Fake news detection using claim specific article attention

Ayush Jain
Indian Institute of Technolgy, Hyderabad
cs17mtech11017@iith.ac.in

G. Sree Datta Sanjay Bharath
Indian Institute of Technology,hyderabad
cs17mtech11016@iith.ac.in

ABSTRACT

Internet and the web is a huge part of people's day to day life in collecting valuable information but fake news is one of the biggest problems in moderns era of civilization. Research on fake news detection is growing in a rapid pace, there are many methods that detect fake news using supervised learning which captures the linguistic styles, stance information of the corresponding article towards the claim etc to asses the credibility of the claim. However, most works do not consider the external evidence to judge the claim, contextual information of the claims, and corresponding evidence.

This paper overcomes these limitations by considering external evidence that supports or renounce the claim and the contextual representation of claim and the corresponding article words. To this extent, we capture the contextual representations of claim texts and article words using Bi-LSTMs, efficient representations of claim and corresponding article sources and use an end to end neural network with claim specific article word attention to access the credibility of the claim. It also derives informative features for generating user-comprehensible explanations that makes the neural network predictions transparent to the end-user

Keywords

Fake news detection; Attention mechanism; Bi-LSTM

1. INTRODUCTION

In this modern era of digitization, the reach of information is very high. The news spread very fast to millions of internet users. Although information reach has it's perks, one of the biggest issues in this process is misinformation which tends to defame famous personalities, Industries, create fake news about celebrities, create fake reviews about popular products etc. So, there is a huge need for detecting false claims and validating the credible claims. Fake news detection is a very challenging problem, in many cases, even humans can't access the credibility of the claim. Moreover in addition to detecting the fake news giving evidence to the credibility assessment is also a crucial job.

In this work we access the credibility of a claim by looking at different articles that are relevent to the claim .In order to get effiecient relevence ,context information of both claim text and article word are important.Using this context information of both claim and article words we produce a ***claim specific article representation*** that contains the weighted average of article words. where weights are assigned to the word representation based on thier contextual relevence with the whole claim text representation.In the following sections we list the works that relate to this paper and describe our proposed approach.The novelty of the paper lies in:

- Efficient representation of claim source using Bi-LSTM [1] using self attention instead of traditional word2vec or glove embeddings.
- Efficient article word representation that constitute complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). These word vectors are learned functions of the internal states of a deep bidirectional language model (Bi-LSTM)[5].
- Efficient claim source and article source representations using ratings labels given by other sources .
- Use of Loung attention [3] on the above given representations.

1.1 RELATED WORK

Our work is closely related to areas of credibiltity analysis of claims, truth discovery, spread of misinformation. Deep learning based approach proposed in DeClarE[6] predict the credibility of any arbitrary textual claim using the related articles and source information. DeClarE [6] completely avoided using hand crafted features and used embeddings for representing claim text and article text but on the negative side they use averaged word embedding to obtain embedding for whole claim text which does not capture the sequence and semantic information. Also the attention mechanism proposed by them use word embedding of article terms which does not capture contextutal information of the word in the article. Rashkin [7]propose neural network based approaches for determining the credibility of a textual claim, but it does not consider externalsources like web evidence and claim sources.

Recently there are works on effective representations which when used for various NLP tasks improved the performance

WOODSTOCK '97 El Paso, Texas USA
© 2018 ACM. ISBN 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

significantly. InferSent [1] sentence embedding method provide semantic representation of an english sentence, it use Bi-LSTM with self attention on hidden states to obtain sentence representation. ELMO [5] is a deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). These word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. They can be easily added to existing models and significantly improve the state of the art across a broad range of challenging NLP problems, including question answering, textual entailment and sentiment analysis.
In our work we used both InferSent and Elmo to obtain effective representations of claim text and article terms overcoming the limitations of DeClarE [6].

2. PROPOSED APPROACH

2.1 Representations

Consider a set of N claims C_n,each claim is made by a respective source CS_n , where n ∈ [1, N]. Each claim C_n has a set of M articles texts $A_{m,n}$ which are extracted from respective article links $AS_{m,n}$,where m ∈ [1, M]. Each training instace is tuple of claim and its source, reporting articles and article links [C_n , CS_n ,A_m, n , AS_m, n] and credibility label of a claim is considered as a target output or ground truth value during network training. Figure 1 gives a pictorial overview of our model.In the following sections,we provide a detailed description of representations all inputs.

