KI - Künstliche Intelligenz (2021) 35:329–341
https://doi.org/10.1007/s13218-021-00714-w

TECHNICAL CONTRIBUTION

Stance Detection Benchmark: How Robust is Your Stance Detection?

Benjamin Schiller[1] · **Johannes Daxenberger**[1] · **Iryna Gurevych**[1]

Received: 25 August 2020 / Accepted: 2 March 2021 / Published online: 26 March 2021

Abstract
Stance detection (StD) aims to detect an author's stance towards a certain topic and has become a key component in applications like fake news detection, claim validation, or argument search. However, while stance is easily detected by humans, machine learning (ML) models are clearly falling short of this task. Given the major differences in dataset sizes and framing of StD (e.g. number of classes and inputs), ML models trained on a single dataset usually generalize poorly to other domains. Hence, we introduce a StD benchmark that allows to compare ML models against a wide variety of heterogeneous StD datasets to evaluate them for generalizability and robustness. Moreover, the framework is designed for easy integration of new datasets and probing methods for robustness. Amongst several baseline models, we define a model that learns from all ten StD datasets of various domains in a multi-dataset learning (MDL) setting and present new state-of-the-art results on five of the datasets. Yet, the models still perform well below human capabilities and even simple perturbations of the original test samples (adversarial attacks) severely hurt the performance of MDL models. Deeper investigation suggests overfitting on dataset biases as the main reason for the decreased robustness. Our analysis emphasizes the need of focus on robustness and de-biasing strategies in multi-task learning approaches. To foster research on this important topic, we release the dataset splits, code, and fine-tuned weights.

Keywords Stance detection · Robustness · Multi-dataset learning

1 Introduction

Stance detection (StD) represents a well-established task in natural language processing and is often described by having two inputs: (1) a topic of a discussion and (2) a comment made by an author. Given these two inputs, the aim is to find out whether the author is in favor or against the topic. For instance, in SemEval-2016 Task 6 [30], the second input is a short tweet and the goal is to detect, whether the author has made a positive or negative comment towards a given controversial topic:

Topic: *Climate Change is a Real Concern*
Tweet: *Gone are the days where we would get temperatures of Min -2 and Max 5 in Cape Town*
Stance: *FAVOR*

While the task has a long tradition in the domain of political and ideological debates [30, 41, 43, 45], in recent years, it has been brought to attention by the uprising debates around fake news, where StD is an important pre-processing step [9, 11, 33], as well as for other downstream tasks like argument search [42] and claim validation [34].

However, while humans are quite capable of assessing correct stances, ML models are often falling short of this task (see Table 1).

As there are numerous domains to which StD can be applied, definitions of this task vary considerably. For instance, the first input can be a short topic, a claim, or may not be given at all. If the first input is not given, the topic or claim has to be inferred from explicit or implicit mentions within the given text. The second input can be another claim, an evidence, or a full argument and may differ in length from a single sentence to a whole document. The number of classes can also vary between 2-class problems (e.g. *for/against*) and more fine-grained 4-class problems (e.g. *comment/support/query/deny*). Moreover, the number of samples varies drasticially between datasets (for our setup: from 2394 to 75,385).

✉ Benjamin Schiller
schiller@ukp.informatik.tu-darmstadt.de
http://www.ukp.tu-darmstadt.de

[1] Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany

Table 1 Inter-annotator agreement (IAA) vs. state-of-the-art results. ARC/FNC-1 in F_1 macro, PERSPECTRUM in F_1 micro

Dataset	State-of-the-art (%)	Agreement (%)
ARC[a] [15]	57.30	77.30
FNC-1 [33]	61.10	75.40
PERSPECTRUM [5]	70.80	90.90

[a]IAA taken from Hanselowski et al. [16]

While these differences are problematic for cross-domain performance, it can also be seen as an advantage, as it concludes in an abundance of datasets from different domains that can be integrated into transfer or multi-task learning approaches. Yet, given the decent human performance on this task, it is hard to grasp why ML models fall short of StD, while they are almost on par for related tasks like Sentiment Analysis[1] and Natural Language Inference[2] (NLI).

Within this work, we provide foundations for answering this question. We empirically assess whether the abundance of differently framed StD datasets from multiple domains can be leveraged by training and evaluating on them collectively in a multi-task fashion. However, as we only have one task but multiple datasets, we henceforth define it as *multi-dataset learning* (MDL). And indeed, our model profits significantly from datasets of the same task via MDL with + 4 percentage points (pp) on average, as well as from related tasks (e.g. NLI or Textual Entailment) via Transfer Learning (TL) with + 3.4pp on average.

We then probe the robustness of our models via three adversarial attacks (perturbations of the original test sets of all StD datasets used) and measure it with a modified version of the *Resilience* score by Thorne et al. [44]. However, while we gain significant performance improvements on the original test sets by using TL and MDL, the expected robustness of these approaches is missing. Surprisingly, the results reveal that MDL models are even less robust than single-dataset learning (SDL) models (i.e. models trained on one dataset only). We investigate this phenomenon through low resource experiments and observe that less training data leads to an improved robustness for the MDL models, narrowing down the gap to the SDL models. We thus assume that lower robustness stems from dataset biases introduced by the vast amount of available training data for the MDL models, leading to overfitting. Consequently, adversarial attacks that cause the data to deviate too much from the learned biases have a more severe impact on these models.

The contributions of this paper are as follows: (1) to the best of our knowledge, within the field of StD we are the first to combine learning from related tasks (via TL) and MDL, designed to capture all facets of StD tasks, and achieve new state-of-the-art results on five of ten datasets. (2) In an in-depth analysis with adversarial attacks, we show that TL and MDL for StD generally improves the performance of ML models, but also drastically reduces their robustness if compared to SDL models. (3) To foster improved analysis of this task, we publish the full benchmark system including model training and evaluation, as well as the means to easily add and evaluate more datasets, adversarial attack sets, and low resource experiments.[3]. All datasets and their fixed splits, the fine-tuned model weights, and the machine translation models[4] can be downloaded for reproducibility.

2 Related Work

Stance detection is a well-established task in natural language processing. Initial work focused on parliamentary debates [43], whereas latest work has shifted to the domain of social media [8], where several shared tasks have been introduced [9, 14, 30]. With the shift in domains, the definition of the task also shifted: more classes were added (e.g. *query* [14] or *unrelated* [33]), the number of inputs has changed (e.g. multiple topics per sample [40]), or the definition of the inputs itself (e.g. tweets [14] or argument components [42]). There also exists a subfield of StD that specializes in classifying the stance towards a given rumour [14, 55]. In addition to stance labels, Sirrianni et al. [39] also predict the intensity of a stance towards a post.

