2019a). However, the advancement in the field of fact verification fails to transfer to the task of fake news detection due to the different nature of input sequences (i.e., fact verification aims to check the reliability of a claim of one or a few sentences while fake news detection aims to check the trustworthy of a long article) and the lack of large scale training data to train the fact verification models for fake news detection.

To address the above issues, in this paper, we propose a simple yet effective approach to connect the dots between fact verification and fake news detection. Our approach exploits off-the-shelf models in two relatively well-studied problems—text classification and fact verification—to tackle the problem of fake news detection. Specifically, our approach first employs a text summarization model pre-trained on news corpora to summarize the long news article into a short claim. Then we use a fact verification model pre-trained on the FEVER dataset to detect whether the input news article is real or fake. Our approach transfers the recent success of fact verification models to enable zero-shot fake news detection,

*Equal Contribution.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1820–1825
Barcelona, Spain (Online), December 8-13, 2020

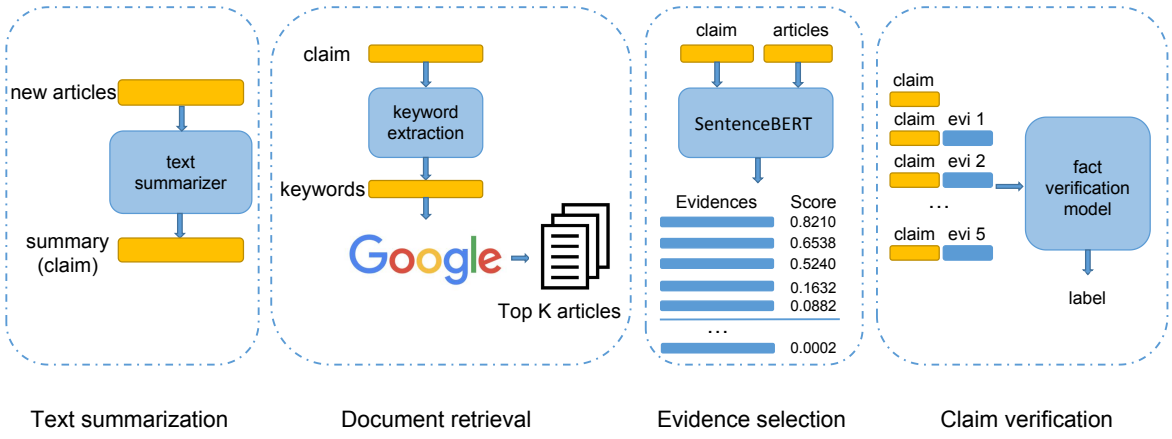

Figure 1: Illustration of the proposed approach. Our approach exploits two off-the-shelf models trained with text summarization and fact verification datasets, which are two relatively well-studied problems to perform zero-shot fake news detection.

alleviating the requirement of large scale datasets to train fake news detection models. Experimental results on FakenewsNet, a benchmark dataset for fake news detection, show that our approach can achieve comparable performance with several competitive supervised content-based fake news detection models in a zero-shot fashion. In addition, when fine-tuning the fact verification model with labeled examples, our approach significantly outperforms the supervised baselines. This demonstrates the effectiveness of our proposed approach.

2 Related Work

The problem of fake news detection has gained much research interests with the widespread of fake news on social media (Zhou and Zafarani, 2018). Existing fake news detection models can be categorized into content-based models (Conroy et al., 2015; Pan et al., 2018) and social context based models (Shu et al., 2017; Qian et al., 2018; Tschiatschek et al., 2018). In this paper, we focus on content-based models that enable early fake news detection before the spread of fake news. Recent content-based fake news detection models (Reddy et al., 2020; Shu et al., 2019; Zhang et al., 2020) generally formulates the problem as a text classification problem. However, many news articles are very long, which hinders the application of state-of-the-art pre-trained language models such as BERT (Devlin et al., 2018) because their maximum context length is generally 512. Moreover, few large scale fake news detection datasets are available and they are generally limited in their domains while human annotation of fake news is time-consuming and expensive, which hinders the application of current supervised content-based fake news detection models. Another recent work related to our paper is an unsupervised fake news detection method proposed by (Yang et al., 2019). However, their approach is based on the social context, which requires additional information and thus less general to content-based approaches. Our approach, in contrast, is the first fake news detection method that enables zero-shot fake news detection without any labeled examples or additional information to the best of our knowledge.

