Neural Computing and Applications
https://doi.org/10.1007/s00521-024-09896-4

ORIGINAL ARTICLE

Is checkworthiness generalizable? Evaluating task and domain generalization of datasets for claim detection

Sami Nenno[1,2]

Received: 23 November 2023 / Accepted: 26 April 2024

Abstract
The spread of misinformation has reached a level at which neither research nor fact-checkers can monitor it only manually anymore. Accordingly, there has been much research on models and datasets for detecting checkworthy claims. However, the research in NLP is mostly detached from findings in communication science on misinformation and fact-checking. Checkworthiness is a notoriously vague concept whose meaning is contested among different stakeholders. Against the background of news value theory, i.e., the study of factors that make an event relevant for journalistic reporting, this is not surprising. It is argued that this vagueness leads to inconsistencies and poor generalization across different datasets and domains. For the experiments, models are trained on one dataset, tested on the remaining, and evaluated against the results on the original performance, against a random baseline, and against the scores when the models are not trained at all. The study finds that there is a drastic reduction in comparison with the performance on the original dataset. Moreover, often the models are outperformed by the random baseline and training on one dataset has no or even a negative impact on the performance on the other datasets. This paper proposes that future research should abandon this task design and instead take inspiration from research in communication science. In the style of news values, Claim Detection should focus on factors that are relevant for fact-checkers and misinformation.

Keywords NLP · Misinformation · Claim detection · Fact-checking

1 Introduction

Misinformation is a major topic in news and research for years now [1], and it is widely recognized as problematic for democratic processes and institutions. Fact-checking is one of the most popular ways to tackle this problem, and research suggests that it is a successful method, too [2]. However, the spread of misinformation has reached a level at which it cannot be monitored only manually anymore. Accordingly, there is a need for computational tools that are specifically tailored to this task [3, 4]. Research in NLP tackled the automation of fact-checking from different perspectives [5]. But when asked, fact-checking practitioners find Claim Detection the most useful subtask [6, 7] and this is also mirrored in the research efforts on it [8]. Claim Detection is the task of retrieving claims that are relevant for fact-checking. It aims at reducing the workload of fact-checkers by providing them with a selection of claims that are checkworthy. This paper argues that while Claim Detection is an important task, current conceptions of checkworthiness render approaches to Claim Detection unrealistic for real-world applications.

For decades, communication science and journalism studies have investigated the factors according to which journalists choose the events they report on [9]. These news values are well documented and empirically researched. Examples are geographical proximity, negativity, prominence, impact, or timeliness. They all contribute to the likelihood of an event being reported about. Moreover, in recent times, scholars have investigated if there are news values that are especially prevalent to misinformation or fact-checking [3, 10]. Yet, this research is entirely detached

✉ Sami Nenno
sami.nenno@hiig.de

[1] ZeMKI, Centre for Media, Communication and Information Research, University of Bremen, Linzer Str. 4, 28359 Bremen, Germany

[2] AI & Society Lab, Humboldt Institute for Internet and Society, Französische Str. 9, 10117 Berlin, Germany

Published online: 13 May 2024

from NLP research on Claim Detection. In this line of research, checkworthiness is understood as a single abstract criterion, for example, as "claims that are interesting to the general public" [11] or "claims that should be checked by a professional fact-checker" [12]. However, beside of lacking an empirical grounding, these definitions of checkworthiness face practical obstacles. Different fact-checking organizations have different selection criteria and do not share a common definition of checkworthiness [13, 14]. But if checkworthiness is not the same across organizations, how can models that are trained on datasets that are annotated according to these abstract definitions be deployed by more than one organization?

This study is motivated by two aims: On the one side, it will be shown that due to these inconsistencies in the concept of checkworthiness, models that are trained on one Claim Detection dataset do not generalize well to others. On the other side, the tension within the concept of checkworthiness is picked up and used for recommendations for future research on Claim Detection. In particular, it is hypothesized that for different data annotation projects different understandings of checkworthiness were implicitly applied. This is not due to laziness or incompetence but because checkworthiness is a contested concept and different actors value different criteria for the selection of misinformation. It is further assumed that because the existing datasets for Claim Detection are labeled according to different implicit rules, models that are trained on the one dataset perform poorly on other datasets for the same task. This is empirically tested in a series of experiments. However, this paper is not only meant as criticism. Instead of abandoning Claim Detection, it is possible to leverage the findings on common characteristics of misinformation and selection criteria of fact-checkers or journalists more general. This paper draws a novel connection between news values and Claim Detection and concludes with possible pathways for how research on Claim Detection can proceed without assuming a unique conception of checkworthiness. This is a significant step toward making Claim Detection match the workflow of fact-checkers and to make it applicable in real-world scenarios.

The contributions are as follows:

- A connection between NLP research on Claim Detection and research in communication science on misinformation and fact-checking is drawn.
- An in-depth analysis of how language models generalize across different datasets and domains for this task is performed.
- An alternative formulation of the Claim Detection task that circumvent the highlighted shortcomings is proposed.

2 Background

2.1 Misinformation and fact-checking

Disinformation is understood as false information that is spread with the intention to deceive and cause harm, while misinformation is false information for which this intention does not matter. Information can be false in many different ways: It can be an entirely false statistic, a misleading statement, an image that is faked, or a video that is presented in a false context. In the following, we focus on misinformation because identifying the intention behind a claim is difficult and often not relevant to automation. Moreover, as fact-checking tends to move away from verifying statements by politicians to debunking content that is posted by anonymous sources on social-media platforms, the intention is often not a relevant selection criterion [15]. However, misinformation is a broader term than disinformation and subsumes cases with malicious intent, too.

Misinformation is a major topic in news and research for years now [1]. And while it is not as far reaching as often depicted [16], it is widely recognized as problematic for democratic processes and institutions. Even misinformation on trivial sounding topics like bed bugs can lead to severe public concern and action.[1] Moreover, citizens across different nations show strong concern for misinformation on topics of major importance like climate change [17].