In past years, the problem of StD has become a cornerstone for downstream tasks like fake news detection [33], claim validation [34], and argument search [42]. Yet, recent work mainly focuses on individual datasets and domains. We, in contrast, concentrate on a higher level of abstraction by aggregating datasets of different domains and definitions to analyze them in a holistic way. To do so, we leverage the idea of TL and multi-task learning (in form of MDL), as they have shown to increase performance and robustness [37, 51], as well as significant support in low resource scenarios [38]. Experiments on StD in a multi-task learning setup have been conducted in the past [2]. Latest frameworks for multi-task learning include the MT-DNN [26], which scored state-of-the-art results on the GLUE Benchmark [47]. In contrast to their work, we will use this framework for MDL, i.e. combining only datasets of the same task to analyze whether StD datasets can benefit from each other by transferring knowledge about their domains. Furthermore, we extend the framework with adversarial attacks to probe

[1] http://nlpprogress.com/english/sentiment_analysis.html.

[2] http://nlpprogress.com/english/natural_language_inference.html.

[3] https://github.com/UKPLab/mdl-stance-robustness

[4] Necessary for one of the adversarial attacks.

Table 2 All datasets, grouped by domain and with examples

Dataset	Domain	Topic	Comment	Stance
ibmcs	Encyclopedia	[...] atheism is the only way	Atheism is a superior basis for ethics	PRO
semeval2019t7	Social media	(Charlie Hebdo)	"[...] #CharlieHebdo gunmen have been killed" yayyy [...]	Support
semeval2016t6		Feminist Movement	[...] every women should have their own rights!! #SemST	Favor
fnc1	News	Hugh Hefner Dead?	Hugh Hefner has denied reports that he is dead [...]	Disagree
snopes		Farmers feed their cattle candy [...]	[...] padding out cow feed with waste candy is nothing new.	Agree
scd	Debating forums	(Obama)	I think Obama has been a great President. [...]	For
perspectrum		School Day Should Be Extended	So much easier for parents!	Support
iac1		existence of god	[...] the Bible tells me that Jesus existed [...]	Pro
arc		Salt should have a place at the table	[...] the iodine in salt is necessary to prevent goiter. [...]	Agree
argmin	Web search	school uniforms	We believe in freedom of choice.	CON

Topics in parentheses signal implicit information

Table 3 Splits, classes, and class distributions for all used datasets

Datasets	# samples				
	Train	Dev	Test	Total	Classes
arc [15, 16]	12,382	1851	3559	17,792	Unrelated (75%), disagree (10%), agree (9%), discuss (6%)
argmin [42]	6845	1568	2726	11,139	Argument_against (56%), argument_for (44%)
fnc1 [33]	42,476	7496	25,413	75,385	Unrelated (73%), discuss (18%), agree (7%), disagree (2%)
iac1 [46]	4227	454	924	5605	Pro (56%), anti (34%), other (10%)
ibmcs [3]	935	104	1355	2394	Pro (55%), con (45%)
perspectrum [5]	6978	2071	2773	11,822	Support (52%), undermine (48%)
scd [18]	3251	624	964	4839	For (60%), against (40%)
semeval2016t6 [30]	2497	417	1249	4163	Against (51%), favor (25%), none (24%)
semeval2019t7 [14]	5217	1485	1827	8529	Comment (72%), support (14%), query (7%), deny (7%)
snopes [17]	14,416	1868	3154	19,438	Support (74%), refute (26%)

the robustness of the learned models and to analyze whether performance increases gained through TL and MDL are in accordance with an increased robustness.

Adversarial attacks describe test sets aimed to discover possible weak points of ML models. While much recent work in adversarial attacks aims to break NLI systems and is especially adapted to this problem [13, 29], these stress tests have been applied to several other tasks, e.g. Question-Answering [49], Machine Translation [4], or Fact Checking [1, 44]. Unfortunately, preserving the semantics of a sentence while automatically generating these adversarial attacks is difficult, which is why some works have defined small stress tests manually [19, 27]. As this is time (and resource) consuming, other work has defined heuristics with controllable outcome to modify existing datasets and to preserve the semantics of the data [31]. In contrast to previous work, we adapt and analyze some of these attacks for the task of StD and probe the robustness of our SDL and MDL models.

3 Stance Detection Benchmark: Setup and Experiments

We describe the dataset and models we use for the benchmark, the experimental setting, and the results of our experiments. For all experiments, we use and adapt the MT-DNN framework[5] [26].

3.1 Datasets

We choose ten StD datasets from five different domains to represent a rich environment of different facets of StD. Datasets within one domain may still vary in some attributes like their number of classes or sample sizes. All datasets are shown with an example and their domain in Table 2. In addition, Table 3 displays the split sizes and the class distributions of each dataset. All code to preprocess and split the

[5] https://github.com/namisan/mt-dnn.

datasets is available online.[6] In the following, all datasets are introduced.

arc We take the version of the Argument Reasoning Corpus [15] that was modified for StD [16]. A sample consists of a claim crafted by a crowdworker, a user post from a debating forum, and its respective class label.

argmin The UKP Sentential Argument Mining Corpus [42] originally contains topic-sentence pairs labelled with *argument_for*, *argument_against*, and *no_argument*. We remove all non-arguments and simplify the original split: we train on the data of five topics, develop on the data of one topic, and test on the data of two topics.

fnc1 The Fake News Challenge dataset [33] contains headline-article pairs from news websites. We take the original data without modifying it.

iac1 The Internet Argument Corpus V1 [46] contains topic-post pairs from political debates on internet forums. We generate a new split without intersection of topics between train, development, and test set.

ibmcs The IBM Debater®—Claim Stance Dataset [3] contains topic-claim pairs. The topics are gathered from a debating database, the claims were manually collected from Wikipedia articles. We take the pre-defined train and test split and split an additional 10% off the train set for development.

perspectrum The PERSPECTRUM dataset [5] contains pairs of claims and related perspectives, which were gathered from debating websites. We take the data they defined for the StD task in their work and keep the exact split.

scd The Stance Classification Dataset [18] contains posts about four topics from an online debate forum with all posts being self-labelled by the post's author. The topics are not part of the actual dataset and have to be inferred from explicit or implicit mentions within a post. We generate a new data split by using the data of two topics for training, the data of one topic for development, and the data of the leftover topic for testing.

semeval2016t6 The SemEval-2016 Task 6 dataset [30] contains topic-tweet pairs, where topics are controversial subjects like politicians, Feminism, or Atheism. We adopt the same split as used in the challenge, but add some of the training data to the development split, as it originally only contained 100 samples.

semeval2019t7 The SemEval-2019 Task 7 [14] contains rumours from reddit posts and tweets towards a variety of incidents like the Ferguson Unrest or the Germanwings crash. Similar to the scd dataset, the topics are not part of the actual dataset.

snopes The Snopes corpus [17] contains data from a fact-checking website,[7] documenting (amongst others) rumours, evidence texts gathered by fact-checkers, and the documents from which the evidence originates. Besides labels for automatic fact-checking of the rumours, the corpus also contains stance annotations towards the rumours for some evidence sentences. We extract these pairs of rumours and evidence sentences and generate a new data split.