3 Methodology

Our approach is a novel content-based fake news detection method that exploits off-the-shelf pre-trained models in well-established text summarization and fact verification problems to enable zero-shot fake news detection without training on labeled datasets. As illustrated in Figure 1, in the first stage, our approach employs a text summarization model pre-trained on large scale news corpus to summarize the input long news article into a short claim which consists of one or a few sentences. Afterward, we use a fact verification model pre-trained on FEVER (Thorne et al., 2018), a large scale fact verification dataset,

to check the trustworthiness of the article. We then describe the details of the two stages in our proposed approach.

For the first stage. we employ an open-sourced BERT-based extractive summarization model[1] (Miller, 2019) to summarize the input news article into a short claim. The length of the output summary is controlled to mimic that of the claims in the FEVER dataset by setting the compression ratio set to 0.1 and selecting the top 2 predicted sentence, which minimizes the inconsistency between the input of the fact verification model used in the second stage during its training and inference procedure. In addition, opinions in long articles are generally scattered, which makes it difficult for a fake news detection model to judge the reliability of long articles. Our approach makes the information more concentrated in a few sentences, making it easier to identify the trustworthiness of the article.

During the second stage, we employ GEAR (Zhou et al., 2019a), a competitive fact verification model based on BERT and graph neural network trained on the FEVER dataset to verify the claim generated during the first stage. However, the model requires multiple support evidence, which is available in the FEVER dataset, to make the prediction. To mitigate this issue, we propose to construct evidence of the claim using a commercial search engine such as Google. Concretely, we first extract the keywords in the claim with AllenNLP[2] (Gardner et al., 2018), an open-sourced NLP toolkit. Afterward, we feed the keywords into a search engine to crawl web texts related to the claim. We filter the urls that contains only pdf or images or the text content is less than 100 tokens, and select support evidence for the claim according to the sentence embedding similarity under SentenceBERT (Reimers and Gurevych, 2019), a pre-trained sentence embedding model. We select the top 5 related sentences as evidences and feed them into the pre-trained GEAR model together with the original claim to predict whether the claim is trustworthy or not. Note that the fact verification model pre-trained on FEVER is a three-way classification model including the "Not Enough Infomation" class. We simply omit this class and compare the output probability of the real and fake class to make the binary classification.

In addition, when labeled examples are available, we can further fine-tune the fact verification model pre-trained on the FEVER dataset in a continual learning fashion with our target dataset to achieve better performance. For the experiments with fine-tuning the fact verification model with labeled fake news detection examples, we initialize the output layer to make the model perform binary classification and use the default hyperparameter of GEAR to perform the fine-tuning procedure.

4 Experiments

4.1 Experimental Settings

Table 1: Dataset Statistics.

Domain	#Fake	#Real
PolitiFact	330	332
GossipCop	4582	14477

Dataset We adopt FakenewsNet (Shu et al., 2018), a recently released benchmark dataset for fake news detection in our experiments. The dataset consists of news articles in two domains including PolitiFact and GossipCop. We run the official code of the paper to obtain the datasets, detailed statistics of the dataset are presented in Table 1. We use 80% of data for training and 20% for testing. For evaluation metrics, we use accuracy, precision, recall, and F1 score following previous work (Shu et al., 2018). In addition, we manually increase the ratio between real and fake news in the test set to simulate the real-world scenario where only a small portion of news articles are fake. Therefore, a random guessing baseline would achieve very poor results (i.e., around 0.2 to 0.3 F1 score).

Compared Models We adopt two settings of baseline to evaluate the effectiveness of the proposed approach. The first is the transfer learning setting where we assume that we have access to labeled examples in another domain but do not have training examples in the target domain. In this setting, we train a supervised fake news detection model with the labeled data in another domain and directly transfer the model for inference in our target domain without parameter updating. The second setting is

[1]https://github.com/dmmiller612/bert-extractive-summarizer

[2]https://github.com/allenai/allennlp

Model	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Zero-Shot Setting								
SVM	33.94	22.31	16.97	19.27	44.45	23.71	**55.10**	33.15
NB	37.74	26.47	19.09	22.18	47.19	24.40	53.02	33.42
LR	35.49	27.30	23.33	25.16	49.09	23.80	47.08	31.61
Ours	**44.10**	**44.82**	**52.42**	**48.32**	**56.49**	**29.09**	52.42	**37.42**
Supervised Setting								
SVM	49.30	49.27	47.22	48.23	52.71	28.53	58.27	38.30
NB	59.72	60.93	54.17	57.35	62.10	32.58	48.23	38.89
LR	49.30	49.20	43.05	45.92	57.47	31.08	57.60	40.37
Ours	**68.75**	**65.17**	**80.56**	**72.50**	**73.74**	**47.64**	**58.77**	**52.50**

