

Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter

Costanza Conforti¹, Jakob Berndt², Mohammad Taher Pilehvar^{1,3},
Chryssi Giannitsarou², Flavio Toxvaerd², Nigel Collier¹

¹ Language Technology Lab, University of Cambridge

² Faculty of Economics, University of Cambridge

³ Tehran Institute for Advanced Studies, Iran

{cc918, jb2088, mp792, cg349, fmot2, nhc30}@cam.ac.uk

Abstract

We present a new challenging stance detection dataset, called Will-They-Won't-They¹ (WT-WT), which contains 51,284 tweets in English, making it by far the largest available dataset of the type. All the annotations are carried out by experts; therefore, the dataset constitutes a high-quality and reliable benchmark for future research in stance detection. Our experiments with a wide range of recent state-of-the-art stance detection systems show that the dataset poses a strong challenge to existing models in this domain. The entire dataset is released for future research².

1 Introduction

Apart from constituting an interesting task on its own, stance detection has been identified as a crucial sub-step towards many other NLP tasks (Mohammad et al., 2017). In fact, stance detection is the core component of fake news detection (Pomerleau and Rao, 2017), fact-checking (Vlachos and Riedel, 2014; Baly et al., 2018), and rumor verification (Zubiaga et al., 2018b).

Despite its importance, stance detection suffers from the lack of a large dataset which would allow for reliable comparison between models. We aim at filling this gap by presenting Will-They-Won't-They (WT-WT), a large dataset of English tweets targeted at stance detection for the rumor verification task. We constructed the dataset based on tweets, since Twitter is a highly relevant platform for rumour verification, which is popular with the public as well as politicians and enterprises (Gorell et al., 2019).

To make the dataset representative of a realistic scenario, we opted for a real-world application

of the rumor verification task in finance. Specifically, we constructed the dataset based on tweets that discuss mergers and acquisition (M&A) operations between companies. M&A is a general term that refers to various types of financial transactions in which the ownership of companies are transferred. An M&A process has many stages that range from informal talks to the closing of the deal. The discussions between companies are usually not publicly disclosed during the early stages of the process (Bruner and Perella, 2004; Piesse et al., 2013). In this sense, the analysis of the evolution of opinions and concerns expressed by users about a possible M&A deal, from its early stage to its closing (or its rejection) stage, is a process similar to rumor verification (Zubiaga et al., 2018a).

Moreover, despite the wide interest, most research in the intersection of NLP and finance has so far focused on sentiment analysis, text mining and thesauri/taxonomy generation (Fisher et al., 2016; Hahn et al., 2018; El-Haj et al., 2018). While sentiment (Chan and Chong, 2017) and targeted-sentiment analysis (Chen et al., 2017) have an undisputed importance for analyzing financial markets, research in stance detection takes on a crucial role: in fact, being able to model the market's perception of the merger might ultimately contribute to explaining stock price re-valuation.

We make the following three contributions. Firstly, we construct and release WT-WT, a large, expert-annotated Twitter stance detection dataset. With its 51,284 tweets, the dataset is an order of magnitude larger than any other stance detection dataset of user-generated data, and could be used to train and robustly compare neural models. To our knowledge, this is the first resource for stance in the financial domain. Secondly, we demonstrate the utility of the WT-WT dataset by evaluating 11 competitive and state-of-the-art stance detection models on our benchmark. Thirdly, we annotate a further

¹<https://en.wiktionary.org/wiki/will-they-won%27t-they>

²<https://github.com/cambridge-wtw/acl2020-wtw-tweets>

M&A	Buyer	Target	Outcome
CVS_AET	CVS Health	Aetna	Succeeded
CI_ESRX	Cigna	Express Scripts	Succeeded
ANTM_CI	Anthem	Cigna	Blocked
AET_HUM	Aetna	Humana	Blocked
DIS_FOXA	Disney	21st Century Fox	Succeeded

Table 1: Considered M&A operations. Note that AET and CI appear both as buyers and as targets.

M&A operation in the entertainment domain; we investigate the robustness of best-performing models on this operation, and show that such systems struggle even over small domain shifts. The entire dataset is released to enable research in stance detection and domain adaptation.

2 Building the WT-WT Dataset

We consider five recent operations, 4 in the healthcare and 1 in the entertainment industry (Table 1).

