# Fake News Detection using Convolutional Neural Network and Attention Layer

Farnoosh Javadi
UBC CS department
fjavadi@cs.ubc.ca

## Abstract

The explosive growth in fake news and its influence in democracy, justice, and public trust has increased the demand for fake news analysis, detection and intervention. There are variety of approaches and methods for tackling this task such as credibility based detection, feature based detection, propagation based detection and feedback based detection [10]. In my project I used text of news and besides meta-data for fake news detection, which is categorized in content and style based detection methods. I proposed a hybrid convolutional-attentive model that takes use of text statements and meta-datas with more attention on the important parts of the text statements to detect fake news. I evaluated my model by reporting accuracy, f1, and other metrics and showed that the proposed model can slightly enhance the performance of the baseline model.

## 1 Task Description

Fake news is intentionally and verifiably false news published by a news outlet for insincere purposes such as earning money or deceiving people for political reasons[21]. Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. It has weakened public trust in governments since its impact in 2016 U.S. presidential election. In our era, due to the rise of social media and its popularity, fake news can be created and published online faster and cheaper. Therefore, nowadays, automating fake news detection is of a high importance. An automated fake news detector tries to predict the likelihood that a given article is fake. This task is usually considered as binary classification, categorizing an article as real or fake. But in my project I considered 6 different levels of truthfulness for an article: true, mostly true, half true, barely true, false, and pant-on-fire (definitely false). I also tested the model merging each two labels and changing the task to a 3-class classification task, and compared the results, because I thought it might be hard even for humans to identify if an article is barely true or half true or false, and 6 different labels maybe misguide the classifier.

## 2 Dataset

I used LIAR dataset [19] because it is relatively large and rich dataset in comparison with other sources for fake news detection. It includes 12,836 short news which are sampled from various contexts, such as news releases, TV or radio interviews, campaign speeches, etc . News are categorized into 6 different classes, and the distribution of labels are roughly equal which is another important positive point of the dataset that prevents the classifier from being biased. Furthermore, LIAR dataset includes a rich set of meta-datas for each news such as speaker, category, job, credit history, party, and etc that I used them to make the model dependent to other features besides text statement and enhance the performance.

## 3 Contributions

My main contributions in this project are:

- Adding an attention layer at the end of the baseline model so that the convolutional-attentive model would take more attention to important parts of the text rather than looking into the sentence as a whole for identifying truthfulness.
- By thoroughly evaluating the proposed convolutional-attentive model using different metrics such as, accuracy, macro-average precision, macro-average recall , and macro-average f1, I showed that it can slightly enhance the performance of the baseline at least for some of the metrics.
- Testing the model in both 6-class fake news detection and 3-class fake news detection framework and comparing the results

## 4 Related Work

Fake news detection methods have been recently classified into two categories - content based methods and social based models - based on their main input sources [15]. Methods belonging in the first category focus on the content of the news, i.e. the body-text, the title, and additional metadata(when available). Methods belonging in the second category focus on social features and signals, such as the engagement and interaction of users with a given news on social media (e.g.“liking” a news on Facebook, “retweeting” it on Twitter, etc.). Recently, most of the techniques are ML based, although other approaches have been proposed, e.g. based on statistical analysis [8] or knowledge graphs [15],[4], [20]. Content-based methods are the traditional approach, as they find application in conventional news media and in all cases in which no social information is available. Historically, these methods have been used for spam detection [18][14]. In the last years, they have also been applied for fake news detection [2] exploited syntactic and semantic features for classifying between real articles and fakes. [12] showed that using a relatively simple approach based on term frequency (TF) and term frequency-inverse document frequency (TF-IDF) could already offer a good baseline accuracy of 88.5% for binary classification; [1] also recently used TF-IDF and six different ML classifiers on a 2000 news dataset, obtaining a 92% accuracy. The main difficulty in applying content based methods for real world fake news detection is that these news are ”intentionally written to mislead consumers, which makes it nontrivial to detect simply based on news content” [16] . Additionally, [11] recently showed that they could use the news' writing style to effectively discriminate hyper partisan news and satire from mainstream news, but they could not claim ”to have solved fake news detection via style analysis alone”. These difficulties are probably the reason behind the rather limited use of content based methods alone for fake news detection on social media. In fact, on platforms such as SNSs, additional information about social context features is available, which can help identifying fake news with higher accuracies compared to content-based approaches, as shown by [13] for the case of Twitter. Social based methods make use of this additional information, and constitute a more recent strategy for fake news detection on social media [15]. Example of features which have been used for this purpose are the characteristics of users (e.g. registration age, number of followers, etc.) - as proposed by [3] and [5] for the case of Twitter - or their opinions and viewpoints, exploited by [6] to assess credibility of content in the same SNS.