Fact-checking is, next to calls for stricter regulation of online platforms and more investment into media literacy, one of the most popular and frequently discussed approaches to tackle misinformation. Currently, Duke reporter's lab counts 425 active outlets and reports that for some years as many as 77 new fact-checking institutions were founded.[2] And while it is no magic bullet, it often helps people to find orientation in the online information environment and to reduce misperceptions [2, 18]. There have been many studies that focus on the use and need of computational tools for fact-checkers, and they find that fact-checkers would like to have tools that are customized to their specific requirements [3, 4, 7]. The need for computational tools is mostly owed to the enormous amount of content that has to be analyzed by fact-checkers. Manual monitoring reaches limits and is difficult to scale. This is also reflected in the news coverage of automated fact-checking [19]. Increasing the quantity of claims that can be checked is a major theme. But it is also emphasized that increased automation allows journalists to focus more

[1] https://www.lemonde.fr/en/france/article/2024/03/01/bedbug-panic-was-stoked-by-russia-says-france_6575870_7.html.

[2] https://reporterslab.org/category/fact-checking/.

on human-driven practices. In sum, there are several reasons why automating (parts of) the fact-checking pipeline is a worthwhile endeavor.

2.2 Automated fact-checking

Beside of generic software tools that have been adapted by fact-checkers (e.g., geolocation or flight-tracking), they already make use of custom tools, as well. For example, a trends tool that shows instances in which other media outlets have mentioned a particular statement or a monitoring tool to keep track of previously verified claims and corresponding fact-check reports [3]. However, NLP research on automated fact-checking even goes above these applications.

Automated fact-checking roughly mirrors the workflow of human fact-checkers [20]: First, a claim is retrieved, then it is matched against a database of previously checked claims, in order to prevent duplication of work, next evidence for or against the claim is retrieved and a verdict and explanation is derived. Analogously, Guo et al. [5] model the automated pipeline as starting with Claim Detection [21, 22], followed by retrieving previously checked claims [23], and finally verdict prediction and optionally explanation generation [24]. Beside of automating parts of the pipeline, there has also been other work, for example, automatically generating ClaimReview files for making fact-checking websites more accessible [25]. Most research in NLP is dedicated to Claim Detection, followed by claim verification and evidence retrieval [8]. This trend is mirrored in the fact-checking practice. While fact-checkers have reservations toward automated claim verification [7], they see a need for Claim Detection and some systems are already finding application in real-world scenarios [3].

2.3 Claim detection

Claim detection is the task to identify individual statements in large corproa of text. The aim is to decrease the workload of fact-checkers by providing them with a list of relevant claims that can then be verified by the journalists. But Claim Detection has not only been approached from the perspective of automated fact-checking. In argument mining Claim Detection matters, too [26–28]. However, due to different aims, this line of Claim Detection is often inconsistent with the same task for fact-checking. For example, in argument mining, a claim is often understood as something that would count as a statement of opinion in the context of fact-checking. This is a problem because for fact-checking statements of opinion are usually not relevant [3].

With regard to automated fact-checking, the central concept to Claim Detection is *checkworthiness*. The most prominent definition of checkworthiness comes from Arslan et al. [11], who define checkworthy claims as "factual claims that the general public will be interested in learning about their veracity." Firoj et al. [12] asked the annotators to label a sentence as checkworthy if they could affirm the following question:" Do you think that a professional fact-checker should verify the claim in the tweet?" (see Table 1).[3] Checkworthy Claim Detection is usually modeled as a binary classification task. There are exceptions which introduce a third class [11] or which model it as a rating task [13]; however, often the task is also released as classification or the classes are projected to a binary format in later publications [21].

A claim is usually understood as a single sentence and in some cases as a set of sentences, for instance in the form of a tweet. Existing datasets vary strongly in their size and range from less than thousand data points to more than 45,000. Often, the label distribution is strongly imbalanced. This is sometimes due to the nature of misinformation: Even though there is much misinformation online and elsewhere, the vast majority of claims do not carry misinformation. In other cases, the imbalance is due to the annotation strategy. Some researchers created datasets by matching sentences drawn from US-presidential debates with articles from fact-checking websites. Each sentence with a corresponding article is labeled as checkworthy, while the remainder is labeled as not checkworthy. This procedure has the advantage of being cheap in terms of labor cost for annotation. However, there are disadvantages to this method. Because fact-checkers usually do not have the resources to check every claim they consider checkworthy, only a small share of the dataset is labeled as such. For example, `CT21` consists of 45,121 negative instances and only 498 positives. And for the same reason, there is a high false negative rate, too.

The currently best performance for Claim Detection has been achieved with a transformer architecture and adversarial training [21]. They reach an F_1 score of.91. Deep Neural Networks have been used for the task before. Jha et al. [34] experiment with CNNs and LSTMs. However, other approaches have also used more traditional architectures like support vector machines in combination with manual feature engineering [13].

2.4 Critique on claim detection

There has been critique on Claim Detection, too. This critique does not focus on the performance of the models,

[3] We will focus on unimodal claims. In recent time, there has been approaches to multimodal Claim Detection, too [29]. However, to limit the scope of this article, we will not focus on this line of research.

Table 1 Datasets for claim detection

Corpus	Rows	Label	Best score	Source
Claimbuster [11]	23,533	CW*	F_1:.91	US-Presidential Debates
CT19 [30]	23,501	CW***	MAP:.17	US-Presidential Debates
CT20 [31]	962	CW**	MAP:.81	Tweets on COVID-19
CT21 [32]	45,619	CW***	MAP:.40	US-Presidential Debates
CT22 [12]	2891	CW**	F_1:.70	Tweets on COVID-19
TATHYA [33]	15,735	CW***	F_1:.26	US-Presidential Debates
Claimrank [13]	7787	CW***	MAP:.43	US-Presidential Debates
IndianClaims [34]	953	CW*	F_1:.70	Indian political Debates

CW is short for checkworthiness. CW* means that checkworthiness is understood as "being of interest to the general public"

CW** means "should be checked by a professional fact-checker," and datasets with CW*** retrieved the label by matching actual fact-checking articles

which is, as mentioned before, going as high as.91 in F_1. Critique on the task mostly focuses on its formulation and design. Different arguments have been brought forward but they all follow a similar line of reasoning: Claim Detection in its current form is not designed in a way that is applicable to real-world fact-checking. Konstantinovskiy et al. [22] point out that determining the importance of a claim is an editorial judgment that is best left to human fact-checkers. Allein and Moens [35] argue that checkworthiness is knowledge-dependent and varies with regard to preexisting knowledge of the annotator. They argue that because of this and other reasons, checkworthiness should be abandoned entirely.

