Adversarial Training for News Stance Detection: Leveraging Signals from a Multi-Genre Corpus.

Costanza Conforti[1], **Jakob Berndt**[2], **Mohammad Taher Pilehvar**[1,3], **Marco Basaldella**[1], **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 constitutes an important issue for news Stance Detection (SD). In this short paper, we investigate *adversarial cross-genre SD*, where knowledge from annotated user-generated data is leveraged to improve news SD on targets unseen during training. We implement a BERT-based adversarial network and show experimental performance improvements over a set of strong baselines. Given the abundance of user-generated data, which are considerably less expensive to retrieve and annotate than news articles, this constitutes a promising research direction.

1 Introduction

Stance Detection (SD) is an important NLP task (Mohammad et al., 2017) with widespread applications, ranging from rumor verification (Derczynski et al., 2017) and fact checking (Hanselowski et al., 2019). Traditionally, research in SD focused on user-generated data, such as Twitter or Reddit (Gorrell et al., 2019): this is mainly due to the abundance of such data, which are usually freely available online; moreover, user-generated data tend to be relatively short and compact, and thus more affordable to annotate and process. Starting from popular shared tasks such as Pomerleau and Rao (2017), SD on complex and articulated input, such as news articles, has gained increasing popularity. Notably, effective news SD would constitute an invaluable tool to enhance the performance of human journalists in rumor and fake news debunking (Thorne and Vlachos, 2018).

In line with the general trend in NLP, deep learning-based models have long since established state-of-the-art results in news SD (Hanselowski et al., 2018). Notably, training neural networks relies heavily on the availability of large labeled datasets, which are especially expensive to obtain for items such as news articles. As a consequence, following research on other text classification tasks such as sentiment analysis (Du et al., 2020), research in SD investigated effective methods for *cross-domain SD*, where the scarcity of data for a specific dataset is supplemented with stance-annotated data from other domains. In this context, preliminary research in adversarial domain adaptation obtained promising results for both Twitter (Wang et al., 2020) and news (Xu et al., 2019) SD.

In this paper, we focus on the new task of *cross-genre SD*: we consider adversarial knowledge transfer from two datasets, WT–WT and STANDER, which collect samples in the *same domain* (i.e. the financial domain), but which belong to *different genres* (i.e. Twitter and news). We show experimentally that improvements in news SD performance can be achieved through cross-genre SD, which constitutes a promising direction for future research.

2 An Aligned Multi-Genre Stance Detection Corpus

In this work, we rely on two recently released datasets for news and Twitter SD: the STANDER corpus for the news genre (Conforti et al., 2020a), and the WT–WT corpus for Twitter (Conforti et al., 2020b). Both corpora collect samples discussing four mergers and acquisition (M&A) operations in the healthcare industry (Table 2): an M&A operation, or *merger*, is the process in which a company (the *buyer*) attempts to acquire the ownership of another company (the *target*). A merger succeeds if ownership of the target is transferred, but can fail at any stage of discussions or can be blocked by authorities due to, e.g., antitrust concerns (Bruner and Perella, 2004).

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 1–7
April 19, 2021

	AET_HUM		ANTM_CI		CI_ESRX		CVS_AET	
Label	tweets	articles	tweets	articles	tweets	articles	tweets	articles
support	1,013	463	959	367	763	207	2,438	372
refute	1,110	537	1,966	313	265	64	530	104
comment	2,776	197	3,101	248	935	70	5,491	294
unrelated	2,930	5	4,995	14	548	5	3,058	31
total	7,829	1,009	11,021	1,199	2,511	376	11,517	831

Table 1: Label distribution in the STANDER News SD corpus and in the WT–WT Twitter SD corpus.

Samples in both STANDER and WT–WT are manually stance-labeled by domain experts using a four-classes annotation schema distinguishing between *support, refute, comment* and *unrelated*, which expresses the sample's orientation about the outcome of the M&A (succeeded or rejected).

