

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.

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 BIBREF0.

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 BIBREF1, BIBREF2, BIBREF3 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 BIBREF4, BIBREF5, BIBREF6 are valuable for the analysis of the fact-checking problem and provide 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 BIBREF2. 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 BIBREF1 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 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 including all annotations. For research purposes, we are allowed to share the original corpus.

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 BIBREF7 and the Fake News Challenge BIBREF8 as well as the Bidirectional Transformer model BERT BIBREF9 on our data. To facilitate the development of future fact-checking systems, we release the code of our experiments.

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.

Related work

Below, we give a comprehensive overview of existing fact-checking corpora, summarized in Table TABREF7. 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 BIBREF4 analyzed the fact-checking problem and constructed a corpus on the basis of the fact-checking blog of Channel 4 and the Truth-O-Meter from PolitiFact. The corpus includes additional evidence, which has been used by fact-checkers to validate the 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 BIBREF4 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 BIBREF⁵. The dataset is based on the project Emergent 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 BIBREF⁸ 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 BIBREF¹⁰ 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 BIBREF⁶ 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 BIBREF². 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 and fictitious people. 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 BIBREF11 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 BIBREF1 is the largest available fact-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.

Corpus construction

This section describes the original data from the Snopes platform, followed by a detailed report on our

corpus annotation methodology.

Corpus construction ::: Source data

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 FIGREF14. 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 (ETs), is marked with a yellow bar. As additional validation support, Snopes fact-checkers provide URLs for original documents (ODs) from which the ETs have been extracted or which provide additional information.

Our crawler extracts the claims, verdicts, ETs, the resolution, as well as ODs along with their URLs, thereby enriching the ETs 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.

Corpus construction ::: Corpus annotation

While ETs 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 ETs 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 ETs and to extract sentence-level evidence from the ETs 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 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 TABREF18 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.

Corpus analysis ::: 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 BIBREF12, i.e. by randomly dividing the workers into two equal groups and determining the aggregate annotation for each group using MACE BIBREF13. 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 $\kappa = 0.7$ BIBREF14, indicating a substantial agreement between the crowd workers BIBREF15. 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 computed gold annotations from the crowd workers is also substantial, $\kappa = 0.683$.

FGE Annotation. Similar to the stance annotation, we used the approach of BIBREF12 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 BIBREF15.

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 TABREF18, 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 BIBREF16, BIBREF17, BIBREF18 that are similar to the annotation of FGE, our framework leads to a better agreement.

Corpus analysis :: Corpus statistics

Table TABREF21 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 TABREF23) 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 ETSs 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.

The distribution of verdicts in Table TABREF22 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.

Table TABREF23 shows the stance distribution for ETSs. Here, supporting ETSs and ETSs that do not express any stance are dominating.

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 TABREF23 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.

Corpus analysis ::: Discussion

In 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 claim's 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 discrete 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.

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 SECREF19).

Experiments and error analysis :: 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.

Experiments and error analysis :: Stance detection :: Models and Results

We report the performance of the following models: AtheneMLP is a feature-based multi-layer perceptron BIBREF19, which has reached the second rank in the Fake News Challenge. DecompAttent BIBREF20 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 BIBREF21. USE+Attent is a model which uses the Universal Sentence Encoder (USE) BIBREF22 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 TABREF27 show that AtheneMLP scores highest. Similar to the outcome of the Fake News Challenge, feature-based models outperform neural networks based on word embeddings BIBREF19. As the comparison to the human agreement bound suggests, there is still substantial room for improvement.

Experiments and error analysis :: Stance detection :: 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 SECREF41). 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 TABREF23 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 TABREF42). 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”.

Experiments and error analysis :: Evidence extraction

We define evidence extraction as the identification of fine-grained evidence (FGE) in the evidence text snippets (ETs). The problem can be approached in two ways, either as a classification problem, where each sentence from the ETs 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.

Experiments and error analysis :: Evidence extraction :: 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 TABREF39 in the Appendix.

To evaluate the performance of the models in the ranking setup, we measure the precision and recall on five highest ranked ET sentences (precision @5 and recall @5), similar to the evaluation procedure used in the FEVER shared task. Table TABREF31 summarizes the performance of several models on our corpus. The rankingESIM BIBREF23 was the best performing model on the FEVER evidence extraction task. The Tf-Idf model BIBREF1 served as a baseline in the FEVER shared task. We also evaluate the performance of DecompAttent and a simple BiLSTM BIBREF24 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 TABREF31 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.

Experiments and error analysis :: Evidence extraction :: 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 [SECREf43](#).

Experiments and error analysis :: 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 [TABREF34](#), 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 ETs.

Experiments and error analysis :: Claim validation :: Experiments

For the claim validation, we consider models of different complexity: BertEmb is an MLP classifier which is based on BERT pre-trained embeddings BIBREF9; DecompAttent was used in the FEVER shared task as baseline; extendedESIM is an extended version of the ESIM model BIBREF23 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 TABREF36 show that BertEmb, USE+MLP, BiLSTM, and extendedESIM reach similar performance, with BertEmb being the best. However, compared to the FEVER claim validation problem, where systems reach up to 0.7 F1 macro, the scores are relatively low. Thus, there is ample opportunity for improvement by future systems.

Experiments and error analysis :: Claim validation :: 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 TABREF45).

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 SECREF44). Although the corpus is biased towards false claims (Table TABREF23), there is a large number of ETSs that support those false claims (Table TABREF22). As discussed in Section SECREF20, 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.

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.

Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1.

Appendix ::: Error analysis ::: Stance detection

Below we give an instance of a misclassified ETS. Even though the ETS supports the claim, the lexical overlap is relatively low. Most likely, for this reason, the model predicts refute.

Example:

Claim: The Reuters news agency has proscribed the use of the word 'terrorists' to describe those who pulled off the September 11 terrorist attacks on America.

ETS: Reuters' approach doesn't sit well with some journalists, who say it amounts to self-censorship.

“Journalism should be about telling the truth. And when you don't call this a terrorist attack, you're not telling the truth,” says Rich Noyes, director of media analysis at the conservative Media Research Center.

...

Appendix ::: Error analysis ::: Evidence extraction

The model wrongly predicts sentences when the topic of the sentences is similar to the topic of the claim, but the sentence is not relevant for the validation of the claim:

Example:

Claim: The Department of Homeland Security uncovered a terrorist plot to attack Black Friday shoppers in several locations.

FGE: Bhakkar Fatwa is a small, relatively unknown group of Islamic militants and fanatics that originated in Bhakkar Pakistan as the central leadership of Al Qaeda disintegrated under the pressures of U.S.

military operations in Afghanistan and drone strikes conducted around the world.

Appendix ::: Error analysis ::: Claim validation

The FGE are contradicting and the classifier predicts refuted instead of supported.

Example:

Gold standard: supported; Prediction: refuted

Claim: As a teenager, U.S. Secretary of State Colin Powell learned to speak Yiddish while working in a Jewish-owned baby equipment store.

FGE: As a boy whose friends and employers at the furniture store were Jewish, Powell picked up a smattering of Yiddish. He kept working at Sickser's through his teens, ... picking up a smattering of Yiddish ... A spokesman for Mr. Powell said he hadn't heard about the spoof ...