3.2 Models

We experiment on all datasets in an SDL setup, i.e. training and testing on all datasets individually, and in an MDL setup, i.e. training on all ten StD datasets jointly but testing on their test splits individually, which allows us to report separate scores for each dataset. We use the MT-DNN framework [26], as it provides the means to do both SDL and MDL.

For SDL, we use the BERT transformer architecture introduced by Devlin et al. [10] and add a classification layer on top. For MDL, we also use the BERT architecture and train it in a multi-task learning fashion as introduced by Liu et al. [26]: all ten datasets share the same BERT model and update it jointly at training time, while dataset-specific classification layers are updated for each dataset individually. For both SDL and MDL, a classification layer is either represented by a single dense layer (in case of the single-input datasets scd and semeval2019t7) or by the stochastic answer network [25] (in case of the eight remaining datasets with input pairs), which has been integrated as part of the MT-DNN framework by its authors and performs additional multi-step reasoning on the BERT-encoded input pairs. All datasets are batched and fed through the architecture in a random order.

As initial weights for BERT, we use either the pre-trained BERT (large, uncased) weights [10] or the MT-DNN (large, uncased) weights [26]. The latter uses the BERT weights and is fine-tuned on all datasets of the GLUE Benchmark [47]. By using the MT-DNN weights, we transfer knowledge from all datasets of the GLUE Benchmark to our models, i.e. we apply TL. Henceforth, we use SDL and MDL to define the model architecture, and BERT and MT-DNN to define the pre-trained weights we use to initialize the model. This leaves us with four combinations of models: $\mathbf{BERT}_{SDL}$, $\mathbf{BERT}_{MDL}$, $\mathbf{MT\text{-}DNN}_{SDL}$, and $\mathbf{MT\text{-}DNN}_{MDL}$ (see Fig. 1).

For all experiments in this section, we mainly keep the MT-DNN Framework's [26] default hyperparameters. To fit the experiments onto our hardware, however, we lower the batch size from 32 to 16 and the maximum sequence length of the samples from 512 to 100 (sub-)words. All other

[6] https://github.com/UKPLab/mdl-stance-robustness.

[7] www.snopes.com.

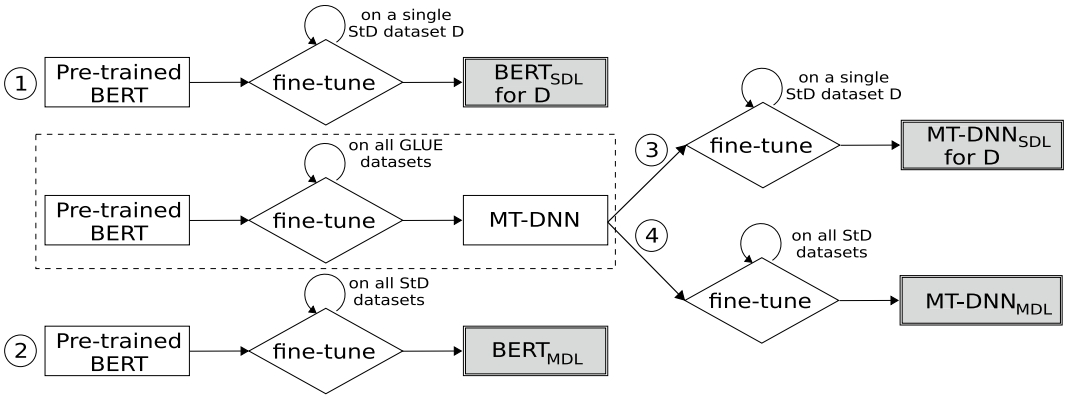


Fig. 1 Overview of the training process. ① represents the training process of the pre-trained BERT model on a single StD dataset, resulting in BERT$_{SDL}$ for that dataset. ② represents the simultaneous training process (MDL) of the pre-trained BERT model on all StD datasets, resulting in BERT$_{MDL}$. ③ and ④ represent the same as ① and ②, respectively, but based on the MT-DNN model. The MT-DNN model was fine-tuned on the datasets of the GLUE benchmark by Liu et al. [26] (shown in the dashed box) and is re-used in this work

hyperparameters are left at the pre-defined default values and are listed in Appendix 1, Table 9. We train all models over 5 different fixed seeds and report the averaged results in F_1 macro (F_1m_+). We run all experiments on a Tesla P-100 with 16 GByte of memory. One epoch with all ten datasets takes around 1.5 h. We use the splits for training and testing as shown in Table 3.

3.3 Results

We report the results of all models and datasets in Table 4. The last column shows the averaged F_1m_+ for a row. We make three observations: (1) TL from related tasks improves the overall performance, (2) MDL with datasets from the same task shows an even larger positive impact, and (3) TL, followed by MDL, can further improve on the individual gains shown by (1) and (2).

We show (1) by comparing the models BERT$_{SDL}$ and MT-DNN$_{SDL}$, where a gain of 3.4 pp due to TL from the GLUE datasets can be observed. While some datasets show a small drop in performance, the performance increases on average. We show (2) by comparing BERT$_{SDL}$ to BERT$_{MDL}$ (+ 4 pp) and MT-DNN$_{SDL}$ to MT-DNN$_{MDL}$ (+ 1.8 pp). The former comparison specifically indicates that learning from similar datasets (i.e. MDL) has a higher impact than TL for StD. The latter comparison shows that, even with TL already applied, MDL can further improve the performance. Lastly, we show (3): combining TL from related tasks (+3.4pp with BERT$_{SDL}$ vs. MT-DNN$_{SDL}$) and MDL on the same task (+4pp with BERT$_{SDL}$ vs. BERT$_{MDL}$) can result in considerable performance gains (+5.1pp with BERT$_{SDL}$ vs. MT-DNN$_{MDL}$). However, as the individual gains from TL and MDL do not add up, it also indicates an information overlap between the datasets of the GLUE benchmark and the StD datasets.

Lastly, while BERT$_{SDL}$ already outperforms five out of six state-of-the-art results, our BERT$_{MDL}$ and MT-DNN$_{MDL}$ are able to add significant performance increases on top.

4 Analysis

As the robustness of an ML model is crucial if applied to other domains or in downstream applications, we analyze this characteristic in more detail. We define adversarial attacks to probe for weaknesses in the models and investigate the reason for observed losses in robustness.

4.1 Adversarial Attacks: Definition

Inspired by stress tests for NLI, we select three adversarial attacks to probe the robustness of the models and modify all samples of all test sets with the following configurations:

Paraphrase Naturally, a model should be able to handle paraphrasing of learned data and, hence, we paraphrase all samples of the test sets. For this, we apply a method by Mallinson et al. [28] and train two machine translation models with OpenNMT [23]: one that translates English originals to German and another one that backtranslates.