As observed in Conforti et al. (2020a), the two corpora present comparable signals, but display different characteristics which reflect the diverse genres they belong to. The Twitter samples are abundant and noisy, as indicated by the high percentage of *unrelated* and *commenting* samples (Figure 1 and Table 1). On the other hand, STANDER collects considerably fewer samples, which are substantially longer and articulated; moreover, news articles in STANDER have been published in high-reputation outlets after careful editorial review, and thus contain a more formal and orthographically correct language with respect to user-generated tweets.

3 Adversarial Training for News Stance Detection

3.1 Motivation

Given the scarcity of news articles in STANDER, which are around one order of magnitude less abundant than tweets in WT–WT, and considering that both corpora collect the same targets in the same domain, *cross-genre* SD from WT–WT to STANDER seems to constitute an interesting research direction. However, due to the consistent genre differences, transferring knowledge from WT–WT to STANDER is non-trivial. To allow the model to capture the stance-specific features from the WT–WT samples which are useful to perform news SD, while ignoring the Twitter-specific features, we propose to treat the task adversarially.

Merger	Buyer	Target	Outcome
AET_HUM	Aetna	Humana	rejected
ANTM_CI	Anthem	Cigna	rejected
CI_ESRX	Cigna	Express Scripts	succeeded
CSV_AET	CVS	Aetna	succeeded

Table 2: Mergers considered in this work. Note that two companies appear both as *Buyer* and as *Target*.

3.2 Models

We propose to consider two classification problems – SD and genre identification (GI) – with a shared BERT-based feature extractor, as shown in Figure 2. To derive genre-invariant features, the GI component is trained adversarially.

The model receives an input sample as:

`[CLS] Target [SEP] Text [SEP]`,

where `Target` is the SD target, expressed as the sentence "*A (a) will merge with B (b)*" (where upper- and lowercase *a* and *b* refers resp. to the buyer's and the target's company names and acronyms); for tweets, `Text` is the entire sample's text, while for news, we concatenate the article's title and its first four sentences into a single string. In this way, the target input is always the same over both genres, and it changes over targets only in the company names.

Feature Extractor. As shared feature extractor, we adopt the pretrained $\text{BERT}_{\text{base}}$ uncased model (Devlin et al., 2019).

Stance Classifier. The stance label is predicted

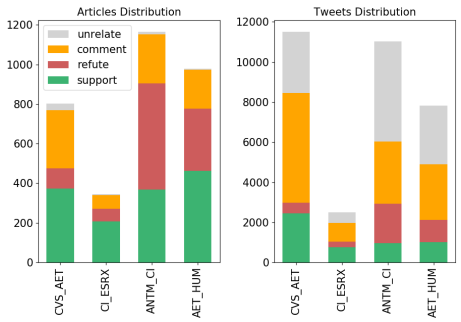

Figure 1: Distribution and number of samples in the STANDER news SD and the WT–WT Twitter SD corpus (Figure taken from Conforti et al. (2020a)).

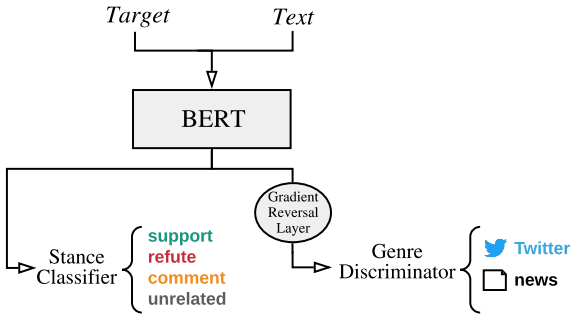

Figure 2: Architecture of our model for news SD.

with a dense layer followed by a softmax operation operating on the hidden state $h_{[CLS]}$ of the special classification embedding `[CLS]`:

$$y_s = \text{softmax}(W_s h_{[CLS]} + b_s) \quad (1)$$

The stance classifier is trained with categorical cross-entropy.