## 5 Model

Convolution neural networks recently have been shown promising results in many different NLP tasks. Since I chose a convolution based baseline for this task proposed in [19], but in order to hopefully improve it I added one attention layer over the text statements to give higher weights to important parts of the text. The final schema of the hybrid convolutional-attentive model is shown in 1. The model consists of two main parts: the text statement model, and meta-data model that I will explain them in the following:

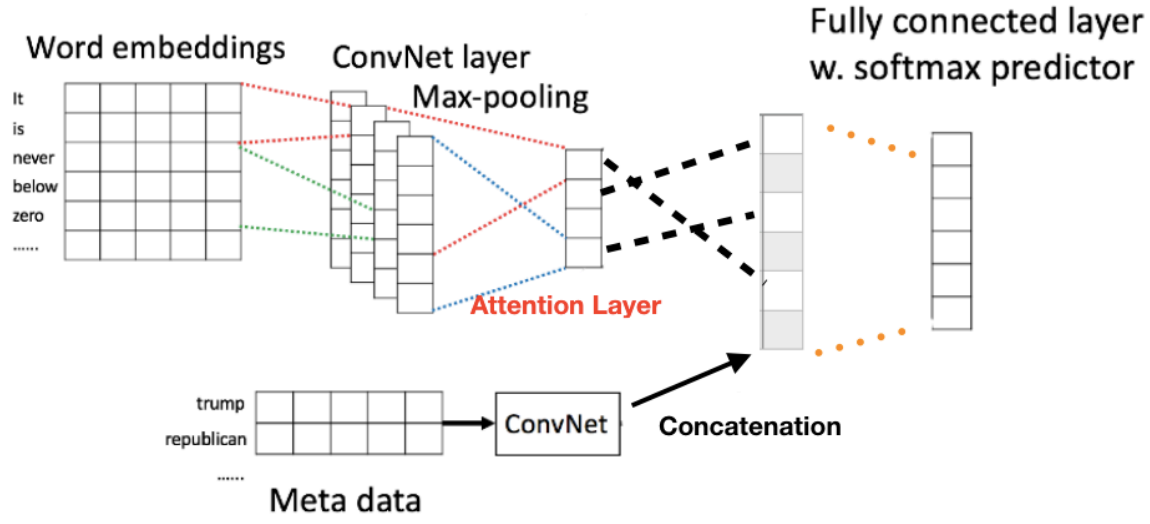


Figure 1: The final hybrid convolutional-attentive model

### 5.1 Text Statement Model

For encoding text statements, I followed CNN-based approach of [7] that trains CNN on top of pretrained word vectors for sentence-level classification tasks. 128 filters with window-size =2,3,5 that I tuned them by 5-fold cross validation are convolved with word embedding matrix to produce a feature map for each filter. Then a max pooling layer is applied over the feature maps to pick the maximum value^for each filter that captures the most important feature for each feature map (filter). This pooling scheme naturally deals with variable sentence lengths.
My main contribution is to apply an attention layer over the convolution results to produce attention weights for each parts of the sentence that determines to which part of the sentence, model should focus during truthfulness detection. The attention layer itself is a 1d convolution layer applied to the output of the CNN to learn different weights for each part of the sentence. The result of this model is a context vector which is fed into the classification layer.

