Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web

Diego Esteves
SDA Research
University of Bonn, Germany
esteves@cs.uni-bonn.de

Aniketh Janardhan Reddy[+,*]
Carnegie Mellon University
Pittsburgh, USA
ajreddy@cs.cmu.edu

Piyush Chawla[*]
The Ohio State University
Ohio, USA
chawla.81@osu.edu

Jens Lehmann
SDA Research / Fraunhofer IAIS
University of Bonn, Germany
jens.lehmann@cs.uni-bonn.de

Abstract

With the growth of the internet, the number of *fake-news* online has been proliferating every year. The consequences of such phenomena are manifold, ranging from lousy decision-making process to bullying and violence episodes. Therefore, fact-checking algorithms became a valuable asset. To this aim, an important step to detect fake-news is to have access to a credibility score for a given information source. However, most of the widely used Web indicators have either been shutdown to the public (e.g., Google PageRank) or are not free for use (Alexa Rank). Further existing databases are short-manually curated lists of online sources, which do not scale. Finally, most of the research on the topic is theoretical-based or explore confidential data in a restricted simulation environment. In this paper we explore current research, highlight the challenges and propose solutions to tackle the problem of classifying websites into a credibility scale. The proposed model automatically extracts source reputation cues and computes a credibility factor, providing valuable insights which can help in belittling dubious and confirming trustful unknown websites. Experimental results outperform state of the art in the 2-classes and 5-classes setting.

1 Introduction

With the enormous daily growth of the Web, the number of *fake-news* sources have also been increasing considerably (Li et al., 2012). This social network era has provoked a communication revolution that boosted the spread of misinformation, hoaxes, lies and questionable claims. The proliferation of unregulated sources of information allows any person to become an opinion provider with no restrictions. For instance, websites spreading manipulative political content or hoaxes can be persuasive. To tackle this problem, different *fact-checking* tools and frameworks have been proposed (Zubiaga et al., 2017), mainly divided into two categories: *fact-checking* over natural language claims (Thorne and Vlachos, 2018) and *fact-checking* over knowledge bases, i.e., triple-based approaches (Esteves et al., 2018). Overall, *fact-checking* algorithms aim at determining the veracity of claims, which is considered a very challenging task due to the nature of underlying steps, from natural language understanding (e.g. *argumentation mining*) to common-sense verification (i.e., humans have prior knowledge that makes far easier to judge which arguments are plausible and which are not). Yet an important underlying fact-checking step relies upon computing the credibility of sources of information, i.e. indicators that allow answering the question: *"How reliable is a given provider of information*?". Due to the obvious importance of the Web and the negative impact that misinformation can cause, methods to demote the importance of websites also become a valuable asset. In this sense the high number of new websites appearing at everyday (Netcraft, 2016), make straightforward approaches - such as *blacklists* and *whitelists* - impractical. Moreover, such approaches are not designed to compute credibility scores for a given website but rather to binary label them. Thus, they aim at detecting mostly "fake" (threatening) websites; e.g., *phishing detection*, which is out of scope of this work. Thus, open credibility models have a great importance, especially due to the increase of fake news being propagated. There is much research into credibility factors. However, they are mostly grouped as follows: (1) theoretical research on psychological aspects of credibility and (2) experiments performed over private

[+]Work was completed while the author was a student at the Birla Institute of Technology and Science, India and was interning at SDA Research.

[*]These two authors contributed equally to this work.

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 50–59
Brussels, Belgium, November 1, 2018.

and confidential users information, mostly from web browser activities (strongly supported by private companies). Therefore, while (1) lacks practical results (2) report *findings* which are not much appealing to the broad open-source community, given the non-open characteristic of the conducted experiments and data privacy. Finally, recent research on credibility has also pointed out important drawbacks, as follows:

1. Manual (human) annotation of credibility indicators for a set of websites is costly (Haas and Unkel, 2017).

2. Search engine results page (SERP) do not provide more than few information cues (URL, title and snippet) and the dominant heuristic happens to be the search engine (SE) rank itself (Haas and Unkel, 2017).