2.1.1 Claim Representation

For claim representation, we referred InferSent[1] model, a universal sentence embedding. where each claim C_n of length T is tokenized into list of T words [w_i] where i ∈ [1...T] and passed through a bidirectional LSTM[1].The bilstm computes a set of T vectors $\{h_t\}_t$. For t ∈[1,. . . , T], $\{h_t\}$ is the concatenation of a forward LSTM and a backward LSTM that read the sentences in two opposite directions:

$$\overrightarrow{h_t} = \overrightarrow{\mathrm{LSTM}}_t\left(w_1, \ldots, w_T\right)$$

$$\overleftarrow{h_t} = \overleftarrow{\mathrm{LSTM}}_t\left(w_1, \ldots, w_T\right)$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h}_t\right]$$

As all the words in the claim are not equally important to describe the claim.So,we have used self attention mechanism of Infersent model [1] to create a representation $\boldsymbol{u}$ of the claim that gives different contextual weights to to different word representations.The attention mechanism is defined as follows :

$$\overline{h}_i = \tanh\left(W h_i + b_w\right)$$

$$\alpha_i = \frac{e^{\overline{h}_i^T u_w}}{\sum_i e^{\overline{h}_i^T u_w}}$$

$$u = \sum_t \alpha_i h_i$$

where $\{h_1, \ldots, h_T\}$ are the output hidden vectors of a BiLSTM. These are fed to an affine transformation (W, b_w) which outputs a set of keys $\left(\overline{h}_1, \ldots, \overline{h}_T\right)$ The $\{\alpha_i\}$ represent the score of similarity between the keys and a learned context query vector u_w. These weights are used to produce the final claim representation$\boldsymbol{u}$, which is a weighted linear combination of the hidden vectors.

2.1.2 Article Representation

An article is considered as a sequence of T tokens. Each token embedding is extracted using ELMO[5] word embeddings,a Bidirectional LSTM based word embedding which incorporates the context information of the word.Given N tokens the forward LSTM computes the probability of the sequence by modeling the probability of to- ken t_k given the history (t_1, . . . , $t_k 1$) :

$$p\left(t_1, t_2, \ldots, t_N\right) = \prod_{k=1}^{N} p\left(t_k | t_1, t_2, \ldots, t_{k-1}\right)$$

A backward LSTM is similar to a forward LSTM, except it runs over the sequence in reverse, predicting the previous token given the future context

$$p\left(t_1, t_2, \ldots, t_N\right) = \prod_{k=1}^{N} p\left(t_k | t_{k+1}, t_{k+2}, \ldots, t_N\right)$$

A biLSTM combines both a forward and backward LSTM. The formulation we used jointly maximizes the log likelihood of the forward and backward directions

$$\begin{aligned} \textstyle\sum_{k=1}^{N} \quad & \left(\log p\left(t_k | t_1, \ldots, t_{k-1}; \Theta_x, \vec{\Theta}_{LSTM}, \Theta_s\right)\right. \\ & \left. + \log p\left(t_k | t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s\right)\right) \end{aligned}$$

The parameters for both the token representation θ_x and Softmax layer θ_s in the forward and backward direction while maintaining separate parameters for the LSTMs in each direction. Thus the concatenation of hidden layers of both forward and backward layer of each token is considered as the article word representation

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h}_t\right]$$

2.1.3 Source Representation

Claim source and article source information provide the framework an extra boost to access the credibility of the claim.Source representation of CS_n,AS_m, n is a list of attribute ratings of the source given by other sources in the blog. Claim source is a 7-dimensional vector and article source is a 12 dimensional vector. All these ratings range between [0-5].This information of the sources is used to represent a particular source.

2.2 Proposed Framework

Each claim is tokenized and passed through a universal sentence encoder[1] to get a claim embedding and The article words are passed through a Bi-LSTM[5] and concatenated hidden layer for each word is considered as it's contextual word representation. Given a claim representation $\boldsymbol{u}$ only a subset of words in an article $w_i \subset$ W can judge the credibility of the claim . So, we aim to give those words an extra weight by attending to article words given claim representation as query vector, resulting in ***claim specific***

