You can simply rely on communities for a robust characterization of stances

Damián Ariel Furman, Santiago Marro, Cristian Cardellino, Diana Nicoleta Popa, Laura Alonso Alemany

You can simply rely on communities for a robust characterization of stances

Damián Ariel Furman
Universidad de Buenos Aires

Santiago Marro
Université Côte d'Azur

Cristial Cardellino
Universidad Nacional de Córdoba

Diana Nicoleta Popa
Université Grenoble Alpes

Laura Alonso Alemany
Universidad Nacional de Córdoba

Abstract

We show that the structure of communities in social media provides robust information for weakly supervised approaches to assign stances to tweets. Using as seed the SemEval 2016 Stance Detection Task annotated data, we retrieved a high number of topically related tweets. We then propagated information from the manually annotated seed to the retrieved tweets and thus obtained a bigger training corpus.

Classifiers trained with this bigger, weakly supervised dataset reach similar or better performance than those trained with the manually annotated seed. In addition, they are more robust with respect to common manual annotator errors or biases and they have arguably more coverage than smaller datasets.

Weakly supervised approaches, most notably self-supervision, commonly suffer from error propagation. Interestingly, communities seem to provide a structure that constrains error propagation. In particular, weakly supervised classifiers that incorporate community structure are more robust with respect to class imbalance.

Additionally, this is a straightforward, transparent approach, using standard tools and pipelines, cheaper and faster than methods like crowd sourcing for manual annotations. Thus it facilitates adoption, interpretability and accountability.

Introduction

The argumentative stance is the position of a speaker with respect to a given claim in an argument, for example "*vaccination is good for your health*" or "*hate speech against immigrants has very negative social effects*". Argumentation is a very complex phenomenon and automated approaches have achieved only mild success (Lippi and Torroni 2016; Cabrio and Villata 2018). It is specially difficult to achieve good performance in social networks (Mohammad et al. 2016; Schiller, Daxenberger, and Gurevych 2020), where it may be useful, for example to counter hate speech, rumours or fake news (Basile et al. 2019; Pomerleau and Rao 2017; Gorrell et al. 2019).

One of the main shortcomings for the improvement of automated stance detection is that annotated datasets are difficult to obtain. On the one hand, the annotation task is difficult and requires that annotators are trained. As we will show, crowd-sourcing annotation has produced datasets with errors in the Semeval-2016 Stance Detection Dataset. Moreover, being a highly interpretative task, it is prone to high inter-annotator variations. On the other hand, the variability of the phenomenon implies that datasets need to be huge to cover the different forms it may take. Therefore, classical supervised approaches fall short, and it is useful to resort to weak supervision.

In social media, Twitter in particular, the structure of the medium itself provides information that is complementary to the purely linguistic content of tweets. Such information seems to be helpful to characterize argumentation as a complex social phenomenon. As discussed in the following Section, different forms of weak supervision have been applied to the problem of stance detection in Twitter. Most of these approaches combine linguistic properties of tweets with their structural properties in the social network.

The work presented here goes in the same direction, focusing on robustness and interpretability. We apply a very simple approach based on a small seed dataset annotated with intuitive stance classes (*favor*, *against* or *neither*) and standard tools to discover communities in social networks. Each community is assigned to one of these stance classes, and every tweet is assigned the label of the community to which it belongs.

With this method, a big dataset of labeled tweets is obtained. These automatically labeled tweets are used to train a classifier. The high number of tweets provides the classifier with much more coverage than a manually annotated dataset. At the same time, the community structure provides robustness to avoid a common pitfall of weak supervision, that is, error propagation in automatically labeled tweets. Indeed, we show that classifiers trained with these automatically labeled tweets perform better than fully supervised classifiers in the SemEval 2016 Stance Detection task.

Relevant work

Weak supervision has been successfully used to improve stance detection in tweets. The best performing approach in the SemEval 2016 Stance Detection Task, Zarrella and Marsh (2016), integrates word and phrase embeddings specific to Twitter and the topics to be modelled. Domain specific embeddings provide a model of the domain that works as a

back-off model for the smaller model inferred from stance-labeled tweets.

