Stance Detection for Cryptocurrency Forums

Toghrul Abbasli
2020280126
Tsinghua University

Malte Meng
2020280441
Tsinghua University

Aurélien Vu Ngoc
2020280219
Tsinghua University

Abstract

In this work, we address the unique challenges of stance detection: the lack of annotated datasets, and the limited work on the application of Transfer Learning using large pre-trained language models. To tackle the former, we propose two novel datasets: **SEthB** (Stance Ethereum Basic) and **SEthC** (Stance Ethereum Clustering), using data sourced from Reddit, one of the most popular online discussion platforms. The data is selected and annotated based on stance directed towards the Cryptocurrency Ethereum. To do so, two annotation strategies are used, standard manual annotation and a clustering approach that benefits from very active 'super-users'. To address the latter, we implement an experimental environment which allows a side-by-side comparison of the most popular pre-trained Language Models. Using this, we create and evaluate a comparison of the model performances on two benchmark datasets in Stance Detection as well as SEthB and SEthC. This work encourages further usage and investigation of clustering approaches for dataset annotation, and highlights the potential of RoBERTa and XL-Net in a Stance Detection scenario.

1 Introduction

Stance Detection is an important area of research in Natural Language Processing with growing interest from the community. By leveraging large amounts of publicly available data from social media, it can aid in tasks ranging from fake news detection to product decision making and review. Although similar to aspect-based Sentiment Analysis, both the amount of annotated datasets and the application of cutting-edge Machine Learning techniques are still lacking. This project contributes to Stance Detection by addressing these main two challenges over the case of Ethereum.

First we create our own datasets SEthB and SEthC which, in contrast to the majority of datasets in Std (and more generally Sentiment Analysis) which are created off data from Twitter, utilizes comments taken from the popular discusison forum Reddit. Our motivation of doing so derives from Reddit being the most popular area of discussion for Cryptocurrencies and being able to provide guidelines and an initial methodology for creating a Stance Detection dataset for Reddit. Furthermore, our approach to dataset creation promotes a recently explored annotation technique based on clustering (Zotova et al., 2021) which possesses significant potential and requires further exploration in future work.

Second, due to the recent performance of large language models on different domains in Natural Language Processing, the application of Transfer Learning has seen many initial successes in Sentiment Analysis. Therefore we decide to take a step forward in this direction and introduce a new comparative study of multiple PLMs of interest in the particular case StD. By transferring knowledge between models with different domains, we approach state-of-the-art results on popular StD datasets with only but minimal human intervention in data preprocessing, model fine-tuning and task description. Transfer Learning is proven to provide robust training and higher accuracy (Pan and Yang, 2010).

As a result, our work proposes the following contributions to the StD research community:

1. Create two annotated Reddit Stance Detection datasets, SEthB and SEthC, which differ from other existing datasets that are mostly Twitter-based (Zotova et al., 2021). [1]

2. Evaluate and compare the potential of trans-

[1] https://www.kaggle.com/malteadrianmeng/reddit-stance-for-ethereum

fer learning with the popular pre-trained language models BERT (Devlin et al., 2019), Albert (Lan et al., 2020), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2020) and XLNet (Yang et al., 2020).

3. Make our code implementation publicly available and supply the research community with a workable framework of transfer learning for stance detection. [2]

2 Background

Stance Detection (StD) can be defined as an automatic classification of the stance of the producer of a piece of text, toward a target, into one of the three classes {*Favor, Against, Neither*} (Küçük and Can, 2020). It can be considered as a sub problem of Opinion Mining and has a lot of applications areas, such as opinion surveys/polling (Lai et al., 2020; Zotova et al., 2020; Luo et al., 2021), trend and market forecast/analysis, consumer health information (Sen et al., 2018) and stance summarization (Jang and Allan, 2018).

Stance detection closely relates to many hot NLP topics, such as target-dependant sentiment analysis, aspect-based sentiment analysis, emotion recognition, sarcasm/irony detection, rumour detection, fake news detection (Hanselowski et al., 2018; Küçük and Can, 2020) and more. Despite the similarities, they are very distinct. In fact, given a stance target, a single sentence can have a positive sentiment overall while having a clear negative stance (e.g. *"Donald Trump is a piece of trash, but I love my life right now, things are fantastic!"* is clearly *Against* Donald Trump but has a general positive sentiment). Such difference makes sentiment analysis advances hardly transferable to StD problems without adaptation.