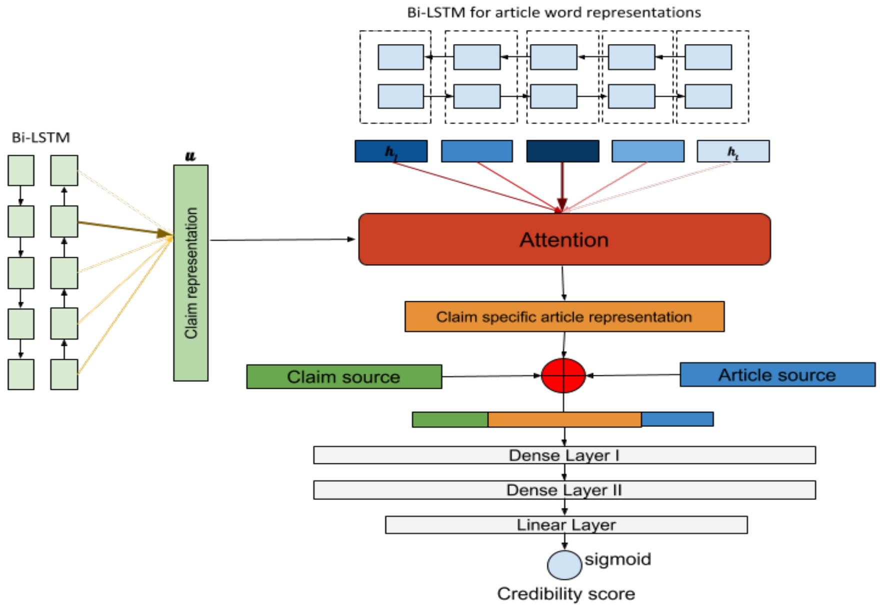

Figure 1: Overall model architecture

article representation.So, the idea of global attention in neural machine translation is applied here.Figure 2 gives a pictorial overview of our attention model

$$\boldsymbol{a}_t(s) = \text{align}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right) = \frac{\exp\left(\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right)\right)}{\sum_{s'} \exp\left(\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_{s'}\right)\right)}$$

where h_t corresponds to the claim representation and $\overline{h_s}$ corresponds words representations of an article,$s \in (1, \ldots, T)$ Here, score is referred as a content-based function for which we consider concat version.

$$\text{score}\left(\boldsymbol{h}_t, \overline{\boldsymbol{h}}_s\right) = \begin{cases} \boldsymbol{h}_t^\top \overline{\boldsymbol{h}}_s & \text{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \overline{\boldsymbol{h}}_s & \text{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a \left[\boldsymbol{h}_t; \overline{\boldsymbol{h}}_s\right]\right) & \text{concat} \end{cases}$$

$$\boldsymbol{a}_t = \text{softmax}\left(\boldsymbol{W}_{\boldsymbol{a}} \boldsymbol{h}_t\right)$$

$$\boldsymbol{A}_c = \frac{1}{T} \sum_k \boldsymbol{a}_t \cdot h_s$$

Given the alignment vector as weights, the claim specific article representation vector A_c is computed as the weighted average over all the article word representations.

2.2.1 *Credibility score prediction*

Claim source representations and article source representations are concatenated with claim specific article representation vector A_c.In order to predict credibility score we pass the concatenated vector through fully connected layers with non-linear activations.

$$d_1 = relu\left(W_c(\boldsymbol{A}_c \oplus C_S \oplus A_S) + b_c\right)$$
$$d_2 = relu\left(W_d d_1 + b_d\right)$$

where, W and b are the corresponding weight ma- trix and bias terms. Finally, to generate the overall credibility label of the article for classification tasks, or credibil- ity score for regression tasks, we process the final representation with a final fully connected layer

$$\text{Classification: } s = \text{sigmoid}\left(d_2\right)$$
$$\text{Regression: } s = linear\left(d_2\right)$$

2.2.2 *Credibility aggregation*

The credibility score in the above step is obtained considering a single reporting article. As previ- ously discussed, we have M reporting articles per claim. Therefore, once we have the per-article credibility scores from our model, we take an av- erage of these scores to generate the overall credi- bility score for the claim.

$$\text{cred}(C) = \frac{1}{M} \sum_m s_m$$

This aggregation is done during testing for all articles corresponding to a claim.

2.2.3 *Claim specific article attention*

3. DATASETS

We evaluate our approach and demonstrate its gen erality by performing experiments on Two differ ent datasets: a general fact-checking website,a news review com- munity