One can add that checkworthiness is not only knowledge- but also value-dependent. What is considered checkworthy depends on the values and ideology of an individual and even different fact-checking organizations have different agendas. For example, Gencheva et al. [13] scraped different fact-checking websites and labeled sentences from US-presidential debates as checkworthy if there was a corresponding fact-checking article. They report that 880 sentences of their corpus were checked by at least one organization, only 388 sentences were checked by at least two organizations, and only one sentence was checked by all nine organizations. Lim [14] compared fact-checks by two organizations of statements made by candidates of the 2016 US-presidential elections. Of 1178 fact-checks in total, only 77 were fact-checked by both organizations. Inconsistencies across different organizations are also often highlighted in interviews with fact-checkers [3, 36]. This indicates that there is no unique definition of checkworthiness that is shared across different institutions, which makes it a normative contested concept.

One conclusion that has been drawn from this critique is to abandon the concept of checkworthiness altogether and focus only on factual or checkable claims, instead [22, 37–39]. These approaches model Claim Detection as the task to classify claims that are factual and/or can be checked. This includes, for example, claims to truth, claims containing external evidence (links, quotes as opposed to internal evidence like personal experience), or causal and statistical claims. However, the major drawback of these approaches is that they lack a criterion of prioritization. Fact-checkers are not interested in just any claim to truth of factual claim. The claim must also have importance to them and the public discourse. Without a criterion for rating a claim's relevance, the resulting selection is too large to decrease the workload of human fact-checkers to a sufficient degree.

2.5 Research objective

The crucial problem of checkworthiness is that it is vague and that there is no definition or understanding on which all actors agree. The core hypothesis of this study is that this has practical impact on the datasets for Claim Detection: Because there is no unique understanding of checkworthiness, datasets for Claim Detection follow the same annotation scheme only in name. Even though, they are all labeled for checkworthiness, the understanding and operationalization are different for the individual datasets. In order to support our hypothesis, we perform tests across datasets. The basic idea is to train models on one dataset and test them on the others. According to the hypothesis, the performance should strongly deteriorate. This is because the datasets are labeled according to different logics, i.e., to different understandings of checkworthiness. In particular, following questions are answered:

- Is the performance of a model higher on the dataset for Claim Detection that it is trained on than on another dataset for Claim Detection?

- Are the predictions of a model that is trained on one Claim Detection dataset more than just guessing on another dataset for Claim Detection?
- Does training on one Claim Detection dataset lower a model's performance on another compared to when not trained at all?

According to the hypothesis of this study, the first question should be answered positively: Model performance does reduce on different datasets. The second question asks if there is any improvement at all and the third if there might even be a negative impact. Assuming that checkworthiness means different things across different datasets, it can be expected that training on one dataset leads to no improvement or even to a lower performance on other datasets. Note that there is a limitation to cross-dataset evaluation. Even if the performance deteriorates, can it be attributed to the conceptual flaws of checkworthiness or is it due to some other (unknown) factor? Methods to answer this question will be explained in the next sections, and it is further discussed in the section on limitations.

3 Method

3.1 Data

Cross-dataset experiments are performed to answer the research questions. To the best of the authors' knowledge, all datasets for checkworthy Claim Detection for the purpose of automated fact-checking are included (Table 1). Since `TATHYA` is not publicly accessible, it is not included. Furthermore, `CT20` and `IndianClaims (ic)` are used only for testing but not for training as they are very small. Moreover, `multifc` [40] is added for testing, even though it was not designed for Claim Detection. `Multifc` consists of claims that were scraped from multiple fact-checking organizations, and it was designed for the automated verification of claims. Even though, it is usually not used for Claim Detection, it is valuable in the present context. Since it consists of real-world claims that were checked by fact-checkers, models for Claim Detection must perform well on it, if they are supposed to work under real-world conditions.

The idea behind cross-dataset evaluation is as follows: For the experiments, a model is trained on a dataset for Claim Detection. This dataset is called the *source*. As usual in machine learning, the source is split into train and test data and the model is fit to the train data and evaluated on the test data. In a second step, the model is—without further training—tested on one or many other datasets for Claim Detection. These datasets are called the *targets*. Evaluation on the target happens on the entire dataset and not just the test split.

3.2 Models and hyperparameters

A broad array of model types and architectures are chosen. As transformer models are known for their state of the art performance in NLP in general and Claim Detection in particular, three different models are used: BERT [41] and RoBERTa [42] in their base and large version and Bloom[4] with 1.7B and 3B parameters. Moreover, as Claim Detection has also been approached with more traditional models, some of them are added, as well: logistic regression and SVM. For the embeddings, the average of the GloVe [43] embeddings of each word of the sentence was used. All transformer models were retrieved from `Huggingface`. For the other models, the `scikit-learn` implementation was used.

