SStance Detection is not Classification: Increasing the Role of Target Entities for Detecting Stance

Ayush Kaushal and **Avirup Saha** and **Niloy Ganguly**
Indian Institute of Technology Kharagpur
{ayushk4, saha.avirup}@gmail.com, niloy@cse.iitkgp.ac.in

Abstract

The stance detection task aims at detecting the stance of a tweet or a text for a target. These targets can be named entities or free-form sentences (claims). Though the task involves reasoning of the tweet with respect to a target, we find that it is possible to achieve high accuracy on several publicly available Twitter stance detection datasets without looking at the target sentence. Specifically, a simple tweet classification model achieved human-level performance on the WT–WT dataset and more than two-third accuracy on various other datasets. We investigate the existence of biases in such datasets to find the potential spurious correlations of sentiment-stance relations and lexical choice associated with the stance category. Furthermore, we propose a new large dataset free of such biases and demonstrate its aptness on the existing stance detection systems. Our empirical findings show much scope for research on the stance detection task and proposes several considerations for creating future stance detection datasets.[1]

1 Introduction

Stance detection is a vital sub-task for fake news detection (Pomerleau and Rao, 2017), automated fact checking (Vlachos and Riedel, 2014; Ferreira and Vlachos, 2016), social media analysis (Zhang et al., 2017), analyzing online debates (Bar-Haim et al., 2017) and rumour verification (Derczynski et al., 2017; Gorrell et al., 2019). Furthermore, it is also an essential measure for progress in Natural Language Understanding, especially in the noisy-text domain.

Over the recent years, several stance detection datasets have been proposed. These datasets, in turn, facilitated progress in stance detection research, with some systems achieving up to 93.7% accuracy (Dulhanty et al., 2019). However, most

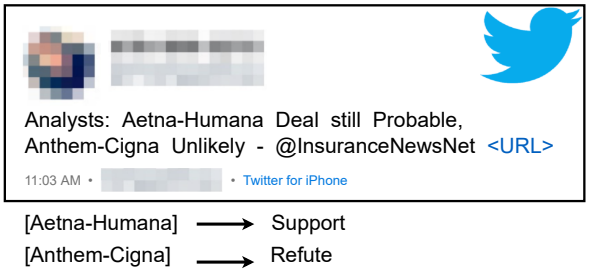


Figure 1: An example tweet from WT–WT dataset with different targets.

of these state-of-the-art systems are complex deep neural networks, making them difficult to interpret. Lack of explainability raises concern since previous works (Gururangan et al., 2018; Goyal et al., 2017; Cirik et al., 2018; Geva et al., 2019) on other tasks demonstrated that superficial dataset biases could result in inflated test-set performance. With this motivation, we carry out the first study analyzing several publicly available Twitter stance detection datasets. Our experiments reveal rampant biases in datasets through which even target-oblivious models can achieve impressive performance.

Various existing works have hinted at the presence of such dataset biases. For example, TAN model (Du et al., 2017) is a very competitive stance detection model.However, Ghosh et al. (2019) recently proved that TAN does not take advantage of target information at all. In RumourEval-2017 (Derczynski et al., 2017), models delivered up to 0.74 accuracy without any knowledge of the target, being only short of 0.004 from the best model considering the context. Similarly, in RumourEval-2019 (Gorrell et al., 2019), the runner-up model (Fajcik et al., 2019) observed a 0.43 *decrease* in accuracy by considering the target information. Schiller et al. (2020) discovered that stance detection models are prone to adversarial attacks of paraphrasing, spelling error and negation similar to other NLP tasks (Ribeiro et al., 2020). However,

[1]Code: https://github.com/Ayushk4/bias-stance
Dataset: https://github.com/Ayushk4/stance-dataset

		Target				Tweets		
Dataset	size	*Number*	*Domain*	*Type*	*# Stance*	*Unique*	*Scrapped*	*DT/T*
WT–WT	51284	5	Finance (M&A)	fixed	4	50210	45865	2%
SE16	4162	5	Various	fixed	3	4162	-	0%
M-T	4455	3	Political	fixed (pairs)	3	4413	2688	0.9%
RE17	5568	-	Rumour-claims	free-form	4	556k	-	0%
RE19	8574	-	Rumour-claims	free-form	4	8574	-	0%
Encryption	2999	1	Encryption-debate	fixed	3	2522	1634	0%

Table 1: Statistics of the Twitter stance detection datasets considered.

this is the first work providing a detailed insight into the **alarmingly impressive performance of target-oblivious models.**

Target plays a crucial role in deciding stance. Consider the example in Figure 1. Here, the tweet stance varies for the two targets. The existing datasets have very few examples with different target labels. Models can pick up on pseudo signals in the tweet content and shortcut the task without looking at the targets. These signals or biases occur due to inherent biases in our language and human nature. For example, certain lexical choices can correlate with their respective stance classes. Upon discovering and studying such correlations, we augment the WT–WT dataset addressing these issues and re-evaluate the stance detection systems.

