RIEHAN: Relevant Information Enhanced Hierarchical Attention Network for Automated Claim Verification

Anonymous ACL submission

Abstract

The spread of misinformation in online media has caused significant societal problems today, underscoring the importance of verifying claims before accepting them as real. In this work, we design a hierarchical attention network based automated claim verification module. This architecture uses latent features from the claim, all articles in the dataset related to the claim as well as the most relevant information extracted from the articles via a gating unit. We show by ablation studies that this trainable method of extracting most relevant information from articles results in better performance compared to using cosine similarity metric. We also show by ablation studies that using the most relevant information from articles explicitly in the model results in better performance metrics. The proposed model, **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN), outperforms the state-of-the-art benchmark architectures on PolitiFact dataset and performs comparably to the state-of-the-art models on the Snopes dataset.

1 Introduction

Online media generates a large number of content containing misinformation, unconfirmed news, or biased claims. Such unverified claims pose a great threat to all aspects of society from personal life to public safety. To reduce the spread of misleading content on the internet, some fact-checking systems and debunking websites have been developed, such as *politifact.com* and *snopes.com*. Online users can debunk claims by manually assessing the credibility along with evidence (e.g. webpages, quotations, etc.). However, this manual verification process is time-consuming. Therefore, it is essential to develop automated claim verification tools. Hence, fake news detection has attracted much attention in recent years and a variety of automatic claim verification methods have been developed. (Castillo et al., 2011) utilized features from user's posting and re-posting to assess the credibility of a given content from social media. (Zhao et al., 2015) detected early rumor information by capturing enquiry patterns, such as 'Is that ture?', 'really?' and 'what?'. Some works like (Rashkin et al., 2017) and (Wang, 2017) labeled claims to construct a new dataset and train neural networks on that. However, these methods only used textual information from claims without external evidence which could provide more insight into the credibility analysis.

Recently claim verification methods using external documents retrieved from the web as evidence have been proposed (Popat et al., 2017; Thorne et al., 2018; Dungs et al., 2018). Following this trend, (Popat et al., 2018) proposed a word-level attention based neural network to design an explainable claim verification model. (Ma et al., 2019a) focused on learning which documents are more crucial to the automatic claim verification process. (Vo and Lee, 2021) proposed a hierarchical attention mechanism combining word-level attention and document-level attention to detect fake claims. Although these methods make use of textual information from evidence and even consider different embedding levels, they do not identify the most relevant information from the evidence and use it more directly in the claim verification process.

In this paper, we propose a novel architecture, called **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN) to address the aforementioned limitation of existing works. Our main contributions are:

- a trainable relevance feature extraction mechanism that captures the relation between the claim and relevant articles and weights them with the relevance metric;
- an architecture that combines latent features from the claim, all related articles and the

most relevant information from the articles explicitly to verify a given claim;

- extensive evaluation on two standard public datasets that show that our method performs better or comparable to SOTA methods;
- ablation studies that show that using relevant information explicitly performs better.

2 Related Work

Existing claim verification methods can be divided into two categories: (i) evidence excluded methods (Castillo et al., 2011; Zhao et al., 2015; Jin et al., 2016; Rashkin et al., 2017; Wang, 2017) and (ii) evidence aided methods (Popat et al., 2017, 2018; Vo and Lee, 2019; Ma et al., 2019a; Vo and Lee, 2021).

The evidence excluded methods only focus on the claim itself without using other supporting or refuting text from outside the claim. (Castillo et al., 2011; Zhao et al., 2015) extract different features from linguistics and textual content to check fake claims. (Jin et al., 2016) improves claim verification performance by mining conflicting viewpoints from social media. (Rashkin et al., 2017; Wang, 2017) train a neural network on labeled claim datasets to verify the unseen claim data sample.

Since extraneous articles can potentially be used to support or refute claims, several researchers have started to focus on evidence aided methods for claim verification. (Popat et al., 2017) propose a method for credibility assessment of textual claims by leveraging the evidence and counter-evidence retrieved online. Using this work as a basis, (Popat et al., 2018) presents an end-to-end framework (DeClarE) for assessing the credibility of claims. The framework combines the article and claims embeddings to predict the credibility score of the claim. (Ma et al., 2019a) propose a neural network to represent the coherent evidence during the claim verification with a hierarchical attention mechanism.

The existing evidence included methods described above use evidence in straightforward ways, like concatenating the claim and related articles together without deeper information mining. Our architecture is different from these works in several ways: (1) we propose a trainable relevance feature extraction mechanism to capture the relation between the claim and relevant articles (2) we then combine latent features from the claim, the articles and the most relevant information from these articles to verify the claim.

