Taking a Stance on Fake News

Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection

Chris Dulhanty {chris.dulhanty@uwaterloo.ca}, **Jason L. Deglint, Ibrahim Ben Daya and Alexander Wong**

Systems Design Engineering, University of Waterloo, Waterloo, ON

Introduction

The exponential rise of social media and digital news in the past decade has had the unfortunate consequence of escalating what the United Nations has called a global topic of concern: **the growing prevalence of disinformation.** [1]

Given the complexity and time-consuming nature of combating disinformation through human assessment, one is motivated to explore harnessing AI solutions to automatically assess news articles for the presence of disinformation. A valuable step towards automatic identification of disinformation is **stance detection.**

Contributions

- Developed a large-scale language model for stance detection via transfer learning of a RoBERTa deep bidirectional transformer model [2] with claim-article pairs via pair encoding with self-attention
- State-of-the-art results on Fake News Challenge, Stage 1 benchmark [3]

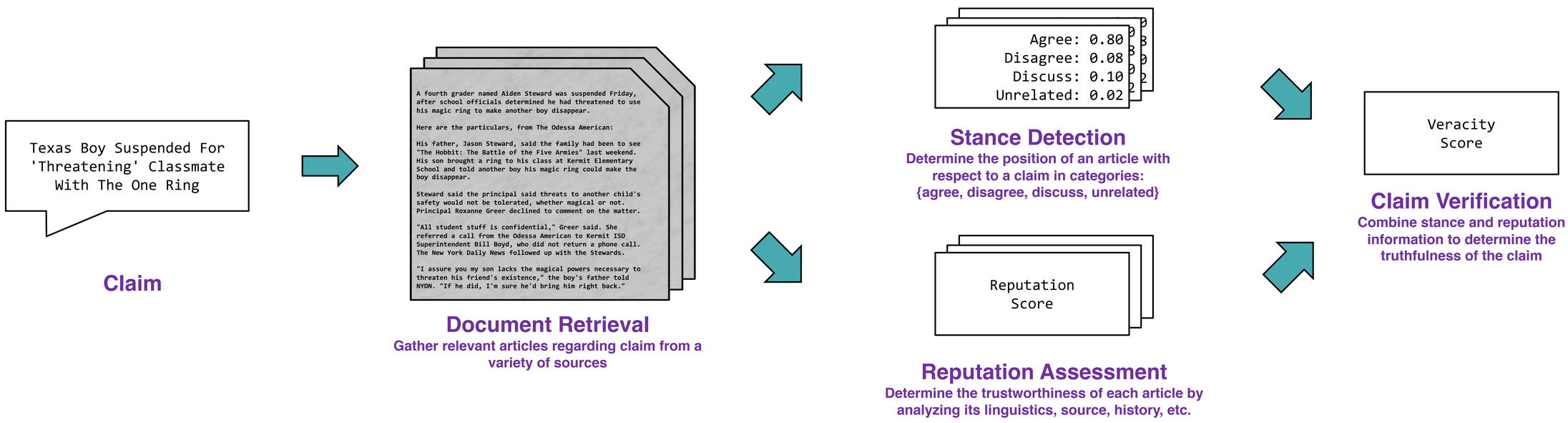

Figure 1: Automated Fact-Checking Process

Dataset

Fake News Challenge, Stage 1 (2017) [3]

Estimate the stance of an article with respect to a claim. Data derived from Emergent dataset [4], sourced from the Emergent Project [5], a real-time rumour tracker created by Tow Center for Digital Journalism at Columbia.

	Training Set	Test Set
Claim-Article Pairs (#)	49,972	25,413
Unrelated (%)	73.13	72.20
Discuss (%)	17.83	17.57
Agree (%)	7.36	7.49
Disagree (%)	1.68	2.74

Table 1: Statistics of the FNC-I Dataset

Methodology

- $\text{RoBERTa}_{\text{BASE}}$ pre-trained model on five English-language corpora (>160GB)
- Tokenize input with byte-level byte-pair-encoding, add special tokens
- Trim or pad claim or article (longest first) to fit **maximum sequence length of 512**
- Train for three epochs with learning rate of 2e-5, weight decay of 0.1, batch size of 8
- Trained on one NVIDIA 1080Ti using HuggingFace's transformers library [6]

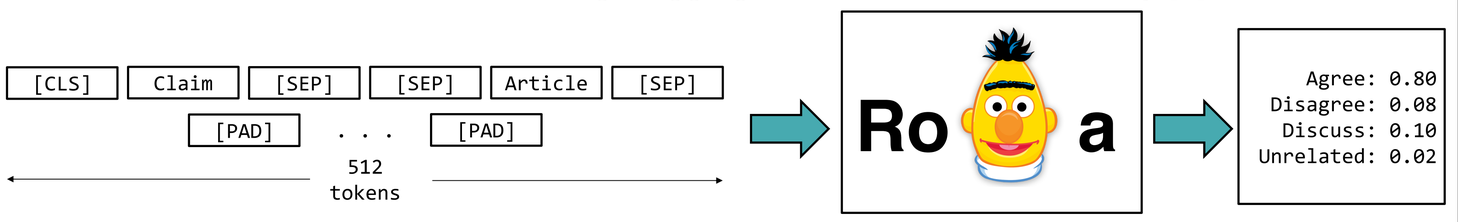

Figure 2: RoBERTa Model Training Setup

Results

Weighted Accuracy (%) = $0.25 \times \text{Acc}_{\text{related}} + 0.75 \times \text{Acc}_{\text{stance}}$

where: $\textbf{Acc}_{\textbf{related}}$ - binary accuracy across related {agree, disagree, discuss} and unrelated claim-article pairs

