MultiFC
A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims
Augenstein, Isabelle; Lioma, Christina; Wang, Dongsheng; Chaves Lima, Lucas; Hansen, Casper; Hansen, Christian; Simonsen, Jakob Grue

Published 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)

DOI:
10.18653/v1/D19-1475

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims

Department of Computer Science
University of Copenhagen
{augenstein,c.lioma,wang,lcl,c.hansen,chrh,simonsen}@di.ku.dk

Abstract
We contribute the largest publicly available dataset of naturally occurring factual claims for the purpose of automatic claim verification. It is collected from 26 fact checking websites in English, paired with textual sources and rich metadata, and labelled for veracity by human expert journalists. We present an in-depth analysis of the dataset, highlighting characteristics and challenges. Further, we present results for automatic veracity prediction, both with established baselines and with a novel method for joint ranking of evidence pages and predicting veracity that outperforms all baselines. Significant performance increases are achieved by encoding evidence, and by modelling metadata. Our best-performing model achieves a Macro F1 of 49.2%, showing that this is a challenging testbed for claim veracity prediction.

1 Introduction
Misinformation and disinformation are two of the most pertinent and difficult challenges of the information age, exacerbated by the popularity of social media. In an effort to counter this, a significant amount of manual labour has been invested in fact checking claims, often collecting the results of these manual checks on fact checking portals or websites such as politifact.com or snopes.com. In a parallel development, researchers have recently started to view fact checking as a task that can be partially automated, using machine learning and NLP to automatically predict the veracity of claims. However, existing efforts either use small datasets consisting of naturally occurring claims (e.g. Mihalcea and Strapparava (2009); Zubiaga et al. (2016)), or datasets consisting of artificially constructed claims such as FEVER (Thorne et al., 2018). While the latter offer valuable contributions to further automatic claim verification work, they cannot replace real-world datasets.

Contributions. We introduce the currently largest claim verification dataset of naturally occurring claims. It consists of 34,918 claims, collected from 26 fact checking websites in English; evidence pages to verify the claims; the context in which they occurred; and rich metadata (see Table 1 for an example). We perform a thorough analysis to identify characteristics of the dataset such as entities mentioned in claims. We demonstrate the utility of the dataset by training state of the art veracity prediction models, and find that evidence pages as well as metadata significantly contribute to model performance. Finally, we propose a novel model that jointly ranks evidence pages and performs veracity prediction. The best-performing model achieves a Macro F1 of 49.2%, showing that this is a non-trivial dataset with remaining challenges for future work.

Table 1: An example of a claim instance. Entities are obtained via entity linking. Article and outlink texts, evidence search snippets and pages are not shown.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClaimID</td>
<td>farg-00004</td>
</tr>
<tr>
<td>Claim</td>
<td>Mexico and Canada assemble cars with foreign parts and send them to the U.S. with no tax.</td>
</tr>
<tr>
<td>Label</td>
<td>distorts</td>
</tr>
<tr>
<td>Reason</td>
<td>None</td>
</tr>
<tr>
<td>Category</td>
<td>the-factcheck-wire</td>
</tr>
<tr>
<td>Speaker</td>
<td>Donald Trump</td>
</tr>
<tr>
<td>Checker</td>
<td>Eugene Kiely</td>
</tr>
<tr>
<td>Tags</td>
<td>North American Free Trade Agreement</td>
</tr>
<tr>
<td>Claim Entities</td>
<td>United States, Canada, Mexico</td>
</tr>
<tr>
<td>Article Title</td>
<td>FactChecking Trump on Trade</td>
</tr>
<tr>
<td>Publish Date</td>
<td>October 3, 2018</td>
</tr>
<tr>
<td>Claim Date</td>
<td>Monday, October 1, 2018</td>
</tr>
</tbody>
</table>

1The dataset is found here: https://copenlu.github.io/publication/2019_emnlp_augenstein/
2 Related Work

2.1 Datasets

Over the past few years, a variety of mostly small datasets related to fact checking have been released. An overview over core datasets is given in Table 2. The datasets can be grouped into four categories (I–IV). Category I contains datasets aimed at testing how well the veracity3 of a claim can be predicted using the claim alone, without context or evidence documents. Category II contains datasets bundled with documents related to each claim – either topically related to provide context, or serving as evidence. Those documents are, however, not annotated. Category III is for predicting veracity; they encourage retrieving evidence documents as part of their task description, but do not distribute them. Finally, category IV comprises datasets annotated for both veracity and stance. Additional labels can then be added to the datasets to better predict veracity, for instance by jointly training stance and veracity prediction models.

Methods not shown in the table, but related to fact checking, are stance detection for claims (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Augenstein et al., 2016a; Kochkina et al., 2017; Augenstein et al., 2016b; Zubiaga et al., 2018; Riedel et al., 2017), satire detection (Rubin et al., 2016), clickbait detection (Karadzhov et al., 2017), conspiracy news detection (Tacchini et al., 2017), rumour cascade detection (Vosoughi et al., 2018) and claim perspectives detection (Chen et al., 2019).

Contributions: We provide a dataset that, uniquely among extant datasets, contains a large number of naturally occurring claims and rich additional meta-information.

2.2 Methods

Fact checking methods partly depend on the type of dataset used. Methods only taking into account claims typically encode those with CNNs or RNNs (Wang, 2017; Pérez-Rosas et al., 2018), and potentially encode metadata (Wang, 2017) in a similar way. Methods for small datasets often use handcrafted features that are a mix of bag of word and other lexical features, e.g. LIWC, and then use those as input to a SVM or MLP (Mihalcea and Strapparava, 2009; Pérez-Rosas et al., 2018; Baly et al., 2018). Some use additional Twitter-specific features (Enayet and El-Beltagy, 2017). More involved methods taking into account evidence documents, often trained on larger datasets, consist of evidence identification and ranking following a neural model that measures the compatibility between claim and evidence (Thorne et al., 2018; Mihaylova et al., 2018; Yin and Roth, 2018).

Contributions: The latter category above is the most related to our paper as we consider evidence documents. However, existing models are not trained jointly for evidence identification, or for stance and veracity prediction, but rather employ a pipeline approach. Here, we show that a joint approach that learns to weigh evidence pages by their importance for veracity prediction can improve downstream veracity prediction performance.

3 Dataset Construction

We crawled a total of 43,837 claims with their metadata (see details in Table 11). We present the data collection in terms of selecting sources, crawling claims and associated metadata (Section 3.1); retrieving evidence pages; and linking entities in the crawled claims (Section 3.3).

3.1 Selection of sources

We crawled all active fact checking websites in English listed by Duke Reporters’ Lab4 and on the Fact Checking Wikipedia page.5 This resulted in

---

3We use veracity, claim credibility, and fake news prediction interchangeably here – these terms are often conflated in the literature and meant to have the same meaning.