The performance of machine learning models strongly depends on the choice of hyperparameters. Beside of manual configurations, there are many algorithms for hyperparameter optimization (HPO). In this study, HPO for the transformer models was performed with population-based learning [44] as implemented in the `ray` library. Population-based learning is an algorithm similar to the family of evolutionary algorithms: A population of models with different hyperparameter settings is trained in parallel. If a model in the population is under-performing, it exploits the rest of the population is replaced by a better performing model and updated hyperparameters. With this strategy, computational resources are focused on the hyperparameter space that has most chance of producing good results. The algorithm was used to optimize learning rate, weight decay, and batch size of the transformer models. For logistic regression and SVM, the simpler random search as implemented in the `scikit-learn` library was used.[5]

3.3 Baselines

In order to answer each research question, three baselines for the evaluation are constructed. Is the performance of a model higher on the dataset for Claim Detection that it is trained on than on another dataset for Claim Detection? To answer this question, performance is measured against the *source baseline*: The source baseline is understood as the difference in the performance on source and target. For example, a model is trained on `cb` and tested on (a) the test

[4] https://huggingface.co/bigscience/bloom.

[5] Code and more details on the training procedure and hyperparameter search spaces can be found on Github: https://github.com/SamiNenno/Domain-Adaptation-of-Claim-Detection/tree/main.

split of cb and (b) cr. The source baseline result is the difference in the performance metrics of (a) and (b).

Are the predictions of a model that is trained on one Claim Detection dataset more than just guessing on another dataset for Claim Detection? To answer this question, a *random baseline* is constructed. This is done by randomly choosing labels for each example in each dataset. In other words, the random baseline simulates guessing the labels instead of predicting them. For example, a model is trained on cb and tested on cr. For the random baseline, the labels for the cr are chosen randomly. The random baseline result is the difference between the model performance and the score that is achieved by the random labels.

Does training on one Claim Detection dataset lower a model's performance on another compared to when not trained at all? To answer this question, the *zero baseline* is constructed. For the zero baseline models with and without training are deployed. For example, a model A is trained on cb and tested on cr. Another model B (of the same architecture) is not trained at all and tested on cr. The zero baseline result is the difference between the performances of A and B.

Note that all results are averaged over tenfold stratified cross-validation. In other words, for each of the baselines, the tests were conducted with 10 different training (and test) splits of the source dataset and the baseline results are the average differences. The F_1 averaged across all target datasets and all models is reported. See Appendix 7 for details on the results on individual datasets and models.

4 Results

The absolute scores of the models when trained and tested on the same dataset are displayed in Table 2. Models performed best when trained on cb. Due to the imbalanced class distribution, scores on ct19 and ct21 are close to zero. Models performed moderately on ct22 and cr. Previous research shows that it is possible to improve the performance on these datasets. Meng et al. [21] reach an F_1 of.91 on cb and Firoj et al. [12] an F_1 of.70 on ct22. However, as this study focuses on the relative performance across datasets, no further improvement on these scores was pursued.

Table 3 displays the results when trained and tested on different datasets and compared to the baselines. For almost all datasets, the performance on the source is above that on the target. The only exception occurs when ct21 is the source. Models trained on ct21 perform on average.04 higher on the target. However, this is because the original performance is already very low. The decrease from source to target is the strongest for ct22 (.29) and cb (.27).

In most of the cases, the random baseline surpasses the target performance. The average performance on the target datasets is up to.27 lower than the random baseline. The only exception is cb as source. Models trained on cb and tested on the targets outperformed the random baseline by.14 points on average.

Only for cb, training on the source leads to significant improvement on the target. For ct19 and ct21, models performed on average worse when being trained than when not being trained. For ct22 there is only a small and for cr no improvement.

Table 2 Absolute F_1 scores including random and zero baseline

Model	cb	ct19	ct21	ct22	cr
LogReg	0.58	0.04	0.00	0.48	0.21
SVM	0.65	0.13	0.06	0.58	0.40
DistilBERT	0.76	0.26	0.02	0.54	0.42
BERT-B	0.76	0.25	**0.09**	0.56	**0.44**
BERT-L	0.76	0.00	0.00	0.60	0.40
Roberta-B	**0.78**	**0.27**	0.00	0.61	**0.44**
Roberta-L	0.69	0.00	0.00	**0.65**	0.22
Bloom1.7b	0.63	0.00	0.00	0.46	0.11
Bloom3b	0.64	0.00	0.00	0.47	0.10
Average	0.69	0.10	0.02	0.55	0.30
Random	0.32	0.05	0.02	0.30	0.22
Zero	0.24	0.04	0.02	0.20	0.16

Bold numbers indicate highest scores

4.1 In-domain errors

Many datasets are drawn from US-presidential debates and share identical sentences. One would assume that model performance is stronger if source and target are both from the same domain. However, this is not the case. One likely explanation is that there are inconsistencies in the labeling of the different datasets. Even though, they all use

Table 3 Cross-dataset experiments

	Target	Source	Random	Zero
cb	.42	+.27	–.14	–.19
ct19	.09	+.02	+.24	+.18
ct21	.06	–.04	+.27	+.21
ct22	.26	+.29	+.03	–.02
cr	.25	+.05	+.05	.00

F_1 scores for the experiments averaged over models and datasets. Target: Absolute scores when trained on source and tested on target. Source, Random, Zero: Relative difference compared to Target. Positive values indicate that the respective baseline outperforms the Target, negative values show that the scores on the Target are higher than on the baselines

Table 4 Overlap between pairs of datasets and χ_2 results

D1	D2	Sentence overlap	Label overlap	Chi	p	Effect size
cb	cr	2680	.79	449.497	$<.001^*$	.41
cb	ct19	2684	.76	165.561	$<.001^*$	.248
cb	ct21	2973	.76	177.665	$<.001^*$	.244
cr	ct19	6587	.83	2.901	$= .088$	.021
cr	ct21	6394	.83	3.989	$= .046^*$	.025
ct19	ct21	10,907	.96	1.189	$= .276$	.01

Sentence overlap indicates the number of sentences that are identical in both datasets. Label overlap indicates the share of labels that these identical sentences have in common. Effect size is measured as Cramer's V. Asterisk indicates statistical significance at $\alpha = .05$

checkworthiness by name, they mean different things, and accordingly, the models learn wrong correlations between the input and the label. This assumption is further supported by the fact that sometimes training on the source does not only have no or just a small effect on the performance on the target but even a negative impact.