We make the following contributions. We empirically demonstrate biases across a variety of Twitter stance detection datasets and carry out a detailed analysis of these datasets. Consequently, we propose a new large scale dataset free of such spurious cues and re-evaluate the stance detection systems to show the usefulness of this dataset.

2 Biases in Stance Detection Datasets

We first discuss the datasets considered (Section §2.1), followed by our experiments (Section §2.2) and analysis (Section §2.3).

2.1 Datasets Considered

We consider a wide variety of publicly available Twitter stance detection datasets including cross-target, multi-target, rumour-claim variants of stance detection. These datasets have a diverse set of targets ranging from free-form sentences to fixed target entities.

Over the past few years, several more variants of this task have been proposed, such as in non-English language (Darwish et al., 2017; Küçük and Can, 2018; Lai et al., 2018) and multi-lingual settings (Zotova et al., 2020; Vamvas and Sennrich, 2020), different learning paradigms of unsupervised (Darwish et al., 2019) semi-supervised (Mohammad et al., 2016b), zero-shot (Allaway and McKeown, 2020) and non-Twitter tasks of debate-argument stance (Bar-Haim et al., 2017) and headline-body stance detection (Pomerleau and Rao, 2017).

Here, however, we only study the English Twitter stance detection tasks in fully supervised learning settings. Specifically, we consider 6 datasets - WT–WT (Conforti et al., 2020), SE16 (task-A) (Mohammad et al., 2016b,a) M-T (Sobhani et al., 2017), RE17 (Derczynski et al., 2017), RE19 (Gorrell et al., 2019) and Encryption (Addawood et al., 2017) with their statistics mentioned in Table 7. This table also reports the percentage of tweets in the entire dataset labelled for different targets (DT) given by DT/T in the last column. We can see that these datasets have very few tweets annotated for different targets. The M-T dataset's targets are a pair of politicians, and for each of its tweet-targets, the label is a pair of stances. We formulate detecting these two stance-pair as separate tasks for the experiments in the following section.

2.2 Performance of Target-Oblivious Models

Method: Given a tuple $(tweet, target, stance)$, a target-oblivious classifier $f(tweet) \rightarrow stance$ is trained in a supervised setting. It is expected that such a classifier would generalize poorly for an unbiased dataset. We set this target-oblivious classifier as the standard Bert classifier (Devlin et al., 2019) pre-trained on Tweets (Nguyen et al., 2020). It receives the input "$[CLS]$ $tweet$ $[SEP]$". Additionally, we train a strong target-aware Bert classifier model for stance detection (Ghosh et al., 2019). This model takes input "$[CLS]$ $tweet$ $[SEP]$ $target$ $[SEP]$". We use PyTorch (Paszke et al., 2019), HuggingFace (Wolf et al., 2019), Wandb (Biewald, 2020) and Scikit-Learn (Pedregosa et al., 2011) for our experiments. We use Adam optimizer (Kingma and Ba, 2014). We elaborate the full experimental settings in the appendix §A.

Models	$F1_{Macro}$ across healthcare merger operations						Entertainment
	CVS_AET	*CI_ESRX*	*ANTM_CI*	*AET_HUM*	$avgF_1$	avg_wF_1	$F1_{Macro}$
Bert (no-target)	0.673	0.703	**0.745**	**0.759**	0.720	0.720	0.347
Human Upperbound	0.753	0.712	0.744	0.737	0.736	0.743	N/A
Bert (with target)	0.668	0.709	**0.746**	**0.756**	0.720	0.719	0.433
Random guessing	0.222	0.237	0.231	0.236	0.230	0.232	0.201
Majority guessing	0.162	0.139	0.155	0.134	0.151	0.148	0.161

Table 2: Results on WT–WT dataset (Conforti et al., 2020). Values higher than Human Upperbound are boldfaced.