4https://reporterslab.org/fact-checking/

5https://en.wikipedia.org/wiki/Fact_checking
<table>
<thead>
<tr>
<th>Dataset</th>
<th># Claims</th>
<th>Labels</th>
<th>metadata</th>
<th>Claim Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Veracity prediction w/o evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang (2017)</td>
<td>12,836</td>
<td>6</td>
<td>Yes</td>
<td>Politifact</td>
</tr>
<tr>
<td>Pérez-Rosas et al. (2018)</td>
<td>980</td>
<td>2</td>
<td>No</td>
<td>News Websites</td>
</tr>
<tr>
<td>II: Veracity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachenko et al. (2008)</td>
<td>275</td>
<td>2</td>
<td>No</td>
<td>Criminal Reports</td>
</tr>
<tr>
<td>Mihalcea and Strapparava (2009)</td>
<td>600</td>
<td>2</td>
<td>No</td>
<td>Crowd Authors</td>
</tr>
<tr>
<td>Mitra and Gilbert (2015)†</td>
<td>1,049</td>
<td>5</td>
<td>No</td>
<td>Twitter</td>
</tr>
<tr>
<td>Ciampaglia et al. (2015)†</td>
<td>10,000</td>
<td>2</td>
<td>No</td>
<td>Google, Wikipedia</td>
</tr>
<tr>
<td>Popat et al. (2016)</td>
<td>5,013</td>
<td>2</td>
<td>Yes</td>
<td>Wikipedia, Snopes</td>
</tr>
<tr>
<td>Shu et al. (2018)†</td>
<td>23,921</td>
<td>2</td>
<td>Yes</td>
<td>Politifact, gossipcop.com</td>
</tr>
<tr>
<td>Datacommons Fact Check2</td>
<td>10,564</td>
<td>2-6</td>
<td>Yes</td>
<td>Fact Checking Websites</td>
</tr>
<tr>
<td>III: Veracity (evidence encouraged, but not provided)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barn-Cedeo et al. (2018)</td>
<td>150</td>
<td>3</td>
<td>No</td>
<td>factcheck.org, Snopes</td>
</tr>
<tr>
<td>IV: Veracity + stance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vlachos and Riedel (2014)</td>
<td>106</td>
<td>5</td>
<td>Yes</td>
<td>Politifact, Channel 4 News</td>
</tr>
<tr>
<td>Zubiaga et al. (2016)</td>
<td>330</td>
<td>3</td>
<td>Yes</td>
<td>Twitter</td>
</tr>
<tr>
<td>Derczynski et al. (2017)</td>
<td>325</td>
<td>3</td>
<td>Yes</td>
<td>Twitter</td>
</tr>
<tr>
<td>Baly et al. (2018)</td>
<td>422</td>
<td>2</td>
<td>No</td>
<td>ara.reuters.com, verify-sy.com</td>
</tr>
<tr>
<td>Thorne et al. (2018)†</td>
<td>185,445</td>
<td>3</td>
<td>No</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>V: Veracity + evidence relevancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiFC</td>
<td>36,534</td>
<td>2-40</td>
<td>Yes</td>
<td>Fact Checking Websites</td>
</tr>
</tbody>
</table>

Table 2: Comparison of fact checking datasets. † indicates claims are not ‘naturally occuring’: Mitra and Gilbert (2015) use events as claims; Ciampaglia et al. (2015) use DBPedia tiples as claims; Shu et al. (2018) use tweets as claims; and Thorne et al. (2018) rewrite sentences in Wikipedia as claims.

There were a number of crawling issues, e.g. security protection of websites with SSL/TLS protocols, time out, URLs that pointed to pdf files instead of HTML content, or unresolvable encoding. In all of these cases, the content could not be retrieved. For some websites, no veracity labels were available, in which case, they were not selected as domains for training a veracity prediction model. Moreover, not all types of metadata (category, speaker, checker, tags, claim date, publish date) were available for all websites; and availability of articles and full texts differs as well.

We performed semi-automatic cleansing of the dataset as follows. First, we double-checked that the veracity labels would not appear in claims. For some domains, the first or last sentence of the claim would sometimes contain the veracity label, in which case we would discard either the full sentence or part of the sentence. Next, we checked the dataset for duplicate claims. We found 202 such instances, 69 of them with different labels. Upon manual inspection, this was mainly due to them appearing on different websites, with labels not differing much in practice (e.g. ‘Not true’, vs. ‘Mostly False’). We made sure that all such duplicate claims would be in the training split of the dataset, so that the models would not have an unfair advantage. Finally, we performed some minor manual merging of label types for the same domain where it was clear that they were supposed to denote the same level of veracity (e.g. ‘distorts’, ‘distorts the facts’).

This resulted in a total of 36,534 claims with their metadata. For the purposes of fact verification, we discarded instances with labels that occur fewer than 5 times, resulting in 34,918 claims. The number of instances, as well as labels per domain, are shown in Table 6 and label names in Table 10 in the appendix. The dataset is split into a training part (80%) and a development and testing part (10% each) in a label-stratified manner. Note that
the domains vary in the number of labels, ranging from 2 to 27. Labels include both straight-forward ratings of veracity (‘correct’, ‘incorrect’), but also labels that would be more difficult to map onto a veracity scale (e.g. ‘grass roots movement!’, ‘mis-attributed’, ‘not the whole story’). We therefore do not postprocess label types across domains to map them onto the same scale, and rather treat them as is. In the methodology section (Section 4), we show how a model can be trained on this dataset regardless by framing this multi-domain veracity prediction task as a multi-task learning (MTL) one.

3.2 Retrieving Evidence Pages

The text of each claim is submitted verbatim as a query to the Google Search API (without quotes). The 10 most highly ranked search results are retrieved, for each of which we save the title; Google search rank; URL; time stamp of last update; search snippet; as well as the full Web page. We acknowledge that search results change over time, which might have an effect on veracity prediction. However, studying such temporal effects is outside the scope of this paper. Similar to Web crawling claims, as described in Section 3.1, the corresponding Web pages can in some cases not be retrieved, in which case fewer than 10 evidence pages are available. The resulting evidence pages are from a wide variety of URL domains, though with a predictable skew towards popular websites, such as Wikipedia or The Guardian (see Table 3 for detailed statistics).