Since there is a limited amount of US-Presidential debates, many of the datasets that are drawn from them share identical sentences. Table 4 displays the number of overlapping sentences between each of them. The Table also shows how many of the overlapping sentences share identical labels. If the labeling was consistent across all datasets, one would expect that identical sentences also have identical labels. However, in many cases, this is not the case. For instance, for the Claimbuster dataset, the overlapping sentences with `ct19`, `ct21`, and `cr` only share 76–79% of their labels. This means that about a quarter is differently annotated. Due to this, the models learn correlations that are flawed with regard to the target datasets. χ_2 tests were performed to test if there is a statistically significant association between the datasets and the labels for identical sentences. In 4 cases, there is a significant association ($p < .05$). This supports the hypothesis that annotators from different labeling projects systematically applied different understandings of checkworthiness to the datasets. This explains part of the poor generalization from one dataset to another.

4.2 Out-of-domain errors

For datasets from different domains, it is more difficult to show that label inconsistencies are the major problem because domain shifts can cause similar reductions in performance. LISA [45] is applied to remove domain-specific spurious correlations from the data. Spurious correlations are understood as features that are correlated with a label within dataset A but not within B. These correlations weaken domain generalization. LISA performs a linear interpolation between training samples. Given samples (x_i, y_i) and (x_j, y_j), whereas $y_i \neq y_j$, and an interpolation ratio $\lambda \in [0, 1]$, mix-up is applied[6]:

$$x_{mix} = \lambda x_i + (1 - \lambda) x_j$$
$$y_{mix} = \lambda y_i + (1 - \lambda) y_j$$

In other words, LISA mixes examples and labels and thereby creates hybrid forms of them. The assumption is that by mixing features of examples with different labels, spurious correlations become associated not only with the one label but also with the other. For the present experiments, this is important in order to find out how much the loss in performance is due to spurious correlations and how much is due to label inconsistency.

The implementation by the authors was followed and a BERT model was used. Only `cb` and `ct22` were used as source datasets since the other datasets are either from the same domain or too small for training.

LISA improves the scores on the target domains in all settings (Table 5). For `cb` as source, the improvement with regard to the non-LISA performance on the targets ranges from.01 to.03. For `ct22`, the improvement is stronger and reaches up to.33 on `multifc`. However, with the exception for `cb` as source and `ic` as target, the performance on the target still reduces strongly when compared to the performance on the source. Furthermore, the strong gains due to LISA happened to the scores that were very low before. In sum, augmentation with LISA did not lead to strong domain generalization. This indicates that label inconsistency has a negative impact here as well.

5 Limitations

Explaining the concrete sources for poor generalization across datasets is difficult because they can be diverse. Data augmentation with LISA was performed in order to

[6] The authors propose two versions of LISA: intra-label and intra-domain. For our experiments, we only used intra-domain LISA.

Table 5 LISA scores for `cb` and `ct22` as source

	cb				ct22		
Target domain	ct20	ct22	ic	multifc	cb	ic	multifc
Absolute scores	0.66	0.36	0.82	0.65	0.38	0.35	0.72
Improvement	0.03	0.01	0.01	0.03	0.05	0.16	0.33
Source baseline	−0.07	−0.37	0.10	−0.12	−0.11	−0.15	−0.08
Random baseline	0.24	0.06	0.32	0.15	0.06	−0.16	0.22
Zero baseline	0.35	0.16	0.43	0.09	0.14	−0.05	0.17

Absolute scores represent the absolute performance on the target when LISA is applied. Improvement denotes the increase when compared to training without LISA. The remaining rows display the relative increase/decrease when compared to the three baseline and if LISA is applied

filter the effect of domain shifts and to identify and isolate the contribution of label inconsistencies to the weak performance. However, eliminating one error source does not necessarily mean that the other is the only cause. There might be other unknown error sources and which also contribute to the poor generalization. Moreover, even though LISA outperforms many other methods for improving domain generalization, it is not perfect and does not always filter all spurious correlations.

The authors acknowledge that it is not fully possible to isolate label inconsistency as a source for errors. However, while being not the only cause, it is likely the main cause for weak generalization across datasets.

6 Discussion

6.1 Checkworthiness generalizes poorly across domains

Claim Detection is meant to assist fact-checkers by spotting claims that potentially carry misinformation and require verification. Most approaches to Claim Detection do this by classifying claims as checkworthy or not. However, it was argued that checkworthiness is an inapt concept for this task. This is not only true on a normative level but also with regard to performance.

The experiments showed that existing datasets for checkworthy Claim Detection are inconsistently labeled and models fail to generalize to other datasets and new domains. It was found that RQ1) high scores on one dataset are not representative for the general performance as they reduce strongly on other datasets, RQ2) random guessing on the target labels is often as good or even better than training on the source, and RQ3) training on the source often has little or even negative impact on the performance on the target.

It was argued that this is not only due to domain shifts but also because of inconsistencies between the labels of the datasets: Even though, all these datasets are designed for checkworthiness, this is only in name. For different datasets, there are different understandings of checkworthiness and accordingly models do not generalize well. This does not come as a surprise, as checkworthiness is a knowledge- and value-dependent concept, which makes it highly subjective and contested.

6.2 From checkworthiness to newsworthiness

Checkworthiness serves as a criterion for prioritization in order to limit the selection of claims, but at the same time, it is value- and knowledge-dependent, untransparent, and contested among fact-checkers. The solution to this dilemma is to redesign checkworthiness such that it serves as a criterion to prioritize claims but without evoking the aforementioned criticism. It is argued that research in communication science and adjacent fields on misinformation and fact-checking provides pathways for future research on Claim Detection.

Based on interviews with fact-checkers, Micallef et al. [3] find that factors, such as virality, timeliness, and importance are relevant for the selection of claims. Humprecht [46] finds that certain topics are more prevalent in fact-checks than others. Tandoc et al. [10] show that the prevalence of timeliness, negativity, and prominence is common to misinformation. In a survey, Damstra et al. [47] list content features of misinformation, like having an ideological bias in favor of the right, being provocative of negative emotions (anger, fear), containing little verifiable information, or making use of fully packed and sensationalist headlines. Other research has shown that misinformation often displays linguistic features, for example capitalization, the use of pronouns, or lexical diversity, that are different from real news [48].