2.1 Data Retrieval

For each operation, we used Selenium³ to retrieve IDs of tweets with one of the following sets of keywords: mentions of both companies’ names or acronyms, and mentions of one of the two companies with a set of merger-specific terms (refer to Appendix A.1 for further details). Based on historically available information about M&As, we sampled messages from one year before the proposed merger’s date up to six months after the merger took place. Finally, we obtain the text of a tweet by crawling for its ID using Tweepy⁴.

2.2 Task Definition and Annotation Guidelines

The annotation process was preceded by a pilot annotation, after which the final annotation guidelines were written in close collaboration with three domain experts. We followed the convention in Twitter stance detection (Mohammad et al., 2017) and considered three stance labels: *support*, *refute* and *comment*. We also added an *unrelated* tag, obtaining the following label set:

1. Support: the tweet is stating that the two companies will merge.
[CI_ESRX] *Cigna to acquire Express Scripts for \$52B in health care shakeup via usatoday*

³www.seleniumhq.org

⁴www.tweepy.org/

2. Refute: the tweet is voicing doubts that the two companies will merge.
[AET_HUM] *Federal judge rejects Aetna’s bid to buy Louisville-based Humana for \$34 billion*
3. Comment: the tweet is commenting on merger, neither directly supporting, nor refuting it.
[CI_ESRX] *Cigna-Express Scripts deal unlikely to benefit consumers*
4. Unrelated: the tweet is unrelated to merger.
[CVS_AET] *Aetna Announces Accountable Care Agreement with Weill Cornell Physicians*

The obtained four-class annotation schema is similar to those in other corpora for news stance detection (Hanselowski et al., 2018; Baly et al., 2018). Note that, depending on the given target, the same sample can receive a different stance label:

Merger hopes for Aetna-Humana remain, Anthem-Cigna not so much.

[AET_HUM] → *support*

[ANTM_CI] → *refute*

As observed in Mohammad et al. (2017), stance detection is different but closely related to targeted sentiment analysis, which considers the emotions conveyed in a text (Alhothali and Hoey, 2015). To highlight this subtle difference, consider the following sample:

[CVS_AET] *#Cancer patients will suffer if @CVSHealth buys @Aetna CVS #PBM has resulted in delfays in therapy, switches, etc – all documented. Terrible!*

While its sentiment towards the target operation is *negative* (the user believes that the merger will be harmful for patients), following the guidelines, its stance should be labeled as *comment*: the user is talking about the implications of the operation, without expressing the orientation that the merger will happen (or not). Refer to Appendix A.2 for a detailed description of the four considered labels.

2.3 Data Annotation

During the annotation process, each tweet was independently labeled by 2 to 6 annotators. Ten experts in the financial domain were employed as annotators⁵. Annotators received tweets in batches of 2,000 samples at a time, and were asked to annotate no more than one batch per week. The entire annotation process lasted 4 months. In case of disagreement, the gold label was obtained through

⁵Two MPhil, six PhD students and two lecturers at the Faculty of Economics of the University of Cambridge

Label	Healthcare								Entertainment	
	CVS_AET		CI_ESRX		ANTM_CI		AET_HUM		DIS_FOXA	
	# samples	%	# samples	%	# samples	%	# samples	%	# samples	%
support	2,469	21.24	773	30.58	970	8.78	1,038	13.14	1,413	7.76
refute	518	4.45	253	10.01	1,969	17.82	1,106	14.00	378	2.07
comment	5,520	47.49	947	37.47	3,098	28.05	2,804	35.50	8,495	46.69
unrelated	3,115	26.80	554	21.92	5,007	45.33	2,949	37.34	7,908	43.46
total	11,622		2,527		11,622		7,897		18,194	

Table 2: Label distribution across different M&A operations (Table 1): four mergers in the healthcare domain (33,090 tweets) and one merger in the entertainment domain. The total number of tweets is: 51,284.

	total twt	avg twt/target
Mohammad et al. (2016b)	4,870	811
Inkpen et al. (2017)	4,455	1,485
Aker et al. (2017)	401	401
Derczynski et al. (2017)	5,568	696
Gorrell et al. (2019) (only Twitter)	6,634	829
WT-WT	51,284	10,256

Table 3: Statistics of Twitter stance detection datasets.

majority vote, discarding samples where this was not possible (0.2% of the total).