Another approach to the SemEval 2016 task, Misra et al. (2016) retrieved tweets with hashtags that were "*stance-bearing on their own*", and labelled them accordingly. These tweets were then used as labelled examples to train a classifier. However, they didn't obtain as good results as Zarrella and Marsh, thus showing that enhancing the dataset just by the textual content of tweets seems to have the same shortcomings as classical self-learning.

Dias and Becker (2016) participated in SemEval 2016 Task B, a stance detection task where no training data was provided for the domain, but for a similar domain: training data targeted Donald Trump, and test data targeted Hillary Clinton. They applied self-learning based on classical bootstrapping, and, similarly to Misra et al., their results came short of performing as well as others.

Weakly supervised approaches tend to suffer from error propagation. Errors in early stages of the learning process tend to be amplified in the resulting model, specially if there is a class imbalance. That is why successful approaches incorporate strategies to constrain such errors, often by integrating complementary information. In the case of social media, the social structure evidenced by the social graph may be useful to complement textual information.

Pamungkas et al. 2019 (2019) show the effectiveness of integrating different information facets of tweets to detect stance. They address the SemEval-2017 Task 8 to detect the stance of a tweet with respect to another tweet that is bringing forward a rumour. They integrate different aspects of tweets, most notably conversation-based and affective-based features, and outperform the systems with best results at the time of the challenge.

Community network structure has been used as a weakly supervised approach to stance identification in tweets by Ebrahimi et al. (2016), who exploit network structure to propagate labels. However, their approach is complex and computation-intensive, involving complex computational machinery to explicitly model their background knowledge through hinge-loss feature functions.

Fraisier et al. (2018) represent social networks as a graph, detect communities in the graph and exploit community structure to propagate stance labels from users with known stance to users with unknown stance. They show the utility of community-based information to identify stances, but is also computer intensive. Lai et al. (2018) exploit different types of communities, based on retweets, quotes or replies, also showing that communities provide useful information to improve stance detection.

Our approach is simpler than previous work. Instead of performing complex learning approaches, or representing information in complex structures like a graph, we rely on a very simple, standard pipeline. We exploit community structure only to automatically assign stance labels to a huge amount of tweets. Then, those tweets are used to train an off-the-shelf classifier. We show that this very simple approach outperforms the approaches with best results at the Semeval 2016 Stance Detection Task.

Relying on communities to automatically label stances in tweets

Our approach to improve automated stance detection by weak supervision based on communities follows these steps:

1. Obtaining tweets for the target topics We targeted four of the topics of the SemEval 2016 Stance Detection on Twitter dataset: "Legalization of Abortion", "Feminist Movement", "Climate Change Is A Real Concern" and "Atheism". We left out "Hillary Clinton" because the debate around the topic is not contemporaneous anymore and it was difficult to obtain a good corpus. Using similar keywords than in SemEval 2016 for each dataset, we retrieved a total of 613,550 tweets for "Legalization of Abortion", 703,486 tweets for "Climate Change Is A Real Concern", 634,383 tweets for "Feminist Movement" and 336,677 tweets for "Atheism", all in English, of which 162,836, 240,718, 186,158 and 152,176 tweets, respectively, were unique[1]. Tweets were collected using the Twitter API within three different periods of 24hs in November and December 2019 and January 2020, and a period of 72 hs in January 2021 for Atheism. These tweets were used both for community detection and as unlabelled instances for baseline self-learning approaches.

2. Building a graph of users From these tweets, for each topic, we built a non-directed and unweighted graph of users representing each user as a node. Edges between users were built when there was a retweet between them. The final graphs had 282,155 nodes for "Legalization of Abortion", 376,589 for "Climate Change Is A Real Concern", 388,309 for "Feminist Movement" and 79,302 nodes for "Atheism". Then we removed nodes with less than five edges, keeping the single biggest connected subgraph.

