A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking

Andreas Hanselowski[†*]**, Christian Stab**[†*]**, Claudia Schulz**[†*]**,**
Zile Li[*]**, Iryna Gurevych**[†*]

[†]Research Training Group AIPHES
https://www.aiphes.tu-darmstadt.de

[*]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
https://www.ukp.tu-darmstadt.de/

[†*] Computer Science Department, Technische Universität Darmstadt

Abstract

Automated fact-checking based on machine learning is a promising approach to identify false information distributed on the web. In order to achieve satisfactory performance, machine learning methods require a large corpus with reliable annotations for the different tasks in the fact-checking process. Having analyzed existing fact-checking corpora, we found that none of them meets these criteria in full. They are either too small in size, do not provide detailed annotations, or are limited to a single domain. Motivated by this gap, we present a new substantially sized mixed-domain corpus with annotations of good quality for the core fact-checking tasks: document retrieval, evidence extraction, stance detection, and claim validation. To aid future corpus construction, we describe our methodology for corpus creation and annotation, and demonstrate that it results in substantial inter-annotator agreement. As baselines for future research, we perform experiments on our corpus with a number of model architectures that reach high performance in similar problem settings. Finally, to support the development of future models, we provide a detailed error analysis for each of the tasks. Our results show that the realistic, multi-domain setting defined by our data poses new challenges for the existing models, providing opportunities for considerable improvement by future systems.

1 Introduction

The ever-increasing role of the Internet as a primary communication channel is arguably the single most important development in the media over the past decades. While it has led to unprecedented growth in information coverage and distribution speed, it comes at a cost. False information can be shared through this channel reaching a much wider audience than traditional means of disinformation (Howell et al., 2013).

While human fact-checking still remains the primary method to counter this issue, the amount and the speed at which new information is spread makes manual validation challenging and costly. This motivates the development of automated fact-checking pipelines (Thorne et al., 2018a; Popat et al., 2017; Hanselowski and Gurevych, 2017) consisting of several consecutive tasks. The following four tasks are commonly included in the pipeline. Given a controversial claim, *document retrieval* is applied to identify documents that contain important information for the validation of the claim. *Evidence extraction* aims at retrieving text snippets or sentences from the identified documents that are related to the claim. This evidence can be further processed via *stance detection* to infer whether it supports or refutes the claim. Finally, *claim validation* assesses the validity of the claim given the evidence.

Automated fact-checking has received significant attention in the NLP community in the past years. Multiple corpora have been created to assist the development of fact-checking models, varying in quality, size, domain, and range of annotated phenomena. Importantly, the successful development of a full-fledged fact-checking system requires that the underlying corpus satisfies certain characteristics. First, training data needs to contain a large number of instances with *high-quality annotations* for the different fact-checking sub-tasks. Second, the training data should not be limited to a particular domain, since potentially wrong information sources can range from official statements to blog and Twitter posts.

We analyzed existing corpora regarding their adherence to the above criteria and identified several drawbacks. The corpora introduced by Vlachos and Riedel (2014); Ferreira and Vlachos (2016); Derczynski et al. (2017) are valuable for the analysis of the fact-checking problem and pro-

vide annotations for stance detection. However, they contain only several hundreds of validated claims and it is therefore unlikely that deep learning models can generalize to unobserved claims if trained on these datasets.

A corpus with significantly more validated claims was introduced by Popat et al. (2017). Nevertheless, for each claim, the corpus provides 30 documents which are retrieved from the web using the Google search engine instead of a document collection aggregated by fact-checkers. Thus, many of the documents are unrelated to the claim and important information for the validation may be missing.

The FEVER corpus constructed by Thorne et al. (2018a) is the largest corpus available for the development of automated fact-checking systems. It consists of 185,445 validated claims with annotated documents and evidence for each of them. The corpus therefore allows training deep neural networks for automated fact-checking, which reach higher performance than shallow machine learning techniques. However, the corpus is based on synthetic claims derived from Wikipedia sentences rather than *natural* claims that originate from heterogeneous web sources.