As a result, many challenges have to be overcome. First, because Stance Detection is a research area that is still in its early stages, there is no general consensus over the task definition, nor an explicit list of existing datasets, nor a clear benchmark of classification methods. Although Küçük and Can (2020) remarkably reviewed the Stance Detection field, there is still much work to be done.

Second, data scarcity. Annotated data is extremely difficult to find since manual annotation is expensive, time consuming and subjective. Researchers are left with but two options: either find a way to automatically annotate data, like in (Zotova et al., 2021; Darwish et al., 2020), or to find a work around avoiding the use of large amounts of data. Recent advances in Transfer Learning can be helpful, but need further investigation (Ghosh et al., 2019).

3 Dataset Creation

To address the issue of dataset scarcity in the area of Stance Detection, we create two novel datasets using comments posted on the popular forum Reddit. The process of doing so consists of gathering target related comments from Reddit, filtering out comment's that don't include stance, and annotating a sub-set of those filtered comments.

3.1 Collection

Although Reddit provides an API to gather content and meta-data for posts and comments, they impose a limitation on the number of requests that can be made per minute. To get past this, we utilised Pushshift, another RESTful API that provides data through an independently maintained, archive of the site. The benefit of doing so, is also the tools Pushshift provides for data aggregation.

Since gathering relevant data from all sub-reddits would be too inefficient, we narrowed down our sourcing to a few relevant subreddits. The first two subreddits we initially decided to source from were r/CryptoCurrency and r/ethereum. After inspecting sampled comments from these forums however, we worried that a dataset only comprising these subreddits would be heavily biased in favor of Ethereum. To counter this we included r/Buttcoin, a subreddit that consists of jokes and meme's making fun of cryptocurrencies and their investors. In addition to this, we also took comments from r/personalfinance and r/investing, both subreddits that provide advice on investing and seemed likewise skeptical of Cryptocurrencies in general.

3.2 Filtering

After splitting each parsed comment into multiple sentences, we used two methods to filter out sentences that were unlikely to contain a stance that was targeted at Ethereum. The first measure we took was a keyword search. Since the target for our dataset was Ethereum, we used a small list of popular names of the Cryptocurrency: Ethereum, Ether and the acronym ETH. Doing so significantly reduced the size of our data to $242,648$ comments.

[2] `https://github.com/meng2468/reddit-sd`

Although there is no automated way to detect whether a sentence contains stance based on linguistic features, Luo et al. (2021) introduces an approach that excludes sentences that don't contain a stance. Following their example we first generated a dependency tree using `spaCy`'s built-in part-of-speech tagging, then searched through the tree to determine whether or not the sentence contained either a complement or propositional clause. Sentences that didn't were removed. Doing so further reduced the size of our dataset down to $86,980$ samples.

3.3 Annotation

We proposed two strategies for the transparent and efficient annotation of our datasets.

3.3.1 Standard Annotation

The basic strategy of annotation is the one traditionally used in the creation of other datasets, not only in stance detection, but also in sentiment analysis (Conforti et al., 2020; Luo et al., 2021; Sen et al., 2018).

Each project member individually annotated comments that were randomly sampled from the filtered dataset. If the annotator wasn't certain about which stance a comment contained, they were given the option to skip it. This ensured that all final annotations were provided with confidence.

3.3.2 Clustered Annotation

The second approach is based on the assumption that user's stance remains relatively similar over time. This is particularly true for comments related to Cryptocurrencies since authors are often directly concerned about Ethereum because of them investing in it (or not). Instead of directly annotating the stance to individual comments, we annotated authors based on their stance.

For each author, we randomly sample a set number of their comments, if the annotator felt confident in attributing stance to the author, they would. If the annotator was still uncertain, they could generate more samples for that specific author. Using these annotations, we then labeled every comment of that specific author to their stance. As a small number of authors were so called 'super users' with very high activity, doing so would provide us with a far larger number of labeled data with a smaller time investment.

Author	**Comments**
coinfeeds-bot	431
MrMoustacheMan	364
cryptolicious501	347
ccModBot	308
ethereumflow	308
frank__costello	269
MONGSTRADAMUS	243
Nomorealcohol2017	215
BicycleOfLife	198

Table 1: Author activity in filtered data

4 Transformer-Based Language Models

In order to address the second challenge being the lack of comparison between PLMs in a Stance Detection context, we decide to run multiple experiments on BERT, RoBERTa, ALBERT, DistilBERT and XLNet.