3. Only around 42.67% of the websites are covered by the credibility evaluation knowledge base, where most domains have a low credibility confidence (Liu et al., 2015)

Therefore, automated credibility models play an important role in the community - although not broadly explored yet, in practice. In this paper, we focus on designing computational models to predict the credibility of a given website rather than performing sociological experiments or experiments with end users (simulations). In this scenario, we expect that a website from a domain such as `bbc.com` gets a higher trustworthiness score compared to one from `wordpress.com`, for instance.

2 Related Work

Credibility is an important research subject in several different communities and has been the subject of study over the past decades. Most of the research, however, focuses on theoretical aspects of credibility and its persuasive effect on different fundamental problems, such as economic theories (Sobel, 1985).

2.1 Fundamental Research

A thorough examination of psychological aspects in evaluating documents credibility has been studied (Fogg and Tseng, 1999; Fogg et al., 2001, 2003), which reports numerous challenges. Apart from sociological experiments, *Web Credibility* - in a more practical perspective - has a different focus of research, described as follows:

Rating Systems, Simulations are mostly platform-based solutions to conduct experiments (mostly using private data) in order to detect credibility factors. Nakamura et al. (2007) surveyed internet users from all age groups to understand how they identified trustworthy websites. Based on the results of this survey, they built a graph-based ranking method which helped users in gauging the trustworthiness of search results retrieved by a search engine when issued a query $\mathcal{Q}$. A study by Stanford University revealed important factors that people notice when assessing website credibility (Fogg et al., 2003), mostly visual aspects (*web site design*, *look* and *information design*). The *writing style* and *bias of information* play a small role as defining the level of credibility (selected by approximately 10% of the comments). However, this process of evaluating the credibility of web pages by users is impacted only by the number of heuristics they are aware of (Fogg, 2003), biasing the human evaluation w.r.t. a limited and specific set features. An important factor considered by humans to judge credibility relies on the search engine results page (SERP). The higher ranked a website is when compared to other retrieved websites the more credible people judge a website to be (Schwarz and Morris, 2011). Popularity is yet another major credibility factor (Giudice, 2010). Liu et al. (2015) proposed to integrate recommendation functionality into a Web Credibility Evaluation System (WCES), focusing on the user's feedback. Shah et al. (2015) propose a full list of important features for credibility aspects, such as 1) the quality of the design of the website and 2) how well the information is structured. In particular, the perceived accuracy of the information was ranked only in 6th place. Thus, superficial website characteristics as heuristics play a key role in credibility evaluation. Dong et al (2015) propose a different method (KBT) to estimate the trustworthiness of a web source based on the information given by the source (i.e., applies fact-checking to infer credibility). This information is represented in the form of triples extracted from the web source. The trustworthiness of the source is determined by the correctness of the triples extracted. Thus, the score is computed based on *endogenous* (e.g., correctness of facts) signals rather then *exogenous* signals (e.g., links).

Unfortunately, this research from Google does not provide open data. It is worth mentioning that - surprisingly - their hypothesis (content is more important than visual) contradicts previous research findings (Fogg et al., 2003; Shah et al., 2015). While this might be due to the dynamic characteristic of the Web, this contradiction highlights the need for more research into the real use of web credibility factors w.r.t. automated web credibility models. Similar to (Nakamura et al., 2007), Singal and Kohli (2016) proposes a tool (dubbed TNM) to re-rank URLs extracted from Google search engine according to the trust maintained by the actual users). Apart from the search engine API, their tool uses several other APIs to collect website usage information (e.g., traffic and engagement info). (Kakol et al., 2017) perform extensive crowdsourcing experiments that contain credibility evaluations, textual comments, and labels for these comments.