Instead of relying on an abstract notion of checkworthiness that cannot be customized for different organizations, it is possible to understand checkworthiness as a set of criteria that are important to misinformation and fact-

checkers. For this version of checkworthiness, it is no problem that different organizations do not agree on a definition because the different definitions can be regarded as different subsets of these criteria.

In communication science, there is a role model for this understanding of checkworthiness: news values [49]. News values are a set of criteria that make an event "newsworthy," i.e., worthy of being published as news. In this aspect, the concept of news values is very similar to checkworthiness. The concept dates back to [9]. They came up with 12 criteria for news selection, e.g., cultural proximity or unexpectedness. Subsequent research has built on these factors and augmented and criticized them [50, 51]. The key difference to checkworthiness is that news values are more nuanced and empirically grounded. Instead of relying on an abstract notion of what is "interesting to the general public," news values break down this notion into individual features that can be empirically investigated. This improves transparency, which benefits engagement and acceptance of the resulting fact-check [52, 53].

6.3 Future research on claim detection

We recommend that future research on Claim Detection focuses on news values. These can be classical news values but also news values that are particular to fact-checking and misinformation.

There is already research on automatically detecting news values [54–56] on which further research can build. Piotrkowicz et al. [55], for example, classify news values like, proximity, prominence, or uniqueness in newspaper headlines and reach competitive results. Future research should focus on combining factual claim detection [22, 37] and news value detection. The result could be a classifier that aims at checkworthy claims but avoids the aforementioned criticism of checkworthiness.

Appendix

Tables 6, 7, 8 display the relative increase/decrease with respect to the three baselines. All tables display the average (weighted) F_1 score over tenfold stratified cross-validation.

Table 6 Source Baseline

Source	cb	ct19	ct21	ct22	cr	ct20	ic	multifc	Avg
cb	–	−0.54	−0.63	−0.35	−0.34	−0.06	0.12	−0.07	−0.27
ct19	−0.00	–	−0.06	−0.04	−0.04	−0.02	0.03	−0.00	−0.02
ct21	0.04	0.01	–	0.02	0.05	0.05	0.06	0.05	0.04
ct22	−0.22	−0.47	−0.51	–	−0.28	−0.00	−0.36	−0.16	−0.29
cr	0.05	−0.17	−0.25	−0.06	–	0.02	0.03	0.03	−0.05

Relative increase/decrease of target-scores compared to the original scores on the source

Table 7 Random Baseline

Source	cb	ct19	ct21	ct22	cr	ct20	ic	multifc	Avg
cb	–	0.11	0.04	0.05	0.13	0.21	0.31	0.12	0.14
ct19	−0.22	–	0.02	−0.23	−0.16	−0.33	−0.37	−0.40	−0.24
ct21	−0.26	−0.02	–	−0.26	−0.16	−0.35	−0.42	−0.43	−0.27
ct22	0.01	0.03	0.02	–	0.05	0.12	−0.31	−0.11	−0.03
cr	0.03	0.09	0.03	−0.06	–	−0.10	−0.17	−0.16	−0.05

Relative increase/decrease of target-scores compared to random guessing with equal probability for each label

Table 8 Zero Baseline

Source	cb	ct19	ct21	ct22	cr	ct20	ic	multifc	Avg	Avg (noMulti)
cb	–	0.12	0.05	0.15	0.19	0.32	0.41	0.06	0.19	0.21
ct19	−0.14	–	0.02	−0.14	−0.10	−0.22	−0.26	−0.46	−0.18	−0.14
ct21	−0.18	−0.01	–	−0.16	−0.10	−0.24	−0.32	−0.49	−0.21	−0.17
ct22	0.09	0.04	0.02	–	0.11	0.24	−0.21	−0.17	0.02	0.05
cr	0.11	0.10	0.04	0.04	–	0.02	−0.07	−0.22	0.00	0.04

Relative increase/decrease of target-scores when models are not trained compared to when they are trained

The only exception is `multifc`. As it consists exclusively of positive examples, we document recall, instead of F_1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets are accessible via their original publications. The code used for the experiments can be accessed here: https://github.com/SamiNenno/Domain-Adaptation-of-Claim-Detection/tree/main.

Declarations

Conflict of interest There are no conflict of interest.

Ethical approval Claim Detection is a less critical endeavor than automating the actual verification step of fact-checking. Nevertheless, it aims at the automation of important parts of online content moderation. This means that not only technical but also societal aspects have to be considered. Therefore, Claim Detection requires accountability. It is part of our argument that checkworthiness is not transparent, which is a precondition for accountability. Our alternative, news values, is meant to fill this gap. By applying a more nuanced and empirically grounded concept, we hope that it becomes more transparent to explain individual choices and general rules according to which Claim Detection works. Furthermore, as news values are modular, i.e., individual criteria can be applied without the others, we hope that our approach can be customized to different cultural contexts.