Genre Discriminator. The genre discriminator aims to predict gender labels of samples (Twitter or news). The feature extractor parameters are optimized to maximize the loss of the genre discriminator, thus encouraging BERT to generate genre-invariant features. In practice, the hidden state $h_{[CLS]}$ is first fed to a Gradient Reversal Layer (GRL, Ganin and Lempitsky (2014)). During the forward propagation, the GRL acts as an identity transformation:

$$GRL_\lambda(x) = x \quad (2)$$

but, during the backpropagation, it multiplies the gradient by a negative factor λ:

$$\frac{\delta GRL_\lambda(x))}{\delta x} = -\lambda I \quad (3)$$

The genre label y_g is finally obtained with a dense layer followed by a sigmoid operation:

$$y_g = \text{sigmoid}(W_g GRL(h_{[CLS]}) + b_g) \quad (4)$$

The genre discriminator is trained with binary cross-entropy.

Joint Learning. The two components are jointly trained, resulting in the total loss:

$$L_{total} = L_{stance} + L_{genre} \quad (5)$$

The GI component is adversarial because it is trained to maximise the loss, while the SD component attempts to minimise it. In this way, the more the GI component is unable to correctly classify the samples, the more the system has learned to extract genre-invariant features.

4 Experimental Setting

Baselines. We report results with the three baseline models proposed in Conforti et al. (2020a): a dummy *random* and *majority vote* baseline, abd *BertEmb*, an MLP leveraging sentence-BERT embeddings (Reimers and Gurevych, 2019); moreover, we also consider two further baselines:

- $BERT_{news}$: A vanilla BERT finetuned on news samples only;
- $BERT_{CoTrain}$: A vanilla BERT finetuned on Tweet and news samples, but without the adversarial component (Blum and Mitchell, 1998; Chen et al., 2011).

Training Setting and Preprocessing. We train in a cross-target setting (train on three mergers, test on the fourth) with the Adam optimizer. For each configuration, we randomly select 20% of the training samples as heldout data. For experiments with adversarial cross-genre SD, we randomly select a number of Twitter samples equal to the news training samples (i.e. we double the size of the training set). This ratio was found to perform best in preliminary experiments (refer to the Appendix for further details). The test set contains news samples only. We lowercase both tweets and news samples.

Hyperparameters. We set 128 as maximum sample length (including special tokens). We initialize our architecture with BERT large uncased[1]. BERT's weights are updated during training. We train using Adam (Kingma and Ba, 2015) on baatches of 23 samples, for a maximum of 70 epochs, with early stopping monitoring the SD loss on the development set.

Evaluation. As in Conforti et al. (2020a,b), and in line with other works on news SD (Hanselowski et al., 2018, 2019), we report on macro-averaged F_1 and consider both per-target operation scores, and average scores weighted by target operation size. For the adversarial cross-genre experiments, we compute accuracy for the binary GI task. For computing the evaluation metrics, we use the sklearn's implementations[2].

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

[1] https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/3

[2] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

Model	per-target F_1				$avgF_1$
	CVS AET	CI ESRX	ANTM CI	AET HUM	
Baselines from Conforti et al. (2020a)					
Majority	12.0	12.0	12.0	12.0	12.0
Random	17.5	17.4	17.1	16.5	17.1
BertEmb	42.5	33.2	46.4	43.9	45.7
BERT_{news}	45.5	60.6	47.7	48.8	48.9
$\text{BERT}_{CoTrain}$	47.1	64.8	48.4	51.5	50.8
BERT_{adv}					
$\lambda = 0.2$	**48.0**	64.5	**52.4**	**52.0**	**52.5**
$\lambda = 0.5$	46.9	**66.6**	51.5	50.8	51.8
$\lambda = 0.7$	45.9	64.8	48.9	50.1	50.2
$\lambda = 1.0$	42.6	60.2	47.2	46.1	47.2

Table 3: Results on the STANDER target operations. Macro F_1 scores are obtained by testing on the target operation while training on the other three. Average scores are weighted by the target operation size.