$\textbf{Acc}_{\textbf{stance}}$ - accuracy for claim-article pairs in related classes only

Method	Weighted Accuracy (%)	Accuracy (%)
Riedel et al. [7]	81.72	88.46
Hanselowski et al. [8]	81.97	89.48
Baird et al. [9]	82.02	89.08
Bhatt et al. [10]	83.08	89.29
Borges et al. [11]	83.38	89.21
Zhang et al. 2018 [12]	86.66	92.00
Wang et al. [13]	86.72	82.91
Zhang et al. 2019 [14]	88.15	93.50
Proposed Method	**90.01**	**93.71**

Table 2: Performance of various methods on the FNC-I benchmark. First and second groups are methods introduced during and after the challenge period, respectively.

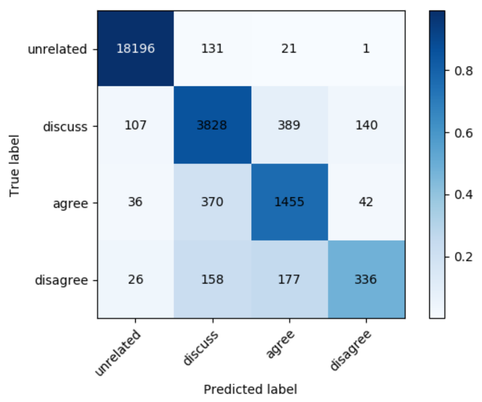

Figure 3: Confusion matrix for proposed method.

Number of Token in Example	Accuracy (%)	Number of Examples
<129	92.05	2,904
129-256	93.90	3,606
257-384	95.07	6,328
385-512	**95.11**	4,763
>512	92.23	7,812
All	93.71	25,413

Table 3: Effect of claim-article pair sequence length of FNC-I test set on classification accuracy of RoBERTa model, with a maximum sequence length of 512.

Maximum Number of Tokens	Weighted Accuracy (%)	Accuracy (%)
128	89.52	93.46
256	89.54	93.48
512	**90.01**	**93.71**

Table 4: Effect of maximum sequence length of RoBERTa model on weighted accuracy and classification accuracy.

Proposed Implementation

- Validate with journalists and professional fact-checkers
- Develop simple browser plug-in to assist individuals to stay informed citizens

Ethical Considerations

Limitations:

- Trained solely on claims and articles in English, from Western-focused media outlets
- Not designed to deal with satire

Risks:

- Codification of unintended biases (gender, racial) into contextual word embeddings through biased pre-training methods, finetuning on FNC-1 dataset [15]
- Prone to adversarial attacks [16]

Unintended Negative Outcomes:

- Interpreted as a definitive answer, rather than an estimate of veracity – individuals defer own judgement to algorithm
- Malicious actors selectively promote claims misclassified by model but adhere to their own agendas

References

[1] https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21287&LangID=E
[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019
[3] Dean Pomerleau and Delip Rao. Fake news challenge, 2017. http://www.fakenewschallenge.org/
[4 William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, pages 1163–1168, 2016]
[5] http://www.emergent.info/
[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, AnthonyMoi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771, 2019.
[7] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. CoRR,abs/1707.03264, 2017.
[8] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, and Felix Caspelherr. Description of the system developed by team athene in the fnc-1, 2017
[9] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory, 2017.
[10] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. Combining neural, statistical and external features for fake news stance identification. In Companion Proceedings of the The Web Conference 2018, WWW '18, pages 1353–1357, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.
[11] Luís Borges, Bruno Martins, and Pável Calado. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. ACM Journal of Data Information and Quality, 2019.
[12] Qiang Zhang, Emine Yilmaz, and Shangsong Liang. Ranking-based method for news stance detection. In Companion Proceedings of the The Web Conference 2018, WWW '18, pages 41–42, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.
[13] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. Relevant document discovery for fact-checking articles. In Companion Proceedings of the The Web Conference 2018, WWW'18, pages 525–533, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.
[14] Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. From stances' imbalance to their hierarchical representation and detection. In The World Wide Web Conference,WWW '19, pages 2323–2332, New York, NY, USA, 2019. ACM.
[15] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In Advances in Neural Information Processing Systems 2019.
[16] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In Proceedings of the 57th Conference of the Association for Computational Linguistics, pages 1520–1529, 2019