Table 2: Test set performance of compared models on two domains in the FakenewsNet benchmark in both the zero-shot and the supervised setting. For the baseline methods in the zero-shot setting, we train a supervised model on the training set of one domain

the supervised setting where we train a fake news detection model in a supervised fashion with training examples in the target domain. We employ doc2vec (Le and Mikolov, 2014), a pre-trained paragraph embedding approach to encode the news articles into a fixed-length vector with 300 dimension. We then apply standard machine learning models that yield state-of-the-art performance on the task of fake new detection including support vector machines (SVM), logistic regression (LR), Naive Bayes (NB) as the classification model. Note that we do not compare against models based on pre-trained language models because the length of news articles are generally longer than their maximum context length and clipping the articles severely affects the model's performance. Conventional CNN or LSTM-based models compare similarly to machine learning based models in our preliminary experiments. We suspect this is because the pre-trained doc2vec is already a good feature extractor.

4.2 Experimental Results

We present the experimental results in both zero-shot and supervised setting in Table 2. We can see that the zero-shot variant of our approach significantly outperforms all compared zero-shot baselines, demonstrating that our approach can successfully connect the dots between a well-trained fact verification model and the task of fake news detection. It is notable that our approach does not require any fake news detection training example while the zero-shot baselines require first training on a fake news detection model in a different domain. It is also remarkable that our approach yields comparable results compared to the baselines trained in the supervised setting. When fine-tuning the fact verification model used in our approach with labeled training examples, our approach outperforms the supervised baselines with a large margin (i.e., over 10 F1 scores). This shows that our approach is also helpful when training examples are available.

4.3 Analysis

Figure 2: Analysis.

To better understand our proposed approach, we conduct an analysis investigating the effectiveness of transferring a well-trained fact verification model for fake news detection. Specifically, we fine-tune the fact verification model with different numbers of training examples to simulate the transition from the zero-shot setting to the supervised setting and compare it to the supervised baselines with the same range of training examples.

The results are presented in Figure 2. We can see that the performance yielded by our approach is already relatively good with only 500 to 2000 training examples. In contrast, the compared supervised baselines need more

training examples to achieve comparable performance. This confirms the effectiveness of our approach to reduce the need of large scale training data for training fake news detection models.

5 Discussion & Conclusion

In this paper, we propose a novel fake news detection approach that exploits well-trained text summarization model and fact verification model. By connecting the dots between fact verification and fake news detection, our approach enables zero-shot fake news detection. There exists previous work that exploit models trained on one task for another related task, such as using machine translation models for paraphrasing (Somers, 2005) and grammatical error correction (Lichtarge et al., 2019; Zhou et al., 2019b). Our approach is similar to this line of research but combines two models in other tasks for another downstream task.

Experiments on a fake news detection benchmark show that our approach can yield comparable performance with competitive supervised content-based fake news detection models in a zero-shot fashion. Our approach can also leverage labeled examples more effectively than conventional supervised methods with continual learning, enabling building fake news detection models more effectively for domains where few labeled data are available. One limitation of our approach is that it is based on pre-trained language models and thus can be computationally expensive to use in real-world applications. Therefore, for future work, we plan to exploit methods to accelerate the inference of pre-trained language models including model compression methods like DistilBERT (Sanh et al., 2019) and BERT-of-Theseus (Xu et al., 2020), as well as adaptive inference methods such as PABEE (Zhou et al., 2020).

Acknowledgments

We thank the anonymous reviewers for their valuable comments.

References

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. *arXiv preprint arXiv:1904.05780*.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, volume 18, pages 3834–3840.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Harita Reddy, Namratha Raj, Manali Gala, and Annappa Basava. 2020. Text-mining-based fake news detection using ensemble methods. *International Journal of Automation and Computing*, pages 1–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

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*, 8.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, pages 517–524.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5644–5651.

Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1826–1829. IEEE.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019a. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. 2019b. Improving grammatical error correction with machine translation pairs. *arXiv preprint arXiv:1911.02825*.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. Bert loses patience: Fast and robust inference with early exit. *arXiv preprint arXiv:2006.04152*.