3.3 Entity Detection and Linking

To better understand what claims are about, we conduct entity linking for all claims. Specifically, mentions of people, places, organisations, and other named entities within a claim are recognised and linked to their respective Wikipedia pages, if available. Where there are different entities with the same name, they are disambiguated. For this, we apply the state-of-the-art neural entity linking model by Kolitsas et al. (2018). This results in a total of 25,763 entities detected and linked to Wikipedia, with a total of 15,351 claims involved, meaning that 42% of all claims contain entities that can be linked to Wikipedia. Later on, we use entities as additional metadata (see Section 4.3). The distribution of claim numbers according to the number of entities they contain is shown in Figure 1. We observe that the majority of claims have one to four entities, and the maximum number of 35 entities occurs in one claim only. Out of the 25,763 entities, 2,767 are unique entities. The top 30 most frequent entities are listed in Table 4. This clearly shows that most of the claims involve entities related to the United States, which is to be expected, as most of the fact checking websites are US-based.

### Table 3: The top 30 most frequently occurring URL domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.snopes.com/">https://www.snopes.com/</a></td>
<td>3.992</td>
</tr>
<tr>
<td><a href="https://www.washingtonpost.com/">https://www.washingtonpost.com/</a></td>
<td>3.025</td>
</tr>
<tr>
<td><a href="https://www.nytimes.com/">https://www.nytimes.com/</a></td>
<td>2.478</td>
</tr>
<tr>
<td><a href="https://www.theguardian.com/">https://www.theguardian.com/</a></td>
<td>1.807</td>
</tr>
<tr>
<td><a href="https://www.youtube.com/">https://www.youtube.com/</a></td>
<td>1.712</td>
</tr>
<tr>
<td><a href="https://www.dailymail.co.uk/">https://www.dailymail.co.uk/</a></td>
<td>1.558</td>
</tr>
<tr>
<td><a href="https://www.usatoday.com/">https://www.usatoday.com/</a></td>
<td>1.279</td>
</tr>
<tr>
<td><a href="https://www.politico.com/">https://www.politico.com/</a></td>
<td>1.241</td>
</tr>
<tr>
<td><a href="http://www.politifact.com/">http://www.politifact.com/</a></td>
<td>1.231</td>
</tr>
<tr>
<td><a href="https://www.pinterest.com/">https://www.pinterest.com/</a></td>
<td>1.169</td>
</tr>
<tr>
<td><a href="https://www.factcheck.org/">https://www.factcheck.org/</a></td>
<td>1.09</td>
</tr>
<tr>
<td><a href="https://www.gossipcop.com/">https://www.gossipcop.com/</a></td>
<td>1.073</td>
</tr>
<tr>
<td><a href="https://www.cnn.com/">https://www.cnn.com/</a></td>
<td>1.065</td>
</tr>
<tr>
<td><a href="https://www.npr.org/">https://www.npr.org/</a></td>
<td>0.957</td>
</tr>
<tr>
<td><a href="https://www.forbes.com/">https://www.forbes.com/</a></td>
<td>0.911</td>
</tr>
<tr>
<td><a href="https://www.vox.com/">https://www.vox.com/</a></td>
<td>0.89</td>
</tr>
<tr>
<td><a href="https://www.theatlantic.com/">https://www.theatlantic.com/</a></td>
<td>0.88</td>
</tr>
<tr>
<td><a href="https://twitter.com/">https://twitter.com/</a></td>
<td>0.767</td>
</tr>
<tr>
<td><a href="https://www.hoax-slayer.net/">https://www.hoax-slayer.net/</a></td>
<td>0.655</td>
</tr>
<tr>
<td><a href="http://time.com/">http://time.com/</a></td>
<td>0.554</td>
</tr>
<tr>
<td><a href="https://www.bbc.com/">https://www.bbc.com/</a></td>
<td>0.551</td>
</tr>
<tr>
<td><a href="https://www.nbcnews.com/">https://www.nbcnews.com/</a></td>
<td>0.515</td>
</tr>
<tr>
<td><a href="https://www.cnbc.com/">https://www.cnbc.com/</a></td>
<td>0.514</td>
</tr>
<tr>
<td><a href="https://www.cbsnews.com/">https://www.cbsnews.com/</a></td>
<td>0.503</td>
</tr>
<tr>
<td><a href="https://www.facebook.com/">https://www.facebook.com/</a></td>
<td>0.5</td>
</tr>
<tr>
<td><a href="https://www.newyorker.com/">https://www.newyorker.com/</a></td>
<td>0.495</td>
</tr>
<tr>
<td><a href="https://www.foxnews.com/">https://www.foxnews.com/</a></td>
<td>0.468</td>
</tr>
<tr>
<td><a href="https://people.com/">https://people.com/</a></td>
<td>0.439</td>
</tr>
<tr>
<td><a href="http://www.cnn.com/">http://www.cnn.com/</a></td>
<td>0.419</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of entities in claims.

4 Claim Veracity Prediction

We train several models to predict the veracity of claims. Those fall into two categories: those that
the multi-task learning (MTL) approach inspired by collaborative filtering presented in Augenstein et al. (2018) (MTL with LEL–multitask learning with label embedding layer) that excels on pairwise sequence classification tasks with disparate label spaces. More concretely, each domain is modelled as its own task in a MTL architecture, and labels are projected into a fixed-length label embedding space. Predictions are then made by taking the dot product between the claim-evidence embeddings and the label embeddings. By doing so, the model implicitly learns how semantically close the labels are to one another, and can benefit from this knowledge when making predictions for individual tasks, which on their own might only have a small number of instances. When making predictions for individual domains/tasks, both at training and at test time, as well as when calculating the loss, a mask is applied such that the valid and invalid labels for that task are restricted to the set of known task labels.

Note that the setting here slightly differs from Augenstein et al. (2018). There, tasks are less strongly related to one another; for example, they consider stance detection, aspect-based sentiment analysis and natural language inference. Here, we have different domains, as opposed to conceptually different tasks, but use their framework, as we have the same underlying problem of disparate label spaces. A more formal problem definition follows next, as our evidence ranking and veracity prediction model in Section 4.2 then builds on it.

4.1 Multi-Domain Claim Veracity Prediction with Disparate Label Spaces

Since not all fact checking websites use the same claim labels (see Table 6, and Table 10 in the appendix), training a claim veracity prediction model is not entirely straight-forward. One option would be to manually map those labels onto one another. However, since the sheer number of labels is rather large (165), and it is not always clear from the guidelines on fact checking websites how they can be mapped onto one another, we opt to learn how these labels relate to one another as part of the veracity prediction model. To do so, we employ

Table 4: Top 30 most frequent entities listed by their Wikipedia URL with prefix omitted