Spelling Spelling errors are quite common, especially in data from social media or debating forums. We add two errors into each input of a sample [31]: (1) we swap two letters of a random word and (2) for a different word, we substitute a letter for another letter close to it on the keyboard. We only consider words with at least four letters, as shorter ones are mostly stopwords.

Table 4 Results of experiments on all datasets in F_1m_+ (F_1 macro) and original paper metrics in parentheses (F_1m_- (F_1 micro), Accuracy (Acc), Fake News Challenge score (FNC1), F_1 macro without class *none* ($F_1m_+ \setminus$ none))

Models	arc 12.4k	argmin 6.8k	fnc1 42.5k	iac1 4.2k	ibmcs 0.9k	perspec-trum 7.0k	scd 3.3k	semeval-2016t6 2.5k	semeval-2019t7 5.2k	snopes 14.4k	Avg.
Metrics (original)	F_1m_+	F_1m_+	F_1m_+ (FNC1)	F_1m_+	F_1m_+ (Acc)	F_1m_+ (F_1m_-)	F_1m_+	F_1m_+ ($F_1m_+ \setminus$ none)	F_1m_+	F_1m_+	F_1m_+
Majority baseline	0.2145	0.3383	0.2096 (39.37)	0.2127	0.3406 (0.5166)	0.3466 (0.5305)	0.3530	0.2427 (0.3641)	0.2234	0.4398	0.2921
Random baseline	0.1907	0.4998	0.1815 (32.09)	0.3374	0.4864 (0.4923)	0.5011 (0.5052)	0.4830	0.3061 (0.3769)	0.1804	0.4652	0.3632
State-of-the-art	0.5730[a]	–	0.6110[b] (86.66)[c]	–	(0.5470)[d]	0.7995[e]	–	(0.7104)[f]	**0.6187**[g]	–	–
$BERT_{SDL}$	0.6480	0.6167	0.7466 (88.57)	0.3167	0.5347 (0.5429)	0.8012 (0.8026)	0.5699	0.6839 (0.7018)	0.5364	0.7274	0.6181 ± 0.0158
$MT\text{-}DNN_{SDL}$	0.6324	0.6019	**0.7690 (88.82)**	0.3329	0.7066 (0.7116)	**0.8480 (0.8486)**	0.6211	0.6882 (0.7080)	0.5649	0.7506	0.6516 ± 0.0104
$BERT_{MDL}$	**0.6583**	0.6157	0.7475 (88.60)	0.3781	0.7211 (0.7240)	0.8093 (0.8102)	0.6444	0.6979 (0.7162)	0.5712	0.7414	0.6585 ± 0.0101
$MT\text{-}DNN_{MDL}$	0.6526	**0.6174**	0.7522 (88.85)	**0.3797**	**0.7772 (0.7787)**	0.8374 (0.8383)	**0.6541**	**0.6979 (0.7181)**	0.5732	**0.7532**	**0.6695 ± 0.0102**
Human Performance	0.7730	–	0.7540	–	–	(0.9090)	–	–	–	–	–

Bold numbers indicate the model with the highest performance for the respective dataset

The last column shows average results of SDL and MDL models over all datasets and five different seeds incl. standard deviation

[a]TalosComb [16]

[b]ESIM w/ GRU + Dropout [20]

[c]Ranking-MLP [54]

[d]Unigrams SVM [3]

[e]$BERT_{CONS}$ [35]

[f]TGMN-CR [50]

[g]GPT-based [53]

Negation We use the negation stress test proposed by Naik et al. [31]. They add the tautology "and false is not true" after each sentence, as they suspect that models might be confused by strong negation words like "not". Recently, this assumption was confirmed by Niven and Kao [32] and we assume the same is also valid for our setup. We add a slightly modified tautology "false is not true and" at the beginning of each sentence, since we truncate all inputs to a maximum length of 100 sub-words.

To measure the effectiveness of each adversarial attack $a \in A$, we calculate the potency score introduced by Thorne et al. [44] as the average reduction from a perfect score and across the systems $s \in S$:

$$Potency(a) = c_a \frac{1}{|S|} \sum_{s \in S} (1 - f(s, a)),$$

with c_a representing the transformation correctness from test to adversarial samples and a function f that returns the performance score for a system s on an adversarial attack set a.

The correctness rate c_a is calculated by taking 25 randomly selected samples from all test sets (i.e. 250 samples in total), comparing them to their adversarial counterparts, and finally dividing the number of correctly transformed samples by the total number of checked samples. The paraphrase attack was manually annotated for semantic equivalence between the original texts and their paraphrased counterparts. Due to the high subjectivity of this task, the annotation was conducted by two human annotators; the first author and a postdoctoral researcher with background in natural language processing (not involved in this work). The inter-annotator agreement was computed with Cohen's Kappa [7] and signals "moderate" agreement [24] with $\kappa = 0.47$ (see Appendix 2 for more information about the annotation process), which is comparable to the inter-annotator agreement in Atanasova et al. [1], where claims generated with GPT-2 were annotated for semantic coherence. The percentage of samples annotated as "semantically equivalent" is 48.4% (average of both annotators), resulting in a correctness ratio c_a of 0.484 for the paraphrase attack.

As the changes through the spelling attack are minor and difficult to evaluate for humans (who easily cope with small typos), we apply the Flesch–Kincaid grade level [21], which is a well-established metric for text readability in the area of education and calculated as follows:

$$0.39 \times \frac{\text{total words}}{\text{total sentences}} + 11.8 \times \frac{\text{total syllables}}{\text{total words}} - 15.59,$$

with the outcome corresponding to a U.S. grade level. We compare the Flesch–Kincaid grade level of an original sample and its adversarial counterpart and label it as incorrectly perturbed if the readability of the adversarial sample requires a higher U.S. grade level. Applying this method, the correctness ratio for the spelling attack is 0.584.

For the negation attack samples, we assume a correctness of 100% ($c_a = 1.0$) as the perturbation adds a tautology and the semantics and grammar are preserved.