Model	Acc	$F1_{Wtd}$	$F1_{Macro}$
SemEval 2016 (SE16)			
Bert (no target)	0.708	0.711	0.675
Bert (target)	**0.738**	**0.737**	**0.695**
Majority Class	0.572	0.416	0.243
Random	0.333	0.353	0.313
M-T Stance Dataset			
Bert (no target)	0.675	0.673	0.654
Bert (target)	**0.691**	**0.681**	**0.657**
Majority Class	0.419	0.247	0.197
Random	0.333	0.336	0.331
RumourEval 2017 (RE17)			
Bert (no target)	**0.783**	**0.766**	**0.543**
Bert (target)	0.769	0.760	**0.543**
Majority Class	0.742	0.632	0.213
Random	0.25	0.310	0.189
RumourEval 2019 (RE19)			
Bert (no target)	**0.840**	0.821	0.577
Bert (target)	0.836	**0.829**	**0.604**
Majority Class	0.808	0.722	0.223
Random	0.25	0.329	0.171
Encryption Debate			
Bert (no target)	**0.916**	**0.903**	**0.778**
Bert (target)	0.907	0.894	0.755
Majority Class	0.863	0.801	0.464
Random	0.500	0.576	0.424

Table 3: Results on other stance detection datasets.

Results and Discussion: The WT–WT dataset is a cross-target dataset containing four in-domain (healthcare) and one out-of-domain (entertainment) target. For the in-domain evaluation, training is done on three health mergers while testing is done on the fourth unseen target. For out of domain, training is on all four health mergers and testing on the entertainment domain. Table 2 shows the performance of target-oblivious Bert, target-aware Bert and the human upper-bound. The human expert upper-bound values were taken from the WT–WT dataset. We observe that the target-oblivious model consistently performs very close to the target-aware model for all the targets. Both these models achieve near-human performance overall on in-domain targets. The target-oblivious Bert surpasses human upper-bound for two mergers individually. Such a feat is alarming, especially because cross-target stance is a more challenging variant (Küçük and Can, 2020; Wang et al., 2020) of the task.

Stance	Lexicons
Support	approves (3.3%), approve (5.1%), **billion (26.2%)**, shareholder (0.7%), close (6.4%)
Refute	urges (3.0%), blocked (5.5%), sues (4.3%), blocks (4.8%), **block (21.8%)**
Comment	ceo (3.7%), healthcare (11.8%), mean (2.3%), **merger (29.3%)**, trial (3.4%)
Unrelated	stocks (3.4%), size (2.6%), merge (11.3%), **bid (19.0%), agreement (16.7%)**

Table 4: Top 5 stance-wise words in WT–WT dataset by $PMI(word, stance)$ across health domain tweets along with percentage of the respective stance-class labelled tweets having each word.

Results on the other datasets are shown in Table 3. We compare these results with **random** guessing, predicting **majority class** and the target-aware Bert. Additionally, RE17, R19, and Encryption datasets are heavily skewed datasets, so Macro-F1 is the proposed metric (Gorrell et al., 2019).

The target-oblivious Bert delivers more than two-third classification accuracy consistently across all these datasets. This model achieves impressive performance for all metrics in SE16 and M-T datasets, while also performing significantly above majority class for datasets with skewed distributions on the Macro-F1 metric. The performance delivered by target oblivious Bert is also very close to the target-aware Bert model on every metric. These surprising numbers across all the datasets indicates the presence of spurious cues that encourages the models to bypass the need for looking at the target.

2.3 Dataset Analysis

After the finding from our previous section, we sought to discover the form in which spurious cues exists and use those findings to create a new dataset. We mainly consider the largest and most recent dataset, WT–WT for analysis. We first discuss target-independent lexical choices associated with stance, followed by target-independent sentiment-stance correlations.

Stance and tweet lexicons: We calculate the smoothed Point-wise Mutual Information (PMI)