3 Proposed Model

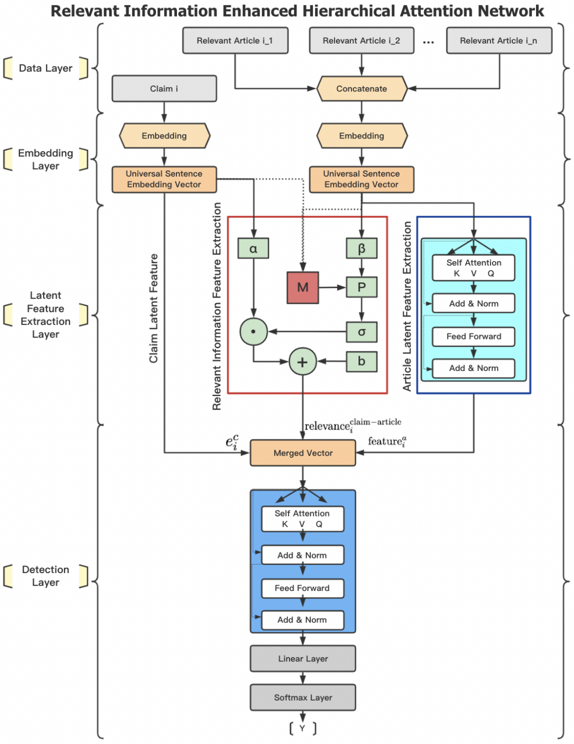

Figure 1: The proposed Relevant Information Enhanced Hierarchical Attention Network (RIEHAN).

We propose a novel model: **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN) (shown in Fig 1), using attention mechanisms (Vaswani et al., 2017) and a gating mechanism that captures the correlation between each claim and the relevant articles. The trainable gate module extracts the most relevant information from the related articles and explicitly uses this as another feature in detecting if the claim is false or true. In the next subsections, we will demonstrate the relevant information feature extraction and article latent feature extraction.

3.1 Relevant Information Feature Extraction

Let C_i and A_i^j be the i^{th} claim and the relevant article corresponding to C_i, where j ranges through all related articles. Let the concatenation of all relevant articles for the i^{th} claim be $A_i^c = \{A_i^0, A_i^1, A_i^2, ..., A_i^j, ...\}$. We use the universal sentence embedding (USE) (Cer et al., 2018) to generate the embeddings of the claim and the concatenated relevant articles, e_i^c and e_i^a, respectively.

In order to capture the information that is most relevant to the claim from all the articles, we define a trainable similarity matrix M, which captures the similarity between the claim embedding vector, e_i^c, and concatenated article embedding vector, e_i^a. We call this the relevant information feature, $\text{relevance}_i^{\text{claim-article}}$ and compute it as follows

$$\text{relevance}_i^{\text{claim-article}} = \sigma(P) \odot \alpha(c_i) + b \quad (1)$$

In the above equation, σ is the sigmoid function, $\odot$ is element-wise multiplication and the rest of the parameters are defined as follows

$$\alpha(c_i) = W_\alpha e_i^c + b_\alpha \quad (2)$$

$$\beta(a_i) = W_\beta e_i^a + b_\beta \quad (3)$$

$$m(c_i, a_i) = e_i^c M e_i^a \quad (4)$$

$$P = m(c_i, a_i)\beta(a_i) \quad (5)$$

3.2 Article Latent Feature Extraction

(Chen et al., 2020) state that the latent feature can also play an important role in misinformation detection. Inspired by it, an h-layer multi head attention (MHA) module is used to extract the article latent feature in this architecture: feature_i^a. This MHA module uses the scaled dot product attention (Vaswani et al., 2017), which is given by

$$\text{feature}_i^a(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where Q, K and V are the query, key and value matrices. In our case, $Q = K = V = e_i^a$ and d_k is the dimension of the query and key vectors.

3.3 RIEHAN: Relevant Information Enhanced Hierarchical Attention Network

The **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN) merges three inputs for the final verification as follows: (1) claim latent feature, e_i^c, (2) relevant information feature, $\text{relevance}_i^{\text{claim-article}}$ and (3) article latent feature, feature_i^a (shown in Fig 1). The three different features are combined using another attention layer followed by a linear layer and a softmax layer for the final result.

4 Experiment

4.1 Experiment Setup

We implement all the models in Pytorch and train them to minimize the cross-entropy loss function of predicting the class label of claims in the training set. The stochastic gradient descent (SGD) algorithm (Ruder, 2016) is used as the optimizer for training. We use 5-fold cross validation to prevent over-fitting the model.