Table 5 Potency of all adversarial attacks

Method	Raw potency (%)	Correctness ratio	Potency (%)
Spelling	43.3	0.584	25.3
Negation	41.1	1.0	41.1
Paraphrase	38.0	0.484	18.4

Table 6 Influence of adversarial attacks, averaged over all datasets, on the $BERT_{SDL}$ and $MT\text{-}DNN_{MDL}$ model (in F_1m_+ and, in parentheses, relative to the score on the test set)

Test	$BERT_{SDL}$	$MT\text{-}DNN_{MDL}$
	0.6181	0.6695
Spelling	0.5568 (– **9.9%**)	0.5767 (– 13.9%)
Negation	0.5914 (– **4.3%**)	0.5871 (– 12.3%)
Paraphrase	0.6012 (– **2.8%**)	0.6380 (– 4.7%)

Bold numbers indicate the model with the smallest relative performance drop (w.r.t the test set performance) for the respective adversarial attack

4.2 Adversarial Attacks: Results and Discussion

We choose to limit the compared systems to $BERT_{SDL}$ and $MT\text{-}DNN_{MDL}$, as the latter uses both TL from related tasks and MDL, whereas the former uses neither. The potencies for all attack sets are shown in Table 5. The raw potency symbolizes the potential strength of an adversarial attack if the automatic creation of adversarial samples would be without errors (i.e. in case of $c_a = 1.0$). The performance results on the adversarial attack sets for both the SDL and MDL model are shown in Table 6.

The paraphrasing attack has the lowest raw potency of all adversarial sets and the average scores only drop by about 2.8–4.7% as compared to the test set performance. Interestingly, on datasets that turned out to be difficult to paraphrase (semeval2019t7, arc, snopes), the score on the $MT\text{-}DNN_{MDL}$ only drops by about 4.3%, 6.4%, and 6.5% (see Appendix 3, Table 11), which is close to average. This confirms Niven and Kao [32] in that the BERT architecture, despite contextualized word embeddings, also primarily focuses on certain cue words and the semantics of the whole sentence is not the main criterion. With raw potencies of 41.1% and 43.3%, the negation and spelling

attacks have the highest negative influence on both SDL and MDL (4.3–13.9% performance loss). We assume this to be another indicator that the models rely on certain key words and fail if the expected occurrence of these words in the seen samples is changed. This is easy to see for the negation attack, as it adds a strong negation word.

For the spelling attack, we look at the following original sample from the perspectrum dataset:

Claim: *School Day Should Be Extended*
Perspective: *So much easier for parents!*
Predict/Gold: *support/support*

And the same sample as spelling attack:

Claim: *School Day Sohuld Be Ectended*
Perspective: *So much esaier for oarents!*
Predict/Gold: *undermine/support*

Since all words of the original sample are in the vocabulary, Google's sub-word implementation WordPiece [52] does not split the tokens into sub-words. However, this is different for the perturbed sentence, as, for instance, the tokens "esaier" and "oarents" are not in the vocabulary. Hence, we get [esa, ##ier] and [o, ##are, ##nts]. These pieces do not carry the same meaning as before the perturbation and the model has not learned to handle them.

However, the most surprising observation represents the much higher relative drop in scores between the test and adversarial attack sets for MT-DNN$_{MDL}$ as compared to BERT$_{SDL}$. For some datasets, even the absolute F_1m_+ of the MDL model drops below that of the SDL model (see Appendix 3, Table 11). MDL should produce more robust models and support them in handling at least some of these attacks, as some of the datasets originate from Social Media and debating forums, where typos and other errors are quite common. On top of that, the model sees much more samples and should be more robust to paraphrased sentences.

We want to further evaluate the robustness of the two systems and, for this, leverage the resilience measure introduced by Thorne et al. [44]:

$$Resilience(s) = \frac{\sum_{a \in A} c_a \times f(s,a)}{\sum_{a \in A} c_a}.$$

It defines the robustness of a model against all adversarial attacks, scaled by the correctness of the attack sets. Surprisingly, the resilience of the MDL (59.6%) and SDL (58.4%) model are almost on par. The score, however, only considers the absolute performance on the adversarial sets, but not the drop in performance when compared to the test set results. If, for instance, model *A* performs better than model *B* on the same test set, but has a higher drop in performance on the same adversarial set, model *A* should show a lower robustness and thus receive a lower resilience score. As the resilience score does not consider this, we adapt the equation by taking the performance on the test set *t* into account. Moreover, we define the highest possible model performance of 1.0 as a common base and subtract the gained relative score from it:

$$Resilience_{rel}(s) = 1.0 - \left| \frac{\sum_{a \in A} c_a \times (f(s,t) - f(s,a))}{\sum_{a \in A} c_a} \right|.$$

Hence, if the performance differences between the test set and its adversarial sets become smaller, the Resilience$_{rel}$ will increase.

Looking at the results, the Resilience$_{rel}$ signals a much higher relative resilience for the SDL model as compared to the MDL model (see Table 7, "Overall"), which is also in accordance with the relative drops shown on the adversarial attacks (see Table 6). We also calculate the Resilience$_{rel}$ for all adversarial attacks separately and observe that the SDL model outperforms the MDL model in each case. While our experiments show that performance-wise we can benefit from MDL, there is a high risk of drastic loss in robustness, which can cancel out the performance gains or, even worse, may render the model inferior in real-world scenarios.

Table 7 Resilience$_{rel}$ of BERT$_{SDL}$ and MT-DNN$_{MDL}$

Method/model	BERT$_{SDL}$ (%)	MT-DNN$_{MDL}$ (%)
Spelling	**96.4**	94.6
Negation	**97.3**	91.8
Paraphrase	**99.2**	98.5
Overall	**96.6**	92.7

Bold numbers indicate the model with the highest Resilience$_{rel}$ for the respective adversarial attack

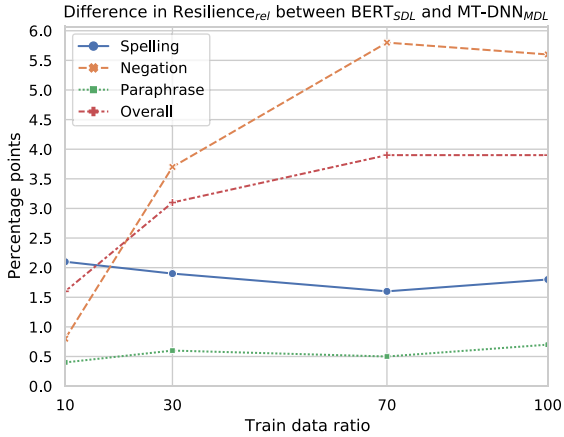
Table 8 Train data ratio performance on the test set in F_1m_+

Model/ratio	10%	30%	70%	100%
MT-*DNN*$_{MDL}$	**0.5855**	**0.6317**	**0.6624**	**0.6695**
Diff.	*− 0.0953*	*− 0.0758*	*− 0.0598*	*− 0.0514*
BERT$_{SDL}$	0.4902	0.5559	0.6026	0.6181

Bold numbers indicate the model with the highest F_1m_+ for the respective train data ratio

4.3 Analysis of Robustness Via Low Resource Experiments

To investigate why the MDL model shows a lower robustness than the SDL models on average, we conduct low resource experiments by training the MDL model and the



(a) Plot with difference ($\text{BERT}_{SDL} - \text{MT-DNN}_{MDL}$) in Resilience_{rel} for all train data ratios and all adversarial attacks separately.