3.1 Snopes

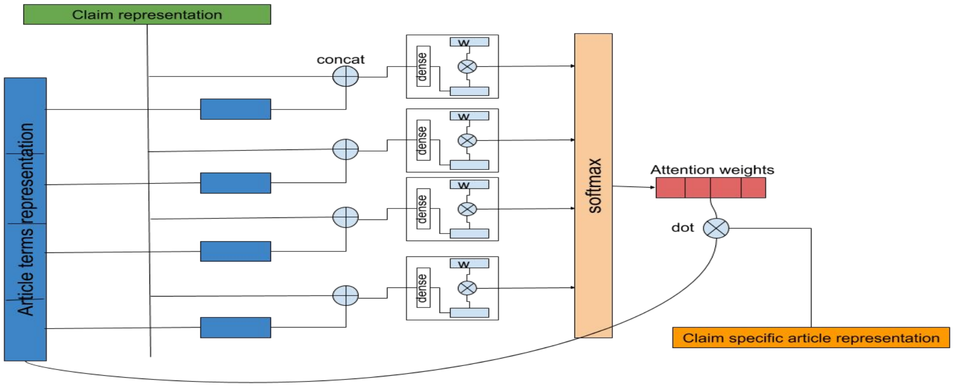
Claim representation
concat
dense
w
Article terms representation
softmax
Attention weights
dot
Claim specific article representation

Figure 2: Attention mechanism used in the overall architecture

Snopes [1] (www.snopes.com) is a general fact checking website where editors manually investi gate various kinds of rumors reported on the In ternet. We used the Snopes dataset provided by DeClarE[6]. This dataset consists of rumors analyzed on the Snopes website along with their credibility labels (true or false), sets of reporting articles, and their respective web sources.

3.2 NewsTrust

NewsTrust [2]is a news review community in which members review the credibility of news articles. We use the NewsTrust dataset made available by Mukherjee and Weikum (2015)[4]. This dataset contains NewsTrust stories from May 2006 to May 2014. Each story consists of a news article along with its source, and a set of reviews and ratings by community members. NewsTrust aggregates these ratings and assigns an overall credibility score (on a scale of 1 to 5) to the posted article. We map the attributes in this data to the inputs expected by our approach are as follows: the title and the web source of the posted (news) article are mapped to the input claim and claim source, respectively. Reviews and their corresponding user identities are mapped to reporting articles and article sources, respectively. We use this dataset for the regression task of pre dicting the credibility score of the posted claim with respect to relevent article words.

4. EXPERIMENTS

We Experimented our model on snopes and newstrust datasets.In the following sections we describe our experimental setup and Results.

4.1 Data Extraction

As snopes datasets contains claims and it's corresponding article links , In order to extract relevant article text from each article link ,we have crawled each article link ,collected the text and broke down the entire article in to snippets of length between 25 - 100 words.And extracted snippet with maximum relevance score sim_{cosine}(tf-idf cosine similarity) with the claim and considered that snippet to represent the entire article. We have not considered the article links which cannot be reached.In snopes dataset we do not have claim source information so during concatenation of vectors we eliminated claim source representation .

In NewsTrust dataset there are news headlines and reviews written by newstrust members about respective news headlines .We have considered news headlines as claims and it's reviews as article texts ,news headline source as claim source ,reviewer identity as article source.Sources are represented as their coresponding rating labels such as quality of the source ,factual consistency of the source , fairness of the source, accuracy of the source ,credibility rating of the source,expertise of the source,popularity of the source and style of the source.During the concatenation of claim specific article representation and source vectors .we concat 7-dimensional claim source representation and 12-dimensional article source representation to the claim specific article representation .

[1] $http : //resources.mpi - inf.mpg.de/impact/web_credibility_analysis/Snopes.tar.gz$

[2] $http : //resources.mpi - inf.mpg.de/impact/credibilityanalysis/data.tar.gz$

Method	True claims accuracy	False claims accuracy	Precision	Recall	F1-score
Declare	78.96	78.32	-	-	0.79
Our approach	**89.04**	**79.94**	82.51	79.4	**80.95**

Table 1: Results on Snopes dataset

4.2 Experimental setup

During experimentation, we tuned various hyper-parameters and checked for an optimal set of parameters for the model.We use the pytorch library to implement our model and used Adam optimizer[2] with a learning rate of 0.0001 to guide the loss function to reach a local minimum. We have used binary cross entropy loss function for classification task of Snopes dataset and mean squared error for regression task of NewsTrust dataset.We have used Xavier's normal weights initialization in attention mechanism which has reduced a lot of training time and loss reached a local minimum quickly.The size of all hidden layers were kept constant(i.e 100) throughout the experimentation in order to easily tune other hyper-parameters.During traning each claim - article pair is considered as a separate training instance.

5. RESULTS

5.1 Classification task

Training on Snopes dataset was done on 4000 samples with 2014 false classes and 1984 true classes .we kept number of false and true claims as close as possible to eliminate dataset bias. Testing was done on 1025 samples of which 404 statements were false . Table1 shows the results obtained on snopes dataset. We compare our results with the state of the art fake news detection model declare [6].As declare does not consider contextual representations of claim and article the F1-score ,True claims accuracy and False claims accuracy are lower than our model.

