Leveraging BERT Encodings for Open-Domain Stance Classification[*]

Lucas Doust Alba[1,*], Tommy Yuan[2]

[1]University of York, Heslington, York YO10 5DD, United Kingdom
[2]University of York, Heslington, York YO10 5DD, United Kingdom

Abstract

This work aims to explore and evaluate a method of generalizing the sub-systems currently used in composite autonomous debating systems. Specifically, it aims to evaluate the feasibility of tackling the task of open domain stance classification with large, language representation models in order to reduce our reliance on domain-specific corpora and training. To this end, novel variants of the BERT architecture inspired by Target Aspect Based Sentiment Analysis (TABSA) techniques are introduced and evaluated. The results show a significant improvement in adversarial sentence classification over the Base BERT model while offering no increase in performance on more straightforward inputs.

Keywords

Open-domain, stance-classification, BERT, autonomous-debate, deep-learning

1. Introduction and Related work

Recent years have seen an emerging interest in “composite AI” systems capable of integrating a wide array of cognitive proficiencies in a holistic manner. One such example is IBM’s project debater (PD), an autonomous debating system made up of smaller, “narrow AI” components [1]. This work aims to explore the feasibility of substituting these “narrow” components with easily fine-tunable and accessible language representation models. Specifically it focuses on the open domain stance classification sub-system due to the fact that PD’s most pervasive errors (local errors) are mainly caused by stance or context miss-classification [1]. To this end, Google’s Bidirectional Encoder Representations from Transformers (BERT), a model reported to reach SOTA results on the GLUE benchmark with minimal fine-tuning [2], as well as various NLP tasks including sentiment analysis [3] [4], stance classification [5] and most influential to this paper, TABSA [6], is benchmarked on this task along with novel variants.

Stance classification is considered a novel research area emerging as a sub-problem of sentiment analysis which aims to determine the stance of a $Perspective(P)$ with respect to a $Topic(T)$, typically into one of three categories {Favour, Against, Neither} [7]. As such, the area overlaps with areas related to sentiment analysis such as TABSA. A frequently employed approach for the task is the use of ensemble systems comprising both feature learning and lexical pattern

CMNA’22: Workshop on Computational Models of Natural Argument, September 12-13, 2022

[*] https://github.com/lda518/cmna_src

[*]Corresponding author.

l.doustalba@gmail.com (L. D. Alba); tommy.yuan@york.ac.uk (T. Yuan)

CEUR Workshop Proceedings (CEUR-WS.org)

matching elements [7] and is used by PD's stance classifier [1]. While effective, the system relies on labour intensive corpora and its lexical matching component introduces a positive bias into the classification process [8]. Novel deep learning approaches for stance classification which alleviate this reliance have begun to emerge with SOTA performances by employing attention mechanisms [9] [10] and may be poised to surpass feature based methods as more, better quality datasets are made publicly available [11].

2. Method and Setup

The novel architectures presented in this paper aim to modify BERT in order to improve its stance classification proficiency by exploiting its intrinsic link to TABSA. The models will attempt to leverage the semantic value of sentence 'target pairs' which consist of the extracted noun phrase pair with the greatest semantic overlap between input sentences *P* and *T*. Semantic overlap is evaluated by a 'Target extractor' in three different ways depending on the variant of the architecture and indicated according to the following naming convention: 'syn' models evaluate the Wordnet Synset path distances [12], 'cos' models evaluate the cosine similarity between the embedded feature vectors (a technique employed in[13]) of each noun produced by the BERT encoder, and 'cosyn' models evaluate a normalized average of the two metrics.

The base BERT model is adapted for stance classification by using the standard sentence pair approach [5] [3]. This results in both a 'Sequential' and a 'Pooled' output where the former represents an array of embedded feature vectors corresponding to each word from the input sentences and the latter to the calculated stance classification probability [3].

The variants build upon this base model by selecting the 'target pairs' from the input sentences depending on the semantic evaluation metric. They then extract and maxpool the embedded vectors corresponding to these 'target pairs' from the 'Sequential' output. Following this, the maxpooled vector along with the 'Pooled output' are combined through either a concatenation or multiplication operation. This is represented in the naming convention as either a 'con' or 'mul' suffix. The combined vector is then passed through a standard dense, dropout and softmax layer resulting in the final classification.

For benchmarking, the base BERT model with the recommended hyperparameters is considered [3] along with the out-of-the-box, most up to date PD stance classifier provided by IBM through an API service[14]. These models are evaluated against both the 'Perspectrum'[15] - a relatively small and adversarial dataset extracted from debate forums - and the 'Multilingual Argument Mining'[16] - a larger, less adversarial dataset mined from a large number of Wikipedia articles - datasets.

3. Results and Discussion

In terms of accuracy, the BERT base model and the novel variants were able to outperform PD in both datasets. However, the novel variants were only able to outperform the BERT model on the more adversarial Perspectrum dataset. This indicates that the additional target features employed by the novel variants provide a benefit to stance classification when inputs are adversarial or ambiguous. One could hypothesize that the noun phrase pairs act almost as a

Table 1
Results on the Perspectrum dataset

Model	Accuracy	Precision	Recall	f1
pd	0.7	0.69	0.77	0.73
bert	0.74	0.76	0.75	0.75
bert_cos_con	0.79	0.8	0.81	0.8
bert_cos_mul	0.78	0.77	0.81	0.79
bert_syn_con	0.77	0.78	0.8	0.79
bert_syn_mul	0.77	0.77	0.79	0.78
bert_cosyn_con	0.78	0.78	0.81	0.79
bert_cosyn_mul	0.77	0.78	0.77	0.77

Table 2
Results on the IBM Multilingual dataset

Model	Accuracy	Precision	Recall	f1
pd	0.77	0.72	0.88	0.79
bert	0.91	0.9	0.92	0.91
bert_cos_con	0.91	0.89	0.93	0.91
bert_cos_mul	0.87	0.88	0.84	0.86
bert_syn_con	0.9	0.9	0.91	0.9
bert_syn_mul	0.88	0.88	0.86	0.87
bert_cosyn_con	0.89	0.88	0.91	0.89
bert_cosyn_mul	0.91	0.9	0.92	0.91

semantic bridge between the two inputs which assists the network in identifying tangentially related topics.