Models / Ratio	10%	30%	70%	100%
MT-DNN_{MDL}	96.9%	94.6%	93.1%	92.7%
BERT_{SDL}	**98.4%**	**97.6%**	**96.9%**	**96.6%**

(b) Resilience_{rel} for both models and all tested train ratios.

Fig. 2 Resilience_{rel} in numbers and plotted over different train data ratios

SDL models on 10, 30, and 70% of the available training data. Development and test sets are kept at 100% of the available data at all times and results are averaged over five seeds.

As is to be expected, the performance gap between BERT_{SDL} and MT-DNN_{MDL} on the test set grows with less training data (see Table 8). Here, the MDL shows its strength in low resource setups [38]. Even more so, while the MDL model showed disencouraging performance with regard to adversarial attacks when trained on 100% of the data, we observe that with less training data, the MT-DNN_{MDL} reduces the difference in Resilience_{rel} to the BERT_{SDL} from 3.9 pp at 100% training data to 1.5 pp at 10% training data (see Fig. 2b). As shown in Fig. 2a, this is mostly due to a lower impact of the negation attack and partly of the paraphrase attack.

Table 9 Hyperparameters for all experiments

Hyperparameter	Value
Batch size	16
Epochs	5
Gradient clipping	max norm: 1.0
Dropout	0.1
Max. sequence length	100
Learning rate	5e−5
Optimizer	Adamax [22]

Our analysis reveals that the amount of training data has a direct negative impact on model robustness. As most (if not all) datasets inevitably inherit the biases of their annotators [12], we assume this negative impact on robustness is due to overfitting on biases in the training data. Hence, less training data leads to less overfitting on these biases, which in turn leads to a higher robustness towards certain attacks that target these biases. For instance, the word “not” in the negation attack can be a bias that adheres to negative class labels [32]. Likewise, an overall shift in the distribution of some words due to the paraphrase attack can interfere with a learned bias. We argue that spelling mistakes are unlikely to be learned as a bias for stance detection classes and the actual reason for the performance drop of the attack is due to the split of ungrammatical tokens into several sub-words (see Sect. 4.2).

5 Discussion

We introduced a StD benchmark system that combines TL and MDL and enables to add and evaluate adversarial attack sets and low resource experiments. We include ten StD datasets of different domains into the benchmark and found the combination of TL and MDL to have a significant positive impact on performance. In five of the ten used datasets, we are able to show new state-of-the-art results. However, our analysis with three adversarial attacks reveals that, contrary to what is expected of TL and MDL, these techniques result in a severe loss of robustness on our StD datasets, with scores often dropping well below SDL performance. We investigate the reasons for this observation by conducting low resource experiments and identify overfitting on biases of vast amounts of training data as a possible issue in our MDL approach.

Reducing the amount of training data for both SDL and MDL models narrows down the robustness anomaly between these two setups, but also lowers the test set performance. Hence, we recommend to develop methods that integrate de-biasing strategies into multi-task learning approaches—for instance, by letting the models learn which samples contain biases and should be penalized or ignored [6] to enhance the robustness, and at the same time being able to leverage more (or all) training data available to maintain the test set performance. Besides de-biasing techniques, in future work, we aim to concentrate on task-specific adversarial attacks and to build defences for the models [36, 48]. We foster the research on StD and model robustness by publishing our benchmark with all dataset splits, models, and experimental code.

Appendix 1: Hyperparameters

We list the important hyperparameters for all experiments in Table 9. With the exception of the batch size and the maximum sequence length of the samples, all hyperparameters are left at the MT-DNN Framework's [26] default values.

Appendix 2: Guidelines: Annotations for the Paraphrase Attack

In order to check the paraphrase attack for correctness, a postdoctoral researcher with background in natural language processing (not involved in this work) and the first author annotated 250 samples from all 10 StD datasets (25 samples per dataset). The annotators checked whether the shown pairs of original and translated samples are semantically equal. Semantic equality was measured binary ("yes" or "no"). The annotation guidelines are as follows:

- The translated sentences need not be grammatically correct, but they have to be comprehensible.
- The meaning of the compared sentences need not be perfectly the same, as different words naturally result in a slightly different meaning.
- Typos or swapped letters in a named entity (e.g. company name, event name, person) should be neglected. If full words or large parts of an entity are replaced, which render it unrecognizable, the sample should be viewed as incorrectly transformed (for example, "Elon Musk" and "Alon Musk" can be seen as a typo whereas "Bill Clinton" and "Invoice Clinton" would be incorrect).
- In case of samples with multiple sentences, semantic equality has to hold for each individual sentence (as the paraphrasing was done on sentence level, the number of sentences for original and translated samples are the same and they must be compared individually).

The following examples were provided:

Example #1
Original: *In particular , school uniforms are often not modest enough in covering the female body to suit Muslims .*
Paraphrased: *In particular, school uniforms are often not modest enough to adapt the female body to Muslims.*
Label: "n"
Reason: "adapt the female body to Muslims" is nonsensical and has not the same meaning as the original sentence.

Table 10 Inter-annotator agreement for the paraphrase attack (on all datasets)

Dataset	Cohen's kappa
arc	0.34
argmin	0.56
fnc1	0.52
iac1	0.48
ibmcs	0.51
perspectrum	0.49
scd	0.49
semeval2016t6	0.47
semeval2019t7	0.45
Snopes	0.43
Total	0.47

Example #2
Original: *Uniforms are certainly easier for administrators to enforce than dress codes .*
Paraphrased: *Uniforms are certainly easier to enforce for administrators than dress codes .*
Label: "y"
Reason: The syntax has changed but not the semantics.

Example #3
Original: *She does n't want to have to wear what everyone else is wearing .*
Paraphrased: *It does not want to carry what all others bear.*
Label: "n"
Reason: The meaning has changed in a way that the paraphrased sentence is talking about "bearing" something (like a burden) and not about wearing a uniform. Also, the personal pronoun has changed and "it" distorts the meaning in that it does not refer to a person anymore.

Example #4
Original: *There 's another thing about uniform though ; even if everybody wears exactly the same , they 're all going to look different , because the same uniform is n't going to suit everybody .*
Paraphrased: *There is another thing about uniform, although even if everyone wears exactly the same, they 'all will look different because the same uniform does not suit everyone.*
Label: "y"
Reason: Some words have changed and a minor typological error has been introduced (apostrophe before "all"), but the meaning is clear and remains the same.