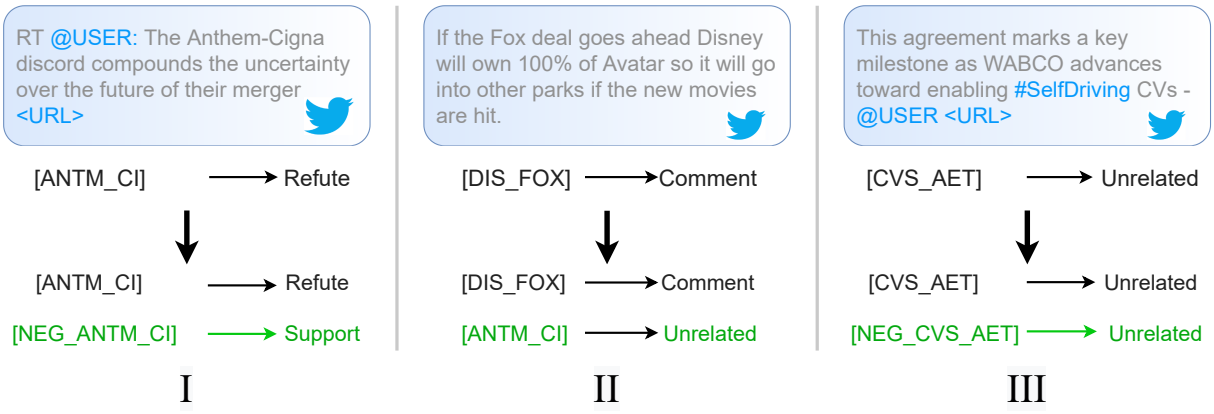


Figure 2: Procedure to create tWT–WT dataset from WT–WT.

between tweet and stance following the exact same procedure as (Gururangan et al., 2018) after removing stopwords. Table 4 shows that top 5 stance-wise words along with the fraction of tweets containing those words. We observer that certain groups of target-independent lexicons are highly correlated with stances in some cases occurring in more 29% of the tweet. For Support and Refute classes respectively, we find the co-occurrence of indicative words for the status of merger, such as 'approves' or 'blocks'. The Comments relating to these health companies' mergers often talk about its impact, leading to the choice of lexicons containing words like 'healthcare' and 'mean' with this stance. Similarly, Unrelated tweets often talk about things related to the companies but unrelated to the merger operation such as 'stocks' or 'bids'.

Sentiment-stance correlation: Stance detection differs from the sentiment analysis task (Mohammad et al., 2016b). However, we observe a strong correlation of sentiment with stance. Formally, we obtain a sentiment score between 0 (negative) and 1 (positive) for each tweet using XLNet model (Yang et al., 2019) trained on SST (Socher et al., 2013; Pang and Lee, 2005) and Imdb (Maas et al., 2011). The average sentiment scores of these tweets across Support, Refute, Comment and Unrelated stances were found to be 0.237, 0.657, 0.492 and 0.485 respectively, while their variance were 0.087, 0.056, 0.110, 0.108. The tweets with Support and Refute stance have strong negative or positive sentiment on average while for the other two is it neutral on average but having a high variance. These serve as strong evidences for stance-sentiment correlations.

Thus sentiment and lexicons together are some of the spurious cues in WT–WT dataset. We found such cues in the remaining datasets, varying with their domains. For example, RE19 has a question mark in more than 75% 'query' stance tweets, while it is present only in 11% of the entire remaining dataset. Similarly 75% of tweets with 'deny' stance have highly negative sentiment of less than 0.1 score. In SE16 dataset, had 91.4% of tweets without any opinion[2] had 'None' stance despite the stance detection task being different from opinion mining task (Mohammad et al., 2016b).

3 The Targeted WT–WT (tWT–WT) dataset

With the understanding from the previous section, we propose a new stance detection dataset on which target-unaware models will not perform well. We use the following reasoning for creating the new dataset. If a tweet in the dataset has different stances depending on different targets, then simple tweet classification models will not be able to perform well. Thus we attempt to increase DT/T ratio from Table 7. Formally, we take the WT–WT dataset, which is the largest dataset of its kind, with high-quality experts labels of 0.88 Cohen-κ (Cohen, 1960), and generate new (tweet, target, stance) triplets in three ways.

First, we attempt to remove the sentiment-stance correlation by making the stance-wise average sentiment neutral. The WT–WT dataset has 5 targets, one target for each merger. We introduce 5 new additional targets which are negations of the original ones. Formally, if the tweet has a Support (Refute) stance to the target CVS_AET, then its stance to the negated target NEG_CVS_AET will be inverted to Refute (Support). This is done only

[2]Tweets have gold labels for opinion-class in the dataset.