References

1. Tandoc EC, Lim ZW, Ling R (2017) Defining fake news. Digital J 6(2):137–153. https://doi.org/10.1080/21670811.2017.1360143
2. Nyhan B, Porter E, Reifler J, Wood TJ (2020) Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. Political Behav 42(3):939–960. https://doi.org/10.1007/s11109-019-09528-x
3. Micallef N, Armacost V, Memon N, Patil S (2022) True or false: studying the work practices of professional fact-checkers. Proc ACM Hum Comput Interact 6(CSCW1):1–44. https://doi.org/10.1145/3512974
4. McClure Haughey M, Muralikumar MD, Wood CA, Starbird K (2020) On the misinformation beat: understanding the work of investigative journalists reporting on problematic information online. Proc ACM Hum Comput Interact 4(CSCW2):133–113322. https://doi.org/10.1145/3415204
5. Guo Z, Schlichtkrull M (2022) A survey on automated fact-checking. Trans Assoc Comput Linguist 10:178–206. https://doi.org/10.1162/tacl_a_00454
6. Arnold P The challenges of online fact checking. Technical report, Full Fact, London (2020)
7. Nakov P, Corney D, Hasanain M, Alam F, Elsayed T, Barrón-Cedeño A, Papotti P, Shaar S, Da San Martino G (2021) Automated fact-checking for assisting human fact-checkers. In: Proceedings of the thirtieth international joint conference on artificial intelligence, pp. 4551–4558. International joint conferences on artificial intelligence organization, Montreal, Canadahttps://doi.org/10.24963/ijcai.2021/619
8. Dierickx L, Lindén C-G, Opdahl A (2023) Automated fact-checking to support professional practices: systematic literature review and meta-analysis. Int J Commun 17:21
9. Galtung J, Ruge MH (1965) The structure of foreign news: the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. J Peace Res 2(1):64–90. https://doi.org/10.1177/002234336500200104
10. Tandoc EC, Thomas RJ, Bishop L (2021) What is (fake) news? Analyzing news values (and more) in fake stories. Media Commun 9(1):110–119
11. Arslan F, Hassan N, Li C (2020) A benchmark dataset of check-worthy factual claims. Proc Int AAAI Conf Web Social Media 14:821–829
12. Firoj A, Dalvi F, Shaar S, Durrani N, Mubarak H, Nikolov A, Martino G.D.S, Abdelali A, Sajjad H, Darwish K, Nakov P (2021) COVID-19 infodemic twitter dataset. In: Proceedings of the fifteenth international AAAI conference on web and social media (ICWSM 2021). Harvard Dataverse, (2021). https://doi.org/10.7910/DVN/XYK2UE . Type: dataset
13. Gencheva P, Nakov P, Màrquez L, Barrón-Cedeño A, Koychev I (2017) A Context-aware approach for detecting worth-checking claims in political debates. In: Proceedings of the international conference recent advances in natural language processing, RANLP, pp 267–276. INCOMA Ltd., Varna, Bulgaria. https://doi.org/10.26615/978-954-452-049-6_037
14. Lim C (2018) Checking how fact-checkers check. Res Politics 5(3):58. https://doi.org/10.1177/2053168018786848
15. Graves L, Bélair-Gagnon V (2023) From public reason to public health: professional implications of the "debunking turn'' in the global fact-checking field. Digital J. https://doi.org/10.1080/21670811.2023.2218454
16. Altay S, Berriche M, Acerbi A (2023) Misinformation on misinformation: conceptual and methodological challenges. Soc Media Soc 9(1):20. https://doi.org/10.1177/20563051221150412
17. Ejaz W, Mukherjee M, Fletcher R (2023) Climate change news audiences: analysis of news use and attitudes in eight countries. Technical report, Oxford Climate Journalism Network
18. Koch TK, Frischlich L, Lermer E (2023) Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. J Appl Soc Psychol 53(6):495–507. https://doi.org/10.1111/jasp.12959
19. Johnson PR (2023) A case of claims and facts: automated fact-checking the future of Journalism's authority. Digital J 5(1):1–24
20. Graves L (2017) Anatomy of a fact check: objective practice and the contested epistemology of fact checking. Commun Culture Critique 10(3):518–537. https://doi.org/10.1111/cccr.12163
21. Meng K, Jimenez D, Arslan F, Devasier J.D, Obembe D, Li C (2020) Gradient-based adversarial training on transformer networks for detecting check-worthy factual claims. arXiv:2002.07725 [cs]
22. Konstantinovskiy L, Price O, Babakar M (2021) Toward automated factchecking: developing an annotation schema and

