Synthetic Examples Improve Cross-Target Generalization: A Study on Stance Detection on a Twitter Corpus

Costanza Conforti[1], **Jakob Berndt**[2], **Mohammad Taher Pilehvar**[1,3], **Chryssi Giannitsarou**[2], **Flavio Toxvaerd**[2], **Nigel Collier**[1]

[1] Language Technology Lab, University of Cambridge
[2] Faculty of Economics, University of Cambridge
[3] Tehran Institute for Advanced Studies, Iran

{cc918,jb2088}@cam.ac.uk

Abstract

Cross-target generalization is a known problem in stance detection (SD), where systems tend to perform poorly when exposed to targets unseen during training. Given that data annotation is expensive and time-consuming, finding ways to leverage abundant unlabeled in-domain data can offer great benefits. In this paper, we apply a weakly supervised framework to enhance cross-target generalization through synthetically annotated data. We focus on Twitter SD and show experimentally that integrating synthetic data is helpful for cross-target generalization, leading to significant improvements in performance, with gains in F_1 scores ranging from $+3.4$ to $+5.1$.

1 Introduction

Stance Detection (SD) is a widely investigated task (Mohammad et al., 2017), which constitutes an important component of many complex NLP problems, ranging from fake news detection to rumour verification (Vlachos and Riedel, 2014; Baly et al., 2018; Zubiaga et al., 2018b). Since from early works (Agrawal et al.), research on SD focused on user-generated content, ranging from blogs and commenting sections on websites (Hercig et al.), to Reddit or Facebook posts (Klenner et al.) and, above all, Twitter data (Inkpen et al., 2017; Zubiaga et al., 2018a).

Recently, Conforti et al. (2020) released Will-They-Won't-They (WT–WT), a very large corpus of stance-annotated tweets discussing five US mergers and acquisitions (M&A) operations spanning over two industries: healthcare and entertainment. M&A is a general term that refers to the process in which the ownership of companies are transferred. Such process has many stages that range from informal talks to the closing of a deal, and discussions may not be publicly disclosed until a formal agreement is signed (Bruner and Perella, 2004): in this sense, the analysis of the evolution of opinions and concerns expressed by users about a possible M&A operation, from early stage discussion to the signing of the merger (or its rejection), is a process similar to rumor verification, a widely studied field (Zubiaga et al., 2018a). Interestingly, Conforti et al. (2020) observed a consistent drop in performance when a system trained on mergers in one industry is tested on data discussing a merger in a different industry. Such a performance drop when testing conditions deviate from training conditions is a known problem in Stance Detection (SD) (Aker et al.).

In this paper, we investigate the impact of using synthetically annotated data to improve zero-shot cross-target generalization in Twitter SD:
(1) We investigate a weakly supervised framework for SD, which integrates synthetically annotated data to improve performance on new targets; as to our knowledge, we are the first to use synthetically annotated data for SD;
(2) We test our framework on Twitter SD and prove that it successfully improves cross-target generalization on new, unseen targets;
(3) We extend the WT–WT corpus with additional annotated tweets discussing M&A operations in one additional domain, which we release for future research on cross-target generalization[1].

2 Cross-Target Generalization with Synthetically Annotated Samples

Given an in-domain (ID) test set and a gold out-of-domain (OOD) train set, we augment the corpus with synthetically labeled ID data (Figure 1):

1. We train a SD system on the gold OOD data.
2. We crawl for a large amount of unlabeled ID data and label it with the system trained in 1, obtaining silver, synthetically annotated data.

[1] https://github.com/cambridge-wtwt/

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 181–187
April 19, 2021.