Models	$F1_{Macro}$ across healthcare merger operations						Entertainment
	CVS_AET	*CI_ESRX*	*ANTM_CI*	*AET_HUM*	avg_wF_1	$avgF_1$	$F1_{Macro}$
Bert (no-target)	0.161	0.258	0.297	0.340	0.264	0.260	0.163
Bert (with target)	**0.460**	**0.386**	**0.596**	**0.598**	**0.510**	**0.527**	**0.365**
SiamNet	0.293	0.292	0.273	0.398	0.312	0.310	0.150
TAN	0.170	0.222	0.308	0.332	0.258	0.260	0.150
Random Guessing	0.233	0.206	0.225	0.223	0.222	0.225	0.205
Majority Guessing	0.145	0.198	0.181	0.177	0.175	0.171	0.169

Table 5: Results on tWT–WT dataset in the same cross-target settings as given by Conforti et al. (2020).

for the two stance classes with non-neutral average sentiment score. Introducing such negated targets reduces their sentiment to near neutral.

Second, we remove lexicon-stance correlations by creating multiple targets with different stances for each tweet. Formally, for each tweet t with only one labelled target tgt, if the tweet-target pair (t, tgt) has the stance $\neq$ 'Unrelated', then pick a target tgt' where $tgt' \neq tgt$ and add the tuple $(t, tgt', Unrelated)$ to the dataset. Due to WT–WT data collection and annotation procedure, this will not generate any wrong labels. This augmentation reduces the lexicon-stance correlations, by having similar sets of lexicons introduced for different stances. Hence, it guarantees target-oblivious shortcuts to result in poor performance.

Last, we balance the target-wise class-distributions. For the tuples with 'Comment' and 'Unrelated' stances, we create a new tuple with inverted target (same as the first step) for 50% and 75% such examples randomly.

The resulting dataset contains 111596 tweet-target pairs each belonging to a stance class. Each merger has at least 10000 data points. The class distribution is also somewhat balanced with more than 10k examples for the least occurring class. Among the tweet-target pairs, the pairs classified as Support, Refute, Comment and Unrelated are distributed in the ratio 1:1:3:5 approximately, having a similar distribution to the WT–WT dataset.

3.1 Re-evaluating stance detection systems

We propose a similar cross-target evaluation setting for tWT–WT as WT–WT. For the in domain (health) mergers, we train on three health merger (total six targets including negated target for each merger) and test on the fourth health merger. For the out-of-domain evaluation, we train on the eight targets corresponding to the 4 health mergers and test on the two targets for entertainment merger.

We re-evaluate the existing stance detection models on *t*WT–WT dataset. We consider Bert (with target), target-oblivious Bert from §2.2, along with the two strongest baselines from the WT–WT paper - SiamNet (Santosh et al., 2019) and TAN (Du et al., 2017). For SiamNet and TAN models, we replace the Glove (Pennington et al., 2014) and LSTM (Hochreiter and Schmidhuber, 1997) features with better features from Bert.

Table 5 shows the performance of these models. Bert (no-target) gives very low performance, showing that target oblivious models perform poorly on this dataset. Similarly, TAN which has been proven to not at take advantage of the target information (Ghosh et al., 2019) also performs very poorly on the dataset. The target aware Bert offers a competitive performance still being only at 0.51 F1 score. SiamNet follows next at 0.31 F1 score. Both these models have their performance reduced significantly from WT–WT dataset.

4 Conclusion and Future Work

In this paper we demonstrated the presence of biases across several Twitter stance detection datasets, which aid simple tweet classifiers to achieve impressive performance. We carried out an investigation for presence of bias for the WT–WT dataset and found correlations of stance-class with sentiment and lexical choice. Consequently, we proposed a new bias-free stance detection dataset - tWT–WT, the largest of its kind. Evaluation of our baselines on this new dataset demonstrates scope for future research on stance detection. The observations are also crucial for the creation of new stance detection datasets. Our future work includes analysing multilingual datasets and exploring explainable target aware stance detection models.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and suggestions. We also thank the Computer Science and Engineering Department at the IIT Kharagpur for providing us with the compute facilities for our research.

References

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th International Conference on Social Media & Society*, #SMSociety17, New York, NY, USA. Association for Computing Machinery.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. 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*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 145–148, New York, NY, USA. Association for Computing Machinery.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2019. Unsupervised user stance detection on twitter.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *Advances in Information Retrieval*, pages 529–536, Cham. Springer International Publishing.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Chris Dulhanty, Jason L. Deglint, Ibrahim Ben Daya, and Alexander Wong. 2019. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection.

Martin Fajcik, Pavel Smrz, and Lukas Burget. 2019. BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional

transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. 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, San Diego, California. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87, Cham. Springer International Publishing.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An svm-based approach.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance Evolution and Twitter Interactions in an Italian Political Debate. In *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 15–27, Cham. Springer International Publishing.

Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained bert model.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2786–2792. AAAI Press.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. Association for Computational Linguistics.

T. Y.S.S. Santosh, Srijan Bansal, and Avirup Saha. 2019. Can siamese networks help in stance detection? In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 306–309, New York, NY, USA. Association for Computing Machinery.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance detection benchmark: How robust is your stance detection?

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. 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, Valencia, Spain. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection.

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

Z. Wang, Q. Wang, C. Lv, X. Cao, and G. Fu. 2020. Unseen target stance detection with adversarial domain generalization. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. "We make choices we think are going to save us": Debate and stance identification for online breast cancer CAM discussions. *Proceedings of the ... International World-Wide Web Conference. International WWW Conference*, 2017:1073–1081.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection: The catalonia independence corpus.

A Appendix

We release our code and pre-training for section §2 at this url- https://github.com/Ayushk4/bias-stance. Our dataset and baselines for section §3 have been released this url - https://github.com/Ayushk4/stance-dataset. The Readme for the respective repositories contain instructions to set up, environment, replicating the codebase and the links to pre-trained models.

In this appendix, we first discuss the baselines §B. Followed by our experimental setup §C and the datasets considered §D

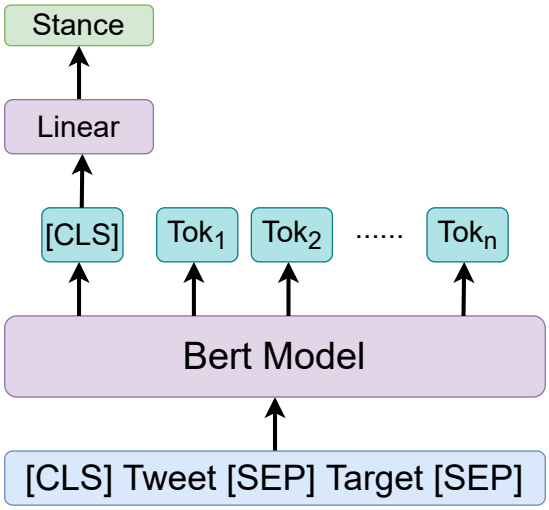
B Baselines

- **Bert for Stance Detection** (Ghosh et al., 2019) is shown in Figure 3a. It takes both tweet and target sentence as input separated by $[SEP]$ token. The final $[CLS]$ token representation is used for stance classification. It delivered state of the art performance across two datasets.

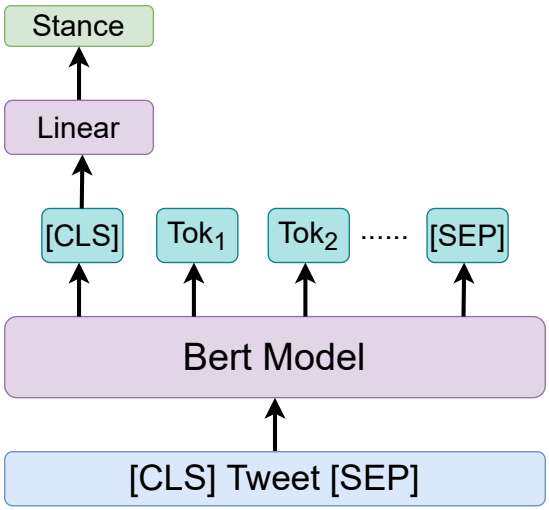
- **Target oblivious Bert** is shown in Figure 3b. It takes only the tweet sentence as input enclose within $[CLS]$ and $[SEP]$ tokens. The final $[CLS]$ token representation is used for stance classification.

- **SiamNet** architecture (Santosh et al., 2019) is shown in Figure 4a. It uses siamese networks (Bromley et al., 1993) to learn representations each for tweet and targets and classify using the bottleneck of a single scalar being the similarity function. Similar to the WT–WT paper (Conforti et al., 2020) we find that the similarity function output scalar alone isn't strong enough feature for a classifier. Hence we concatenate the tweet and target representation vectors of the similarity function (inverse exponential of Manhattan distance) following (Mueller and Thyagarajan, 2016). We replaced the Glove and BiLstm with Bert embedding, where we obtain the tweet and target representations we taking the '[CLS]' vector representations from bert for those sentences.

- **TAN** (Du et al., 2017) is shown in Figure 4b. It uses a target-specific attention extraction over the tweet features obtained from BiLstm similar to (Dey et al., 2018). We use the same approach as them except that we substitute the LSTM features for better features from Bert.