3. Detecting Communities To detect communities, we used the Louvain's Algorithm for Modularity Maximization (Gach and Hao 2013). We incremented progressively the time-scale parameter as described in (Lambiotte, Delvenne, and Barahona 2008), which makes the Louvain's algorithm achieve stability with partitions of bigger size, until more than 50% of the nodes were contained in the two biggest communities, kept those and discarded the rest. Note that we were targeting the two biggest communities, because we were assuming binarized positions (*favor* - *against*) in stances, as defined by the Semeval Task, and assumed that *none* stances would not create a community of their own. As will be discussed in the Analysis of Results, this assumption worked for most of the topics but not for Feminism. Another underlying assumption is that users that generate content (with at least 5 retweets from others) on polemic topics have a defined stance and retweet content from people with the same stance.

4. Labelling Communities Our goal was to associate each of the two remaining communities for each topic to a class

[1]The dataset with the unlabelled tweets will be made public at the time of publication

(*against* or *favor*). We used the BERT Supervised Classifier with the best result on the validation dataset from Semeval and classified all tweets in both communities. If the majority of tweets in each community were of different classes then each community was assigned to their majority class. Otherwise, the community with the highest percentage of tweets in its majority class was assigned to it while the other was assigned to the remaining class. As an independent assessment of the stances in the communities, two annotators labelled 50 tweets from each community in each topic, with 20 overlapping tweets per topic, that is, 180 tweets per topic. The Cohen's kappa coefficient for inter-annotator agreement was 0.68. Then, communities were assigned as with automatically labelled tweets. Both manual and automatic approaches assigned communities to the same classes.

5. Training a classifier All tweets in a community were labelled as belonging to the class to which the community had been assigned and were used as training examples for the class. Instances from a different topic were included as examples of the *neither* class, making 20% of the total training instances.A Bert classifier was trained with these labelled tweets.

Experimental settings

Following the evaluation methodology proposed in the SemEval 2016 Stance Detection Task A, we used the average F1 score of *favor* and *against* classes, not considering the *neither* class.

SemEval Manually Labeled Dataset

We worked on four topics of the SemEval 2016 Stance Detection on Twitter dataset: "Legalization of Abortion" (933 tweets, 653 for training), "Feminist Movement" (949 tweets, 664 for training), "Climate Change Is A Real Concern" (564 tweets, 395 for training) and "Atheism" (733 tweets, 513 for training). We used 20% of the training portions as development datasets, for hyperparameter optimization.

In close inspection we found that both in the Feminist Movement and the Abortion datasets, some of the tweets were incorrectly classified. Some examples of this misclassified tweets can be seen in Figure 1.

We relabelled the corpus only on those cases where the misclassification was evident. This was the case for 68 examples (10.2% of the dataset) from Feminist Movement's training portion, 48 examples (16.8% of the dataset) from Feminist Movement's test corpus and 79 examples (28% of the dataset) from Legalization of Abortion's test. Table 1 shows the distribution of the classes after relabelling.

It can be seen that annotator's errors have a bias towards the *against* class regardless of the topic. The newly annotated datasets will be released at the time of publication.

Community-based dataset

For the "Legalization of Abortion" topic, the automatically labelled dataset based on communities has 59,876 tweets, "Climate Change Is A Real Concern" has 60,800, "Feminist Movement" has 44,275 and "Atheism" has 24,920 tweets.

The corpus presents a set of tweets that could be difficult to classify for a manual annotator because of the context information needed to interpret it. As it can be seen in figure 2, this tweet references the campaign's slogans for legalizing abortion in Ireland therefore, it stance is in *favor* of the topic, something that a non-expert annotator may miss.