4.1 BERT

Introduced by (Devlin et al., 2019), BERT is a popular pre-trained language model which, when released, achieved SoTa on multiple tasks throughout the NLP research field. The Bidirectional Encoder Representation from Transformers model introduces a novel pre-training task of Masked Language Model and Next Structure Prediction.

4.2 RoBERTa

A variant of BERT, RoBERTa introduced by Liu et al. (2019b), is a Robustly Optimised BERT model. Authors increased the size of training data, computational power and improved training methodology to outperform BERT on multiple different tasks. More specifically they introduced a dynamic masking strategy during training so that masked tokens change from one epoch to the other, helping the model better comprehend complex word relations.

4.3 DistilBERT

Aimed at reducing the memory complexity of BERT, (Sanh et al., 2020) worked on DistilBERT to reduce training time and its inference latency. Authors approximate BERT using Model Distillation to achieve 97% of BERT's performance while retaining only half its number of parameters. More specifically, they rely on the assumption that a fully trained (large) language model has output distributions which can be approximated using smaller networks through an optimisation technique based on the Kullback–Leibler divergence.

4.4 ALBERT

ALBERT, developed by Lan et al. (2020), also tries to lighten BERT and speed up the training process by showing evidence that parameter-reduction techniques can greatly lower the memory consumption and training time of BERT while maintaining most of its capabilities. First, factorized embedding parametrization helps reduce the parameter complexity induced when growing the hidden size. Second, cross-layer parameter sharing drastically decreases BERT's number of parameters. The authors achieved comparable results to BERT's, while having a smaller model, at the cost of increasing the computational complexity of training.

4.5 XLNet

XLNet, developed by Yang et al. (2020) belongs to the Autoregressive family of pretrained language models. Introducing an improved training task of Permutation Language Modeling, this bidirectional transformer improves BERT results by using relationships from both directions, thus enhancing its ability to handle dependencies between words.

5 Experimental Setup

In this section, we describe organization of banchmarking for our datasets and Transfer Learning method procedures.

5.1 Benchmark Datasets

As means of comparison to our dataset, we used two popular Stance Detection datasets, SemEval2016Task6 and ARC.

5.1.1 SemEval2016Task6

The SemEval2016Task6 dataset (Mohammad et al., 2016) consists of statements taken from Twitter that provide stance on one of the following topics: Atheism, Climate Change, Feminism, Hillary Clinton, and Abortion. Each tweet is its own datapoint, and the stances for tweet-topic pairs are annotated as Favor, Against or Neither. That dataset consists of a total of $4,060$ entries, which are evenly split between the different stance targets / topics. For our experiments, we utilised the split that was provided in the dataset, which followed a 70/30 train test split. The split also followed this distribution for each individual target.

5.1.2 ARC

Although ARC (Argument Reasoning Corpus) was initially designed for natural language argumentation in (Habernal et al., 2018), we use the version modified by Hanselowski et al. (2018) for Stance Detection. In contrast to the SemEval dataset, the targets that are used consist of headline statements. The sentences that contain stance are commenting these headlines and are labeled as one of the following categories: Agree, Unrelated, Disagree, and Discuss. The dataset has a total of $17,792$ entries, which are split up to roughly 100 entries per target. We use the provided train test split, which followed an 80/20 train test split. In contrast to the prior dataset, 3 target's were not included in the test-split, which we excluded.

5.2 Benchmark Models

As a means to determine whether the Transfer Learning approach with different Pre-trained Language Models is effective, we implemented three standard models to compare with.

5.3 Random- and Majority- Baseline

The first two models are those of a Random- and Majority- Baseline. The Random Baseline, provides a uniformly-distributed guess of one of the dataset labels to each data point in the test-set. The Majority Baseline first determines the target for the specific sentence, then uses the most common label for that class as an annotation.

5.4 Support Vector Machine

As the third method, we implemented the SVM (Support Vector Machine) that was used as the baseline method in (Mohammad et al., 2016). The model consists of a SVM with a linear kernel. The input features we feed into the model are word level 1-gram, 2-gram and 3-gram's, in addition to character level 2-gram, 3-gram and 5-gram's. A different SVM is trained for each of the different targets.

