

What do our stance detection systems learn? On dataset biases exploited by the systems.

Ayush Kaushal

Indian Institute of Technology, Kharagpur
ayushkaushal@iitkgp.ac.in

Abstract

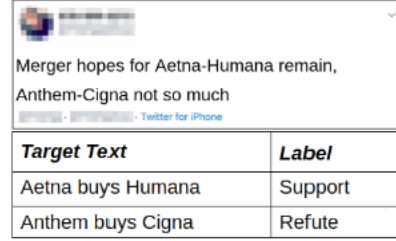
The stance detection task aims at automating the task to identify whether a tweet is in favor of the given target, against the given target, or whether neither inference is likely. Our experiments show that the target-oblivious models give human performance with a decrease of only about 2-3% F1 against their target aware counterparts. The empirical findings suggest the existing biases in the stance detection datasets and that the success of stance detection systems have been over-estimated. My contributions are proposing the project idea, formulating the experiments, implementing and carrying out all the experiments.

Introduction

In the recent times there has been a proliferation of misinformation and rumours in social media, COVID pandemic is a prime example of this (Alam et al. 2020). Stance Detection is the first step to misinformation detection (Pomerleau and Rao 2017) and fact checking (Vlachos and Riedel 2014). Since its introduction in Semeval 2016, task 6: (Mohammad et al. 2016), numerous stance detection systems have been proposed with impressive performance upto 89.9% Accuracy (Santosh, Srijan, and Saha 2019) on the FNC dataset (Pomerleau and Rao 2017).

Stance Detection, being an empirical field of research is heavily dependent on datasets. These datasets not only provide the training data, but also a means to compare various systems. However, due to the nature of dataset collection process and bias in our language, the datasets often have some biases crept into them. Gururangan et al. (2018) demonstrated biases in the NLI dataset using which a premise-unaware text classification models achieved impressive performance. A similar bias was observed in referring expression datasets, where the state of the art systems (Cirik, Morency, and Berg-Kirkpatrick 2018) could achieve promising accuracy by completely ignoring the language syntax. Biases in datasets lead us to question the progress in a field.

Here, we demonstrated the biases in stance detection datasets by consider two mis-interpretations of the task. Findings show that the systems trained on incorrect target



Target Text	Label
Aetna buys Humana	Support
Anthem buys Cigna	Refute

Figure 1: An example tweet from wtwt-dataset with multiple-targets. Only 2% tweets have more than 1 targets.

or even target-unaware systems can give near state of the art performance on the WT-WT dataset.

Dataset and Problem statement

Conforti et al. (2020) released the Will-They-Won't-They (wt-wt) stance detection dataset of 51,000+ tweets (8 times larger than the previous largest), annotated by experts. The dataset consists of tweets that discuss recent Merger and Acquisitions events (M & A) between companies. Each tweet is labeled with the stance of a text towards merger from the label set- $\{Support, Refute, Comment, Unrelated\}$.

A data-point in the dataset is a tuple (x, t, y) , where y is the stance of a tweet x with respect to a target t . About 2% tweet x may be labelled for multiple targets. An example tweet with multiple targets is given in figure 1. The stance detection model $f(x, t) \rightarrow y$ aims to classify the stance y of x with respect to a target t .

Mis-interpretations of Stance Detection task

Consider 2 mis-interpretations of stance detection tasks:

- **Incorrect Targets:** Given a tuple (x, t, y) in the dataset, the model is trained on wrong target text (x, t', y) , where the label of x with respect to some other target t' is not y . Thus, here the model is training on incorrect tweet-target-stance tuple.
- **No Targets:** Given a tuple (x, t, y) in the dataset, a target-oblivious model is trained for stance detection $f(x, t, y) = g(x, y)$.

For an unbiased dataset, we would expect both the interpretations to generalize poorly to the test set.

Model	CVS-AET <i>MacroF₁</i>	CI-ESRX <i>MacroF₁</i>	ANTM-CI <i>MacroF₁</i>	AET-HUM <i>MacroF₁</i>	<i>avgF₁</i>	<i>avg_wF₁</i>
Bert	0.708	0.724	0.769	0.752	0.738	0.740
Bert incorrect	0.674	0.701	0.747	0.739	0.715	0.716
Bert no target	0.679	0.696	0.752	0.738	0.716	0.719
Bi-LSTM	0.675	0.687	0.739	0.732	0.708	0.711
Bi-LSTM incorrect	0.658	0.677	0.724	0.731	0.699	0.698
Bi-LSTM no target	0.663	0.674	0.733	0.720	0.700	0.697
Human Performance	0.753	0.712	0.744	0.737	0.747	0.752

Table 1: Results on the healthcare M&A operations in the WT-WT dataset each averaged across 2 seeds. Bold values denote better than human performance. Macro F1 scores are obtained by testing on target operation while training on the other three. $avgF_1$ and avg_wF_1 are, respectively, unweighted and weighted (by operations size) average of the 4 M&A operations.

Experiments and Results

A Bert baseline (Devlin et al. 2019) is considered, with input of the form “[CLS] tweet [SEP] target [SEP]”, followed by a fully connected classification layer over the features of [CLS]. The **Bert incorrect** is the Bert but based on the Incorrect Target mis-interpretations, and thus is trained with incorrect targets. The **Bert no target** is based on no-target misinterpretation, with input of the form “[CLS] tweet [SEP]”. Similarly, **BiLSTM**, **BiLSTM incorrect** and **BiLSTM no target** are defined with a two-layer bidirectional LSTMs (Hochreiter and Schmidhuber 1997).¹

Table 1 shows the results on the dataset. The human performance is obtained from the dataset paper (Conforti et al. 2020). **Bert** baseline model performance better than human performance on three targets owing to its pre-training on large corpora. **BiLSTM** performs worse than Bert and yet falls short of human performance by just 0.04 $avgF_1$. **Bert incorrect** and **Bert no target**, performs unexpectedly very well with surpassing human performance on two target while being only about 0.025 $avgF_1$ lesser than Bert. Similar pattern is observed for **BiLSTM incorrect** and **BiLSTM no target** while performing very close to **BiLSTM**. This demonstrates the dataset biases for wtwt-dataset, which can be easily exploited by the target-unaware models to reach human level performance.

Conclusion and Future Work

The paper presented two mis-interpretations for stance detection task. The empirical findings demonstrate bias in the stance detection datasets using which a target-unaware model can achieve human level performance. Our work also question on what our stance detection systems learn.

My future work includes de-biasing the dataset by including at least two targets for each tweet (i.e. multi-target labelling), followed by a thorough evaluation of the existing state of the art stance detection systems on the new unbiased dataset to check the progress in the field.

References

Alam, F.; Dalvi, F.; Shaar, S.; Durrani, N.; Mubarak, H.; Nikolov, A.; Martino, G. D. S.; Abdelali, A.; Sajjad, H.; Dar-

wish, K.; and Nakov, P. 2020. Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms.

Cirik, V.; Morency, L.-P.; and Berg-Kirkpatrick, T. 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)*, 781–787.

Conforti, C.; Berndt, J.; Pilehvar, M. T.; Giannitsarou, C.; Toxvaerd, F.; and Collier, N. 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*, 1715–1724.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41.

Pomerleau, D.; and Rao, D. 2017. Fake news challenge. URL <http://www.fakenewschallenge.org/>.

Santosh, T.; Srijan, B.; and Saha, A. 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*, 306–309.

Vlachos, A.; and Riedel, S. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. doi:10.3115/v1/W14-2508.

¹Unlike the baseline, here the word embeddings are trainable.