Fully Supervised baselines

We establish as baselines the performance of fully supervised approaches:

- the baseline proposed by the SemEval task (Mohammad et al. 2016), a Support Vector Machine with tweets as bags of n-grams: 1-, 2-, 3-, 4- and 5-grams taking only the n-grams with a probability higher than 0.0075. An RBF kernel was used.
- fastText linear classifier (Joulin et al. 2016) with learning rate 0.001 and embedding of size 500.
- a logistic regression classifier with L2 norm and a balanced class weight.
- a UMLFIT fine-tuned classifier following (Howard and Ruder 2018), namely being gradual unfreezing, discriminative learning rates, and one-cycle training. The pre-trained model used was the wikitext-103 model.
- BERT (Devlin et al. 2018), encoding each tweet with the 12 layer cased model, with 12 attention heads and embedding of 768. The output embedding for the entire sequence is further passed to a softmax layer to obtain class probabilities. All weights are fine-tuned. The maximum sequence length is set to 128, padding shorter sequences. We use dropout keep of 0.9, a learning rate of 2e-5 10 epochs, with early stopping with a patience of 40 steps.

Self-learning baselines

Self-learning is a very simple approach to incorporate unlabelled data with a model obtained from manually labelled data, so we included some variants in our experiments to better assess the benefits of community structure, by assessing the benefits of merely incorporating unlabelled data, without community structure.

We trained a base classifier with the SemEval training corpus, then applied it to classify the tweets used for the Community approach. From these automatically classified tweets, the 200 instances classified with the highest confidence, keeping the proportion of classes in the training set, were taken and added to the training dataset. Given the baseline nature of self learning approaches, self-learning iterations were limited to 3, but runs were repeated 5 times with different random seeds to assess the stability of results.

We have applied this method with three base classifiers: an SVM (the SemEval baseline), fasText (Joulin et al. 2016) and BERT (Devlin et al. 2018).

Discussion

A summary of results can be seen in Table 2. We can see that the community-based approach beats the performance of the best SemEval-2016 systems (Mohammad et al. 2016) and

Tweet	Target	Semeval	Correction
Those who deny women who've been raped abortion are the same ppl who tell rape victims they asked for it. #rape	Abortion	AGAINST	FAVOR
I hope you all either enjoy the rugby or enjoy not enjoying the rugby	Abortion	AGAINST	NONE
No one has 5he right to tell any person what they should do with their body	Abortion	AGAINST	FAVOR
A bundle of cells feels more pain than a fully grown women? no.	Abortion	AGAINST	FAVOR
Wish I could be at the #rallyforlife counter protest. Women are people. Fetuses are clumps of cells. End of story. #ireland	Abortion	AGAINST	FAVOR
Feminists can TOTALLY wear makeup and don't tell me otherwise. #choices	Feminism	AGAINST	FAVOR
I searched for posts with "feminist" tag and saw how much hatred is against feminism. #misogyny #tumblr #hatred #sadness #tears	Feminism	AGAINST	FAVOR
how many were young women? Were there any black young women?	Feminism	AGAINST	NONE
Ask her father for her hand in marriage. How language reinforces patriarchy. She is his property. And mom? #equality	Feminism	AGAINST	FAVOR

Figure 1: Some examples of tweets erroneously labeled on SemEval 2016 dataset.

	Legalization of Abortion			Climate Change		Atheism		Feminist Movement			
	test		train	test	train	test	train	test		train	
	orig	corr						orig	corr	orig	corr
against	67%	40%	54%	7%	4%	72%	59%	64%	47%	49%	39%
favor	16%	31%	19%	73%	53%	15%	18%	20%	29%	32%	37%
neither	16%	29%	27%	20%	43%	13%	23%	15%	23%	19%	24%

Table 1: Distribution of classes in the original and corrected SemEval manually annotated dataset.

```
Well done for writing it and well done to
all involved in #TogetherForYes #ItsAYes
#RepealedThe8th #abortion
```

Figure 2: Example taken from the "Legalization of Abortion" Favor community

outperforms the rest of approaches except in the Feminist Movement topic. As we will discuss, our underlying assumptions for community building do not seem to hold for this topic, which would explain the bad results.

We also evaluated the community-based approach with the newly retrieved tweets that we manually labelled (described in step 4. of the methodology), and results were consistent with the evaluation based on SemEval, as can be seen below:

Test Set	Average	Abort	Climate	Atheism	Feminism
SemEval	.72	.69	73	.71	.60
Community	.70	.72	.60	.82	.64

Fully supervised or self-learning approaches consistently perform below community-based and the best SemEval approaches, except for Feminism, where fastText outperforms the rest of approaches. As could be expected, embedding-based classifiers like fastText or Bert tend to perform better than those without embeddings. It is noteworthy that there are not important losses in performance in the self-learning, even when half of the examples used to train the resulting classifiers have been automatically annotated. Thus, it seems that weak supervision without constraints for error propagation works quite well for this problem.