<table>
<thead>
<tr>
<th>Entity</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>United_States</td>
<td>2810</td>
</tr>
<tr>
<td>Barack_Obama</td>
<td>1598</td>
</tr>
<tr>
<td>Republican_Party_(United_States)</td>
<td>783</td>
</tr>
<tr>
<td>Texas</td>
<td>665</td>
</tr>
<tr>
<td>Democratic_Party_(United_States)</td>
<td>560</td>
</tr>
<tr>
<td>Donald_Trump</td>
<td>556</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>471</td>
</tr>
<tr>
<td>United_States_Congress</td>
<td>354</td>
</tr>
<tr>
<td>Hillary_Rodham_Clinton</td>
<td>306</td>
</tr>
<tr>
<td>Bill_Clinton</td>
<td>292</td>
</tr>
<tr>
<td>California</td>
<td>285</td>
</tr>
<tr>
<td>Russia</td>
<td>275</td>
</tr>
<tr>
<td>Ohio</td>
<td>239</td>
</tr>
<tr>
<td>China</td>
<td>229</td>
</tr>
<tr>
<td>George_W_Bush</td>
<td>208</td>
</tr>
<tr>
<td>Medicare_(United_States)</td>
<td>206</td>
</tr>
<tr>
<td>Australia</td>
<td>186</td>
</tr>
<tr>
<td>Iran</td>
<td>183</td>
</tr>
<tr>
<td>Brad_Pitt</td>
<td>180</td>
</tr>
<tr>
<td>Islam</td>
<td>178</td>
</tr>
<tr>
<td>Iraq</td>
<td>176</td>
</tr>
<tr>
<td>Canada</td>
<td>174</td>
</tr>
<tr>
<td>White_House</td>
<td>166</td>
</tr>
<tr>
<td>New_York_City</td>
<td>164</td>
</tr>
<tr>
<td>Washington_D.C.</td>
<td>164</td>
</tr>
<tr>
<td>Jennifer_Aniston</td>
<td>163</td>
</tr>
<tr>
<td>Mexico</td>
<td>158</td>
</tr>
<tr>
<td>Ted_Cruz</td>
<td>152</td>
</tr>
<tr>
<td>Federal_Bureau_of_Investigation</td>
<td>146</td>
</tr>
<tr>
<td>Syria</td>
<td>130</td>
</tr>
</tbody>
</table>

only consider the claims themselves, and those that encode evidence pages as well. In addition, claim metadata (speaker, checker, linked entities) is optionally encoded for both categories of models, and ablation studies with and without that metadata are shown. We first describe the base model used in Section 4.1, followed by introducing our novel evidence ranking and veracity prediction model in Section 4.2, and lastly the metadata encoding model in Section 4.3.

4.1 Multi-Domain Claim Veracity Prediction with Disparate Label Spaces

The base model is a classic deep neural network MTL model (Caruana, 1993) that shares its parameters across tasks and has task-specific softmax output layers that output a probability distribution \( p^T_i \) for task \( T_i \):

\[
p^T_i = \text{softmax}(W^T_i h + b^T_i) \tag{1}
\]

where \( \text{softmax}(x) = e^x / \sum_i |x_i| e^x_i \), \( W^T_i \in \mathbb{R}^{L_i \times h} \), \( b^T_i \in \mathbb{R}^{L_i} \) is the weight matrix and bias term of the output layer of task \( T_i \) respectively, \( h \in \mathbb{R}^h \) is the jointly learned hidden rep-
presentation, $L_t$ is the number of labels for task $T_t$, and $h$ is the dimensionality of $h$. The MTL model is trained to minimise the sum of individual task losses $L_1 + \ldots + L_T$ using a negative log-likelihood objective.

Label Embedding Layer. To learn the relationships between labels, a Label Embedding Layer (LEL) embeds labels of all tasks in a joint Euclidian space. Instead of training separate softmax output layers as above, a label compatibility function $c(\cdot, \cdot)$ measures how similar a label with embedding $l$ is to the hidden representation $h$:

$$c(l, h) = 1 \cdot h$$

where $\cdot$ is the dot product. Padding is applied such that $l$ and $h$ have the same dimensionality. Matrix multiplication and softmax are used for making predictions:

$$p = \text{softmax}(Lh)$$

where $L \in \mathbb{R}^{(\sum_t L_t) \times l}$ is the label embedding matrix for all tasks and $l$ is the dimensionality of the label embeddings. We apply a task-specific mask to $L$ in order to obtain a task-specific probability distribution $p^{T_t}$. The LEL is shared across all tasks, which allows the model to learn the relationships between labels in the joint embedding space.

4.2 Joint Evidence Ranking and Claim Veracity Prediction

So far, we have ignored the issue of how to obtain claim representation, as the base model described in the previous section is agnostic to how instances are encoded. A very simple approach, which we report as a baseline, is to encode claim texts only. Such a model ignores evidence for and against a claim, and ends up guessing the veracity based on surface patterns observed in the claim texts.

We next introduce two variants of evidence-based veracity prediction models that encode 10 pieces of evidence in addition to the claim. Here, we opt to encode search snippets as opposed to whole retrieved pages. While the latter would also be possible, it comes with a number of additional challenges, such as encoding large documents, parsing tables or PDF files, and encoding images or videos on these pages, which we leave to future work. Search snippets also have the benefit that they already contain summaries of the part of the page content that is most related to the claim.

4.2.1 Problem Definition

Our problem is to obtain encodings for $N$ examples $X_{T_t} = \{x_{T_t}^1, \ldots, x_{T_t}^N\}$. For simplicity, we will henceforth drop the task superscript and refer to instances as $X = \{x_1, \ldots, x_N\}$, as instance encodings are learned in a task-agnostic fashion. Each example further consists of a claim $a_i$ and $k = 10$ evidence pages $E_k = \{e_{10}, \ldots, e_{N10}\}$.

Each claim and evidence page is encoded with a BiLSTM to obtain a sentence embedding, which is the concatenation of the last state of the forward and backward reading of the sentence, i.e. $h = BiLSTM(\cdot)$, where $h$ is the sentence embedding.

Next, we want to combine claims and evidence sentence embeddings into joint instance representations. In the simplest case, referred to as model variant crawled_avg, we mean average the BiLSTM sentence embeddings of all evidence pages (signified by the overline) and concatenate those with the claim embeddings, i.e.

$$s_{gi} = [h_{a_i}; \bar{h}_{E_i}]$$

where $s_{gi}$ is the resulting encoding for training example $i$ and $[\cdot; \cdot]$ denotes vector concatenation.
However, this has the disadvantage that all evidence pages are considered equal.

Evidence Ranking  The here proposed alternative instance encoding model, *crawled_ranked*, which achieves the highest overall performance as discussed in Section 5, learns the compatibility between an instance’s claim and each evidence page. It ranks evidence pages by their utility for the veracity prediction task, and then uses the resulting ranking to obtain a weighted combination of all claim-evidence pairs. No direct labels are available to learn the ranking of individual documents, only for the veracity of the associated claim, so the model has to learn evidence ranks implicitly.