5.5 Transfer Learning

5.5.1 Architecture

We take from the Sentiment Analysis research community (Liu et al., 2019a) as well as from similar attempts of using PLMs on Stance Detection (Schiller et al., 2020; Ghosh et al., 2019; Vamvas and Sennrich, 2020) to create a basic classifier architecture which consists of a PLM with a classifier head, i.e. a single linear layer. Models thus take text data as input, then use the language model outputs to classify between the different stance classes, usually {*Favor, Against, Neither*} depending on the dataset specifications. Moreover models are target-dependent, meaning that each model is dedicated to a single target. Such design is motivated by a concern for simplicity since we only have limited computational resources at our disposal, and also influenced by potential industrial applications. In most industry use-cases for stance detection, companies are often interested in just a few targets relative to their product, objective and mission. Therefore we believe their main concern lies in the model performance rather than in its ability to generalise to other targets.

5.5.2 Training

We use PyTorch and the Transformers library with custom functions to process different datasets, load pretrained language model weights, run training sessions and evalute models. After multiple attempts we found that fine-tuning the whole model without 'freezing the body' works best compared to tuning the classifier only. Each and every trainable layer is thus tuned from the start, including the language models, in order to achieve the best results.
Hyperparameter selection is done partially using preliminary work on the SemEval2016t6 dataset and following general guidelines given by the models' authors.

6 Results & Discussion

6.1 SEthB

Using the standard annotation procedure outlined earlier in Section 3.3.1, we ended up with 774 annotated data points. As expected the number of sentences that displayed a stance favoring Ethereum outnumbered those against 3 to 1.

Despite our filtering operations, around 362 sentences ended up being unrelated to Ethereum. These sentences either mentioned Ethereum but didn't offer stance towards it, or were explanatory and informative in nature (e.g. "*There are several services that will allow you to start staking with just .1 ETH .*").

6.2 SethC

As a result of the clustering annotation method outlined in Section 3.3.2, we got more than 7608 data points. Although the number of datapoints is significantly higher than that of SethB, the dataset is also significantly more imbalanced, with 5403 in favor, 308 against and 1897 as neither. This was definitely within expectations, as intuitively, users that are more active in discussions related to ETH are more likely to favor it.

Furthermore, the increase in datapoints come at a possible cost of a decrease in accuracy. Although we sampled the sentences randomly over a large time-frame, user stance does not necessarily remain unchanged. In addition to this, users that favor Ethereum, can write sentences that, when taken out of context, can be against.

6.3 Model Performance

Investigating various PLMs performance in a Stance Detection context on these datasets led to the following findings.

To begin with, given their convincing performance on GLUE tasks (Liu et al., 2019b; Yang et al., 2020), we expect RoBERTa and XLNet to outperform the others by far. Interestingly, results reported in Table 2 prove otherwise: although RoBERTa has a clear advantage on SemEval2016, outperforming BERT by 1% and XLNet by almost 30%, it is overtaken by XLNet on the ARC dataset. There are many possible explanations to this phenomenon, the first being that state-of-the-art results in Sentiment Analysis are achieved by RoBERTa (Liu et al., 2019a), so such behaviour is to be expected given the similarity between StD and SA. The other explanation relies on data composition difference between SemEval2016 and ARC: tweets have nothing in common with posts taken from a debate forum in terms of complexity. Therefore, thanks to its powerful capability to handle very compound word relationships, XLNet can understand arguments and their connection to the stance target far better than RoBERTa.

Furthermore, we can anticipate the less convincing results of approximate models (DistilBERT) and simplified models (ALBERT) on these StD tasks. Eventually, they achieve decent performance, although not matching RoBERTa's nor BERT's. However ALBERT shows a strange behaviour with extremely poor performance, sometimes even worse than baseline results (on SEthB). Our intuition to this happening is threefold:

1. With reduced number of parameters, its comprehension of word dependencies as complex as stance is diminished

2. Because ALBERT architecture relies a lot on parameter sharing and similar distributions across layers, we suggest that utilizing the output of the last layer has direct impact on the performance and that earlier layers output can yield better results.

3. We believe the training setup, especially hyperparameter configuration, is not optimal for ALBERT and it requires more precise tuning, given its extensively optimised structure.

