Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection

Chris Dulhanty, Jason L. Deglint, Ibrahim Ben Daya and Alexander Wong
Systems Design Engineering, University of Waterloo
{chris.dulhanty, jdeglint, ibendaya, a28wong}@uwaterloo.ca

Abstract

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 first step towards automatic identification of disinformation is stance detection, where given a claim and a news article, the aim is to predict if the article agrees, disagrees, takes no position, or is unrelated to the claim. Existing approaches in literature have largely relied on hand-engineered features or shallow learned representations (e.g., word embeddings) to encode the claim-article pairs, which can limit the level of representational expressiveness needed to tackle the high complexity of disinformation identification. In this work, we explore the notion of harnessing large-scale deep bidirectional transformer language models for encoding claim-article pairs in an effort to construct state-of-the-art stance detection geared for identifying disinformation. Taking advantage of bidirectional cross-attention between claim-article pairs via pair encoding with self-attention, we construct a large-scale language model for stance detection by performing transfer learning on a RoBERTa deep bidirectional transformer language model, and were able to achieve state-of-the-art performance (weighted accuracy of 90.01%) on the Fake News Challenge Stage 1 (FNC-I) benchmark. These promising results serve as motivation for harnessing such large-scale language models as powerful building blocks for creating effective AI solutions to combat disinformation.

1 Introduction

Disinformation presents a serious threat to society, as the proliferation of *fake news* can have a significant impact on an individual's perception of reality. Fake news is a claim or story that is fabricated, with the intention to deceive, often for a secondary motive such as economic or political gain [1]. In the age of digital news and social media, fake news can spread rapidly, impacting large amounts of people in a short period of time [2]. To mitigate the negative impact of fake news on society, various organizations now employ personnel to verify dubious claims through a manual fact-checking procedure, however, this process is very laborious. With a fast-paced modern news cycle, many journalists and fact-checkers are under increased stress to be more efficient in their daily work. To assist in this process, *automated fact-checking* has been proposed as a potential solution [3, 4, 5, 6, 7].

[1] https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21287&LangID=E

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

Automated fact-checking systems aim to assess the veracity of claims through the collection and assessment of news articles and other relevant documents pertaining to the claim at hand. These systems have the potential to augment the work of professional fact-checkers, as well as provide a tool to the public to verify claims they come across online or in their daily lives. An automated fact-checking system consists of several sub-tasks that, when combined, can predict if a claim is truthful [8]. *Document retrieval* aims to gather relevant articles regarding the claim from a variety of sources. *Stance detection* aims to determine the position of each article with respect to the claim. *Reputation assessment* aims to determine the trustworthiness of each article by analyzing its linguistics and source. *Claim verification* aims to combine stance and reputation information to determine the truthfulness of the claim.

In this paper, we focus on stance detection; given a proposed claim and article, predict if the article agrees, disagrees, has no stance, or is unrelated to the claim. Within the natural language processing (NLP) community, research in stance detection has been catalyzed by the organization of competitions [9, 10, 11] and the collection of benchmark datasets [12, 13, 14]. Prominent methods addressing stance detection largely differ in terms of their feature representation (e.g., *n*-grams, TF-IDF, word embeddings, etc.) and algorithms (e.g., decision trees, multi-layer perceptions, LSTM networks, etc.); retrospectives on recent challenges [9, 10, 15] provide a comprehensive overview of NLP methods in stance detection. While results have been promising, recent developments in NLP hold the potential for significant improvement. Whereas pre-trained word embeddings such as word2vec [16] and GloVe [17] encode language into shallow numerical representations for input to machine learning models, deep bidirectional transformer language models [18, 19, 20, 21, 22] train on large, unlabelled datasets to learn deeper hierarchical representations of language. The result has been a significant improvement on multi-task NLP benchmarks [23], akin to an "ImageNet moment"[2] for the field.