Model	GI $avgAcc$	SD $avgF_1$	Avg. per-class F_1			
			sup	*ref*	*com*	*unr*
BERT_{news}		48.9	70.4	65.8	43.6	18.2
$\text{BERT}_{CoTrain}$		50.8	70.5	67.6	45.4	18.5
BERT_{adv}						
$\lambda = 0.2$	65.8	**52.5**	**72.5**	70.4	**48.0**	19.2
$\lambda = 0.5$	65.3	51.8	69.3	**71.0**	46.5	**20.4**
$\lambda = 0.7$	43.6	50.2	68.9	68.5	44.0	16.4
$\lambda = 1.0$	13.7	47.2	69.3	68.2	34.5	12.0

Table 4: Per-target averaged accuracy for Gender Identification (GI) and per-target averaged F_1 score for Stance Detection (SD), along with single-label per-target averaged F_1 scores.

5 Experiments and Discussion

In this Section, we report on our cross-genre SD experiments. The discussion is organized around three research questions.

RQ1 *What is the effect on news SD of including annotated data from a different genre?*
As shown in Table 3, $\text{BERT}_{CoTrain}$ performs better than all baselines, including the BERT_{news} model, which was trained on in-genre data only. This seems to suggest that exposing the model to in-domain stance-annotated data, even if from a completely different genre, improves the generalizability over unseen targets.

RQ2 *Is adversarial training effective to improve cross-genre SD?*
Adding an adversarial component to the $\text{BERT}_{CoTrain}$ model leads to gains in performance over all considered targets, with improvements ranging from $+0.5$ (AET_HUM) and $+4.0$ (ANTM_CI) in macro-averaged F_1 score. (Table 3). Such performance gains are driven by improved performance over all stance labels(Table 4), with *refute* samples benefiting most from the adversarial component.

Such results suggest that a model which is punished for identifying the input's genre can still perform SD.

RQ3 *How does a decrease in GI through adversarial training correlate with SD performance?*
To understand to which extent genre-invariant representations are useful for SD, we experiment with different values of λ, the GRL hyperparameter in Equation 3. As expected, GI performance lowers with increasing λ (Table 4), reaching 13.7 GI accuracy for the model with $\lambda = 1$; this proves the GRL efficacy in forcing the model to learn genre-independent features. However, this also correlates with a steady decrease in SD performance, which holds true all target operations (Table 3), with the only exception of the relatively small CI_ESRX target, which also exhibit very strong label unbalancy (Table 1).

Moving to single-label classification, higher losses in performance are observed for *comment* and *unrelated* samples (resp. -13.5 and -8.3 in weighted accuracy), while *support* and *refute* label seem to be more robust to changes in the values of λ. A possible explanation for this might be in the stylistic differences between the two corpora: *unrelated* and *comment* samples in the Twitter WT–WT corpus were often retrieved because of keywords homonymy[3], and, as such, they tend to discuss completely different topics; on the contrary, such samples in the news STANDER corpus are actually covering the target companies. For this reason, completely genre-unaware knowledge transfer might not be optimal for those stance labels.

This is in line with previous work by McHardy et al. (2019) on satire detection, and seems to indicate that, while learning partially genre-invariant features is beneficial for cross-target performance, features which are completely opaque with respect to the genre component are not ideal for SD.

[3]For example, 'cvs' not referring to the company 'CVS Health', but used as as plural of 'resume'.

6 Conclusions

In this paper, we discussed the new task of *cross-genre SD*: our experiments with a range of BERT-based architectures show that partially obfuscating the genre component through adversarial training leads to better generalization, especially considering low-frequency labels. Cross-genre SD thus constitutes a promising future research direction. Future work might include experiments using different underlying feature extractors, such as RoBERTa, or with adapters, to study the robustness of cross-genre SD over modeling choices. The integration of cross-genre and cross-domain adaptation, possibly in a multi-task setting as in Conforti et al. (2020a), also offers interesting ideas for future investigation.

Acknowledgements

We thank the anonymous reviewers of this paper for their efforts and for the constructive comments and 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

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998.*, pages 92–100.

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

Minmin Chen, Kilian Q. Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2456–2464.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. STANDER: an expert-annotated dataset for news stance detection and evidence retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4086–4101. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020b. 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, ACL 2020, Online, July 5-10, 2020*, pages 1715–1724. Association for Computational Linguistics.