### 5.2 Meta-data Model

For taking use of meta-data which consists of speaker, speaker job, party, context, subject and credit history, I used another CNN network with the same approach defined above but with 10 filters and window-size=3,5 and random initialization of meta-data matrix, because nearly no meaningful embedding exists for proper nouns like Donald Trump or FoxNews that formed the meta-datas.

### 5.3 Hybrid model

At the end, text statements vector resulted from the attention layer and meta-data vector came from the CNN are concatenated to form a 404-d vector and the result is fed into a linear layer that maps the 404-d vector to a 6-d vector containing the probability for each class of truthfulness.

## 6 Implementation and Experiments

For implementing all the models I used PyTorch. At first, all trained the models with random initialization for both text statements and meta-data, but it performed poorly (results are shown in section7), however the performance enhanced considerably when I Initialized word embedding matrix with the publicly available word2vec vectors that were trained on 100 billion words from GoogleNews. The vectors have dimensionality of 300 and were trained using the continuous bag-of-words architecture [9]. Words not present in the set of pretrained words are initialized randomly. I tuned all the hyper parameters on the validation dataset using 5-fold cross validation. The best filter sizes for the CNN model for text statements was (2,3,4) and for meta-data (3,5). For text-statement CNN, each size has 128 filters, however for meta-data CNN each size has 10. I used dropout-rate=0.5,

Adam optimizer, cross entropy loss, learning rate=0.001, and trained baseline model for 10 epochs and the final hybrid-attentive model for 13 epochs.

## 7 Results

I outlined my quantitative results in Table 1. First, I compared the results of baseline model [19] and the proposed convolutional-attentive (CONV-ATTN) model in a 6-level fake news detection with random initialization and with initializing with word2vec vectors. As it is clear from results, initialization with pretrained weights is an important step to improve performance particularly in the absence of a large supervised training set. Both baseline and CONV-ATTN model obtained considerably better results with pretrained initialization. The results shows that for 6-class classification, the accuracy and precision of the CONV-ATTN model is always slightly better than the baseline, however that is not correct for recall and f1.
Afterwards, since I thought that considering 6 different levels for fakeness might not be really realistic and the boundary of some of them are not clear even for humans, I merged each pair of 6 original labels to end up with 3 labels: true (by merging true and mostly true labels), half-true (by merging half-true and barely-true labels) , and false (by merging false and pants-fire labels). I repeated the experiments with pretrained initialization and tested the models again using the whole dataset but in the 3-level of truthfulness framework. The results again showed a slightly better accuracy for CONV-ATNN model. This result was a bit surprising for me bacause I expected the model with attention perform considerably well in comparison with the baseline. I thought the similar performance rather considerably better one has two main reasons:

1. More data is needed for learning attention layer weights. Since I added a layer, the parameters of the model increased that either needs more data or more training epochs to be learned. Because I didn't have more data I trained the model for 3 more epochs in comparison with the baseline, but I early stopped due to not overfitting. So I think more data is needed to throughly asses the CONV-ATTN model.

2. Text statements and meta-data are not enough for predicting if an article is fake or real. By looking at predictions of the model I figured out that for the failure cases even by looking at meta-data such as name of the speaker, job and party, it is even hard for me to predict the correct label. Therefore, it is a high expectation of the model to predict them correctly. Hence, I think more features that take reliability of resource,or credit history of the speaker into account is needed for the task.

In addition some successful and failure cases for 3-class CONV-ATNN model is represented in table 2. As it is obvious from the table, predicting the correct labels only from the content and style of the text even with taking use of some meta-data without external knowledge are difficult even for humans.