Motivated by recent advances in NLP and the potential of this technology to meaningfully impact society by addressing the United Nation's Sustainable Development Goals of "Quality Education" and "Peace, Justice, and Strong Institutions"[3], we explore the notion of harnessing large-scale deep bidirectional transform language models for achieving state-of-the-art stance detection. Our major contributions are: (1) constructing a large-scale language model for stance detection by performing transfer learning on a RoBERTa deep bidirectional transformer language model by taking advantage of bidirectional cross-attention between claim-article pairs via pair encoding with self-attention, and (2) state-of-the-art results on the Fake News Challenge Stage 1 (FNC-I)[4] benchmark.

2 Methodology

The RoBERTa (Robustly Optimized BERT Approach) model, released in July 2019 by Liu *et al.* [24], is an open-source language model that achieves state-of-the-art results on benchmark NLP multi-task General Language Understanding Evaluation (GLUE) benchmark [23]. RoBERTa is built upon the BERT (Bidirectional Encoder Representations from Transformers) model, released by Devlin *et al.* in October 2018 [20]. RoBERTa and BERT achieve high performance by pretraining a transformer model, initially proposed by Vaswani *et al.* [18], in a bidirectional manner on a very large corpus of unlabelled text, and fine-tuning the model on a relatively small amount task-specific labelled data. These models are well-suited for use in stance detection as the pretrained model can be leveraged to perform transfer learning on the target task. Using deep bidirectional transformer language models, RoBERTa and BERT have the ability to gain a deeper understanding of language and context when compared to earlier unidirectional transformer architectures [20]. In addition, RoBERTa demonstrates great results on sentence-pair classification tasks of GLUE, such as Multi-Genre Natural Language Inference [25] and Question Natural Language Inference [26, 23], tasks very similar in nature to the claim-article classification of stance detection. Following RoBERTa's method of fine-tuning on GLUE tasks, we include both claim and article, separated by a special token, in each example during training and inference.

[2] http://ruder.io/nlp-imagenet/

[3] https://sustainabledevelopment.un.org

[4] http://www.fakenewschallenge.org/

Table 1: Statistics of the FNC-I Dataset

	Training Set	Test Set
# of 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

3 Experiments and Analysis

3.1 Dataset

To investigate the task of stance detection in the context of fake news detection, we use data released for the Fake News Challenge, Stage 1 (FNC-I). The challenge was organized by Pomerleau and Rao in 2017, with the goal of estimating the stance of an article with respect to a claim. Data is derived from the Emergent dataset [12], sourced from the Emergent Project[5], a real-time rumour tracker created by the Tow Center for Digital Journalism at Columbia University. The stance takes one of four labels: **Agree** if the article agrees with the claim, **Disagree** if the article disagrees with the claim, **Discuss** if the article is related to the claim, but the author takes no position on the subject, and **Unrelated** if the content of the article is unrelated to the claim. There are approximately 50k claim-article pairs in the training set and 25k pairs in the test set; Table 1 summarizes the data distribution.

3.2 Metrics

To evaluate the performance of our method, we report standard accuracy as well as weighted accuracy, suggested by the organizers of the Fake News Challenge, as it provides a more objective metric for comparison given the class imbalance in the dataset. The weighted accuracy, Acc_w, is expressed as:

$$Acc_w = 0.25 \times Acc_{r,u} + 0.75 \times Acc_{a,d,d} \tag{1}$$

where $Acc_{r,u}$ is the binary accuracy across related {agree, disagree, discuss} and unrelated article-claim pairs, and $Acc_{a,d,d}$ is the accuracy for pairs in related classes only.

3.3 Model

We construct our large-scale language model via transfer learning on a pretrained RoBERTa$_{\text{BASE}}$ deep transformer model, consisting of 12-layers of 768-hidden units, each with 12 attention heads, totalling 125M parameters. We leverage the Transformers library by Hugging Face for implementation [27]. To perform transfer learning, we train for three epochs and follow hyperparameter recommendations by Liu *et al.* [24] for fine-tuning on GLUE tasks, namely, a learning rate of 2e-5 and weight decay of 0.1. We train on one NVIDIA 1080Ti GPU with a batch size of 8.