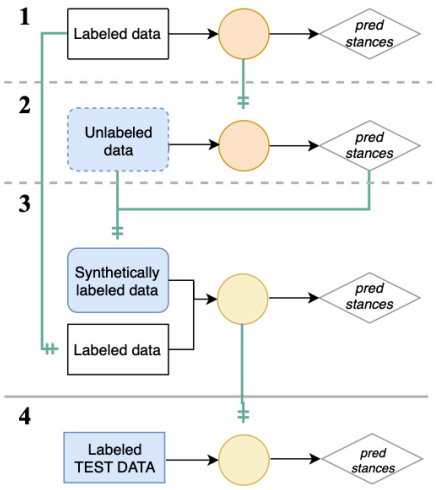

Figure 1: Pipeline of the framework. Rectangular boxes: gold annotations; cornered boxes: unlabeled/synthetic annotations; green lines: elements which are passed from different stages of the pipeline.

3. We train a new system on both gold OOD and synthetic ID data: in this way, the system is exposed to a gold signal from the OOD data and to a noisy but ID signal from silver data.
4. We predict the ID test data with the system trained in 3.

Comparison with previous work on Data Augmentation and Domain Adaptation.
Note that this framework differs from data augmentation (DAug) strategies adopted to supply for small training data, like in question answering (Kafle et al.), machine translation (Fadaee et al.) distillation (Tang et al., 2019), or for adversarial sample generation (Jia and Liang, 2017). Such techniques, inspired by DAug in speech recognition and computer vision (Chatfield et al., 2014), work by deformating gold samples to generate new artificial samples (for example, by random token masking, or POS- or semantics-based token replacement). Our approach differs in a number of aspects:

1. In DAug the goal is to enlarge a set of initial ID data; here, we assume we don't have *any* ID training data, but only OOD;
2. For this reason, while DAug helps to cope with data sparsity, our approach is also useful for domain shifts;
3. In DAug, sample generation might introduce two kinds of noise: it can lead to mismatches between the new samples and the associated labels, and also produce ungrammatical samples; in our approach, the system is always exposed to well-structured input: the only noise are potential errors in synthetic labeling.

Our approach fits into the broad family of weakly- and semi-supervised frameworks which have been adopted to tackle domain adaptation (DAda) problems (Søgaard, 2013). In recent literature, such methods have been applied with mixed success to many tasks, ranging from named entity recognition (Fries et al., 2017) to relation extraction (Mintz et al., 2009), tagging (Plank et al., 2014), parsing (McClosky et al., 2010), and sentiment analysis (Blitzer et al., 2007; Ruder and Plank, 2018; Ratner et al., 2020). In this paper, we propose to apply weakly supervision to SD, by adopting the extremely simple and inexpensive framework described above.

3 Related Work on Stance Detection

SD is a widely investigated field in NLP. Starting from Mohammad et al. (2017), research in SD focused on the analysis of Twitter posts. Another research direction explored the classification of Twitter users with respect to given topics, like political independence (Darwish et al., 2019). Work on other types of user-generated data includes SD on parenting blogs (Skeppstedt et al., 2017), political posts on newspapers websites (Hanselowski et al., 2018), posts on online debate forums on various topics (Hasan and Ng, 2014) and posts on wordpress blogs (Simaki et al., 2017). SD has been also integrated into Fake News Detection (Pomerleau and Rao, 2017) and constitutes an important step in the rumor verification pipeline (Zubiaga et al., 2018b): in this framework, popular shared tasks focused on SD of rumorous tweets (Gorrell et al., 2018) and Reddit posts (Gorrell et al., 2018). These works analyze tweets in a tree-shaped stream (Zubiaga et al., 2015). Note that SD constitutes a related but different task than sentiment analysis (Mohammad et al., 2017): the latter focuses on the *polarity* expressed w.r.t. a topic, while the former aims to determine the text's *orientation* w.r.t. the topic. Consider the following tweet:

#Cancer patients will suffer if CVSHealth buys Aetna CVS #PBM has resulted in delays in therapy, switches, etc all documented. Terrible!