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@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 69–76.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

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 SemEval 2019*.

Andreas Hanselowski, Avinesh P. V. S., 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 the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1859–1874. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*

(CoNLL), pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 660–665. Association for Computational 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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3346–3359. Association for Computational Linguistics.

Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. 2020. Unseen target stance detection with adversarial domain generalization. *2020 International Joint Conference on Neural Networks (IJCNN)*.

Brian Xu, Mitra Mohtarami, and James Glass. 2019. Adversarial domain adaptation for stance detection. *CoRR*, abs/1902.02401.

A Appendix

In this Appendix, we report on the results of preliminary experiments which were run in order to find the best proportion between news and Twitter samples. We consider six settings, with increasing proportion of Twitter to news samples: 50%, 100%, 150%, 200%, 250% and a last setting in which all Twitter samples were included. For each setting, experiments were run in a cross-validation setting (training on data on three operations, and testing on the fourth).

% of Twitter data	per-target F_1 CVS AET	CI ESRX	ANTM CI	AET HUM	$avgF_1$
BERT$_{news}$	43.3	62.2	48.5	50.9	49.4
= 50	42.5	63.5	49.5	51.2	49.8
= 100	48.0	64.5	52.4	52.0	52.5
= 150	48.9	65.5	47.3	52.5	51.2
= 200	56.1	63.6	47.3	51.7	52.3
= 250	51.4	61.1	51.6	51.8	52.5
all	54.6	62.1	48.3	47.7	51.2

Table 5: Stance Detection performance with the genre-adversarial model ($\lambda = 0.2$), by adding different proportion of Twitter data from the WT–WT corpus to the STANDER news samples. A % of 100 corresponds to the proportion used in the experiments reported in the paper. *all* corresponds to all the samples in the WT–WT corpus (33,668).

Interestingly, we observe gains in overall performance w.r.t. the BERT$_{news}$ baseline, which is trained on news data only (Table 5), with all adversarial models. This holds true even in the case of the model trained on the union of STANDER (2945) with *all* tweets from the WT–WT corpus (30711): the model's considerable performance gain with respect to the BERT$_{news}$ model testifies the ability of the adversarial model to learn partially genre-invariant features even when exposed to extremely unbalanced training data. Single-label results (Table 6) show that increasing the ratio of Twitter samples included in the training data tends to correlate with performance gains in recall, at the expense of losses in performance. The best news-to-tweets ratio lies between 100 and 250, with small differences between target operations and stance labels. Thus, our adversarial cross-genre models seem to be relatively robust over the exact amount of out-of-genre samples which are included during training.

% of	SD			Avg. per-class performance											
Twitter				*sup*			*ref*			*com*			*unr*		
data	avgP	avgR	$avgF_1$	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
50	52.2	49.9	49.8	69.5	73.5	70.4	69.1	70.3	69.5	47.5	42.6	44.9	15.3	20.8	15.4
100	54.6	55.2	**52.5**	**75.3**	70.9	**72.2**	67.8	73.6	70.5	49.8	51.6	49.0	25.9	24.1	**18.8**
150	**56.4**	52.4	51.2	68.7	74.1	69.8	**73.4**	67.3	69.9	48.8	48.8	46.7	**34.6**	19.4	18.1
200	54.9	54.1	52.3	73.7	70.8	71.8	72.6	67.9	70.0	49.0	**57.7**	**52.0**	24.2	20.1	16.5
250	55.2	**55.3**	**52.5**	71.3	**73.6**	71.8	68.9	**75.3**	**71.9**	**54.0**	46.0	48.9	26.1	**27.2**	18.6
all	52.1	53.5	51.1	72.0	70.5	70.1	70.5	71.8	71.1	49.6	45.4	46.7	16.2	25.7	17.0

Table 6: Per-label detailed performance when adding different percentage of tweets to the STANDER news samples. A % of 100 corresponds to the proportion used in the experiments reported in the paper. All results are obtained with the genre-adversarial model ($\lambda = 0.2$). Considering macro-averaged Precision, Recall and F_1 measures, weighted according to the target operation's size.