To combine claim and evidence representations, we use the matching model proposed for the task of natural language inference by Mou et al. (2016) and adapt it to combine an instance’s claim representation with each evidence representation, i.e.

\[
s_{r_{ij}} = [h_{a_i}; h_{e_{ij}}; h_{a_i} - h_{e_{ij}}; h_{a_i} \cdot h_{e_{ij}}]
\]

where \(s_{r_{ij}}\) is the resulting encoding for training example \(i\) and evidence page \(j\), \([::]\) denotes vector concatenation, and \(\cdot\) denotes the dot product.

All joint claim-evidence representations \(s_{r_0}, \ldots, s_{r_{10}}\) are then projected into the binary space via a fully connected layer FC, followed by a non-linear activation function \(f\), to obtain a soft ranking of claim-evidence pairs, in practice a 10-dimensional vector,

\[
o_i = [f(FC(s_{r_0})); \ldots; f(FC(s_{r_{10}}))]
\]

where \([::]\) denotes concatenation.

Scores for all labels are obtained as per (6) above, with the same input instance embeddings as for the evidence ranker, i.e. \(s_{r_{ij}}\). Final predictions for all claim-evidence pairs are then obtained by taking the dot product between the label scores and binary evidence ranking scores, i.e.

\[
p_i = \text{softmax}(c(1, s_{r_i}) \cdot o_i)
\]

Note that the novelty here is that, unlike for the model described in Mou et al. (2016), we have no direct labels for learning weights for this matching model. Rather, our model has to implicitly learn these weights for each claim-evidence pair in an end-to-end fashion given the veracity labels.

<table>
<thead>
<tr>
<th>Model</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>claim-only</td>
<td>0.469</td>
<td>0.253</td>
</tr>
<tr>
<td>claim-only_embavg</td>
<td>0.384</td>
<td>0.302</td>
</tr>
<tr>
<td>crawled-docavg</td>
<td>0.438</td>
<td>0.248</td>
</tr>
<tr>
<td>crawled_ranked</td>
<td>0.613</td>
<td>0.441</td>
</tr>
<tr>
<td>claim-only + meta</td>
<td>0.494</td>
<td>0.324</td>
</tr>
<tr>
<td>claim-only_embavg + meta</td>
<td>0.418</td>
<td>0.333</td>
</tr>
<tr>
<td>crawled-docavg + meta</td>
<td>0.483</td>
<td>0.286</td>
</tr>
<tr>
<td>crawled_ranked + meta</td>
<td>0.625</td>
<td>0.492</td>
</tr>
</tbody>
</table>

Table 5: Results with different model variants on the test set, ‘meta’ means all metadata is used.

5 Experiments

5.1 Experimental Setup

The base sentence embedding model is a BiLSTM over all words in the respective sequences with randomly initialised word embeddings, following Augenstein et al. (2018). We opt for this strong baseline sentence encoding model, as opposed to engineering sentence embeddings that work particularly well for this dataset, to showcase the dataset. We would expect pre-trained contextual encoding models, e.g. ELMo (Peters et al., 2018), ULMFit (Howard and Ruder, 2018), BERT (Devlin et al., 2018), to offer complementary performance gains, as has been shown for a few recent papers (Wang et al., 2018; Rajpurkar et al., 2018).

For claim veracity prediction without evidence documents with the MTL with LEL model, we use the following sentence encoding variants: *claim-
only, which uses a BiLSTM-based sentence embedding as input, and claim-only_embavg, which uses a sentence embedding based on mean averaged word embeddings as input.

We train one multi-task model per task (i.e., one model per domain). We perform a grid search over the following hyperparameters, tuned on the respective dev set, and evaluate on the corresponding test set (final settings are underlined): word embedding size \([64, 128, 256]\), BiLSTM hidden layer size \([64, 128, 256]\), number of BiLSTM hidden layers \([1, 2, 3]\), BiLSTM dropout on input and output layers \([0.0, 0.1, 0.2, 0.5]\), word-by-word-attention for BiLSTM with window size 10 (Bahdanau et al., 2014) \([\text{True, False}]\), skip-connections for the BiLSTM \([\text{True, False}]\), batch size \([32, 64, 128]\), label embedding size \([16, 32, 64]\). We use ReLU as an activation function for both the BiLSTM and the CNN. For the CNN, the following hyperparameters are used: number filters \([32, 64, 128]\), kernel size \([32, 64]\). We train using cross-entropy loss and the RMSProp optimiser with initial learning rate of 0.001 and perform early stopping on the dev set with a patience of 3.

5.2 Results

For each domain, we compute the Micro as well as Macro F1, then mean average results over all domains. Core results with all vs. no metadata are shown in Table 5. We first experiment with different base model variants and find that label embeddings improve results, and that the best proposed models utilising multiple domains outperform single-task models (see Table 8). This corroborates the findings of Augenstein et al. (2018). Per-domain results with the best model are shown in Table 6. Domain names are from hereon after abbreviated for brevity, see Table 11 in the appendix for correspondences to full website names. Unsurprisingly, it is hard to achieve a high Macro F1 for domains with many labels, e.g. tron and snes. Further, some domains, surprisingly mostly with small numbers of instances, seem to be very easy – a perfect Micro and Macro F1 score of 1.0 is achieved on ranz, bove, buca, fani and thal. We find that for those domains, the verdict is often already revealed as part of the claim using explicit wording.

Claim-Only vs. Evidence-Based Veracity Prediction. Our evidence-based claim veracity prediction models outperform claim-only veracity prediction models by a large margin. Unsurprisingly, claim-only_embavg is outperformed by claim-only. Further, crawled_ranked is our best-performing model in terms of Micro F1 and Macro F1, meaning that our model captures that not every piece of evidence is equally important, and can