The sentiment of the tweet w.r.t. the target is *negative*: the user believes that the merger would harm patients; however, its stance is *comment*, as it is

M&A	buyer	target	industry	crawl start	end dates	samples	outcome	labels
ABT_STJ	Abbott Lab.	St. Jude	pharma	15/11/15	05/02/17	756	success	no
AVGO_QCOM	Broadcom	Qualcomm	broadband	01/02/16	01/07/19	7,211	failure	no
BMY_CELG	B-M-S	Celgene	pharma	04/03/17	01/07/19	3,940	tbd	no
CHTR_TWC	Charter Com	Time W. Cable	broadband	01/01/14	30/06/16	4,248	success	no
CLN_HUN	Clariant	Huntsman	chemicals	01/01/17	30/11/17	836	failure	no
CMCSA_TWC	Comcast	Time W. Cable	broadband	01/04/13	01/01/14	23,672	failure	no
CTL_LVLT	CenturyLink	Level 3	technology	01/04/16	30/11/17	1,524	success	no
DELL_EMC	Dell	EMC	technology	11/06/14	29/03/17	7,978	success	no
HAL_BHGE	Halliburton	Baker Hughes	oil industry	01/12/13	06/05/17	6,386	failure	no
IBM_RHT	IBM	Red Hat	technology	01/03/18	31/08/19	16,106	success	no
MDT_COV	Medtronic	Covidien	pharma	01/05/14	31/03/19	5,608	success	no
MSFT_LNKD	Microsoft	LinkedIn	technology	16/01/16	06/01/17	15,107	success	no
TMUS_S	T-Mobile	Sprint	broadband	01/04/17	31/08/19	24,559	success	no
VIAB_CBS	Viacom	CBS Corp	entertainment	01/09/16	10/12/19	12,934	success	no
WATS_AGN	Actavis	Allergan	pharma	01/09/12	30/04/15	2,740	success	no
WATS_WCRX	Actavis	Warner Chilcott	pharma	01/04/13	31/12/13	613	success	no
AET_HUM*	Aethna	Humana	healthcare	01/09/14	23/01/17	7,829	failure	yes
ANTM_CI*	Anthem	Cigna	healthcare	01/04/14	28/04/17	11,021	failure	yes
CVS_AET*	CVS Health	Aetna	healthcare	15/02/17	17/12/18	11,517	success	yes
CI_ESRX$^{\diamond\star}$	Cigna	Express Scripts	healthcare	27/05/17	17/09/18	2,511	success	yes
DIS_FOXA$^{\diamond\star}$	Disney	21 Century Fox	entertainment	09/07/17	18/04/18	18,428	success	yes
UTX_COL$^{\diamond}$	United Tech.	Rockwell Col.	conglomerate	23/12/16	04/09/17	535	success	yes

Table 1: M&A operations considered in this work. Operations before the horizontal line are unlabeled. Operations followed by: * are part of the WT–WT corpus; $^{\diamond}$ are used for testing. Note that some companies (WATS, AETNA and CI) appear in different operations.

not stating that the merger is going to happen or to be rejected, but is talking about its consequences.

4 Experimental Setup

Data. We consider the following data (Table 1-2):

- *Annotated data.* The WT–WT corpus constitutes our primary source of labeled data, which we extend with gold-annotated tweets discussing a merger in the defense industry, following the same procedure as in Conforti et al. (2020). Each {`tweet, merger`} sample is annotated with a label from *support, comment, refute* and *unrelated*, which expresses its stance w.r.t the likelihood of the merger to happen.
- *Unlabeled data.* We crawl for 16 additional mergers, obtaining 134,922 unlabeled tweets.

We consider 3 healthcare mergers as gold train data (AET_HUM, ANTM_CI, CVS_AET) and 3 test sets: CI_ESRX (healthcare, ID), DIS_FOXA (entertainment, OOD) and UTX_COL (defense, OOD).

Models and Hyperparameters. We employ a multi-layer perceptron (MLP) classifier, which takes as input the concatenation of the tweet's and the target's TF-IDF representations and their cosine similarity. This simple model achieved good results on SD (Riedel et al., 2017) and is relatively stable over parameter selection. Hyperparameters used are listed in Table 6 (Appendix B) for replication.