In order to address the drawbacks of existing datasets, we introduce a new corpus based on the Snopes[1] fact-checking website. Our corpus consists of 6,422 validated claims with comprehensive annotations based on the data collected by Snopes fact-checkers and our crowd-workers. The corpus covers multiple domains, including discussion blogs, news, and social media, which are often found responsible for the creation and distribution of unreliable information. In addition to validated claims, the corpus comprises over 14k documents annotated with evidence on two granularity levels and with the stance of the evidence with respect to the claims. Our data allows training machine learning models for the four steps of the automated fact-checking process described above: document retrieval, evidence extraction, stance detection, and claim validation.

The contributions of our work are as follows:

1) We provide a substantially sized mixed-domain corpus of natural claims with annotations for different fact-checking tasks. We publish a web crawler that reconstructs our dataset includ-

ing all annotations[2]. For research purposes, we are allowed to share the original corpus[3].

2) To support the creation of further fact-checking corpora, we present our methodology for data collection and annotation, which allows for the efficient construction of large-scale corpora with a substantial inter-annotator agreement.

3) For evidence extraction, stance detection, and claim validation we evaluate the performance of high-scoring systems from the FEVER shared task (Thorne et al., 2018b)[4] and the Fake News Challenge (Pomerleau and Rao, 2017)[5] as well as the Bidirectional Transformer model BERT (Devlin et al., 2018) on our data. To facilitate the development of future fact-checking systems, we release the code of our experiments[6].

4) Finally, we conduct a detailed error analysis of the systems trained and evaluated on our data, identifying challenging fact-checking instances which need to be addressed in future research.

2 Related work

Below, we give a comprehensive overview of existing fact-checking corpora, summarized in Table 1. We focus on their key parameters: fact-checking sub-task coverage, annotation quality, corpus size, and domain. It must be acknowledged that a fair comparison between the datasets is difficult to accomplish since the length of evidence and documents, as well as the annotation quality, significantly varies between the corpora.

PolitiFact14 Vlachos and Riedel (2014) analyzed the fact-checking problem and constructed a corpus on the basis of the fact-checking blog of Channel 4[7] and the Truth-O-Meter from PolitiFact[8]. The corpus includes additional evidence, which has been used by fact-checkers to validate the

---

	claims	docs.	evid.	stance	sources	rater agr.	domain
PolitiFact14	106	no	yes	no	no	no	political statements
Emergent16	300	2,595	no	yes	yes	no	news
PolitiFact17	12,800	no	no	no	no	no	political statements
RumourEval17	297	4,519	no	yes	yes	yes	Twitter
Snopes17	4,956	136,085	no	no	yes	no	Google search results
CLEF-2018	150	no	no	no	no	no	political debates
FEVER18	185,445	14,533	yes	yes	yes	yes	Wikipedia
Our corpus	6,422	14,296	yes	yes	yes	yes	multi domain

Table 1: Overview of corpora for automated fact-checking. docs: documents related to the claims; evid.: evidence in form of sentence or text snippets; stance: stance of the evidence; sources: sources of the evidence; rater agr.: whether or not the inter-annotator agreement is reported; domain: the genre of the corpus

claims, as well as metadata including the speaker ID and the date when the claim was made. This is early work in automated fact-checking and Vlachos and Riedel (2014) mainly focused on the analysis of the task. The corpus therefore only contains 106 claims, which is not enough to train high-performing machine learning systems.

Emergent16 A more comprehensive corpus for automated fact-checking was introduced by Ferreira and Vlachos (2016). The dataset is based on the project Emergent[9] which is a journalist initiative for rumor debunking. It consists of 300 claims that have been validated by journalists. The corpus provides 2,595 news articles that are related to the claims. Each article is summarized into a headline and is annotated with the article's stance regarding the claim. The corpus is well suited for training stance detection systems in the news domain and it was therefore chosen in the Fake News Challenge (Pomerleau and Rao, 2017) for training and evaluation of competing systems. However, the number of claims in the corpus is relatively small, thus it is unlikely that sophisticated claim validation systems can be trained using this corpus.

PolitiFact17 Wang (2017) extracted 12,800 validated claims made by public figures in various contexts from Politifact. For each statement, the corpus provides a verdict and meta information, such as the name and party affiliation of the speaker or subject of the debate. Nevertheless, the corpus does not include evidence and thus the models can only be trained on the basis of the claim, the verdict, and meta information.