Prior to training, the dataset is pre-processed by initializing each example with a start token to signify the beginning of a sequence, followed by the claim, two separator tokens, the article and an additional separator token. The sequence is then tokenized by RoBERTa's byte-level byte-pair-encoding and trimmed or padded to fit a maximum sequence length of 512. We explore the effects of claim-article pair sequence length and maximum sequence length on classification accuracy in the Appendix.

3.4 Results & Discussion

Results of our proposed method, the top three methods in the original Fake News Challenge, and the best-performing methods since the challenge's conclusion on the FNC-I test set are displayed in Table 2. A confusion matrix for our method is presented in the Appendix. To the best of our knowledge, our method achieves state-of-the-art results in weighted-accuracy and standard accuracy on the dataset. Notably, since the conclusion of the Fake News Challenge in 2017, the weighted-accuracy error-rate has decreased by 8%, signifying improved performance of NLP models and innovations in the domain ~~of stance detection, as well as a con~~tinued interest in combating the spread of disinformation.

[5] `http://www.emergent.info/`

Table 2: Performance of various methods on the FNC-I benchmark. The first and second groups are methods introduced during and after the challenge period, respectively. Best results are in **bold**.

Method	Acc_w	Acc
Riedel *et al.* [28]	81.72	88.46
Hanselowski *et al.* [29]	81.97	89.48
Baird *et al.* [30]	82.02	89.08
Bhatt *et al.* [31]	83.08	89.29
Borges *et al.* [32]	83.38	89.21
Zhang *et al.* 2018 [33]	86.66	92.00
Wang *et al.* [34]	86.72	82.91
Zhang *et al.* 2019 [35]	88.15	93.50
Proposed Method	**90.01**	**93.71**

4 Ethical Considerations

Implementation and potential end-users: The implementation of our stance detection model into a real-world system is predicated on the development of solutions to the document retrieval, reputation assessment and claim verification elements of an automated fact-checking system. While this is an active field of research, it is imperative to note that the reputation assessment sub-task is difficult, as the trustworthiness of an individual or media source may be interpreted differently by different individuals due to personal bias. Provided these elements can be developed, the first intended end-users of an automated fact-checking system should be journalists and fact-checkers. Validation of the system through the lens of experts of the fact-checking process is something that the system's performance on benchmark datasets cannot provide. The implementation of such a system into the daily workflow of these individuals is likely a field of research onto itself. Ultimately, the development of a simple user interface for the general public, such as a browser plug-in, is the goal of this system, assisting individuals to stay informed citizens.

Limitations: The model proposed in this work is limited by the fact that it was trained solely on claims and articles in English, from western-focused media outlets. Further work is necessary to extend this work to other languages, where differences in writing style and cultural norms and nuances may lead to differences in performance. In addition, this model is not designed to deal with satire, where the stance of an article with respect to a claim may appear on the surface to be one way, but the underlying intention of its author is to exploit humor to demonstrate an opposing viewpoint.

Risks and potential unintended negative outcomes: A major risk of a stance detection model or automated fact-checking system is the codification of unintended biases into the model through biased training data. In the field of NLP, gender and racial biases have been reported in word embeddings [36, 37] and captioning models [38]; the extent to which such social biases are encoded in recently developed language models is only beginning to be studied [39, 40]. A secondary risk to the roll-out of these systems for adversarial attacks. Early work by Hsieh *et al.* to investigate the robustness of self-attentive architectures has demonstrated that adversarial examples that could mislead neural language models but not humans are capable of being developed for sentiment analysis, entailment and machine translation [41]. In addition, the development of such a system may be interpreted by some as to provide a definitive answer with respect to the truthfulness of a claim, rather than a predictive estimate of its veracity. A potential unintended negative outcome of this work is for people to take the outputs of an automated fact-checking system as the definitive truth, without using their own judgement, or for malicious actors to selectively promote claims that may be misclassified by the model but adhere to their own agenda.