2.4 Quality Assessment

The average Cohen’s κ between the annotator pairs⁶ 0.67, which is *substantial* (Cohen, 1960). To estimate the quality of the obtained corpus, a further domain-expert labeled a random sample of 3,000 tweets, which were used as human upperbound for evaluation (Table 4). Cohen’s κ between those labels and the gold is 0.88. This is well above the agreement obtained in previously released datasets where crowd-sourcing was used (the agreement scores reported, in terms of percentage, range from 63.7% (Derczynski et al., 2017) to 79.7% (Inkpen et al., 2017)).

Support-comment samples constitute the most common source of disagreement between annotators: this might indicate that such samples are the most subjective to discriminate, and might also contribute to explain the high number of misclassifications between those classes which have been observed in other research efforts on stance detection (Hanselowski et al., 2018). Moreover, w.r.t. stance datasets where unrelated samples were randomly generated (Pomerleau and Rao, 2017; Hanselowski et al., 2018), we report a slightly

⁶The average κ was weighted by the number of samples annotated by each pair. The standard deviation of the κ scores between single annotator pairs is 0.074.

higher disagreement between *unrelated* and *comment* samples, indicating that our task setting is more challenging.

2.5 Label Distribution

The distribution of obtained labels for each operation is reported in Table 2. Differences in label distribution between events are usual, and have been observed in other stance corpora (Mohammad et al., 2016a; Kochkina et al., 2018). For most operations, there is a clear correlation between the relative proportion of *refuting* and *supporting* samples and the merger being approved or blocked by the US Department of Justice. *Commenting* tweets are more frequent than *supporting* over all operations: this is in line with previous findings in financial microblogging (Žnidaršič et al., 2018).

2.6 Comparison with Existing Corpora

The first dataset for Twitter stance detection collected 4,870 tweets on 6 political events (Mohammad et al., 2016a) and was later used in SemEval-2016 (Mohammad et al., 2016b). Using the same annotation schema, Inkpen et al. (2017) released a corpus on the 2016 US election annotated for multi-target stance. In the scope of PHEME, a large project on rumor resolution (Derczynski and Bontcheva, 2014), Zubiaga et al. (2015) stance-annotated 325 conversational trees discussing 9 breaking news events. The dataset was used in RumourEval 2017 (Derczynski et al., 2017) and was later extended with 1,066 tweets for RumourEval 2019 (Gorrell et al., 2019). Following the same procedure, Aker et al. (2017) annotated 401 tweets on mental disorders (Table 3).

This makes the proposed dataset by far the largest publicly available dataset for stance detection on user-generated data. In contrast with Mohammad et al. (2016a), Inkpen et al. (2017) and

Encoder	Macro F_1 across healthcare operations				Average per-class accuracy					
	CVS_AET	CI_ESRX	ANTM_CI	AET_HUM	$avgF_1$	avg_wF_1	<i>sup</i>	<i>ref</i>	<i>com</i>	<i>unr</i>
SVM	51.0	51.0	65.7	65.0	58.1	58.5	54.5	43.9	41.2	88.4
MLP	46.5	46.6	57.6	59.7	52.6	52.7	55.7	40.3	48.6	68.1
EmbAvg	50.4	51.9	50.4	58.9	52.9	52.3	55.2	50.5	52.7	67.4
CharCNN	49.6	48.3	65.6	60.9	56.1	56.8	55.5	44.2	41.6	82.1
WordCNN	46.3	39.5	56.8	59.4	50.5	51.7	62.9	37.0	31.0	71.7
BiCE	56.5	52.5	64.9	63.0	59.2	60.1	61.0	48.7	45.1	79.9
CrossNet	59.1	54.5	65.1	62.3	60.2	61.1	63.8	48.9	50.5	75.8
SiamNet	58.3	54.4	68.7	67.7	62.2	63.1	67.0	48.0	52.5	78.3
CoMatchAtt	54.7	43.8	50.8	50.6	49.9	51.6	71.9	24.4	33.7	65.9
TAN	56.0	55.9	66.2	66.7	61.2	61.3	66.1	49.0	51.7	74.1
HAN	56.4	57.3	66.0	67.3	61.7	61.7	67.6	52.0	55.2	69.1
<i>mean</i>	53.1	50.5	61.6	62.0	—	—	61.9	44.2	45.8	74.6
<i>upperbound</i>	75.3	71.2	74.4	73.7	74.7	75.2	80.5	89.6	71.8	84.0

Table 4: Results on the healthcare operations in the WT–WT dataset. Macro F_1 scores are obtained by testing on the target operation while training on the other three. $avgF_1$ and avg_wF_1 are, respectively, the unweighted and weighted (by operations size) average of all operations.