RumourEval17 Derczynski et al. (2017) organized the RumourEval shared task, for which they provided a corpus of 297 rumourous threads from Twitter, comprising 4,519 tweets. The shared task

was divided into two parts, *stance detection* and *veracity prediction* of the rumors, which is similar to claim validation. The large number of stance-annotated tweets allows for training stance detection systems reaching a relatively high score of about 0.78 accuracy. However, since the number of claims (rumours) is relatively small, and the corpus is only based on tweets, this dataset alone is not suitable to train generally applicable claim validation systems.

Snopes17 A corpus featuring a substantially larger number of validated claims was introduced by Popat et al. (2017). It contains 4,956 claims annotated with verdicts which have been extracted from the Snopes website as well as the Wikipedia collections of proven hoaxes[10] and fictitious people[11]. For each claim, the authors extracted about 30 associated documents using the Google search engine, resulting in a collection of 136,085 documents. However, since the documents were not annotated by fact-checkers, irrelevant information is present and important information for the claim validation might be missing.

CLEF-2018 Another corpus concerned with political debates was introduced by Nakov et al. (2018) and used for the CLEF-2018 shared task. The corpus consists of transcripts of political debates in English and Arabic and provides annotations for two tasks: identification of check-worthy statements (claims) in the transcripts, and validation of 150 statements (claims) from the debates. However, as for the corpus **PolitiFact17**, no evidence for the validation of these claims is available.

FEVER18 The FEVER corpus introduced by Thorne et al. (2018a) is the largest available fact-

---

[9] http://www.emergent.info/

[10] https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxe

[11] https://en.wikipedia.org/wiki/List_of_fictitious_people

checking corpus, consisting of 185,445 validated claims. The corpus is based on about 50k popular Wikipedia articles. Annotators modified sentences in these articles to create the claims and labeled other sentences in the articles, which support or refute the claim, as evidence. The corpus is large enough to train deep learning systems able to retrieve evidence from Wikipedia. Nevertheless, since the corpus only covers Wikipedia and the claims are created synthetically, the trained systems are unlikely to be able to extract evidence from heterogeneous web-sources and validate claims on the basis of evidence found on the Internet.

As our analysis shows, while multiple fact-checking corpora are already available, no single existing resource provides full fact-checking sub-task coverage backed by a substantially-sized and validated dataset spanning across multiple domains. To eliminate this gap, we have created a new corpus as detailed in the following sections.

3 Corpus construction

This section describes the original data from the Snopes platform, followed by a detailed report on our corpus annotation methodology.
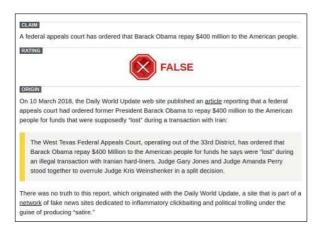
3.1 Source data

Figure 1: Snopes fact-checking data example

Snopes is a large-scale fact-checking platform that employs human fact-checkers to validate claims. A simple *fact-checking instance* from the Snopes website is shown in Figure 1. At the top of the page, the *claim* and the *verdict* (rating) are given. The fact-checkers additionally provide a *resolution* (origin), which backs up the verdict. Evidence in the resolution, which we call *evidence text snippets* (ETSs), is marked with a yellow bar. As additional validation support, Snopes

fact-checkers provide URLs[12] for *original documents* (ODCs) from which the ETSs have been extracted or which provide additional information.

Our crawler extracts the *claims*, *verdicts*, *ETSs*, the *resolution*, as well as *ODCs* along with their URLs, thereby enriching the ETSs with useful contextual information. Snopes is almost entirely focused on claims made on English speaking websites. Our corpus therefore only features English fact-checking instances.

3.2 Corpus annotation

While ETSs express a stance towards the claim, which is useful information for the fact-checking process, this stance is not explicitly stated on the Snopes website. Moreover, the ETSs given by fact-checkers are quite coarse and often contain detailed background information that is not directly related to the claim and consequently not useful for its validation. In order to obtain an informative, high-quality collection of evidence, we asked crowd-workers to label the stance of ETSs and to extract sentence-level evidence from the ETSs that are directly relevant for the validation of the claim. We further refer to these sentences as *fine grained evidence* (FGE).