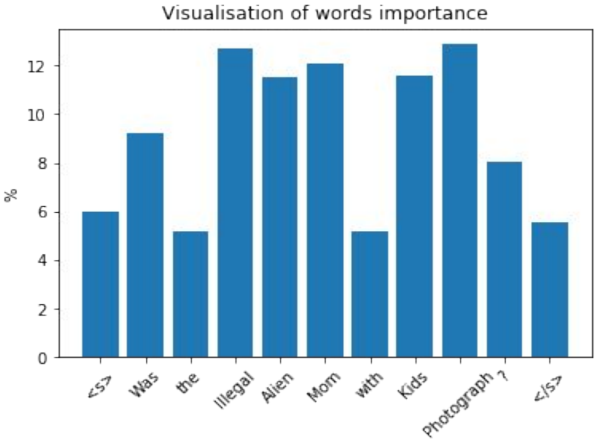

Figure 3: Visualisation of self attention weights of Infersent

5.2 Regression task

Training on NewsTrust dataset was done on same settings as above ,Table 2 show the testing Mean square error loss comparision between our approach and declare model.NewsTrust

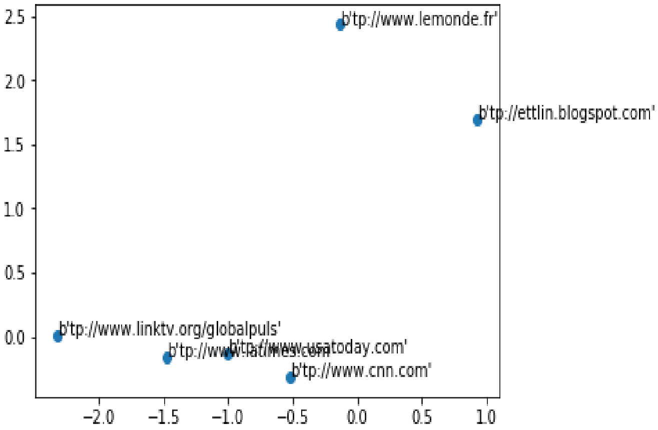

Figure 4: Effectiveness of claim source representation

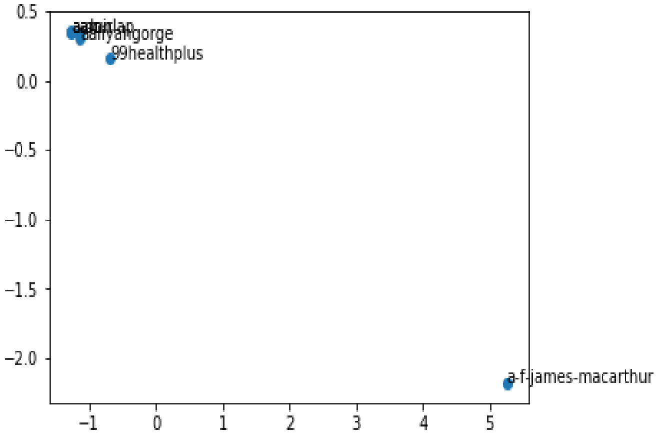

Figure 5: Effectiveness of article source representation

Method	MSE loss
Declare	**0.29**
our approach	0.70

Table 2: Results on NewsTrust dataset

dataset is a highly ambiguous dataset as multiple headlines have have same content .The results vary for different settings and different models .Upon using the above setting and comparing with the existing model our approach ranks 2^nd on the list.

6. DISCUSSION

6.1 Claim and article representation

In order to show that contextual information is being used in the Bi-LSTM[1] representation we plot words in the claim text and their importance in final claim representation .Figure3 demonstrates the claim representation importance.

6.2 Claim source and article source representation

To show that the rating labels of the sources we considered can contribute in accessing the credibility we have plotted the PCA outputs of source representations .In the Figure5 and Figure4 you can spot the clear seperation of sources that are credible and sources that contribute to the fake news.

7. FUTURE WORK AND CONCLUSION

Although the novelty of this work lies in representations and attention mechanism there is a high scope of improvement in terms of network achitecture and ablation study over network parameters and attention mechanism.In the future work we plan to improve the overall network architecture to consider other factors that contribute to fake news detection such as time stamp of the claims,credibility of the article that we are considering etc.

8. REFERENCES

[1] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680, 2017.

[2] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[3] M. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[4] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *CIKM*, 2015.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[6] K. Popat, S. Mukherjee, A. Yates, and G. Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 22–32, 2018.

[7] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, 2017.