benchmark for consistent automated claim detection. Digital Threats Res Pract 2(2):1–16. https://doi.org/10.1145/3412869
23. Shaar S, Babulkov N, Da San Martino G, Nakov P (2020) That is a known lie: detecting previously fact-checked claims. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 3607–3618. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.332
24. Atanasova P, Simonsen J.G, Lioma C, Augenstein I (2020) Generating fact checking explanations. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 7352–7364. Association for computational linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.656
25. Jiang S, Baumgartner S, Ittycheriah A, Yu C (2020) Factoring fact-checks: ctructured information extraction from fact-checking articles. In: Proceedings of the web conference 2020. WWW '20, pp. 1592–1603. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3366423.3380231
26. Biran O, Rambow O (2011) Identifying Justifications in Written Dialogs. In: 2011 IEEE fifth international conference on semantic computing, pp. 162–168. IEEE, Palo Alto, CA, USA (2011). https://doi.org/10.1109/ICSC.2011.41 . http://ieeexplore.ieee.org/document/6061427/
27. Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Comput Linguist 43(3):619–659. https://doi.org/10.1162/COLI_a_00295
28. Iskender N, Schaefer R, Polzehl T, Möller S (2021) Argument mining in tweets: comparing crowd and expert annotations for automated claim and evidence detection. In: Métais E, Meziane F, Horacek H, Kapetanios E (eds) Natural language processing and information systems. Lecture Notes in Computer Science. Springer, Cham, pp 275–288. https://doi.org/10.1007/978-3-030-80599-9_25
29. Cheema G.S, Hakimov S, Sittar A, Müller-Budack E, Otto C, Ewerth R MM-Claims: a dataset for multimodal claim detection in social media. arXiv. arXiv:2205.01989 [cs] (2022).https://doi.org/10.48550/arXiv.2205.01989
30. Elsayed T, Nakov P, Barrón-Cedeño A, Hasanain M, Suwaileh R, Da San Martino G, Atanasova P (2019) Overview of the CLEF-2019 CheckThat! Lab: automatic identification and verification of claims. In: Crestani F, Braschler M, Savoy J, Rauber A, Müller H, Losada DE, Heinatz Bürki G, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Lecture Notes in Computer Science. Springer, Cham, pp 301–321. https://doi.org/10.1007/978-3-030-28577-7_25
31. Shaar S, Nikolov A, Babulkov N, Alam F, Barron-Cedeno A, Elsayed T, Hasanain M, Suwaileh R, Haouari F (2020) Overview of CheckThat! 2020 English: automatic identification and verification of claims in social media
32. Shaar S, Hasanain M, Hamdan B, Ali ZS, Haouari F, Nikolov A, Kutlu M, Kartal YS, Alam F, Martino GDS, Barŕon-Cedẽno, A, Miguez R, Beltŕan J, Elsayed T, Nakov P (2021) Overview of the clef-2021 checkthat lab task 1 on check-worthiness estimation in tweets and political debates, pp. 369–392 (2021). http://ceur-ws.org/Vol-2936/#paper-28
33. Patwari A, Goldwasser D, Bagchi S TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In: Proceedings of the 2017 ACM on conference on information and knowledge management. CIKM '17, pp 2259–2262. Association for computing machinery, New York, NY, USA (2017).https://doi.org/10.1145/3132847.3133150
34. Jha R, Motwani E, Singhal N (2023) Towards automated check-worthy sentence detection using gated recurrent unit. Neural Comput Appl. https://doi.org/10.1007/s00521-023-08300-x
35. Allein L, Moens M-F (2020) Checkworthiness in automatic claim detection models: definitions and analysis of datasets. In: Duijn M, Preuss M, Spaiser V, Takes F, Verberne S (eds) Disinformation in open online media. Lecture Notes in Computer Science. Springer, Cham, pp 1–17
36. Vinhas O, Bastos M The WEIRD governance of fact-checking and the politics of content moderation. New Media & Society, 14614448231213942 (2023)https://doi.org/10.1177/14614448231213942 . Publisher: SAGE Publications. Accessed 2023-12-06
37. Risch J, Stoll A, Wilms L, Wiegand M (2021) Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In: Proceedings of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments, pp 1–12. Association for Computational Linguistics, Duesseldorf, Germany. https://aclanthology.org/2021.germeval-1.1 Accessed 2022-10-07
38. Gupta S, Singh P, Sundriyal M, Akhtar M.S, Chakraborty T Lesa (2021) Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume, pp 3178–3188. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.eacl-main.277
39. Sundriyal M, Kulkarni A, Pulastya V, Akhtar M.S, Chakraborty T (2022) Empowering the fact-checkers! automatic identification of claim spans on twitter. arXiv. arXiv:2210.04710 [cs] . https://doi.org/10.48550/arXiv.2210.04710
40. Augenstein I, Lioma C, Wang D, Chaves Lima L, Hansen C, Hansen C, Simonsen J.G (2019) MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 4685–4697. Association for computational linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1475
41. Devlin J, Chang M.-W, Lee K, Toutanova K Bert (2019) Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North, pp 4171–4186 (2019) https://doi.org/10.18653/v1/N19-1423 . Conference Name: proceedings of the 2019 conference of the North place: Minneapolis, Minnesota Publisher: Association for computational linguistics. Accessed 2023-01-23
42. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv. arXiv:1907.11692 [cs] (2019). https://doi.org/10.48550/arXiv.1907.11692
43. Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543. Association for computational linguistics, Doha, Qatar. https://doi.org/10.3115/v1/D14-1162
44. Jaderberg M, Dalibard V, Osindero S, Czarnecki W.M, Donahue J, Razavi A, Vinyals O, Green T, Dunning I, Simonyan K, Fernando C, Kavukcuoglu K (2017) Population based training of neural networks. arXiv. arXiv:1711.09846 [cs]
45. Yao H, Wang Y, Li S, Zhang L, Liang W, Zou J (2022) Finn C Improving out-of-distribution robustness via selective augmentation. arXiv. arXiv:2201.00299 [cs]
46. Humprecht E (2019) Where 'fake news' flourishes: a comparison across four Western democracies. Inf Commun Soc 22(13):1973–1988 https://doi.org/10.1080/1369118X.2018.1474241
47. Damstra A, Boomgaarden HG, Broda E, Lindgren E, Strömbäck J, Tsfati Y (2021) What does fake look like? A review of the literature on intentional deception in the news and on social

media. J Stud 22(14):1947–1963. https://doi.org/10.1080/1461670X.2021.1979423
48. Horne B, Adali S (2017) This Just In: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the International AAAI Conference on Web and Social Media 11(1), 759–766 https://doi.org/10.1609/icwsm.v11i1.14976
49. Caple H (2018) News Values and Newsworthiness. Oxford research encyclopedia of communication. Oxford University Press, Oxford
50. Strömbäck J, Karlsson M, Hopmann DN (2012) Determinants of news content: comparing journalists' perceptions of the normative and actual impact of different event properties when deciding what's news. J Stud 13(5–6):718–728
51. Bednarek M (2014) Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. Discourse Soc 25(2):135–158. https://doi.org/10.1177/0957926513516041
52. Curry AL, Stroud NJ (2021) The effects of journalistic transparency on credibility assessments and engagement intentions. Journalism 22(4):901–918. https://doi.org/10.1177/1464884919850387
53. Kim HS, Suh YJ, Kim E-M, Chong E, Hong H, Song B, Ko Y, Choi JS (2022) Fact-checking and audience engagement: a study of content analysis and audience behavioral data of fact-checking coverage from news media. Digital J 10(5):781–800. https://doi.org/10.1080/21670811.2021.2006073
54. Potts A, Bednarek M (2015) How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. Discourse Commun 9(2):149–172. https://doi.org/10.1177/1750481314568548
55. Piotrkowicz A, Dimitrova V, Markert K (2017) Automatic extraction of news values from headline text. In: Proceedings of the student research workshop at the 15th conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics, Valencia, Spain. pp 64–74. https://doi.org/10.18653/v1/E17-4007
56. Bednarek M, Caple H (2021) Computer-based analysis of news values: a case study on national day reporting. J Stud 22(6):702–722. https://doi.org/10.1080/1461670X.2020.1807393