PHEME, where crowd-sourcing was used, only highly skilled domain experts were involved in the annotation process of our dataset. Moreover, previous work on stance detection focused on a relatively narrow range of mainly political topics: in this work, we widen the spectrum of considered domains in the stance detection research with a new financial dataset.

For these reasons, the WT–WT dataset constitutes a high quality and robust benchmark for the research community to train and compare performance of models and their scalability, as well as for research on domain adaptation. Its large size also allows for pre-training of models, before moving to domain with data-scarcity.

3 Experiments and Results

We re-implement 11 architectures recently proposed for stance detection. Each system takes as input a tweet and the related target, represented as a string with the two considered companies. A detailed description of the models, with references to the original papers, can be found in Appendix B.1. Each architecture produces a single vector representation h for each input sample. Given h , we predict \hat{y} with a softmax operation over the 4 considered labels.

3.1 Experimental Setup

We perform common preprocessing steps, such as URL and username normalization (see Appendix B.2). All hyper-parameters are listed in Appendix B.1 for replication. In order to allow for a fair

comparison between models, they are all initialized with Glove embeddings pretrained on Twitter⁷ (Pennington et al., 2014), which are shared between tweets and targets and kept fixed during training.

3.2 Results and Discussion

Results of experiments are reported in Table 4. Despite its simple architecture, SiamNet obtains the best performance in terms of both averaged and weighted averaged F_1 scores. In line with previous findings (Mohammad et al., 2017), the SVM model constitutes a very strong and robust baseline. The relative gains in performance of CrossNet w.r.t. BiCE, and of HAN w.r.t. TAN, consistently reflect results obtained by such models on the SemEval 2016-Task 6 corpus (Xu et al., 2018; Sun et al., 2018).

Moving to single labels classification, analysis of the confusion matrices shows a relevant number of misclassifications between the *support* and *comment* classes. Those classes have been found difficult to discriminate in other datasets as well (Hanselowski et al., 2018). The presence of linguistic features, as in the HAN model, may help in spotting the nuances in the tweet’s argumentative structure which allow for its correct classification. This may hold true also for the *refute* class, the least common and most difficult to discriminate. *Unrelated* samples in WT–WT could be about the involved companies, but not about their merger: this makes classification more challenging than in datasets containing randomly generated *unre-*

⁷<https://nlp.stanford.edu/projects/>

lated samples (Pomerleau and Rao, 2017). SVM and CharCNN obtain the best performance on *unrelated* samples: this suggests the importance of character-level information, which could be better integrated into future architectures.

Concerning single operations, CVS_AET and CI_ESRX have the lowest average performance across models. This is consistent with higher disagreement among annotators for the two mergers.

3.3 Robustness over Domain Shifts

We investigate the robustness of SiamNet, the best model in our first set of experiments, and BiCE, which constitutes a simpler neural baseline (Section 3.2), over domain shifts with a cross-domain experiment on an M&A event in the entertainment business.

Data. We collected data for the Disney-Fox (DIS_FOXA) merger and annotated them with the same procedure as in Section 2, resulting in a total of 18,428 tweets. The obtained distribution is highly skewed towards the *unrelated* and *comment* class (Table 2). This could be due to the fact that users are more prone to digress and joke when talking about the companies behind their favorite shows than when considering their health insurance providers (see Appendix A.2).

train → test	BiCE		SiamNet	
	acc	F_1	acc	F_1
health → health	77.69	76.08	78.51	77.38
health → ent	57.32	37.77	59.85	40.18
ent → ent	84.28	74.82	85.01	75.42
ent → health	46.45	33.62	48.99	35.25

Table 5: Domain generalization experiments across entertainment (ent) and healthcare datasets. Note that the data partitions used are different than in Table 4.

Results. We train on all healthcare operations and test on DIS_FOXA (and the contrary), considering a 70-15-15 split between train, development and test sets for both sub-domains. Results show SiamNet consistently outperforming BiCE. The consistent drop in performance according to both accuracy and macro-avg F_1 score, which is observed in all classes but particularly evident for *commenting* samples, indicates strong domain dependency and room for future research.