Stance annotation. We asked crowd workers on Amazon Mechanical Turk[13] to annotate whether an ETS *agrees* with the claim, *refutes* it, or has *no stance* towards the claim. An ETS was only considered to express a stance if it explicitly referred to the claim and either expressed support for it or refuted it. In all other cases, the ETS was considered as having *no stance*.

FGE annotation. We filtered out ETSs with *no stance*, as they do not contain supporting or refuting FGE. If an ETS was annotated as supporting the claim, the crowd workers selected only supporting sentences; if the ETS was annotated as refuting the claim, only refuting sentences were selected. Table 2 shows two examples of ETSs with annotated FGE. As can be observed, not all information given in the original ETS is directly relevant for validating the claim. For example, sentence (1c) in the first example's ETS simply provides additional background information and is therefore not considered FGE.

---

[12]underlined words in the resolution are hyperlinks
[13]https://www.mturk.com/

ETS stance: support
Claim: The Fox News will be shutting down for routine maintenance on 21 Jan. 2013.
Evidence text snippet: (1a) *Fox News Channel announced today that it would shutdown for what it called "routine maintenance".* (1b) *The shutdown is on 21 January 2013.* (1c) Fox News president Roger Ailes explained the timing of the shutdown: "We wanted to pick a time when nothing would be happening that our viewers want to see."
ETS stance: refute
Claim: Donald Trump supported Emmanuel Macron during the French election.
Evidence text snippet: (2a) In their first meeting, the U.S. President told Emmanuel Macron that he had been his favorite in the French presidential election saying "You were my guy". (2b) *In an interview with the Associated Press, however, Trump said he thinks Le Pen is stronger than Macron on what's been going on in France.*

Table 2: Examples of FGE annotation in supporting (top) and refuting (bottom) ETSs, sentences selected as FGE in italic.

4 Corpus analysis

4.1 Inter-annotator agreement

Stance annotation. Every ETS was annotated by at least six crowd workers. We evaluate the inter-annotator agreement between groups of workers as proposed by Habernal et al. (2017), i.e. by randomly dividing the workers into two equal groups and determining the aggregate annotation for each group using MACE (Hovy et al., 2013). The final inter-annotator agreement score is obtained by comparing the aggregate annotation of the two groups. Using this procedure, we obtain a Cohen's Kappa of $\kappa = 0.7$ (Cohen, 1968), indicating a substantial agreement between the crowd workers (Artstein and Poesio, 2008). The gold annotations of the ETS stances were computed with MACE, using the annotations of all crowd workers. We have further assessed the quality of the annotations performed by crowd workers by comparing them to expert annotations. Two experts labeled 200 ETSs, reaching the same agreement as the crowd workers, i.e. $\kappa = 0.7$. The agreement between the experts' annotations and the com-

puted gold annotations from the crowd workers is also substantial, $\kappa = 0.683$.

FGE Annotation. Similar to the stance annotation, we used the approach of Habernal et al. (2017) to compute the agreement. The inter-annotator agreement between the crowd workers in this case is $\kappa = 0.55$ Cohen's Kappa. We compared the annotations of FGE in 200 ETSs by experts with the annotations by crowd workers, reaching an agreement of $\kappa = 0.56$. This is considered as *moderate* inter-annotator agreement (Artstein and Poesio, 2008).

In fact, the task is significantly more difficult than stance annotation as sentences may provide only partial evidence for or against the claim. In such cases, it is unclear how large the information overlap between sentence and claim should be for a sentence to be FGE. The sentence (1a) in Table 2, for example, only refers to one part of the claim without mentioning the time of the shutdown. We can further modify the example in order to make the problem more obvious: *(a) The channel announced today that it is planing a shutdown. (b) Fox News made an announcement today.*

As the example illustrates, there is a gradual transition between sentences that can be considered as essential for the validation of the claim and those which just provide minor negligible details or unrelated information. Nevertheless, even though the inter-annotator agreement for the annotation of FGE is lower than for the annotation of ETS stance, compared to other annotation problems (Zechner, 2002; Benikova et al., 2016; Tauchmann et al., 2018) that are similar to the annotation of FGE, our framework leads to a better agreement.