(a) Bert Stance Detection

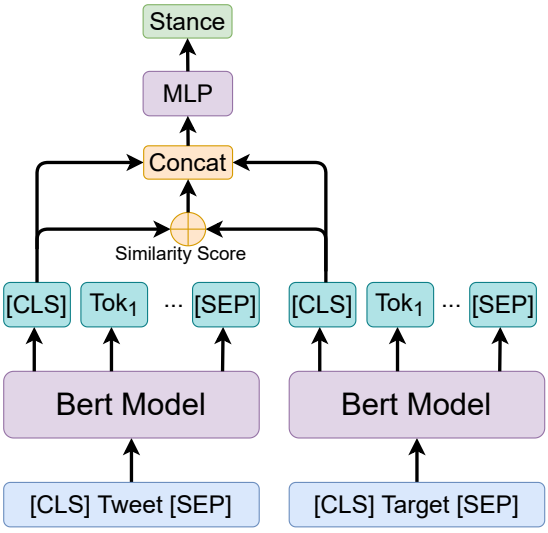


(b) Target oblivious Bert

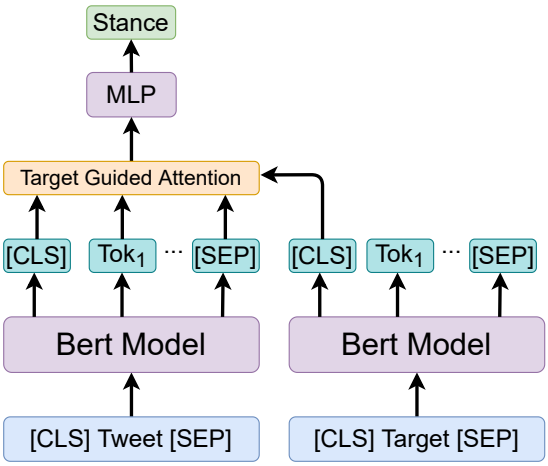
Figure 3: Bert Models

C Experimental Setup

All our experiments were performed using Pytorch (Paszke et al., 2019), wandb (Biewald, 2020) and Huggingface (Wolf et al., 2019). The optimization algorithm used was the Adam optimizer (Kingma and Ba, 2014). We keep Bert layers and embedding trainable. In case of SiamNet and TAN, the Bert parameters are hard-shared. Experiments takes less than 10 minute per epoch and less than 5 GB GPU memory on a Tesla P100 GPU. The total model parameters for the Bert models are the same as the Bert (including being approximately same for SiamNet and TAN). Following the previous works demonstrating that domain-specific weights



(a) SiamNet Model



(b) TAN Model

Figure 4: SiamNet and TAN Baselines

result in improved performance (Gururangan et al., 2020; Lee et al., 2019; Beltagy et al., 2019; Lee and Hsiang, 2019; Liu et al., 2020; Müller et al., 2020), we use the Bert weights fine tuned on tweets (Nguyen et al., 2020) with the exception of the target aware Bert in tWT–WT where it was found to be unstable. So, we used *bert-base-cased* instead.

C.1 Hyperparameters

We use the Huggingface's Bert default config for our experiments. For the experiments, we tuned the learning rate from 5 values - {1e-6, 3e-6, 1e-5, 3e-5, 1e-4} and number of epochs from 3 values - {2, 5, 10} on the development set. The batch-size was fixed to 16. For the datasets with no development split, we use 5-fold cross validation. We evaluated and trained our model in the same settings as proposed for their respective datasets. Table 6 lists the hyper-parameters.

All Models	
Batch size	16
Num epochs	2, 5, 10
Optimizer	Adam
Bert dropout	0.1
Max tokens	99
Classifier dropout	0.1
Bert trainable	Yes
Learning rate	{1e-6, 3e-6, 1e-5, 3e-5, 1e-4}
5-fold cross-valid	Macro-F1
SiamNet	
Final mlp hidden	786
Distance metric	Inverse exponential of Manhattan distance
TAN	
Final mlp hidden	786

Table 6: Hyperparameters for Bert.

C.2 Preprocessing

We use ekphrasis library (Baziotis et al., 2017) to perform the preprocessing. We do word tokenization and spelling correction. We also remove URLs, Emoji, non-ascii characters and do normalization to limit the length of inputs. For the RumourEval2019 dataset we trimmed the input to 99 tokens, since Reddit text can even cross 500 characters length.