4 Conclusions

We presented WT-WT, a large expert-annotated dataset for stance detection with over 50K labeled tweets. Our experiments with 11 strong models indicated a consistent ($>10\%$) performance gap between the state-of-the-art and human upperbound, which proves that WT-WT constitutes a strong challenge for current models. Future research directions might explore the usage of transformer-based models, as well as of models which exploit not only linguistic but also network features, which have been proven to work well for existing stance detection datasets (Aldayel and Magdy, 2019).

Also, the multi-domain nature of the dataset enables future research in cross-target and cross-domain adaptation, a clear weak point of current models according to our evaluations.

Acknowledgments

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

- Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, Anna Kolliakou, Rob Procter, and Maria Liakata. 2017. *Stance classification in out-of-domain rumours: A case study around mental health disorders*. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II*, volume 10540 of *Lecture Notes in Computer Science*, pages 53–64. Springer.
- Abeer Aldayel and Walid Magdy. 2019. *Your stance is exposed! analysing possible factors for stance detection on social media*. *PACMHCI*, 3(CSCW):205:1–205:20.
- Areej Alhothali and Jesse Hoey. 2015. *Good news or bad news: Using affect control theory to analyze readers’ reaction towards news articles*. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1548–1558, Denver, Colorado. Association for Computational Linguistics.
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. *Stance detec-*

- tion with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 876–885. The Association for Computational Linguistics.
- 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 the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics.
- Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors. 2018. *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics.
- Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. 2016. *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. The Association for Computer Linguistics.
- Robert F Bruner and Joseph R Perella. 2004. *Applied mergers and acquisitions*, volume 173. John Wiley & Sons.
- Samuel WK Chan and Mickey WC Chong. 2017. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64.
- Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. [NLG301 at semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 847–851. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Leon Derczynski and Kalina Bontcheva. 2014. [PHEME: Veracity in digital social networks](#). In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, volume 1181 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [Semeval-2017 task 8: Rumoureal: 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. 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 - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 529–536. Springer.
- 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.
- Mahmoud El-Haj, Paul Rayson, and Andrew Moore. 2018. Proceedings of the first financial narrative processing workshop. In *Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan*.
- Ingrid E. Fisher, Margaret R. Garnsey, and Mark E. Hughes. 2016. [Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research](#). *Int. Syst. in Accounting, Finance and Management*, 23(3):157–214.
- 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.
- Iryna Gurevych and Yusuke Miyao, editors. 2018. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. Association for Computational Linguistics.
- Udo Hahn, Véronique Hoste, and Ming-Feng Tsai. 2018. Proceedings of the first workshop on economics and natural language processing. In *The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia*.
- 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 (Bender et al., 2018), pages 1859–1874.
- Kazi Saidul Hasan and Vincent Ng. 2013. [Stance classification of ideological debates: Data, models, features, and constraints](#). In *Sixth International Joint*

- Conference on Natural Language Processing, *IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1348–1356. Asian Federation of Natural Language Processing / ACL.
- Diana Inkpen, Xiaodan Zhu, and Parinaz Sobhani. 2017. [A dataset for multi-target stance detection](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 551–557. Association for Computational Linguistics.
- Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. [Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm](#). *CoRR*, abs/1704.07221.
- Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. [All-in-one: Multi-task learning for rumour verification](#). In (Bender et al., 2018), pages 3402–3413.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan 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 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016b. [Semeval-2016 task 6: Detecting stance in tweets](#). In (Bethard et al., 2016), pages 31–41.
- Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. [Stance and sentiment in tweets](#). *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.
- Jonas Mueller and Aditya Thyagarajan. 2016. [Siamese recurrent architectures for learning sentence similarity](#). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2786–2792. AAAI Press.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Jenifer Piesse, Cheng-Few Lee, Lin Lin, and Hsien-Chang Kuo. 2013. Merger and acquisition: Definitions, motives, and market responses. *Encyclopedia of Finance*, pages 411–420.
- Dean Pomerleau and Delip Rao. 2017. [Fake news challenge](#).
- 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.
- 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, COMAD/CODS 2019, Kolkata, India, January 3-5, 2019*, pages 306–309. ACM.
- Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. [Stance detection with hierarchical attention network](#). In (Bender et al., 2018), pages 2399–2409.
- Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. [Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns](#). In (Bethard et al., 2016), pages 413–419.
- 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, Baltimore, MD, USA, June 26, 2014*, pages 18–22. Association for Computational Linguistics.
- Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. [A co-matching model for multi-choice reading comprehension](#). In (Gurevych and Miyao, 2018), pages 746–751.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. [Recognizing contextual polarity in phrase-level sentiment analysis](#). In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 347–354. The Association for Computational Linguistics.
- Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. [Cross-target stance classification with self-attention networks](#). In (Gurevych and Miyao, 2018), pages 778–783.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. [Hierarchical attention networks for document classification](#). In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.
- Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič, and Senja Pollak. 2018. Trust and doubt terms in financial tweets and periodic reports. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- 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, and Michal Lukasik. 2016. *Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations*. In *COLING 2016, 26th International Conference on Computational Linguistics, December 11-16, 2016, Osaka, Japan*, pages 2438–2448. ACL.