4.2 Corpus statistics

Table 3 displays the main statistics of the corpus. In the table, *FGE sets* denotes groups of FGE extracted from the same ETS. Many of the ETSs have been annotated as *no stance* (see Table 5) and, following our annotation study setup, are not used for FGE extraction. Therefore, the number of FGE sets is much lower than that of ETSs. We have found that, on average, an ETS consists of 6.5 sentences. For those ETS that have support/refute stance, on average, 2.3 sentences are selected as FGE. For many of the ETSs, no original documents (ODCs) have been provided (documents from which they have been extracted). On

the other hand, in many instances, links to ODCs are given that provide additional information, but from which no ETSs have been extracted.

entity:	claims	ETSs	FGE sets	ODCs
count:	6,422	16,509	8,291	14,296

Table 3: Overall statistics of the corpus

The distribution of verdicts in Table 4 shows that the dataset is unbalanced in favor of false claims. The label *other* refers to a collocation of verdicts that do not express a tendency towards declaring the claim as being false or true, such as *mixture, unproven, outdated, legend,* etc.

verdict:	false	true	most. false	most. true	other
count	2,943	659	334	93	2,393
%	45.8	10.3	5.2	1.4	37.3

Table 4: Distribution of verdicts for claims

Table 5 shows the stance distribution for ETSs. Here, supporting ETSs and ETSs that do not express any stance are dominating.

stance:	support	refute	no stance
ETSs:			
count	6,734	2,266	7,508
%	40.8	13.7	45.5
FGE sets:			
count	6,178	2,113	–
%	74.5	25.5	–

Table 5: Class distribution of ETSs the FGE sets

For supporting and refuting ETSs annotators identified FGE sets for 8,291 out of 8,998 ETSs. ETSs with a stance but without FGE sets often miss a clear connection to the claim, so the annotators did not annotate any sentences in these cases. The class distribution of the FGE sets in Table 5 shows that supporting ETSs are more dominant.

To identify potential biases in our new dataset, we investigated which topics are prevalent by grouping the fact-checking instances (claims with their resolutions) into categories defined by Snopes. According to our analysis, the four categories *Fake News*, *Political News*, *Politics* and *Fauxtography* are dominant in the corpus ranging from more than 700 to about 900 instances. A significant number of instances are present in the categories *Inboxer Rebellion (Email hoax)*, *Business*, *Medical*, *Entertainment* and *Crime*.

We further investigated the sources of the collected documents (ODCs) and grouped them into a number of classes. We found that 38% of the articles are from different news websites ranging from mainstream news like CNN to tabloid press and partisan news. The second largest group of documents are false news and satirical articles with 30%. Here, the majority of articles are from the two websites *thelastlineofdefense.org* and *worldnewsdailyreport.com*. The third class of documents, with a share of 11%, are from social media like Facebook and Twitter. The remaining 21% of documents come from diverse sources, such as debate blogs, governmental domains, online retail, or entertainment websites.

4.3 Discussion

I this subsection, we briefly discuss the differences of our corpus to the FEVER dataset as the most comprehensive dataset introduced so far. Due to the way the FEVER dataset was constructed, the claim validation problem defined by this corpus is different compared to the problem setting defined by our corpus. The verdict of a claim for FEVER depends on the stance of the evidence, that is, if the stance of the evidence is agree the claim is necessarily true, and if the stance is disagree the claim is necessarily false. As a result, the claim validation problem can be reduced to stance detection. Such a transformation is not possible for our corpus, as the evidence might originate from unreliable sources and a claim may have both supporting and refuting ETSs. The stance of ETSs is therefore not necessarily indicative of the veracity of the claim. In order to investigate how the stance is related to the verdict of the claim for our dataset, we computed their correlation. In the correlation analysis, we considered how a claims' verdict, represented by the classes *false*, *mostly false*, *other*, *mostly true*, *true*, correlates with the number of supporting ETSs minus the number of refuting ETSs. More precisely, the verdicts of the claims are considered as one variable, which can take 5 discreet values ranging from *false* to *true*, and the stance is considered as the other variable, which is represented by the difference between the number of supporting versus the number of refuting evidence. We found that the verdict is only weakly correlated with the stance, as indicated by the Pearson correlation coefficient of 0.16. This illustrates that the fact-checking problem setting

for our corpus is more challenging than for the FEVER dataset.