D Datasets

For the datasets that released only the tweet ids, we obtain the tweet text using the Twitter API.[3] However, some tweets are not accessible over time as accounts or tweets get banned/blocked/deleted etc. Table lists the full statistics 7 for the datasets.

- **Will-They-Won't-They (WT–WT) Dataset**: (Conforti et al., 2020) is a cross-target stance detection dataset. It has 50k tweet-target pairs from financial domain. It has 5 targets, each a fixed Merger and Acquisition (M&A) operation. Four of the five M&A targets are from health domain - {CVS-Aetna, Cigna-Esrx, Anthem-Cigna, Aetna-Humana} and one from entertainment - {Disney-Fox} serving as an out-of-domain target. For the experimental settings, the model is trained on three health mergers and tested on the fourth. For out of domain

[3] https://developer.twitter.com/

Dataset	Download Link
WT–WT	https://github.com/cambridge-wtwt/acl2020-wtwt-tweets
SE16	https://alt.qcri.org/semeval2016/task6/
M-T	https://www.site.uottawa.ca/~diana/resources/stance_data/
RE17	https://alt.qcri.org/semeval2017/task8/
RE19	https://competitions.codalab.org/competitions/19938
Encryption	https://github.com/aseelad/The-Encryption-Debate

Table 7: Links to download the datasets.

merger, the model is trained on four health mergers and tested on the entertainment domain. Its stance label set is - {Support, Refute, Comment, Unrelated}. As part of pre-processing, we normalize the health merger company names /acronyms , for example 'Anthem Inc' → 'Anthem', 'Antm' → 'Anthem'.

- **SemEval 2016** Task 6-A (Mohammad et al., 2016b) is based on a stance detection dataset (Mohammad et al., 2016a) of 4163 tweets. The targets are fixed entities (politicians, movements, policy etc.). Specifically, these are from the set - {Atheism, Climate-Change, Hillary Clinton, Feminism, Legalizing Abortion}. Each of the tweet-target pair is labelled for one of the 3 stance from {Against, None, Favor}. The dataset has somewhat balanced class distribution and the metrics considered for this were Accuracy and F_1. In this task, the models were evaluated on same targets as they were trained.

- **Multi-target** (M-T) stance dataset (Sobhani et al., 2017) contains 4455 tweets from political domain. Target was a fixed entity pair containing two political entities - Hillary-Sanders, Hillary-Trump, Cruz-Trump. This tweet-target pair has two stances - one for each target from - {Against, None, Favor}. Thus, in pairs of two, it leads to 9 total possible combinations of the 3 labels. We treat the pair as two separate problems, and trained two separate models for it. The dataset has somewhat balanced class distribution and the models were evaluated on the same targets (pairs) as they were trained using the Accuracy and F_1 metrics.

- **RumourEval 2017** (Derczynski et al., 2017), was a rumour-stance detection task that proposed a new dataset for the task. The dataset consisted of 285 rumoured tweet threads with a total of 4519 tweets. The root node of each thread was the rumour target, for which users replied and created a response thread exhibiting a tree structure. The tweet-targets pairs were labelled from one of the four stance classes being - {Support, Query, Comment, Deny}. The dataset has a very skewed distribution with the majority class (Comment) having about 80% examples. So, Macro-Averaged F_1 score is a suitable metric. Here the models were evaluated on different threads (and hence different targets) as they were trained.

- **RumourEval 2019** (Gorrell et al., 2019) was similar to the RumourEval 2017 task. It extended the dataset to include Reddit threads from selected sub-reddits. The resulting dataset has a total of 8574 datapoints. The tweet-targets pairs were labelled from the same labelset from one of the four stance classes being - {Support, Query, Comment, Deny}. This dataset has a very skewed distribution with the majority class (Comment) having about 80% examples. So, Macro-Averaged F_1 score is a suitable metric for the dataset.

- **Encryption Debate** dataset (Addawood et al., 2017) consists of 2999 tweets labelled for three stances - {For, Against, Neutral} on one encryption debate topic. We observed repeated entries in the dataset, including some having conflicting labels for the same tweet-target pair; we excluded such tweets for our experiments. Additionally, only 5 tweets from the dataset belonged to the 'against' class. Since 5 examples is a very small number for most machine learning models to learn, we exclude this class for our analysis. The dataset has a very skewed distribution with the majority class (neutral) having about 86% examples. So, Macro-Averaged F_1 score is a suitable metric. The dataset has only one target for training and evaluating the models.