5 Conclusions

We have presented a state-of-the-art large-scale language model for stance detection based upon a RoBERTa deep bidirectional transformer. Our promising results motivate efforts to develop additional sub-components of a fully automated fact-checking system such that AI can effectively be harnessed to combat disinformation and allow citizens and democratic institutions to thrive.

References

[1] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[2] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[3] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600, 2014.

[4] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[5] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully automated fact checking using external sources. *arXiv preprint arXiv:1710.00341*, 2017.

[6] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812. ACM, 2017.

[7] Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. Fakta: An automatic end-to-end fact checking system. *arXiv preprint arXiv:1906.04164*, 2019.

[8] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, 2014.

[9] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016.

[10] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[11] Dean Pomerleau and Delip Rao. Fake news challenge, 2017.

[12] 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.

[13] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, 2017.

[14] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, 2019.

[15] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*, 2018.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

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

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[22] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[23] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[24] 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.

[25] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[26] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, 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.

[28] 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.

[29] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, and Felix Caspelherr. Description of the system developed by team athene in the fnc-1, 2017.

[30] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory, 2017.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[37] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[38] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

[39] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring Bias in Contextualized Word Representations. *arXiv e-prints*, page arXiv:1906.07337, Jun 2019.

[40] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13209–13220. Curran Associates, Inc., 2019.

[41] 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.

A Claim-Article Pair Sequence Length

Table 3 presents the results of the RoBERTa model on the FNC-I test set, based on the length of claim-article pair. The model has a maximum sequence length of 512 tokens, so any examples longer than this are trimmed. We find that the model performs best for examples that utilize the full capacity of the input sequence (385 to 512 tokens). Very short sequences (<129 tokens) provide the least amount of information to the model, and the model performs poorly. Long sequences (>512 tokens) have some of their context removed from their input, and these examples also perform relatively poor.

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.

Number of Tokens in Example	Acc	Number of Examples
<129	92.05	2904
129-256	93.90	3606
257-384	95.07	6328
385-512	**95.11**	4763
>512	92.23	7812
All	93.71	25413

B Maximum Sequence Length

Table 4 presents the results of RoBERTa models of varying maximum sequence lengths on the FNC-I test set. We find an increase in accuracy with a longer maximum sequence length, as more context is provided to the model. We cannot increase the length of the input sequence beyond 512 tokens without training the RoBERTa model from scratch, which is not feasible for us.

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

Maximum Number of Tokens	Acc_w	Acc
128	89.52	93.46
256	89.54	93.48
512	**90.01**	**93.71**

C Confusion Matrices

Figures 1 and 2 present confusion matrices for the previous best method and our proposed method on the FNC-I test set.

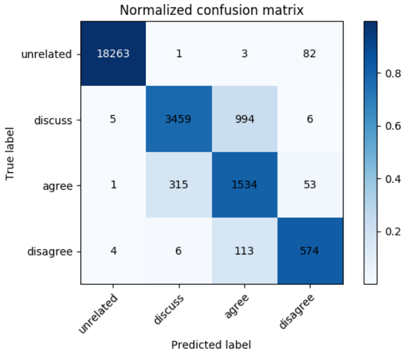

Figure 1: Confusion matrix for Zhang *et al.* [35].

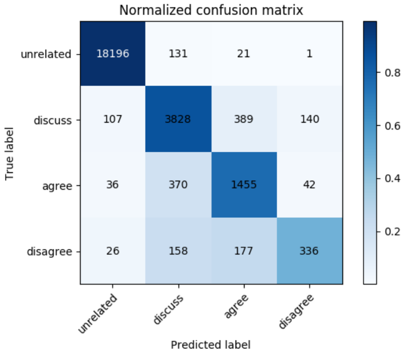

Figure 2: Confusion matrix for proposed method.