### 7.1 Lessons Learned

The project was really interesting, although I spent a lot of time on it and might have not get the predicted result, since I learned many valuable things through it.

- Convolution Networks for text: I have seen many interesting applications of convolution neural networks for images, but not for texts. Through this project by thoroughly studying and implementing CNNs for text statements and meta-data I figured out how it works and how useful it can be in many cases in comarison with simple RNNs, or LSTMs.

- Attention Layer for text: Using attention layer was not a plan that I had from the first point, because I had not used them over texts, but during the project I understood that convolutional architecture that I was using was very similar to the architecture for image-classification, so I thought applying an attention layer to assign higher weights to important parts of the sentence, might be useful for the task. Therefore, I spent the final week learning how attention can be applied over sentences and implementing and testing it, which I am very satisfied by doing that.

| Model | Accuracy | Macro-avg Precision | Macro-avg Recall | Macro-avg F1 |
|---|---|---|---|---|
| baseline-RandInit6 | 22.8 | 20.6 | 20.1 | 0.196 |
| CONV-ATTN-RandInit6 | 23.2 | 20.8 | 20.1 | 0.198 |
| baseline-PreInit6 | 27.3 | 25.4 | 26.2 | 0.248 |
| CONV-ATTN-PreInit6 | 27.9 | 26.0 | 25.9 | 0.247 |
| baseline-PreInit3 | 46.8 | 45.2 | 45.4 | 0.447 |
| CONV-ATTN-PreInit3 | 47.3 | 46.6 | 45.9 | 0.452 |

Table 1: The results of quantitative evaluation of the two trained models, the baseline and convolutional-attentive model, with different initialization and different set of labels.

| Text News | Target | Predicted |
|---|---|---|
| We cut business taxes. so today 70 percent of businesses don't pay tax. | half-true | true |
| Under President George W. Bush, we added 4.9% trillion to the debt. | true | false |
| Hillary Clinton in 2005 co-sponsored legislation that would jail flag burners. | true | false |
| Obamacare will provide insurance to all non-U.S. residents, even if they are here illegally. | false | false |
| In the past 20 years, Egypt has made great strides in political and democratic reform. | false | false |
| The economy is creating jobs at the fastest pace since 1999. | true | true |

Table 2: three successful and three failure cases for the convolutional-attentive model.

- Content is not enough! : An important lesson that I learned from this project is that only text and even some meta-datas that represent content and style of an article are not enough for solving fake news detection. Because the task is difficult inherently , and some extra features to take external information into account are needed to make a fake news detector more reliable.

- Importance of using pretrained weights: Through running the baseline and CONV-ATTN models with different initialization settings, I figured out the importance of using pretrained text embeddings as the warm-start especially when the dataset and vocabulary size is relatively small. However extra effort is needed to make the dataset's words compatible with the ones whose pretrained embedding is available (e.g. stop-word elimination, stemming, casting to lowercase), but it completely worths it.

- Importance of hyper-parameter tuning: Without hyper-parameter I achieved poor results in any case. But after extensive and tiring 5-fold cross validation, I figured out the importance and benefits of this subtask on the final results.

### 7.2 Evaluation

I personally think my project is successful, because I implemented the baseline model from scratch and obtained similar results as it is reported in [19] , although I expected that adding an attention layer to the baseline would increase the accuracy significantly , which it didn't. But I think there is a potential for more experiments and study. I will list some strengths and weaknesses of my project.
Strengths

- Coding and implementing the baseline model and CONV-ATTN from scratch

- Using CNNs for texts

- Taking use of meta-data besides text statements that can help the performance.

- Using attention layer over words in the sentence to give the model flexibility and enable it to focus on important parts of the sentence instead of looking at the sentence as a whole.

- Reporting other metrics besides accuracy which makes the comparison of two models more fair. By just reporting the accuracy one might conclude CONV-ATTN

achieve better results for any metric, however we can see in some cases macro-average recall for the baseline model is better.