Synthetic Label Generation. We train a system on the gold train set (total 30,367 samples). We use early stopping with a patience of 5 over the

		Sup	Rel	Com	Unr
Train Set	–	14.52	11.87	37.43	36.16
Test Sets					
CI_ESRX	S	30.39	10.55	37.24	21.82
DIS_FOXA	S	7.67	2.05	46.09	42.91
UTX_COL	S	36.99	7.30	18.13	44.13
Synthetic Data					
ABT_STJ	S	12.70	1.06	8.73	77.51
AVGO_QCOM	F	9.22	7.79	23.58	59.41
BMY_CELG	T	7.67	1.68	26.45	46.20
CHTR_TWC	S	9.37	1.15	10.92	78.55
CLN_HUN	F	9.69	7.54	13.64	69.13
CMCSA_TWC	F	4.27	3.30	19.07	73.35
CTL_LVLT	S	9.25	0.79	20.28	69.69
DELL_EMC	S	7.65	1.42	28.92	62.02
HAL_BHGE	F	7.62	8.75	21.77	61.68
IBM_RHT	S	4.20	0.21	12.97	82.62
MDT_COV	S	7.99	0.64	14.30	77.07
MSFT_LNKD	S	2.68	1.22	18.66	77.43
TMUS_S	S	4.91	4.83	19.54	70.73
VIAB_CBS	S	4.67	1.72	21.75	71.86
WATS_AGN	S	12.37	0.55	10.88	76.20
WATS_WCRX	S	20.23	2.12	11.42	66.23

Table 2: Label distribution of: the training set, the test sets and the synthetically labeled data. The second column reports the merger's outcome (*Success*/*Fail*/*Tbd*). See Appendix A for a complete list of companies.

	dataset	synth data	prec	rec	F_1	acc	SUP	REF	COM	UNR
health *ID*	CI_ESRX	*none*	**56.80**	**54.19**	**54.99**	**59.52**	**71.17**	**37.70**	**62.48**	45.39
	CI_ESRX	*related merger*	52.24	51.81	52.12	56.59	63.52	28.57	53.82	61.03
	CI_ESRX	*succeeded merger*	52.05	49.60	49.50	55.64	62.41	18.65	53.69	63.65
	CI_ESRX	*all merger*	53.94	50.94	50.94	56.59	63.25	21.83	53.95	**64.74**
entertain *OOD*	DIS_FOXA	*none*	39.61	35.10	34.55	55.34	**83.36**	7.41	14.43	34.85
	DIS_FOXA	*related merger*	39.34	**37.93**	**37.69**	55.56	60.38	**15.87**	**17.12**	59.33
	DIS_FOXA	*succeeded merger*	38.80	35.55	35.92	54.75	46.86	6.61	16.34	**72.39**
	DIS_FOXA	*all merger*	**40.99**	36.16	36.95	**57.42**	54.99	6.35	12.87	70.44
defense *OOD*	UTX_COL	*none*	35.18	27.16	21.91	44.02	8.08	0.00	**8.54**	92.00
	UTX_COL	*related merger*	**46.91**	**38.98**	**24.09**	45.54	15.15	0.00	6.53	94.22
	UTX_COL	*succeeded mergers*	41.68	29.19	23.99	**45.73**	**16.16**	0.00	5.03	**95.56**
	UTX_COL	*all merger*	37.62	28.52	23.67	44.97	14.14	0.00	7.04	92.89

Table 3: Results of SD on the three test sets (one ID and two OOD), when selecting synthetic data of different types; as recommended when dealing with unbalanced class distribution (Hanselowski et al., 2018), we report on macro-averaged precision, recall and F_1 score; the last four columns report on single label accuracy.

