End-to-End Joint Modeling for Fake News Detection

Munshi Mahbubur Rahman **James R. Foulds**
Department of Information Systems, University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250
{mrahman4, jfoulds}umbc.edu

The rapid spread of misinformation, including misleading and manipulative content, is a current and urgent threat to our society and to our democracy (Starbird, 2017). Several fact-checking websites (e.g., Snopes.com and PolitiFact.com) have been formed to manually verify/falsify claims, but this process is expensive and lacks scalability. An automated process to verify these claims is in high demand so that we can keep up with the speed that misinformation spreads.

Early notable works in this domain fit a model (e.g. a neural network) to labeled training instances from sites like PolitiFact.com to predict a claim's veracity (Rashkin et al., 2017). More recent methods also consider external evidence to verify the claims' authenticity. Popat et al. (2017, 2018a,b) leverage articles (retrieved from the Web via a search engine) which confirm or refute a claim and jointly assesses the language style (using subjectivity lexicons), the trustworthiness of the sources, and the credibility of the claims which are provided in natural language form, such as news headlines, quotes from speeches, blog posts, etc.

In this work, we propose to jointly model articles' *relevance to a claim* together with their support of the claim, etc., thereby making these inferences mutually informing. Our approach is the first to use a single unified model to verify user-generated claims in an end-to-end fashion. Our approach builds on the DeClarE DNN architecture proposed by Popat et al. (2018b), later developed and deployed in an online portal named DeepEye (Popat et al., 2018a). First, we construct claim and article representations $\bar{c}_i$, $\bar{a}_j$ by averaging GloVe (Pennington et al., 2014) word embeddings.

$$\bar{c}_i = \frac{1}{L}\sum_{l=1}^{L} c_{i,l} \quad \text{and} \quad \bar{a}_j = \frac{1}{L}\sum_{l=1}^{L} a_{j,l} \qquad (1)$$

Here, L is the number of words per article or claim. We then retrieve a set of all the potentially relevant articles based on cosine similarity of the embeddings with the target claim (e.g. the top 200 articles). Next, we predict the relevance of these candidate articles using an attention mechanism. We concatenate each candidate article j's representation with the target claim i's representation and then apply a neural network layer (e.g. with a *tanh* activation), followed by a softmax transformation to calculate a normalized attention score $\alpha_{j,i}$ which represents the relevance to the target claim i.

$$\hat{a}_{j,i} = \bar{a}_j \oplus \bar{c}_i \,, \quad a'_{j,i} = f(W_a \hat{a}_{j,i} + b_a) \qquad (2)$$

$$\alpha_{j,i} = \frac{exp(a'_{j,i})}{\sum_k exp(a'_{k,i})} \qquad (3)$$

Next, we sum up the article representations, weighted by their attention weights $\alpha_{j,i}$, to construct the relevance-weighted claim representation z_i that feeds into a deep neural network to predict the overall credibility label y_i.

$$z_i = \sum_{k=1}^{N} \alpha_k \bar{a}_k \oplus \bar{c}_i \,, y_i = \text{DNN}(z_i) \qquad (4)$$

The credibility prediction DNN and the attention parameters are trained jointly via back-propagation. We use the dataset published by Popat et al. (2018b) based on Snopes.com, where there are one or more relevant articles snippets per claim and the labels are manually generated. We have tested our initial implementation of the model on 20% held-out data and found out it was 73% accurate. This does not yet outperform a simpler neural network (without article attention), but we are currently working to refine our implementation and we expect that we will be able to present improved results at the symposium.

References

Jeffrey Pennington, Richard Socher, and Christopher D 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.

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, Jannik Strötgen, and Gerhard Weikum. 2018a. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, pages 155–158.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018b. 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*, pages 22–32.

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.

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Eleventh AAAI International Conference on Web and Social Media (ICWSM)*.