As for DistilBERT, it is only but normal to observe similar results to BERT since the initial objective of Sanh et al. (2020) is to approximate the layers distribution.

6.4 Future Work

Following our findings in Section 6.3, both XLNet and RoBERTa matched our expectations and outperformed BERT. While currently achieving state-of-the in a range of tasks in NLP, they still haven't done so in Stance Detection. We are convinced that these models yield great potential, and that the StD research community would benefit from looking into them. Further work dedicated to make improvements is thus required: enhancing the hyperparameter search aiming for the best configuration, perform a model-specific optimised fine-tuning in order to overtake the current SoTa system implemented by Schiller et al. (2020).

On the newly created dataset side, careful evaluation of these datsets using dedicated metrics and in-depth analysis can reveal interesting findings that we might have missed. Our datasets are groundbreaking in the sense that they use data collected from Reddit forums about cryptocurrencies and that SEthC uses largely unexplored annotation technique based on clustering. Being the first of their kind, they require we emphasize the evaluation process so that future work will have simple criterions to compare with. In spite of the current results we obtain using baseline approaches, SVM and PLMs which are decent, we must verify more cautiously the SEthC dataset contain information that is valid, insightful and complex enough to let room for future model improvements.

In general, we recommend researchers to invest further efforts into Reddit as it is the most popular user forum in the world. Doing so would make new data extremely accessible, as it is clearly structured into categories called subreddits that each research community can take from for their own purpose (not only Cryptocurrencies but also health issues, political stance, financial advice and more). Additionally the high user activity allows for complex natural language interaction between users, hence the rich potential of collected data.

7 Related Work

Our research work is related to dataset creation and application of novel Transfer Learning techniques for Stance Detection.

7.1 Dataset Creation for StD

In a recent study, stance detection is used to analyze argumentation on Global Warming. Luo et al. (2021) created a hand-annotated dataset sourced from 56k Newspaper Articles, and they used specific APIs to obtain the data using keywords filtering. For annotation, they used Amazon Mechanical Turk (AMT) crowd-sourcing platform by defining the labels as "*agree*", "*neutral*" or "*disagree*". Annotators' agreement is considered for label decision.

Another study proposed the stance classification dataset for Multi-Perspective Consumer Health Information (MPCHI) (Sen et al., 2018). The data is collected by retrieving the first 50 results of different search engines for each of the 5 MPCHI queries. The authors also used a crowdsourced Human Intelligent Task (HIT) by AMT for annotations. Used labels were "*supports*", "*opposes*" or "*neutral*" and an annotator agreement decided the final label annotation. Different from the existing works, we developed a web crawler that retrieves comments from Reddit given the specific keywords and subreddit names.

In (Vamvas and Sennrich, 2020) authors propose X-Stance, a dataset about Swiss politics and elections. The data is extracted from the voting advice application *Smartvote*. Candidates responded and commented on questions in categorical form

Datasets	SemEval2016		ARC		SEthB		SEthC	
Models	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**
Random Baseline	0.3363	0.2910	0.2413	0.1663	0.3355	0.3179	0.3127	0.2526
Majority Baseline	0.5833	0.2377	0.7607	0.3684	0.4581	0.2094	0.7523	0.2862
SVM	0.6765	0.5451	0.8059	**0.4794**	0.5677	**0.4532**	0.7584	0.5385
ALBERT	0.6335	0.3419	0.7737	0.4035	0.4710	0.2135	0.7828	0.4576
BERT	0.7343	0.6366	0.7903	0.4331	**0.6065**	0.4281	0.7907	**0.6139**
DistilBERT	0.7004	0.5958	0.7935	0.4235	0.5806	0.4436	0.7946	0.5932
RoBERTa	**0.7470**	**0.6413**	0.7920	0.4359	0.6000	0.4292	**0.8071**	0.5809
XLNet	0.6887	0.5037	**0.8099**	0.4780	0.5742	0.4371	0.8038	0.5942
SotA	**-**	**0.6979**	**-**	**0.6583**	**0.6065**	**0.4532**	**0.8071**	**0.6139**

Table 2: Benchmark results on popular stance detection datasets (SemEval & ARC) and on ours (SEthB & SEthC)

(yes / rather yes / rather no / no). Then the labels were processed in a way that "yes" and "rather yes" were combined into "*favor*"; "rather no" or "no" into "*against*". Also, the authors compared the performance of their dataset with the MPCHI dataset and benchmarking datasets that we also used for evaluation which we describe next.