	dataset	synth data	prec	rec	F_1	acc	SUP	REF	COM	UNR
entertainment (out of domain)	DIS_FOXA	*none*	39.61	35.10	34.55	55.34	**83.36**	7.41	14.43	34.85
	DIS_FOXA	UNR	39.33	33.27	32.52	**58.35**	63.60	1.06	3.04	63.91
	DIS_FOXA	COM-UNR	40.28	34.79	34.33	55.15	47.39	12.17	5.09	**74.49**
	DIS_FOXA	COM	37.06	34.24	33.28	54.75	50.82	11.09	4.17	70.06
	DIS_FOXA	SUP-REF-COM	37.08	34.40	35.98	56.97	76.54	8.47	4.24	57.36
	DIS_FOXA	SUP-REF	**40.79**	35.42	35.18	58.42	60.47	8.47	4.53	68.23
	DIS_FOXA	*all stances*	39.34	**37.93**	**37.69**	55.56	60.38	**15.87**	**17.12**	59.33
defense (out of domain)	UTX_COL	*none*	35.18	27.16	21.91	44.02	8.08	0.00	8.54	92.00
	UTX_COL	UNR	**45.38**	28.27	22.00	44.97	13.13	0.00	3.52	**96.44**
	UTX_COL	COM-UNR	36.93	28.30	23.21	44.21	16.16	0.00	5.03	92.00
	UTX_COL	COM	41.42	**30.94**	**27.04**	**47.06**	**22.22**	0.00	**9.55**	92.00
	UTX_COL	SUP-REF-COM	34.48	28.14	23.24	43.45	18.18	0.00	5.03	89.33
	UTX_COL	SUP-REF	39.91	29.04	24.46	45.35	16.16	0.00	7.54	92.44
	UTX_COL	*all stances*	40.99	28.84	25.87	45.73	20.20	0.00	8.40	91.11

Table 4: Results of SD on the OOD test sets, selecting synthetic data annotated with different stances (3^{rd} col).

heldout data. The system achieved an F_1 score of 78.33 on the heldout data. Then, the unlabeled data is annotated using the trained system. The predicted label distribution reflects the actual merger output (Table 2). Refer to Table 5 (Appendix A) for qualitative examples of correctly and wrongly synthetically annotated samples.

5 Experiments and Discussion

Baseline. Table 3 reports on results without using any synthetic data. As expected, we observe a notable gap in generalization performance between the ID healthcare test set and the OOD test sets.

Experiment I. To understand the impact of including different types of synthetic data during training, we consider three settings:
(1) *related mergers*: adding synthetic data from mergers which are ID w.r.t. the considered test set (we select ID mergers for each test set according to similarities between industries, see Appendix A);
(2) *succeeded mergers*: adding data from mergers which were successful: such mergers tend to better match the distribution of the test mergers, as all of them succeeded;
(3) *all mergers*: adding data from all synthetically annotated mergers: this last setting was implemented to test whether synthetically annotated data, even if not perfectly ID w.r.t. the test-set, could have a positive regularization function beyond DA (as hypothesized by Sennrich et al. (2016) in the context of Machine Translation).

For experiments, we randomly add synthetic samples with a proportion of 50% w.r.t. the train set size; to account for uncertainty, we use sample weighting for synthetic samples: *sup*, *ref* and *com* are weighted 0.6, while *unr* are weighted 0.2 (after qualitative analysis, we found them to be noisier).

Results in Table 3 show that adding synthetic samples leads to improvements in generalization over OOD test sets in all considered settings (up

to +3.4 in F_1 score for FOXA_DIS and up to +5.1 for UTX_COL; note that results on UTX_COL without synthetic data were significantly lower than on FOXA_DIS). This is in line with previous results on semi-supervised learning investigating other tasks, such as sentiment analysis (Blitzer et al., 2007) or text categorization (Ando and Zhang, 2005). Interestingly, synthetic samples didn't bring any improvement to the ID test set; moreover, best results overall were obtained with the *related merger* setting: this seems to indicate that synthetic data act as a powerful domain adaptation technique rather than as a regularizer alone, this is in line with findings in Machine Translation (Edunov et al., 2018).

Experiment II. We consider the best performing setting, *related mergers*, and perform a second set of experiments to understand the impact of adding synthetic samples belonging to different stances; we consider: only *unr*; only *com*; only *unr+com*; *sup+ref+com*; *sup+ref*; and finally adding samples from all stances. Differences in performance between settings are negligible (Table 4). Concerning single labels, synthetic samples had the most significant impact on *unr* not only for OOD testsets (up to +39.7 in accuracy for FOXA_DIS and +4.5 for UTX_COL), but even for ID (+18.44).