Statistics	PolitiFact	Snopes
True Claims	1,867	1,164
False Claims	1,701	3,177
Related Articles	29,556	29,242

Table 1: Dataset Statistics

4.2 Data

We test the RIEHAN on two publicly available datasets released by (Popat et al., 2018). In Politifact dataset, there are originally six labels: true, mostly true, half true, false, mostly false, pants on fire. We merge false, mostly false and pants on fire into false claims and the rest are into true claims. Each Snopes claim is labeled as true or false. Details of the datasets are shown in Table 1.

4.3 Models

We provide the performance (in terms of accuracy and F1 score) of several benchmark architectures on the datasets. The comparisons are shown in Table 2 and Table 3. The benchmark architectures are described below.

- **DeClare (Popat et al., 2018):** is a completely automated end-to-end neural network model for evidence-aware credibility assessment. It captures information from external evidence articles and models joint interactions between various factors.
- **HAN (Ma et al., 2019b):** is a hierarchical attention network for claim verification with the representations of relevant articles and evidences. It focuses on the evidence representations by attending on the sentences instead of each word.
- **NSMN (Nie et al., 2019):** is a model designed to predict credibility score of claim by using stance of the document with respect to claim.

To inspect the influence of each component in our model, we conduct ablation studies by removing or replacing different modules in our model.

- **RIEAN:** is the proposed network without article latent feature extraction.
- **HAttN:** is the proposed network without relevant information feature extraction.
- **RIEHAN:** is the proposed Relevant Information Enhanced Hierarchical Attention Network.

Model	ACC	Macro F1	Micro F1
HAN	-	0.5866	0.5912
NSMN	-	0.6021	0.6043
DeClare	-	**0.6521**	**0.6535**
RIEAN	0.6584	0.6566	0.6584
HAttN	0.6444	0.6422	0.6444
RIEHAN	**0.6708**	**0.6686**	**0.6708**
RIECosAN	0.6413	0.6358	0.6413
RIECosHAN	0.6382	0.6379	0.6382

Table 2: Performance on PolitiFact Dataset

Model	ACC	Macro F1	Micro F1
HAN	-	0.6251	0.7280
NSMN	-	0.6801	0.7613
DeClare	-	**0.7245**	**0.7881**
RIEAN	0.7560	0.6693	0.7560
HAttN	0.7273	0.6118	0.7273
RIEHAN	**0.7779**	**0.7127**	**0.7779**
RIECosAN	0.7434	0.6612	0.7434
RIECosHAN	0.7583	0.6869	0.7583

Table 3: Performance on Snopes Dataset

- **RIECosAN:** is the RIEAN with the learnt relevance, M, replaced by cosine similarity.
- **RIECosHAN:** is obtained by replacing M by the cosine similarity in the proposed RIEHAN.

4.4 Performance

We show experimental results of our model and baselines in Table 2 and Table 3.

The detection results on PolitiFact dataset are shown in Table 2, we can see that the **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN) performs best compared with the other architectures on the PolitiFact dataset using the same experiment setup with an accuracy, Macro F1 and Micro F1 of 0.6708, 0.6686 and 0.6708 respectively. The percentage increase over the best baseline model is 2.53% for Macro F1 and 2.65% of Micro F1. The performance of RIEAN in terms of Macro F1 and Micro F1 are 0.6566 and 0.6584, which is also better than all the benchmark architectures.

The performance of RIEAN and RIEHAN point to the fact that extracting the most relevant information from all articles related to a claim and directly using it to verify a claim can lead to better overall performance.

We also note that, the performance of RIECosAN and RIECosHAN in terms of F1 drop compared with RIEAN and RIEHAN. This indicates that the trainable similarity matrix, M, is more efficient in capturing the relevance of articles and claims compared with the static cosine similarity.

The detection results on Snopes dataset are shown in Table 3, the performance of RIEHAN is comparable to DeClare. The percentage difference is: 1.63% for Macro F1 and 1.30% for Micro F1 compared with the best baseline architecture. The comparison between proposed architectures with relevant information feature extraction and those without relevant information feature extraction shows that the relevant information feature extraction is meaningful to improve the performance of claim verification.

5 Conclusion

We proposed the **R**elevant **I**nformation **E**nhanced **H**ierarchical **A**ttention **N**etwork (RIEHAN), using attention mechanism and a gate module that captures relevant information from articles related to the claim. We find, through ablation studies, that explicitly including this relevant information in the architecture works better than using latent features derived only from the entire articles. We also show that using a trainable gate architecture works better than using a cosine similarity metric to capture relevance information.

References

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Mingxuan Chen, Ning Wang, and KP Subbalakshmi. 2020. Explainable rumor detection using inter and intra-feature attention networks. *TrueFact KDD 2020*.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019a. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019b. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

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.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

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.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344.

Nguyen Vo and Kyumin Lee. 2021. Hierarchical multi-head attentive network for evidence-aware fake news detection. *arXiv preprint arXiv:2102.02680*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405.