Weaknesses

- Not using resource credibility or some other external features
- Reporting the results based on one run for each model with different settings. However, the results would be more reliable if I ran each model several times and averaged the results. But unfortunately I didn't have enough time for that.
- Not testing the model for other datasets rather than LIAR dataset. For proving that models (baseline and CONV-ATTN) are general and they are not overfitting to LIAR dataset, it was better to train them over more datasets. But unfortunately in a 2-month period I couldn't do that.

## 8 Future works

I think there is a high potential for extending this project either by doing more experiments on LIAR dataset for fake news detection or using and testing the CONV-ATTN model for other NLP tasks. I summarized the further works and potential directions in the following.

- Generalizing the CONV-ATTN model and using it in other similar NLP tasks such as stance classification, rumor detection, documents classification, etc.
- Testing the CONV-ATTN model over different datasets such as KaggleFN and FacebookHoax
- Combining the content based CONV-ATTN model with social based models so that they can complete each other as proposed in [17]
- Visualizing attention weights so that we can understand what is happening inside the network and observing which parts of the sentence are important for detecting it as fake, which is really intuitive and beneficial
- Testing a different model rather than CNN for meta-data representing like graph convolution neural networks and comparing the result with CNN. Because GCNs recently has been shown the promising results in deep learning particularly when input is unordered, which is the case for meta-data.

## References


[1] H. Ahmed, I. Traore, and S. Saad. Detection of online fake news using n-gram analysis and machine learning techniques. pages 127–138, 10 2017.

[2] S. Badaskar, S. Agarwal, and S. Arora. Identifying real or fake articles: Towards better language modeling. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II, 2008.

[3] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. Internet Research, 23(5):560–588, 2013.

[4] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. CoRR, abs/1501.03471, 2015.

[5] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In International Conference on Social Informatics, pages 228–243. Springer, 2014.

[6] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[7] Y. Kim. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[8] A. Magdy and N. Wanas. Web-based statistical fact checking of textual documents. In Proceedings of the 2nd international workshop on Search and mining user-generated contents, pages 103–110. ACM, 2010.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[10] S. Mohseni, E. Ragan, and X. Hu. Open issues in combating fake news: Interpretability as an opportunity. arXiv preprint arXiv:1904.03016, 2019.

[11] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638, 2017.

[12] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. arXiv preprint arXiv:1707.03264, 2017.

[13] H. Rodríguez, E. L. Quarantelli, R. R. Dynes, W. A. Andersson, P. A. Kennedy, and E. Ressler. Handbook of disaster research. Springer, 2006.

[14] M. Sharifi, E. Fink, and J. G. Carbonell. Detection of internet scam using logistic regression. In 2011 IEEE International Conference on Systems, Man, and Cybernetics, pages 2168–2172. IEEE, 2011.

[15] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl., 19(1):22–36, Sept. 2017.

[16] K. Shu, S. Wang, and H. Liu. Exploiting tri-relationship for fake news detection. CoRR, abs/1712.07709, 2017.

[17] M. D. Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro. Automatic online fake news detection combining content and social signals. In Proceedings of the 22st Conference of Open Innovations Association FRUCT, FRUCT'22, pages 38:272–38:279, Helsinki, Finland, Finland, 2018. FRUCT Oy.

[18] M. Vuković, K. Pripužić, and H. Belani. An intelligent automatic hoax detection system. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pages 318–325. Springer, 2009.

[19] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. CoRR, abs/1705.00648, 2017.

[20] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. Proceedings of the VLDB Endowment, 7(7):589–600, 2014.

[21] X. Zhou and R. Zafarani. Fake news: A survey of research, detection methods, and opportunities. CoRR, abs/1812.00315, 2018.

# A Appendix

- The data set can be downloaded via this link: https://sites.cs.ucsb.edu/ william/data/liar_dataset.zip.
- All the source codes are available in my github: https://github.com/fjavadi/FakeNewsDetection.