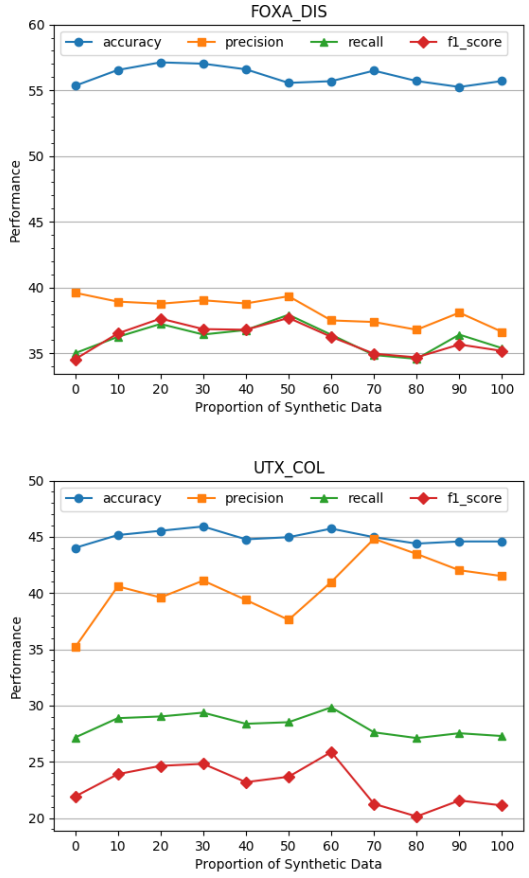

Figure 2: Performance of models trained on different amounts of synthetic data (percentage with respect to the train set size).

Experiment III. We run a final set of experiments to investigate the relation between performance and the amount of synthetic data considered. For both operations (Figure 2), we observe that improvements in F_1 score are supported by a rise in recall which reaches a pleateau around 30% and, for UTX_COL, in precision.

6 Conclusions and Future Work

We investigated an inexpensive framework to integrate unlabeled ID data to improve cross-target SD. We studied Twitter SD and showed, through a comprehensive set of experiments, that it is a promising strategy. We reserve to study its applicability to other domains in future work.

Acknowledgments

We thank the anonymous reviewers for their efforts and for the constructive suggestions. We gratefully acknowledge funding from the Keynes Fund, University of Cambridge (grant no. JHOQ). CC is grateful to NERC DREAM CDT (grant no. 1945246) for partially funding this work. CG and FT are thankful to the Cambridge Endowment for Research in Finance (CERF).

References

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of WWW 2003*. ACM.

Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. Stance classification in out-of-domain rumours: A case study around mental health disorders. In *Social Informatics - 9th International Conference*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of NAACL 2018*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and

blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*.

Robert F Bruner and Joseph R Perella. 2004. *Applied mergers and acquisitions*, volume 173. John Wiley & Sons.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC 2014*. BMVA Press.

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 ACL 2020*.

Kareem Darwish, Peter Stefanov, Michaël J. Aupetit, and Preslav Nakov. 2019. Unsupervised user stance detection on twitter. *CoRR*, abs/1904.02000.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP 2018*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of ACL 2017*.

Jason A. Fries, Sen Wu, Alexander Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *CoRR*, abs/1704.06360.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. Rumoureval 2019: Determining rumour veracity and support for rumours. *CoRR*, abs/1809.06683.

Andreas Hanselowski, Avinesh P., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of COLING 2018*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP 2014*, pages 751–762.

Tomáš Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. Detecting stance in czech news commentaries. In *Proceedings of SloNLP 2017*.

Diana Inkpen, Xiaodan Zhu, and Parinaz Sobhani. 2017. A dataset for multi-target stance detection. In *Proceedings of EACL 2017*, pages 551–557. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP 2017*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Kushal Kafle, Mohammed A. Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of INLG 2017*.

Manfred Klenner, Don Tuggener, and Simon Clematide. Stance detection in facebook posts of a german right-wing party. In *Proceedings of LSDSem@EACL 2017*.