<table>
<thead>
<tr>
<th>Domain</th>
<th># Insts</th>
<th># Labs</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranz</td>
<td>21</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>bove</td>
<td>295</td>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>abbc</td>
<td>436</td>
<td>3</td>
<td>0.463</td>
<td>0.453</td>
</tr>
<tr>
<td>huca</td>
<td>34</td>
<td>3</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>mpws</td>
<td>47</td>
<td>3</td>
<td>0.667</td>
<td>0.583</td>
</tr>
<tr>
<td>peck</td>
<td>65</td>
<td>3</td>
<td>0.627</td>
<td>0.472</td>
</tr>
<tr>
<td>faan</td>
<td>111</td>
<td>3</td>
<td>0.682</td>
<td>0.679</td>
</tr>
<tr>
<td>clot</td>
<td>38</td>
<td>3</td>
<td>0.833</td>
<td>0.619</td>
</tr>
<tr>
<td>fani</td>
<td>20</td>
<td>3</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>chet</td>
<td>355</td>
<td>4</td>
<td>0.550</td>
<td>0.513</td>
</tr>
<tr>
<td>obry</td>
<td>59</td>
<td>4</td>
<td>0.417</td>
<td>0.268</td>
</tr>
<tr>
<td>vees</td>
<td>504</td>
<td>4</td>
<td>0.721</td>
<td>0.425</td>
</tr>
<tr>
<td>faly</td>
<td>111</td>
<td>5</td>
<td>0.278</td>
<td>0.5</td>
</tr>
<tr>
<td>goop</td>
<td>2948</td>
<td>6</td>
<td>0.822</td>
<td>0.387</td>
</tr>
<tr>
<td>pose</td>
<td>1361</td>
<td>6</td>
<td>0.438</td>
<td>0.328</td>
</tr>
<tr>
<td>thal</td>
<td>79</td>
<td>6</td>
<td>0.55</td>
<td>0.37</td>
</tr>
<tr>
<td>auck</td>
<td>433</td>
<td>7</td>
<td>0.357</td>
<td>0.259</td>
</tr>
<tr>
<td>hoer</td>
<td>1310</td>
<td>7</td>
<td>0.694</td>
<td>0.549</td>
</tr>
<tr>
<td>para</td>
<td>222</td>
<td>7</td>
<td>0.375</td>
<td>0.311</td>
</tr>
<tr>
<td>wasp</td>
<td>201</td>
<td>7</td>
<td>0.344</td>
<td>0.214</td>
</tr>
<tr>
<td>vogo</td>
<td>654</td>
<td>8</td>
<td>0.594</td>
<td>0.297</td>
</tr>
<tr>
<td>point</td>
<td>15390</td>
<td>9</td>
<td>0.321</td>
<td>0.276</td>
</tr>
<tr>
<td>snes</td>
<td>6455</td>
<td>12</td>
<td>0.551</td>
<td>0.097</td>
</tr>
<tr>
<td>farg</td>
<td>485</td>
<td>11</td>
<td>0.500</td>
<td>0.140</td>
</tr>
<tr>
<td>tron</td>
<td>3423</td>
<td>27</td>
<td>0.429</td>
<td>0.046</td>
</tr>
<tr>
<td>avg</td>
<td>7.17</td>
<td>0.625</td>
<td>0.492</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Total number of instances and unique labels per domain, as well as per-domain results with model crawled_ranked + meta, sorted by label size

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.627</td>
<td>0.441</td>
</tr>
<tr>
<td>Speaker  + Tags</td>
<td>0.602</td>
<td>0.435</td>
</tr>
<tr>
<td>Tags</td>
<td>0.585</td>
<td>0.461</td>
</tr>
<tr>
<td>Entity   + Speaker + Tags</td>
<td>0.569</td>
<td>0.427</td>
</tr>
<tr>
<td></td>
<td>0.607</td>
<td>0.477</td>
</tr>
<tr>
<td></td>
<td>0.625</td>
<td>0.492</td>
</tr>
</tbody>
</table>

Table 7: Ablation results with base model crawled_ranked for different types of metadata

<table>
<thead>
<tr>
<th>Model</th>
<th>Micro F1</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>0.527</td>
<td>0.388</td>
</tr>
<tr>
<td>MTL</td>
<td>0.556</td>
<td>0.448</td>
</tr>
<tr>
<td>MTL + LEL</td>
<td>0.625</td>
<td>0.492</td>
</tr>
</tbody>
</table>

Table 8: Ablation results with crawled_ranked + meta encoding for STL vs. MTL vs. MTL + LEL training
utilise this for veracity prediction.

**Metadata.** We perform an ablation analysis of how metadata impacts results, shown in Table 7. Out of the different types of metadata, topic tags on their own contribute the most. This is likely because they offer highly complementary information to the claim text of evidence pages. Only using all metadata together achieves a higher Macro F1 at similar Micro F1 than using no metadata at all. To further investigate this, we split the test set into those instances for which no metadata is available vs. those for which metadata is available. We find that encoding metadata within the model hurts performance for domains where no metadata is available, but improves performance where it is. In practice, an ensemble of both types of models would be sensible, as well as exploring more involved methods of encoding metadata.

6 Analysis and Discussion

An analysis of labels frequently confused with one another, for the largest domain 'pomt' and best-performing model crawled_ranked + meta is shown in Figure 3. The diagonal represents when gold and predicted labels match, and the numbers signify the number of test instances. One can observe that the model struggles more to detect claims with labels ‘true’ than those with label ‘false’. Generally, many confusions occur over close labels, e.g. ‘half-true’ vs. ‘mostly true’.

We further analyse what properties instances that are predicted correctly vs. incorrectly have, using the model crawled_ranked meta. We find that, unsurprisingly, longer claims are harder to classify correctly, and that claims with a high direct token overlap with evidence pages lead to a high evidence ranking. When it comes to frequently occurring tags and entities, very general tags such as ‘government-and-politics’ or ‘tax’ that do not give away much, frequently co-occur with incorrect predictions, whereas more specific tags such as ‘brisbane-4000’ or ‘hong-kong’ tend to co-occur with correct predictions. Similar trends are observed for bigrams. This means that the model has an easy time succeeding for instances where the claims are short, where specific topics tend to co-occur with certain veracities, and where evidence documents are highly informative. Instances with longer, more complex claims where evidence is ambiguous remain challenging.

7 Conclusions

We present a new, real-world fact checking dataset, currently the largest of its kind. It consists of 34,918 claims collected from 26 fact checking websites, rich metadata and 10 retrieved evidence pages per claim. We find that encoding the metadata as well evidence pages helps, and introduce a new joint model for ranking evidence pages and predicting veracity.

Acknowledgments

This research is partially supported by QUARTZ (721321, EU H2020 MSCA-ITN) and DABAI (5153-00004A, Innovation Fund Denmark).

References


Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In ACL (2). The Association for Computer Linguistics.


<table>
<thead>
<tr>
<th>Websites (Sources)</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mediabiasfactcheck</td>
<td>Website that checks other news websites</td>
</tr>
<tr>
<td>CBC</td>
<td>No pattern to crawl</td>
</tr>
<tr>
<td>apnews.com/APFactCheck</td>
<td>No categorical label and no structured claim</td>
</tr>
<tr>
<td>weeklystandard.com/tag/fact-check</td>
<td>Mostly no label, and they are placed anywhere</td>
</tr>
<tr>
<td>ballotpedia.org</td>
<td>No categorical label and no structured claim</td>
</tr>
<tr>
<td>channel3000.com/news/politics/reality-check</td>
<td>No categorical label, lack of structure, and no clear claim</td>
</tr>
<tr>
<td>npr.org/sections/politics-fact-check</td>
<td>No label and no clear claim (only some titles are claims)</td>
</tr>
<tr>
<td>dailycaller.com/buzz/check-your-fact</td>
<td>Is a subset of checkyourfact which has already been crawled</td>
</tr>
<tr>
<td>sacbee.com</td>
<td>Contains very few labelled articles, and without clear claims</td>
</tr>
<tr>
<td>TheGuardian</td>
<td>Only a few websites have a pattern for labels.</td>
</tr>
</tbody>
</table>

Table 9: The list of websites that we did not crawl and reasons for not crawling them.