5 Experiments and error analysis

The annotation of the corpus described in the previous section provides supervision for different fact-checking sub-tasks. In this paper, we perform experiments for the following sub-tasks: (1) detection of the stance of the ETSs with respect to the claim, (2) identification of FGE in the ETSs, and (3) prediction of a claim's verdict given FGE.

There are a number of experiments beyond the scope of this paper, which are left for future work: (1) retrieval of the original documents (ODCs) given a claim, (2) identification of ETSs in ODCs, and (3) prediction of a claim's verdict on the basis of FGE, the stance of FGE, and their sources.

Moreover, in this paper, we consider the three tasks independent of each other rather than as a pipeline. In other words, we always take the gold standard from the preceding task instead of the output of the preceding model in the pipeline. For the three independent tasks, we use recently suggested models that achieved high performance in similar problem settings. In addition, we provide the human agreement bound, which is determined by comparing expert annotations for 200 ETSs to the gold standard derived from crowd worker annotations (Section 4.1).

5.1 Stance detection

In the stance detection task, models need to determine whether an ETS *supports* or *refutes* a claim, or expresses *no stance* with respect to the claim.

5.1.1 Models and Results

We report the performance of the following models: AtheneMLP is a feature-based multi-layer perceptron (Hanselowski et al., 2018a), which has reached the second rank in the Fake News Challenge. DecompAttent (Parikh et al., 2016) is a neural network with a relatively small number of parameters that uses decomposable attention, reaching good results on the Stanford Natural Language Inference task (Bowman et al., 2015). USE+Attent is a model which uses the Universal Sentence Encoder (USE) (Cer et al., 2018) to extract representations for the sentences of the ETSs and the claim. For the classification of the stance, an attention mechanism and a MLP is used.

The results in Table 6 show that AtheneMLP scores highest. Similar to the outcome of the

Fake News Challenge, feature-based models outperform neural networks based on word embeddings (Hanselowski et al., 2018a). As the comparison to the human agreement bound suggests, there is still substantial room for improvement.

model	recall	precision	F1m
agreement bound	0.770	0.837	0.802
random baseline	0.333	0.333	0.333
majority vote	0.150	0.333	0.206
AtheneMLP	**0.585**	**0.607**	**0.596**
DecompAttent	0.510	0.560	0.534
USE+Attent	0.380	0.505	0.434

Table 6: Stance detection results (F1m = F1 macro)

5.1.2 Error analysis

We performed an error analysis for the best-scoring model AtheneMLP. The error analysis has shown that *supporting* ETSs are mostly classified correctly if there is a significant lexical overlap between the claim and the ETS. If the claim and the ETSs use different wording, or if the ETS implies the validity of the claim without explicitly referring to it, the model often misclassifies the snippets (see example in the Appendix A.2.1). This is not surprising, as the model is based on bag-of-words, topic models, and lexica.

Moreover, as the distribution of the classes in Table 5 shows, *support* and *no stance* are more dominant than the *refute* class. The model is therefore biased towards these classes and is less likely to predict *refute* (see confusion matrix in the Appendix Table 11). An analysis of the misclassified *refute* ETSs has shown that the contradiction is often expressed in difficult terms, which the model could not detect, e.g. "the myth originated", "no effect can be observed", "The short answer is no".

5.2 Evidence extraction

We define evidence extraction as the identification of fine-grained evidence (FGE) in the evidence text snippets (ETSs). The problem can be approached in two ways, either as a *classification problem*, where each sentence from the ETSs is classified as to whether it is an evidence for a given claim, or as a *ranking problem*, in the way defined in the FEVER shared task. For FEVER, sentences in introductory sections of Wikipedia articles need to be ranked according to their relevance for the validation of the claim and the 5 highest ranked sentences are taken as evidence.

5.2.1 Models and Results

We consider the task as a ranking problem, but also provide the human agreement bound, the random baseline and the majority vote for evidence extraction as a *classification problem* for future reference in Table 10 in the Appendix.