In terms of the Precision, Recall and F1 values (tables:1, 2), the PD model consistently scores a higher Recall than Precision metric confirming the bias issue referenced in its paper [8]. In contrast, the precision and recall values of the BERT based models indicate an absence of bias.

Despite the BERT based models achieving these results with minimal fine-tuning, the added computational cost of running the models obscures whether or not their benefits outweigh their cost. In terms of more straightforward, explicit inputs, the BERT model provides a stronger argument for its adoption than on the adversarial dataset set. If current trends continue, however, we may see these models make up for their computational complexity relatively soon.

Due to a lack of resources, this paper serves more as a proof of concept than a presentation of a novel model which should be adopted. More investigation should be made into adapting and modifying these larger language representation models in order to tackle more semantically adversarial tasks, an approach which has proven to be effective again and again in the literature [5] [6] [4]. Improvements to the target extractor module should also be investigated, perhaps via the use of a logistic regression classifier leveraging BERT encoded feature vectors.

References

[1] N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein, L. Ein-Dor, R. Friedman-Melamed, A. Gavron, A. Gera, M. Gleize, S. Gretz, D. Gutfreund, A. Halfon, D. Hershcovich, R. Hoory, Y. Hou, S. Hummel, M. Jacovi, C. Jochim, Y. Kantor, Y. Katz, D. Konopnicki, Z. Kons, L. Kotlerman, D. Krieger, D. Lahav, T. Lavee, R. Levy, N. Liberman, Y. Mass, A. Menczel, S. Mirkin, G. Moshkowich, S. Ofek-Koifman, M. Orbach, E. Rabinovich, R. Rinott, S. Shechtman, D. Sheinwald, E. Shnarch, I. Shnayderman, A. Soffer, A. Spector, B. Sznajder, A. Toledo, O. Toledo-Ronen, E. Venezian, An autonomous debating system, Nature 591 (2021) 379–384. doi:10.1038/s41586-021-03215-w.

[2] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, 2019. arXiv:1804.07461.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (2019). arXiv:1810.04805.

[4] C. Sun, L. Huang, X. Qiu, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, arXiv:1903.09588 [cs] (2019). arXiv:1903.09588.

[5] K. Popat, S. Mukherjee, A. Yates, G. Weikum, STANCY: Stance Classification Based on Consistency Cues, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6412–6417. doi:10.18653/v1/D19-1675.

[6] Z. Gao, A. Feng, X. Song, X. Wu, Target-Dependent Sentiment Classification With BERT, IEEE Access 7 (2019) 154290–154299. doi:10.1109/ACCESS.2019.2946594.

[7] D. Küçük, F. Can, Stance Detection: A Survey, ACM Computing Surveys 53 (2021) 1–37. doi:10.1145/3369026.

[8] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, N. Slonim, Stance Classification of Context-Dependent Claims, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 251–261. doi:10.18653/v1/E17-1024.

[9] P. Wei, W. Mao, D. Zeng, A Target-Guided Neural Memory Model for Stance Detection in Twitter, in: 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8. doi:10.1109/IJCNN.2018.8489665.

[10] Y. Zhou, A. I. Cristea, L. Shi, Connecting Targets to Tweets: Semantic Attention-Based Model for Target-Specific Stance Detection, in: A. Bouguettaya, Y. Gao, A. Klimenko, L. Chen, X. Zhang, F. Dzerzhinskiy, W. Jia, S. V. Klimenko, Q. Li (Eds.), Web Information Systems Engineering – WISE 2017, volume 10569, Springer International Publishing, Cham, 2017, pp. 18–32. doi:10.1007/978-3-319-68783-4_2.

[11] D. Küçük, F. Can, Stance Detection: A Survey, ACM Computing Surveys 53 (2021) 1–37. doi:10.1145/3369026.

[12] G. A. Miller, WordNet: A lexical database for English, Communications of the ACM 38

(1995) 39–41. doi:10.1145/219717.219748.

[13] H. Karande, R. Walambe, V. Benjamin, K. Kotecha, T. Raghu, Stance detection with BERT embeddings for credibility analysis of information on social media, PeerJ Computer Science 7 (2021) e467. doi:10.7717/peerj-cs.467.

[14] Project Debater API, https://early-access-program.debater.res.ibm.com/early-access-program.debater.res.ibm.com, ????

[15] S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, D. Roth, Seeing Things from a Different Angle:Discovering Diverse Perspectives about Claims, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 542–557. doi:10.18653/v1/N19-1053.

[16] O. Toledo-Ronen, M. Orbach, Y. Bilu, A. Spector, N. Slonim, Multilingual Argument Mining: Datasets and Analysis, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 303–317. doi:10.18653/v1/2020.findings-emnlp.29.