David McClosky, Eugene Charniak, and Mark Johnson. When is self-training effective for parsing? In *Proceedings of COLING2008*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proceedings of NAACL-HLT 2010*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL 2009*. The Association for Computer Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.

Barbara Plank, Dirk Hovy, Ryan T. McDonald, and Anders Søgaard. 2014. Adapting taggers to twitter with not-so-distant supervision. In *Proceedings of COLING 2014*.

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

Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason A. Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *VLDB J.*

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of ACL 2018*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL 2016*.

Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. 2017. Annotating speaker stance in discourse: the brexit blog corpus. *Corpus Linguistics and Linguistic Theory*.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of DDDSM@IJCNLP 2017*, pages 1–8.

Anders Søgaard. 2013. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014*. Association for Computational Linguistics.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *CoRR*, abs/1511.07487.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.*, 54(2):273–290.

Appendix A: Details on Data

Table 5 reports examples of correctly and wrongly synthetically labeled samples.

Appendix B: Details on Modeling

For each test set, we include synthetically annotated tweets from a number of related mergers. Related mergers have been manually selected by an expert in the Economics domain, based on industry similarity.

- CI_ESRX (health): TMUS_S, VIAB_CBS, BMY_CELG, MSFT_LNKD, MDT_COV, IBM_RHT
- DIS_FOXA (entertainment): TMUS_S, VIAB_CBS, CMCSA_TWC, MSFT_LNKD, IBM_RHT, CHTR_TWC, MDT_COV, BMY_CELG, DELL_EMC
- UTX_COL (defense): TMUS_S, MDT_COV, VIAB_CBS, CHTR_TWC, MSFT_LNKD, BMY_CELG, CMCSA_TWC, CTL_LVLT

Preprocessing. We perform the following steps on all tweets: lowercasing, tokenization; digits/URL normalization; stripping of the # sign from hashtags; normalization of low-frequency users.

Hyperparameters Specifications Hyperparameters are reported in Table 6. When possible, we follow Riedel et al. (2017) for parameter selection. Note that we perform minimal parameter tuning: the goal of this paper is to investigate the efficacy of synthetically annotated data for SD, independently from the chosen architecture.

batch size	32
epochs	70
optimizer	*Adam* ($\lambda = 0.001$)
BoW vocabulary size	3000
dense hidden layer size	100
hidden layer dropout	0.2

Table 6: Network hyperparameters

Computing Infrastructure. We run experiments on an NVIDIA GeForce GTX 1080 GPU.

Evaluation Specifications. We use the sklearn's implementation of macro-averaged precision, recall and F_1 score.

M&A	correct	predicted	tweet text
IBM_RHT	support	support	*IBM Completes Red Hat Deal the Largest Software Acquisition Ever <URL>via Barronsonline*
AVGO_QCOM	refute	refute	*EU's $1.2-Billion Fine Against Qualcomm Might Complicate Broadcom's Bid <URL>*
MSFT_LNKD	comment	comment	*Bill Gates believes #Microsoft Can Make #LinkedIn as Successful as #Facebook <URL>$MSFT $LNKD $FB*
DELL_EMC	unrelated	unrelated	*Synnex reaches agreement with Dell for Canadian distribution <url >*
DELL_EMC	comment	support	*The largest tech deal in history is like mating elephants? Really. #dellemc <URL>*
IBM_RHT	comment	unrelated	*IBM and Red Hat Explained #ibm – <URL>*
MDT_COV	support	refute	*Medtronic's proposed $43B acquisition of Covidien has cleared all anti-trust hurdles worldwide <URL>*
MSFT_LNKD	comment	unrelated	*How Microsofts bid for LinkedIn sets a standard for every business to copy <URL>*
DELL_EMC	support	unrelated	*Dell to acquire EMC in $67 billion record tech deal <URL>#CloudComputing*
MDT_COV	support	comment	*Medtronic is still in on Covidien buyout — MassDevice <URL>*

Table 5: Examples of correctly and wrongly synthetically annotated tweets.