To evaluate the performance of the models in the *ranking* setup, we measure the precision and recall on five highest ranked ETS sentences (precision @5 and recall @5), similar to the evaluation procedure used in the FEVER shared task. Table 7 summarizes the performance of several models on our corpus. The `rankingESIM` (Hanselowski et al., 2018b) was the best performing model on the FEVER evidence extraction task. The `Tf-Idf` model (Thorne et al., 2018a) served as a baseline in the FEVER shared task. We also evaluate the performance of `DecompAttent` and a simple `BiLSTM` (Hochreiter and Schmidhuber, 1997) architecture. To adjust the latter two models to the ranking problem setting, we used the hinge loss objective function with negative sampling as implemented in the `rankingESIM` model. As in the FEVER shared task, we consider the recall @5 as a metric for the evaluation of the systems.

The results in Table 7 illustrate that, in terms of recall, the neural networks with a small number of parameters, `BiLSTM` and `DecompAttent`, perform best. The `Tf-Idf` model reaches best results in terms of precision. The `rankingESIM` reaches a relatively low score and is not able to beat the random baseline. We assume this is because the model has a large number of parameters and requires many training instances.

model	precision @5	recall @5
random baseline	0.296	0.529
BiLSTM	0.451	**0.637**
DecompAttent	0.420	0.627
Tf-Idf	**0.627**	0.601
rankingESIM	0.288	0.507

Table 7: Evidence extraction: ranking setting

5.2.2 Error analysis

We performed an error analysis for the `BiLSTM` and the `Tf-Idf` model, as they reach the highest recall and precision, respectively. `Tf-Idf` achieves the best precision because it only predicts a small set of sentences, which have lexical overlap with the claim. The model therefore misses FGE that paraphrase the claim. The `BiLSTM` is

better able to capture the semantics of the sentences. We believe that it was therefore able to take related word pairs, such as "*Israel*" - "*Jewish*", "*price*"-"*sold*", "*pointed*"-"*pointing*", "*broken*"-"*injured*", into account during the ranking process. Nevertheless, the model fails when the relationship between the claim and the potential FGE is more elaborate, e.g. if the claim is not paraphrased, but reasons for it being true are provided. An example of a misclassified sentence is given in the Appendix A.2.2.

5.3 Claim validation

We formulate the claim validation problem in such a way that we can compare it to the FEVER *recognizing textual entailment* task. Thus, as illustrated in Table 8, we compress the different verdicts present on the Snopes webpage into three categories of the FEVER shared task. In order to form the *not enough information* (NEI) class, we compress the three verdicts *mixture*, *unproven*, and *undetermined*. We entirely omit all the other verdicts like *legend, outdated, miscaptioned*, as these cases are ambiguous and difficult to classify. For the classification of the claims, we provide only the FGE as they contain the most important information from ETSs.

FEVER	Snopes
refuted:	false, mostly false
supported:	true, mostly true
NEI:	mixture, unproven, undetermined

Table 8: Compression of Snopes verdicts

5.3.1 Experiments

For the claim validation, we consider models of different complexity: `BertEmb` is an MLP classifier which is based on BERT pre-trained embeddings (Devlin et al., 2018); `DecompAttent` was used in the FEVER shared task as baseline; `extendedESIM` is an extended version of the ESIM model (Hanselowski et al., 2018b) reaching the third rank in the FEVER shared task; `BiLSTM` is a simple BiLSTM architecture; `USE+MLP` is the Universal Sentence Encoder combined with a MLP; `SVM` is an SVM classifier based on bag-of-words, unigrams, and topic models.

The results illustrated in Table 9 show that `BertEmb`, `USE+MLP`, `BiLSTM`, and `extendedESIM` reach similar performance, with `BertEmb` being the best. However, compared to the FEVER claim validation problem,

Labeling method	recall m	prec. m	F1 m
random baseline	0.333	0.333	0.333
majority vote	0.198	0.170	0.249
BertEmb	0.477	**0.493**	**0.485**
USE+MLP	0.483	0.468	0.475
BiLSTM	0.456	0.473	0.464
extendedESIM	**0.561**	0.503	0.454
featureSVM	0.384	0.396	0.390
DecompAttent	0.336	0.312	0.324

Table 9: Claim validation results (m = macro)

where systems reach up to 0.7 F1 macro, the scores are relatively low. Thus, there is ample opportunity for improvement by future systems.