Table 11 Comparison of MT-DNN$_{MDL}$ (all datasets with subscript *MDL*) and BERT$_{SDL}$ (all datasets with subscript *SDL*)

Datasets	arc 12.4k	argmin 6.8k	fnc1 42.5k	iac1 4.2k	ibmcs 0.9k	perspectrum 7.0k	scd 3.3k	seme-val2016t6 2.5k	seme-val2019t7 5.2k	snopes 14.4k	Avg.
Test$_{SDL}$	0.6480	0.6167	0.7466	0.3167	0.5347	0.8012	0.5699	0.6839	0.5364	0.7274	0.6182
Test$_{MDL}$	**0.6526**	**0.6174**	**0.7522**	**0.3797**	**0.7772**	**0.8374**	**0.6541**	**0.6979**	**0.5732**	**0.7532**	**0.6695**
Negation$_{SDL}$	**0.6463**	**0.6205**	**0.7233**	0.3055	0.5365	**0.7854**	**0.5962**	**0.6799**	**0.4266**	**0.5942**	**0.5914** (**−4.3%**)
Negation$_{MDL}$	0.6398	0.5832	0.7017	**0.3424**	**0.6841**	0.7497	0.5901	0.6550	0.3358	0.5896	0.5871 (−12.3%)
Spelling$_{SDL}$	0.4767	**0.5863**	0.6988	**0.3492**	0.4980	0.6665	0.5886	0.5034	0.5092	0.6912	0.5568 (**−9.9%**)
Spelling$_{MDL}$	**0.4973**	0.5403	**0.7046**	0.3311	**0.6412**	**0.6796**	**0.6348**	**0.5049**	**0.5197**	**0.7132**	**0.5767** (−13.9%)
Paraphrase$_{SDL}$	0.6043	**0.6097**	0.7019	0.3477	0.5320	0.7649	0.5735	0.6589	0.5137	**0.7049**	0.6012 (**−2.8%**)
Paraphrase$_{MDL}$	**0.6110**	0.6031	**0.7093**	**0.3682**	**0.7489**	**0.7941**	**0.6321**	**0.6611**	**0.5483**	0.7040	**0.6380** (−4.7%)

All absolute scores are in F_1 macro

Bold numbers indicate the model with the highest performance for the respective dataset and test set, as well as the model with the highest average performance

The inter-annotator agreement (computed with Cohen's kappa [7]) between the annotators is 0.47, which signals "moderate" agreement [24]. This is comparable to the inter-annotator agreement in Atanasova et al. [1], where claims generated with GPT-2 were annotated for semantic coherence. Table 10 shows the Cohen's kappa for each dataset separately.

Appendix 3: Adversarial Attacks on Stance Detection Models

Table 11 shows the absolute performance scores of models MT-DNN$_{MDL}$ (all datasets with subscript *MDL*) and BERT$_{SDL}$ (all datasets with subscript *SDL*). All absolute scores are in F_1 macro. The numbers in parentheses in the *Avg.* column represent the relative drop to the respective score on the test set. Bold numbers in a column represent the best score between the *MDL* and *SDL* on an adversarial attack set.

Acknowledgements We thank Chris Stahlhut for his role as an annotator for the paraphrase attack (Sect. 4.1). This work has been supported by the German Research Foundation within the project "Open Argument Mining" (GU 798/25-1), associated with the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999), and by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Atanasova P, Wright D, Augenstein I (2020) Generating label cohesive and well-formed adversarial claims. In: EMNLP'20, Online. pp 3168–3177. https://doi.org/10.18653/v1/2020.emnlp-main.256
2. Augenstein I, Ruder S, Søgaard A (2018) Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In: NAACL'18, Volume 1 (Long Papers), New Orleans, Louisiana. pp 1896–1906. https://doi.org/10.18653/v1/N18-1172
3. Bar-Haim R, Bhattacharya I, Dinuzzo F, Saha A, Slonim N (2017) Stance classification of context-dependent claims. In: EACL'17.