<table>
<thead>
<tr>
<th>Domain</th>
<th># Insts</th>
<th># Labels</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>abbc</td>
<td>436</td>
<td>3</td>
<td>in-between, in-the-red, in-the-green</td>
</tr>
<tr>
<td>afck</td>
<td>433</td>
<td>7</td>
<td>correct, incorrect, mostly-correct, unproven, misleading, understate, exaggerated</td>
</tr>
<tr>
<td>bove</td>
<td>295</td>
<td>2</td>
<td>none, rating: false</td>
</tr>
<tr>
<td>chct</td>
<td>355</td>
<td>4</td>
<td>verdict: true, verdict: false, verdict: unsubstantiated, none</td>
</tr>
<tr>
<td>clck</td>
<td>38</td>
<td>3</td>
<td>incorrect, unsupported, misleading</td>
</tr>
<tr>
<td>faan</td>
<td>111</td>
<td>3</td>
<td>factscan score: false, factscan score: true, factscan score: misleading</td>
</tr>
<tr>
<td>faly</td>
<td>71</td>
<td>5</td>
<td>true, none, partly true, unverified, false</td>
</tr>
<tr>
<td>fani</td>
<td>20</td>
<td>3</td>
<td>conclusion: accurate, conclusion: false, conclusion: unclear</td>
</tr>
<tr>
<td>farg</td>
<td>485</td>
<td>11</td>
<td>false, none, distorts the facts, misleading, spins the facts, no evidence, not the whole story, unsupported, cherry picks, exaggerates, out of context</td>
</tr>
<tr>
<td>goop</td>
<td>2943</td>
<td>6</td>
<td>0, 1, 2, 3, 4, 10</td>
</tr>
<tr>
<td>huer</td>
<td>1310</td>
<td>7</td>
<td>facebook scams, true messages, bogus warning, statistcal reports, fake news, unsubstantiated messages, misleading recommendations</td>
</tr>
<tr>
<td>huca</td>
<td>34</td>
<td>3</td>
<td>a lot of baloney, a little baloney, some baloney</td>
</tr>
<tr>
<td>mpws</td>
<td>47</td>
<td>3</td>
<td>accurate, false, misleading</td>
</tr>
<tr>
<td>obry</td>
<td>59</td>
<td>4</td>
<td>mostly_true, verified, unobservable, mostly_false</td>
</tr>
<tr>
<td>para</td>
<td>222</td>
<td>7</td>
<td>mostly false, mostly true, half-true, false, true, pants on fire!, half flip</td>
</tr>
<tr>
<td>peck</td>
<td>65</td>
<td>3</td>
<td>false, true, partially true</td>
</tr>
<tr>
<td>pont</td>
<td>15390</td>
<td>9</td>
<td>half-true, false, mostly true, mostly false, true, pants on fire!, full flop, half flip, no flip</td>
</tr>
<tr>
<td>pose</td>
<td>1361</td>
<td>6</td>
<td>promise kept, promise broken, compromise, in the works, not yet rated, stalled</td>
</tr>
<tr>
<td>ranz</td>
<td>21</td>
<td>2</td>
<td>fact, fiction</td>
</tr>
<tr>
<td>snes</td>
<td>6455</td>
<td>12</td>
<td>false, true, mixture, unproven, mostly false, mostly true, miscaptioned, legend, outdated, misattributed, scam, correct attribution</td>
</tr>
<tr>
<td>thet</td>
<td>79</td>
<td>6</td>
<td>none, mostly false, mostly true, half true, false, true</td>
</tr>
<tr>
<td>thal</td>
<td>74</td>
<td>2</td>
<td>none, we rate this claim false</td>
</tr>
<tr>
<td>tron</td>
<td>3423</td>
<td>27</td>
<td>fiction!, truth!, unproven!, truth! &amp; fiction!, mostly fiction!, none, disputed!, truth! &amp; misleading!, authorship confirmed!, mostly truth!, incorrect attribution!, scam!, investigation pending!, confirmed authorship!, commentary!, previously true! now resolved!, outdated!, truth! &amp; outdated!, virus!, fiction! &amp; satire!, truth! &amp; unproven!, misleading!, grass roots movement!, opinion!, correct attribution!, truth! &amp; disputed!, inaccurate attribution!</td>
</tr>
<tr>
<td>vees</td>
<td>504</td>
<td>4</td>
<td>none, fake, misleading, false</td>
</tr>
<tr>
<td>wast</td>
<td>201</td>
<td>7</td>
<td>4 pinnochios, 3 pinnochios, 2 pinnochios, false, not the whole story, needs context, none</td>
</tr>
</tbody>
</table>