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: Dataset-related Specifications

A.1 Crawling Specifications

- M&A-specific terms used for crawling: one of merge, acquisition, agreement, acquire, takeover, buyout, integration + mention of a given company/acronym.
- Crawl start and end dates:

CVS_AET	15/02/2017	→	17/12/2018
CI_ESRX	27/05/2017	→	17/09/2018
ANTM_CI	01/04/2014	→	28/04/2017
AET_HUM	01/09/2014	→	23/01/2017
DIS_FOXA	09/07/2017	→	18/04/2018

A.2 Description and Examples of the Considered Labels

This is an extract from the annotation guidelines sent to the annotators.

The annotation process consists of choosing one of four possible labels, given a tweet and an M&A operation. The four labels to choose from are *Support*, *Comment*, *Refute*, and *Unrelated*.

Label 1: Support – If the tweet is supporting the theory that the merger is happening. Supporting tweets can be, for example, one of the following:

1. Explicitly stating that the deal is happening:
→ [CI_ESRX] *Cigna to acquire Express Scripts for \$52B in health care shakeup via usatoday*

2. Stating that the deal is likely to happen:
→ [CVS_AET] *CVS near deal to buy Aetna (Via Boston Herald)* <URL>
3. Stating that the deal has been cleared:
→ [CVS_AET] *#Breaking DOJ clears #CVS \$69Billion deal for #Aetna*.

Label 2: Comment – If the tweet is commenting on the merger. The tweet should neither directly state that the deal is happening, nor refute this. Tweets that state the merger as a fact and then talk about, e.g. implications or consequences of the merger, should also be labelled as commenting. Commenting tweets can be, for example, one of the following:

1. Talking about implications of the deal:
→ [CI_ESRX] *Cigna-Express Scripts deal unlikely to benefit consumers*
2. Stating merger as fact and commenting on something related to the deal:
→ [CVS_AET] *#biotechnology Looking at the CVSAetna Deal One Academic Sees Major Disruptive Potential*
3. Talking about changes in one or both of the companies involved:
→ [CVS_AET] *Great article about the impact of Epic within the CVS and Aetna Merge* <URL>

Label 3: Refute – This label should be chosen if the tweet is refuting that the merger is happening. Any tweet that voices doubts or mentions potential roadblocks should be labelled as refuting. Refuting tweets can be, for example, one of the following:

1. Explicitly voicing doubts about the merger:
→ [ANTM_CI] *business: JUST IN: Cigna terminates merger agreement with Anthem*
2. Questioning that the companies want to move forward:
→ [CI_ESRX] *Why would \$ESRX want a deal with \$CI?*
3. Talking about potential roadblocks for the merger:
→ [CI_ESRX] *Why DOJ must block the Cigna-Express Scripts merger* <URL>

Label 4: Unrelated – If the tweet is unrelated to the given merger. Unrelated tweets can be, for example, one of the following:

1. Talking about something unrelated to the companies involved in the merger:
→ [DIS_FOXA] *I'm watching the Disney version of Robin Hood someone tell me how I have a crush on a cartoon fox*

2. Talking about the companies involved in the merger, however not about the merger:
→ [CVS_AET] *CVS and Aetna’s combined revenue in 2016 was larger than every U.S. company’s other than Wal Mart* <URL>
3. Talking about a different merger:
→ [CVS_AET] *What are the odds and which one do you think it will be? Cigna or Humana? Aetna acquisition rumor*