When we have a closer inspection of per class precision and recall figures (see Table 3), we can see that all approaches except community-based suffer from sensitivity to class imbalance, systematically labelling instances as belonging to the majority class. This is more acute with bigger class imbalance, as is the case of the Climate Change topic, where only 3,8% of the tweets (15) were labeled as *against* (see Table 1). Indeed, most approaches perform worse in this topic, probably precisely because of oversensitivity to the class imbalance. In contrast, the community-based approach is more robust in that case, probably because communities provide more examples for the minority class. We could hypothesize that those examples are spurious, but in fact we see that performance is not damaged in the SemEval test set, on the contrary, it is one of the best.

The Feminist Movement topic deserves special attention. In this case, the community-based approach does not outperform SemEval best result or the best fully supervised approach. During the process of manual annotation of tweets belonging to communities, we could see that they were defined around stances over subtopics inside the topic of Feminist Movement (e.g. Sander's statements about women not being able to win a presidential campaign) that didn't reflect stances over the topic itself (e.g. people that were both in favor and against Sander's statements stated to be in favor of the "Feminist Movement" and used the Feminism hashtag used to retrieve the corpus). This is arguably the reason why results for this topic are lower with a community-based approach: because the stances proposed by SemEval, *for* and *against*, are not prevalent in the communities. Thus the assumptions underlying the method to build communities (that

	Average	Abortion	Climate	Atheism	Feminism
Best SemEval Approach					
	.68	.66	.55	.67	.62
Semi-Supervised SemEval 2016 Approaches					
Zarrella and Marsh (2016)	.68	.57	.42	.61	.62
Misra et al. (2016)	.59	.62	.42	.57	.49
Fully Supervised					
SVM	.43	.44	.38	.43	.48
LR	.47	.56	.34	.43	.52
UMLFIT	.50	.54	.41	.53	.54
fastText	.56	.59	.41	.49	**.67**
Bert	.57	.68	.44	.69	.60
Self-learning					
SVM	.44	.47	.37	.43	.48
fastText	.51	.59	.42	.49	.52
Bert	.55	.63	.44	.54	.58
Community	**.72**	**.69**	**.73**	**.71**	.60

Table 2: Results of different approaches for fully supervised, self-supervised and community-based to stance detection in three topics of SemEval-2016, with F1 macro average for *favor* and *against* on the corrected SemEval test set. We also include the reference of the F1 for the best SemEval approach for each topic. Boldface is marking the best results.

		Favor		Against	
	F1	Precision	Recall	Precision	Recall
SVM	.38	.8	.7	.0	.0
LR	.34	.8	.6	.0	.0
UMLFIT	.41	.8	.7	.0	.0
fastText	.41	.9	.7	.0	.0
Bert	.44	.9	.9	.0	.0
Community	.73	.8	.9	.5	.7

Table 3: F1, Precision and Recall for some of the approaches to stance detection for the Climate Change domain.

there are two big communities, each representing one of the stances) do not hold and the whole approach is flawed.

Finally, it can be argued that classifiers trained with more examples model the target better. We believe this is the case because results are in a smaller range of values for the performance for community-based approaches in comparison with the rest of approaches. Moreover, we also noticed a significant decrease in the standard deviation of the metrics of performance (F1) when exploring different hyper-parameters from the Bert classifier training with the SemEval data only (**0.18**, **0.12**, **0.14** and **0.10** for Legalization of Abortion, Climate Change, Feminist Movement and Atheism, respectively) to the Bert classifier training with automatically labelled examples from the community-based approach (**0.11**, **0.07**, **0.02** and **0.10** for the same topics).