Table 10: Number of instances, and labels per domain sorted by number of occurrences
<table>
<thead>
<tr>
<th>Website</th>
<th>Domain</th>
<th>Claims</th>
<th>Labels</th>
<th>Category</th>
<th>Speaker</th>
<th>Checker</th>
<th>Tags</th>
<th>Article</th>
<th>Claim date</th>
<th>Publish date</th>
<th>Full text</th>
<th>Outlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>abc</td>
<td>abbc</td>
<td>436</td>
<td>436</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>436</td>
<td>-</td>
<td>436</td>
<td>436</td>
<td>436</td>
<td>7676</td>
</tr>
<tr>
<td>africacheck</td>
<td>afck</td>
<td>436</td>
<td>436</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>436</td>
<td>-</td>
<td>436</td>
<td>436</td>
<td>436</td>
<td>2325</td>
</tr>
<tr>
<td>altnews</td>
<td>-</td>
<td>496</td>
<td>-</td>
<td>-</td>
<td>496</td>
<td>-</td>
<td>496</td>
<td>-</td>
<td>496</td>
<td>496</td>
<td>496</td>
<td>6389</td>
</tr>
<tr>
<td>boomlive</td>
<td>-</td>
<td>302</td>
<td>302</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>302</td>
<td>-</td>
<td>302</td>
<td>302</td>
<td>302</td>
<td>6054</td>
</tr>
<tr>
<td>checkyourfact</td>
<td>chht</td>
<td>358</td>
<td>358</td>
<td>-</td>
<td>358</td>
<td>-</td>
<td>358</td>
<td>-</td>
<td>358</td>
<td>358</td>
<td>358</td>
<td>5271</td>
</tr>
<tr>
<td>climatefeedback</td>
<td>click</td>
<td>45</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45</td>
<td>-</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>489</td>
</tr>
<tr>
<td>crikey</td>
<td>-</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>-</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>212</td>
</tr>
<tr>
<td>factcheckni</td>
<td>-</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>-</td>
<td>-</td>
<td>36</td>
<td>-</td>
<td>36</td>
<td>-</td>
<td>36</td>
<td>151</td>
</tr>
<tr>
<td>factcheckorg</td>
<td>farg</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>8282</td>
</tr>
<tr>
<td>factly</td>
<td>-</td>
<td>77</td>
<td>77</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>77</td>
<td>-</td>
<td>77</td>
<td>-</td>
<td>77</td>
<td>658</td>
</tr>
<tr>
<td>factscan</td>
<td>-</td>
<td>115</td>
<td>115</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>115</td>
<td>-</td>
<td>115</td>
<td>115</td>
<td>115</td>
<td>1138</td>
</tr>
<tr>
<td>fullfact</td>
<td>-</td>
<td>336</td>
<td>336</td>
<td>336</td>
<td>-</td>
<td>-</td>
<td>336</td>
<td>-</td>
<td>336</td>
<td>336</td>
<td>336</td>
<td>3838</td>
</tr>
<tr>
<td>gossipcop</td>
<td>goop</td>
<td>2947</td>
<td>2947</td>
<td>-</td>
<td>2947</td>
<td>2947</td>
<td>2947</td>
<td>-</td>
<td>2947</td>
<td>2947</td>
<td>2947</td>
<td>12583</td>
</tr>
<tr>
<td>hoaxslayer</td>
<td>hoer</td>
<td>1310</td>
<td>1310</td>
<td>-</td>
<td>1310</td>
<td>1310</td>
<td>1310</td>
<td>-</td>
<td>1310</td>
<td>1310</td>
<td>1310</td>
<td>14499</td>
</tr>
<tr>
<td>huffingtonpostca</td>
<td>luca</td>
<td>38</td>
<td>38</td>
<td>-</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>-</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>78</td>
</tr>
<tr>
<td>leadstories</td>
<td>-</td>
<td>1547</td>
<td>1547</td>
<td>-</td>
<td>1547</td>
<td>1547</td>
<td>1547</td>
<td>-</td>
<td>1547</td>
<td>1547</td>
<td>1547</td>
<td>12015</td>
</tr>
<tr>
<td>mprnews</td>
<td>mpws</td>
<td>49</td>
<td>49</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>49</td>
<td>-</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>319</td>
</tr>
<tr>
<td>nytimes</td>
<td>-</td>
<td>17</td>
<td>17</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>271</td>
</tr>
<tr>
<td>observatory</td>
<td>obry</td>
<td>60</td>
<td>60</td>
<td>-</td>
<td>60</td>
<td>-</td>
<td>60</td>
<td>-</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>592</td>
</tr>
<tr>
<td>pandora</td>
<td>para</td>
<td>225</td>
<td>225</td>
<td>225</td>
<td>225</td>
<td>225</td>
<td>225</td>
<td>-</td>
<td>225</td>
<td>225</td>
<td>225</td>
<td>114</td>
</tr>
<tr>
<td>politico</td>
<td>-</td>
<td>102</td>
<td>102</td>
<td>-</td>
<td>102</td>
<td>-</td>
<td>102</td>
<td>-</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>150</td>
</tr>
<tr>
<td>politifact_promise</td>
<td>pose</td>
<td>1361</td>
<td>1361</td>
<td>1361</td>
<td>-</td>
<td>-</td>
<td>1361</td>
<td>-</td>
<td>1361</td>
<td>1361</td>
<td>1361</td>
<td>6279</td>
</tr>
<tr>
<td>politifact_stmt</td>
<td>point</td>
<td>15390</td>
<td>15390</td>
<td>15390</td>
<td>-</td>
<td>-</td>
<td>15390</td>
<td>-</td>
<td>15390</td>
<td>15390</td>
<td>15390</td>
<td>78543</td>
</tr>
<tr>
<td>politifact_story</td>
<td>-</td>
<td>5460</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5460</td>
<td>-</td>
<td>5460</td>
<td>5460</td>
<td>5460</td>
<td>24836</td>
</tr>
<tr>
<td>radionz</td>
<td>ranz</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>-</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>44</td>
</tr>
<tr>
<td>snopes</td>
<td>snes</td>
<td>6457</td>
<td>6457</td>
<td>6457</td>
<td>6457</td>
<td>-</td>
<td>6457</td>
<td>-</td>
<td>6457</td>
<td>6457</td>
<td>6457</td>
<td>46735</td>
</tr>
<tr>
<td>swissinfo</td>
<td>-</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>-</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>theferret</td>
<td>thet</td>
<td>81</td>
<td>81</td>
<td>81(1)</td>
<td>-</td>
<td>-</td>
<td>81</td>
<td>81(1)</td>
<td>81</td>
<td>81(1)</td>
<td>81</td>
<td>885</td>
</tr>
<tr>
<td>theguardian</td>
<td>-</td>
<td>155</td>
<td>155</td>
<td>155</td>
<td>-</td>
<td>-</td>
<td>155</td>
<td>-</td>
<td>155</td>
<td>155</td>
<td>155</td>
<td>2600</td>
</tr>
<tr>
<td>thejournal</td>
<td>thal</td>
<td>179</td>
<td>179</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>179</td>
<td>-</td>
<td>179</td>
<td>179</td>
<td>179</td>
<td>2375</td>
</tr>
<tr>
<td>truthorfiction</td>
<td>tron</td>
<td>3674</td>
<td>3674</td>
<td>3674</td>
<td>-</td>
<td>-</td>
<td>3674</td>
<td>-</td>
<td>3674</td>
<td>3674</td>
<td>3674</td>
<td>8268</td>
</tr>
<tr>
<td>verafiles</td>
<td>vees</td>
<td>509</td>
<td>509</td>
<td>-</td>
<td>509</td>
<td>-</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>23</td>
</tr>
<tr>
<td>voicesofsandiego</td>
<td>vogo</td>
<td>660</td>
<td>660</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>660</td>
<td>-</td>
<td>660</td>
<td>660</td>
<td>660</td>
<td>2352</td>
</tr>
<tr>
<td>washingtonpost</td>
<td>wast</td>
<td>227</td>
<td>227</td>
<td>227</td>
<td>227</td>
<td>-</td>
<td>227</td>
<td>227</td>
<td>227</td>
<td>227</td>
<td>227</td>
<td>2470</td>
</tr>
<tr>
<td>wral</td>
<td>-</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>-</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>355</td>
</tr>
<tr>
<td>zimfact</td>
<td>-</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>-</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>179</td>
</tr>
</tbody>
</table>

Total 43837 43837 43837 43837 43837 43837 43837 43837 43837 43837 43837 43837 260330

Table 11: Summary statistics for claim collection. ‘Domain’ indicates the domain name used for the veracity prediction experiments, ‘–’ indicates that the website was not used due to missing or insufficient claim labels, see Section 3.2.