Appendix B: Models-related Parameters

B.1 Encoder’s Architectures

- **SVMs**: linear-kernel SVM leveraging bag of n -grams (over words and characters) features. A similar simple system outperformed all 19 teams in the SemEval-Task 6 (Mohammad et al., 2017).
- **MLP**: a multi-layer perceptron (MLP) with one dense layer, taking as input the concatenation of tweet’s and target’s TF-IDF representations and their cosine similarity score (similar to the model in Riedel et al. (2017)).
- **EmbAvg**: a MLP with two dense layers, taking as input the average of the tweet’s and the target’s word embeddings. Averaging embeddings was proven to work well for Twitter data in previous papers by Zubiaga et al. (2016); Kochkina et al. (2017), who - differently than in this paper - classified stream of tweets in a conversation tree.
- **CharCNN** and **WordCNN**: two CNN models, one over character and one over words, following the work by Vijayaraghavan et al. (2016).
- **BiCE**: a similar Bidirectional Conditional Encoding model to that of Augenstein et al. (2016): the tweet is processed by a BiLSTM whose forward and backward initial states are initialized with the last states of a further BiLSTM which processed the target.
- **CrossNet**: a BiCE model augmented with self-attention and two dense layers, as in the cross-target stance detection model (Xu et al., 2018).
- **SiamNet**: siamese networks have been recently used for fake news stance detection (Santosh et al., 2019). Here we implement a siamese network based on a BiLSTM followed by a self-attention layer (Yang et al., 2016). The obtained tweet and target vector representations are concatenated with their similarity score (following Mueller and Thyagarajan (2016), we used the inverse exponential of the Manhattan distance).
- **Co-MatchAtt**: we use a similar co-matching attention mechanism as in Wang et al. (2018) to

connect the tweet and the target, encoded with two separated BiLSTM layers, followed by a self-attention layer (Yang et al., 2016).

- **TAN**: a model combining a BiLSTM and a target-specific attention extractor over target-augmented embeddings (Du et al., 2017; Dey et al., 2018), similarly as in Du et al. (2017).
- **HAN**: we follow Sun et al. (2018) to implement a Hierarchical Attention Network, which uses two levels of attention to leverage the tweet representation along with linguistic information (sentiment, dependency and argument).

<i>SVM model</i>	
Word NGrams	1, 2, 3
Char NGrams	2, 3, 4
<i>Common to all neural models</i>	
max tweet len	25
batch size	32
max epochs	70
optimizer	Adam
Adam learning rate	0.001
word embedding size	200
embedding dropout	0.2
<i>TFIDF-MLP model</i>	
BOW vocabulary size	3000
dense hidden layer size	100
<i>EmbAvg model</i>	
dense hidden layers size	128
<i>WordCNN model</i>	
window size	2, 3, 4
no filters	200
dropout	0.5
<i>CharCNN model</i>	
no of stacked layers	5
window size	7, 7, 3, 3, 3
no filters	256
dropout	0.2
<i>BiCE, CrossNet, SiamNet and TAN model</i>	
BiLSTM hidden size	265*2
BiLSTM recurrent dropout	0.2
<i>HAN model</i>	
max sentiment input len	10
max dependency input len	30
max argument input len	25
BiLSTM hidden size	128

Table 6: Hyperparameters used for training. Whenever reported, we used the same as in the original papers.

B.2 Preprocessing Details

After some preliminary experiments, we found the following preprocessing steps to perform the best:

1. Lowercasing and tokenizing using NLTK’s TwitterTokenizer⁸.
2. Digits and URL normalization.

⁸<https://www.nltk.org/api/nltk.tokenize.html>

3. Low-frequency users have been normalized; high frequency users have been kept, stripping the "@" from the token. Such users included the official Twitter accounts of the companies involved in the mergers (like @askanthem), media (@wsj), official accounts of US politicians (@potus, @thejusticedept, ...)
4. The # signs have been removed from hashtags.

We keep in the vocabulary only tokens occurring at least 3 times, resulting in 19,561 entries considering both healthcare and entertainment industry.

We use `gensim` to extract the TF-IDF vectors from the data⁹, which are used in the TFIDF-MLP model. For the HAN model, following Sun et al. (2018), we use the MPQA subjective lexicon (Wilson et al., 2005) to extract the sentiment word sequences and the Stanford Parser¹⁰ to extract the dependency sequences. We train an SVM model to predict argument labels on Hasan and Ng (2013)'s training data, and we predict the argument sentences for the WT-WT dataset, as discussed in Sun et al. (2018).

⁹<https://radimrehurek.com/gensim/models/tfidfmodel.html>

¹⁰<https://nlp.stanford.edu/software/lex-parser.html>