Conclusion

We have presented a very simple method to enhance automatic stance detection in Twitter, based on standard tools and a transparent pipeline. We augment a small labelled dataset by finding coarse-grained communities in social media and assigning each community to one of the stances, using a classifier trained on the initial dataset. Then, tweets in each of the communities are assigned the corresponding label, and are used to train a model with this bigger corpus. Such approach produces more robust models than using manually labelled tweets alone, with better performance and less bias toward the majority class, even in cases of acute class imbalance, which is one of the important caveats for using machine learning systems for societal problems.

We argue that this approach improves reproducibility, because results are more stable than with less training examples. In addition, this community-based approach is highly scalable to huge sources of data (Louvain's algorithm is O(n*log(n))), updatable at a very low cost, and it can be a good complement to the dilemma of the cost of fully manual annotation, which may produce low quality results if the budget for annotation is tight. Indeed, we believe that low-paid, poorly-trained annotators are liable to produce unreliable annotations. A robust data-driven method like community detection can complement such annotations and help detect human annotation errors or subjectivities.

For this approach to succeed, it is necessary to have adequate assumptions about the number and distribution of stances. We have shown that in the case of the Feminism topic, where communities are not in *favor* or *against* feminism, but in favor or against different subtopics within feminism, the approach does not yield good results.

Future Work

Future work includes assessing how increasing the number of tweets impacts on the performance of this community approach. We will also explore classifiers and embeddings that are tailored to the task, instead of off-the-shelf.

Communities can be also used as a system for manual annotation's double check: all examples where the community doesn't match the annotator's labeling can be re-checked by a third annotator.

This finding opens a very interesting line of research: assessing whether the proposed stances prevail in the universe of tweets being analyzed and what happens with communities when they don't, like in the case of "Feminist Movement". We believe that topics can be characterized from different perspectives that recognize different numbers and types of stances. Communities can help evaluate the definition of a stance regarding a particular corpus and provide valuable information to assess whether a more sophisticated and complex definition is needed.

Different ways for defining how to build graphs using tweet's metadata is also to be explored. We used information about user's retweets following (Lai et al. 2018), who show that users usually retweet from other users having the same stance. Frasier et al(Fraisier et al. 2018) propose a proximity concept that generalizes to generic models involving different criteria based on keywords, references to pieces of information and social relations like retweets or citations. We want to see how different relations between users impact on the building of communities for stance recognition.

References

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Cabrio, E., and Villata, S. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Dias, M., and Becker, K. 2016. INF-UFRGS-OPINION-MINING at SemEval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 378–383. San Diego, California: Association for Computational Linguistics.

Ebrahimi, J.; Dou, D.; and Lowd, D. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1012–1017. Austin, Texas: Association for Computational Linguistics.

Fraisier, O.; Cabanac, G.; Pitarch, Y.; Besançon, R.; and Boughanem, M. 2018. Stance classification through proximity-based community detection. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, 220–228. New York, NY, USA: Association for Computing Machinery.

Gach, O., and Hao, J. 2013. Improving the louvain algorithm for community detection with modularity maximization. In *Artificial Evolution*, volume 8752 of *Lecture Notes in Computer Science*, 145–156. Springer.

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 845–854. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems*, volume 10859 of *Lecture Notes in Computer Science*, 15–27. Springer.

Lambiotte, R.; Delvenne, J.-C.; and Barahona, M. 2008. Laplacian dynamics and multiscale modular structure in networks.

Lippi, M., and Torroni, P. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.* 16(2):10:1–10:25.

Misra, A.; Ecker, B.; Handleman, T.; Hahn, N.; and Walker, M. 2016. NLDS-UCSC at SemEval-2016 task 6: A semi-supervised approach to detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 420–427. San Diego, California: Association for Computational Linguistics.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2019. Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure.

Pomerleau, D., and Rao, D. 2017. Fake news challenge. www.fakenewschallenge.org.

Schiller, B.; Daxenberger, J.; and Gurevych, I. 2020. Stance detection benchmark: How robust is your stance detection?

Zarrella, G., and Marsh, A. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 458–463. San Diego, California: Association for Computational Linguistics.