pp 251–261. https://www.aclweb.org/anthology/E17-1024. Accessed 10 Mar 2021
4. Belinkov Y, Bisk Y (2017) Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv:171102173
5. Chen S, Khashabi D, Yin W, Callison-Burch C, Roth D (2019) Seeing things from a different angle:discovering diverse perspectives about claims. In: NAACL'19. pp 542–557. http://dx.doi.org/10.18653/v1/N19-1053
6. Clark C, Yatskar M, Zettlemoyer L (2019) Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: EMNLP-IJCNLP'19. pp 4067–4080. https://doi.org/10.18653/v1/D19-1418
7. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104
8. Conforti C, Berndt J, Pilehvar MT, Giannitsarou C, Toxvaerd F, Collier N (2020) Will-they-won't-they: a very large dataset for stance detection on Twitter. In: ACL'20, Online. pp 1715–1724. https://doi.org/10.18653/v1/2020.acl-main.157
9. Derczynski L, Bontcheva K, Liakata M, Procter R, Hoi GWS, Zubiaga A (2017) Semeval-2017 task 8: Rumoureval: determining rumour veracity and support for rumours. arXiv preprint arXiv:170405972
10. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805
11. Ferreira W, Vlachos A (2016) Emergent: a novel data-set for stance classification. In: NAACL'16. pp 1163–1168. https://doi.org/10.18653/v1/N16-1138
12. Geva M, Goldberg Y, Berant J (2019) Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In: EMNLP-IJCNLP'19. pp 1161–1166. https://doi.org/10.18653/v1/D19-1107
13. Glockner M, Shwartz V, Goldberg Y (2018) Breaking nli systems with sentences that require simple lexical inferences. In: ACL'18. Short papers, vol 2, pp 650–655. https://doi.org/10.18653/v1/P18-2103
14. Gorrell G, Aker A, Bontcheva K, Derczynski L, Kochkina E, Liakata M, Zubiaga A (2019) Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours. In: SemEval-2019, pp 845–854. https://doi.org/10.18653/v1/S19-2147
15. Habernal I, Wachsmuth H, Gurevych I, Stein B (2018) The argument reasoning comprehension task: identification and reconstruction of implicit warrants. In: NAACL'18, pp 1930–1940. https://doi.org/10.18653/v1/N18-1175
16. Hanselowski A, PVS A, Schiller B, Caspelherr F, Chaudhuri D, Meyer CM, Gurevych I (2018) A retrospective analysis of the fake news challenge stance-detection task. In: COLING'18. pp 1859–1874. https://www.aclweb.org/anthology/C18-1158. Accessed 10 Mar 2021
17. Hanselowski A, Stab C, Schulz C, Li Z, Gurevych I (2019) A richly annotated corpus for different tasks in automated fact-checking. In: CoNLL'19, pp 493–503. https://doi.org/10.18653/v1/K19-1046
18. Hasan KS, Ng V (2013) Stance classification of ideological debates: data, models, features, and constraints. In: IJCNLP'13, pp 1348–1356. https://www.aclweb.org/anthology/I13-1191. Accessed 10 Mar 2021
19. Isabelle P, Cherry C, Foster G (2017) A challenge set approach to evaluating machine translation. arXiv preprint arXiv:170407431
20. Jiang Y (2019) Using machine learning for stance detection. Master's thesis, The University of Texas at Austin. https://repositories.lib.utexas.edu/handle/2152/72801. Accessed 10 Mar 2021
21. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. University of Central Florida, Institute for Simulation and Training. 56. https://stars.library.ucf.edu/istlibrary/56. Accessed 10 Mar 2021
22. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. arXiv:1412.6980
23. Klein G, Kim Y, Deng Y, Senellart J, Rush AM (2017) OpenNMT: open-source toolkit for neural machine translation. In: ACL'17. pp 67–72. https://doi.org/10.18653/v1/P17-4012
24. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. http://www.jstor.org/stable/2529310. Accessed 10 Mar 2021
25. Liu X, Shen Y, Duh K, Gao J (2018) Stochastic answer networks for machine reading comprehension. In: ACL'18, Volume 1: long papers, pp 1694–1704. https://doi.org/10.18653/v1/P18-1157
26. Liu X, He P, Chen W, Gao J (2019) Multi-task deep neural networks for natural language understanding. In: ACL'19, pp 4487–4496. https://doi.org/10.18653/v1/P19-1441
27. Mahler T, Cheung W, Elsner M, King D, de Marneffe MC, Shain C, Stevens-Guille S, White M (2017) Breaking NLP: using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. EMNLP'17, pp 33–39. https://doi.org/10.18653/v1/W17-5405
28. Mallinson J, Sennrich R, Lapata M (2017) Paraphrasing revisited with neural machine translation. In: EACL'17, pp 881–893. https://www.aclweb.org/anthology/E17-1083. Accessed 10 Mar 2021
29. Minervini P, Riedel S (2018) Adversarially regularising neural nli models to integrate logical background knowledge. In: CoNLL'18, pp 65–74. https://doi.org/10.18653/v1/K18-1007
30. Mohammad S, Kiritchenko S, Sobhani P, Zhu X, Cherry C (2016) Semeval-2016 task 6: detecting stance in tweets. In: SemEval-2016, pp 31–41. https://doi.org/10.18653/v1/S16-1003
31. Naik A, Ravichander A, Sadeh N, Rose C, Neubig G (2018) Stress test evaluation for natural language inference. In: COLING'18. pp 2340–2353. https://www.aclweb.org/anthology/C18-1198. Accessed 10 Mar 2021
32. Niven T, Kao HY (2019) Probing neural network comprehension of natural language arguments. In: ACL'19. pp 4658–4664. https://doi.org/10.18653/v1/P19-1459
33. Pomerleau D, Rao D (2017) The fake news challenge: exploring how artificial intelligence technologies could be leveraged to combat fake news. http://www.fakenewschallenge.org/. Accessed 06 Jan 2020
34. Popat K, Mukherjee S, Strötgen J, Weikum G (2017) Where the truth lies: explaining the credibility of emerging claims on the web and social media. In: WWW'17, pp 1003–1012. https://doi.org/10.1145/3041021.3055133
35. Popat K, Mukherjee S, Yates A, Weikum G (2019) STANCY: stance classification based on consistency cues. In: EMNLP-IJCNLP'19, pp 6412–6417. https://doi.org/10.18653/v1/D19-1675
36. Pruthi D, Dhingra B, Lipton ZC (2019) Combating adversarial misspellings with robust word recognition. In: ACL'19, pp 5582–5591. https://doi.org/10.18653/v1/P19-1561
37. Ruder S (2017) An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:170605098
38. Schulz C, Eger S, Daxenberger J, Kahse T, Gurevych I (2018) Multi-task learning for argumentation mining in low-resource settings. In: NAACL'18, pp 35–41. https://doi.org/10.18653/v1/N18-2006
39. Sirrianni J, Liu X, Adams D (2020) Agreement prediction of arguments in cyber argumentation for detecting stance polarity and intensity. In: ACL'20, Online, pp 5746–5758. https://doi.org/10.18653/v1/2020.acl-main.509

40. Sobhani P, Inkpen D, Zhu X (2017) A dataset for multi-target stance detection. In: EACL'17. pp 551–557. https://www.aclweb.org/anthology/E17-2088. Accessed 10 Mar 2021
41. Somasundaran S, Wiebe J (2010) Recognizing stances in ideological on-line debates. In: NAACL-HLT'10, pp 116–124. https://www.aclweb.org/anthology/W10-0214. Accessed 10 Mar 2021
42. Stab C, Miller T, Schiller B, Rai P, Gurevych I (2018) Cross-topic argument mining from heterogeneous sources. In: EMNLP'18, pp 3664–3674. https://doi.org/10.18653/v1/D18-1402
43. Thomas M, Pang B, Lee L (2006) Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: EMNLP'06, pp 327–335. https://www.aclweb.org/anthology/W06-1639. Accessed 10 Mar 2021
44. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2019) Evaluating adversarial attacks against multiple fact verification systems. In: EMNLP-IJCNLP'19. pp 2944–2953. https://doi.org/10.18653/v1/D19-1292
45. Walker MA, Anand P, Abbott R, Tree JEF, Martell C, King J (2012a) That is your evidence? Classifying stance in online political debate. Decis Support Syst 53(4):719–729. https://doi.org/10.1016/j.dss.2012.05.032
46. Walker MA, Tree JEF, Anand P, Abbott R, King J (2012b) A corpus for research on deliberation and debate. In: LREC'12, pp 812–817. https://www.aclweb.org/anthology/L12-1643/. Accessed 10 Mar 2021
47. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: EMNLP'18 Workshop BlackboxNLP. pp 353–355. https://doi.org/10.18653/v1/W18-5446
48. Wang D, Li C, Wen S, Xiang Y, Zhou W, Nepal S (2018) Defensive collaborative multi-task training-defending against adversarial attack towards deep neural networks. arXiv preprint arXiv:180305123
49. Wang Y, Bansal M (2018) Robust machine comprehension models via adversarial training. In: NAACL'18, pp 575–581. https://doi.org/10.18653/v1/N18-2091
50. Wei P, Mao W, Zeng D (2018) A target-guided neural memory model for stance detection in twitter. In: IJCNN'18, pp 1–8. https://doi.org/10.1109/IJCNN.2018.8489665
51. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3(1):9. https://doi.org/10.1186/s40537-016-0043-6
52. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144
53. Yang R, Xie W, Liu C, Yu D (2019) BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In: SemEval-2019, pp 1090–1096. https://doi.org/10.18653/v1/S19-2191
54. Zhang Q, Yilmaz E, Liang S (2018) Ranking-based method for news stance detection. In: WWW'18, pp 41–42. https://doi.org/10.1145/3184558.3186919
55. Zubiaga A, Kochkina E, Liakata M, Procter R, Lukasik M, Bontcheva K, Cohn T, Augenstein I (2018) Discourse-aware rumour stance classification in social media using sequential classifiers. Inf Process Manag 54(2):273–290. https://doi.org/10.1016/j.ipm.2017.11.009