5.3.2 Error analysis

We performed an error analysis for the best-scoring model `BertEmb`. The class distribution for claim validation is highly biased towards *refuted* (false) claims and, therefore, claims are frequently labeled as *refuted* even though they belong to one of the other two classes (see confusion matrix in the Appendix in Table 12).

We have also found that it is often difficult to classify the claims as the provided FGE in many cases are contradicting (e.g. Appendix A.2.3). Although the corpus is biased towards false claims (Table 5), there is a large number of ETSs that support those false claims (Table 4). As discussed in Section 4.2, this is because many of the retrieved ETSs originate from false news websites.

Another possible reason for the lower performance is that our data is heterogeneous and, therefore, it is more challenging for a machine learning model to generalize. In fact, we have performed additional experiments in which we pre-trained a model on the FEVER corpus and fine-tuned the parameters on our corpus and vice versa. However, no significant performance gain could be observed in both experiments

Based on our analysis, we conclude that heterogeneous data and FGE from unreliable sources, as found in our corpus and in the real world, make it difficult to correctly classify the claims. Thus, in future experiments, not just FGE need to be taken into account, but also additional information from our newly constructed corpus, that is, the stance of the FGE, FGE sources, and documents from the Snopes website which provide additional information about the claim. Taking all this information into account would enable the system to

find a consistent configuration of these labels and thus potentially help to improve performance. For instance, a claim that is *supported* by evidence coming from an *unreliable* source is most likely *false*. In fact, we believe that modeling the meta-information about the evidence and the claim more explicitly represents an important step in making progress in automated fact-checking.

6 Conclusion

In this paper, we have introduced a new richly annotated corpus for training machine learning models for the core tasks in the fact-checking process. The corpus is based on heterogeneous web sources, such as blogs, social media, and news, where most false claims originate. It includes validated claims along with related documents, evidence of two granularity levels, the sources of the evidence, and the stance of the evidence towards the claim. This allows training machine learning systems for document retrieval, stance detection, evidence extraction, and claim validation.

We have described the structure and statistics of the corpus, as well as our methodology for the annotation of evidence and the stance of the evidence. We have also presented experiments for stance detection, evidence extraction, and claim validation with models that achieve high performance in similar problem settings. In order to support the development of machine learning approaches that go beyond the presented models, we provided an error analysis for each of the three tasks, identifying difficulties with each.

Our analysis has shown that the fact-checking problem defined by our corpus is more difficult than for other datasets. Heterogeneous data and evidence from unreliable sources, as found in our corpus and in the real world, make it difficult to correctly classify the claims. We conclude that more elaborate approaches are required to achieve higher performance in this challenging setting.

7 Acknowledgements

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. 2016. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1039–1050, Osaka, Japan.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, pages 1163–1168, San Diego, CA, USA.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task. *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*.

Andreas Hanselowski and Iryna Gurevych. 2017. A framework for automated fact-checking for real-time validation of emerging claims on the web. *Proceedings of the NIPS Workshop on Prioritising Online Content (WPOC2017)*.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018a. A retrospective analysis of the fake news challenge stance detection task. *Proceedings of the 2018 International Committee on Computational Linguistics*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018b. Ukp-athene: Multi-sentence textual entailment for claim verification. *Proceedings of the EMNLP 2018 First Workshop on Fact Extraction and Verification*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, pages 1120–1130, Atlanta, GA, USA.

Lee Howell et al. 2013. Digital wildfires in a hyper-connected world. *WEF Report*, 3:15–94.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Avignon, France. Springer.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Dean Pomerleau and Delip Rao. 2017. The Fake News Challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. http://www.fakenewschallenge.org/. Accessed: 2019-4-20.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee.

Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M Meyer, and Margot Mieskes. 2018. Beyond generic summarization: A

multi-faceted hierarchical summarization corpus of large heterogeneous data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

William Yang Wang. 2017. ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Klaus Zechner. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, 28(4):447–485.