ARC (Habernal et al., 2018) is an Argument Reasoning Corpus consists of conversational topics taken from the *Room for Debate* section of the New York Times[3].The authors manually selected 188 debates and crawled all comments from each debate. The annotations were obtained using the AMT crowdsourcing platform and an inter-annotator agreement was also considered.

SemEval2016T6 (Mohammad et al., 2016) is SemEval-2016 [4] dataset for the task of Stance Detection in Tweets. The authors created a hashtag list to query the tweets using Tweeter API. They discarded retweets and tweets with URLs. For the actual annotation, the authors proposed a questionnaire and used the CrowdFlower crowdsourcing platform with the participation of at least eight respondents. Inter-annotator agreement for this dataset is 81.85%.

7.2 Transfer Learning techniques for StD

The benchmarking of different models on Stance Detection datasets is implemented in a recent research work (Schiller et al., 2020). Using transfer learning techniques the authors provide a Stance Detection benchmark of 6 state-of-the-art Pretrained Language Models with baselines. The authors used testing accuracy and F1 score as performance metric. In (Ghosh et al., 2019) authors explore the reproducibility of several existing stance detection models. They analyzed the reproducibility of BERT, Target-Specific Attention NN, LSTM, CNN, and SVM, over two datasets: MCPHI and SemEval2016T6. In our work, we benchmark our proposed datasets and two Stance Detection datasets by implementing novel Transfer Learning techniques with five SoTA models alongside three baselines. They also verify the effectiveness of existing SoTA models.

8 Conclusion

In this work, we introduced two novel datasets: Stance Ethereum Basic (SEthB) and Stance Ethereum Clustering (SEthC). Both of which regard studying stance towards Ethereum in Cryptocurrency related subreddits. SEthB and SEthC contain 774 and about 7608 data points respectively.

We evaluated the performance of multiple SoTA Pretrained Language Models, comparing their performance with baseline models. Also, we benchmarked the proposed datasets with SemEval2016 and ARC datasets. The results of our analysis showed that RoBERTa XLNet are effective PLMs for TF. We also found that the clustering method that we used for SEthC has potential, as it showed good benchmark.

Hence, future work should focus evaluating the quality of datasets created with the clustered approach. We believe our research work is fundamental for preventing spread of hoax news and scam projects in Blockchain field.

[3]https://www.nytimes.com/roomfordebate

[4]SemEval-2016 International Workshop on Semantic Evaluation https://alt.qcri.org/semeval2016/index.php?id=tasks

References

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. STANDER: An Expert-Annotated Dataset for News Stance Detection and Evidence Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101, Online. Association for Computational Linguistics.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised User Stance Detection on Twitter. *arXiv:1904.02000 [cs]*. ArXiv: 1904.02000.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance Detection in Web and Social Media: A Comparative Study. *arXiv:2007.05976 [cs]*, 11696:75–87. ArXiv: 2007.05976.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance Detection Task. *arXiv:1806.05180 [cs]*. ArXiv: 1806.05180.

Myungha Jang and James Allan. 2018. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1221–1224.

Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1):12:1–12:37.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. ArXiv: 1909.11942.

Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. 2019a. A Survey of Sentiment Analysis Based on Transfer Learning. *IEEE Access*, 7:85401–85412. Conference Name: IEEE Access.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2021. Detecting Stance in Media on Global Warming. *arXiv:2010.15149 [cs]*. ArXiv: 2010.15149.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance Detection Benchmark: How Robust Is Your Stance Detection? *arXiv:2001.01565 [cs]*. ArXiv: 2001.01565.

Anirban Sen, Manjira Sinha, Sandya Mannarswamy, and Shourya Roy. 2018. Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '18, pages 273–281, New York, NY, USA. Association for Computing Machinery.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. *arXiv:2003.08385 [cs]*. ArXiv: 2003.08385.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. ArXiv: 1906.08237.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual Stance Detection: The Catalonia Independence Corpus. *arXiv:2004.00050 [cs]*. ArXiv: 2004.00050.

Elena Zotova, Rodrigo Agerri, and German Rigau. 2021. Semi-automatic Generation of Multilingual Datasets for Stance Detection in Twitter. *Expert Systems with Applications*, 170:114547. ArXiv: 2101.11978.