SPAM/phishing detection: Abbasi et al. (2010) propose a set of design guidelines which advocated the development of SLT-based classification systems for fraudulent website detection, i.e., despite seeming credible - websites that try to obtain private information and defraud visitors. PhishZoo (Afroz and Greenstadt, 2011) is a phishing detection system which helps users in identifying phishing websites which look similar to a given set of protected websites through the creation of profiles.

2.2 Automated Web Credibility

Automated Web Credibility models for website classification are not broadly explored, in practice. The aim is to produce a predictive model given training data (annotated website ranks) regardless of an input query $\mathcal{Q}$. Existing gold standard data is generated from surveys and simulations (see *Rating Systems, Simulations* related work). Currently, state of the art (SOTA) experiments rely on the Microsoft Credibility dataset[1] (Schwarz and Morris, 2011). Recent research use the website label (Likert scale) released in the Microsoft dataset as a gold standard to train automated web credibility models, as follows:

Olteanu et al. (2013) proposes a number of properties (37 linguistic and textual features) and applies machine learning methods to recognize trust levels, obtaining 22 relevant features for the task. Wawer et al. (2014) improve this work using psychosocial and psycholinguistic features (through The General Inquirer (GI) Lexical Database (Stone and Hunt, 1963)) achieving state of the art results.

Finally, another resource is the Content Credibility Corpus (C3) (Kakol et al., 2017), the largest Web credibility Corpus publicity available so far. However, in this work authors did not perform experiments w.r.t. *automated credibility models* using a standard measure (i.e., Likert scale), such as in (Olteanu et al., 2013; Wawer et al., 2014). Instead, they rather focused on evaluating the theories of web credibility in order to produce a much larger and richer corpus.

3 Experimental Setup

3.1 State-of-the-art (SOTA) Features

Recent research on credibility factors for web sites (Olteanu et al., 2013) have initially divided the features into the following logical groups:

1. **Content-based** (25 features): number of special characters in the text, spelling errors, web site category and etc..
   (a) **Text** (20 features)
   (b) **Appearance** (4 features)
   (c) **Meta-information** (1 feature)
2. **Social-based** (12 features): Social Media Metadata (e.g., Facebook shares, Tweets pointing to a certain URL, etc.), Page Rank, Alexa Rank and similar.
   (a) **Social Popularity** (9 features)
   (b) **General Popularity** (1 feature)
   (c) **Link structure** (2 features)

According to (Olteanu et al., 2013), a resultant number of 22 features (out of 37) were selected as most significant (10 for **content-based** and all **social-based** features). Surprisingly (but also following (Dong et al., 2015)), none from the subgroup **Appearance**, although studies have systematically shown the opposite, i.e., that visual aspects are one of the most important features (Fogg et al., 2003; Shah et al., 2015; Haas and Unkel, 2017).

In this picture, we claim the most negative aspect is the reliance on **Social-based** features. This dependency not only affects the final performance

[1] It is worth mentioning that this survey is mostly based on confidential data and it is not available to the open community (e.g., overall popularity, popularity among domain experts, geo-location of users and number of awards)

of the credibility model, but also implies in financial costs as well as presenting high discriminative capacity, adding a strong bias to the performance of the model[2]. The computation of these features relies heavily on external (e.g., Facebook API[3] and AdBlock[4]) and commercial libraries (Alchemy[5], PageRank[6], Alexa Rank[7]. Thus, engineering and financial costs are a must. Furthermore, popularity on Facebook or Twitter can be measured only by data owners. Additionally, vendors may change the underlying algorithms without further explanation. Therefore, also following Wawer et al. (2014), in this paper we have excluded **Social-based** features from our experimental setup.

On top of that, (Wawer et al., 2014) incremented the model, adding features extracted from the General Inquirer (GI) Lexical Database, resulting in a vector of 183 extra categories, apart from the selected 22 base features, i.e. total of 205 features (However, this is subject to contradictions. Please see Section 4.1 for more information).

3.2 Datasets

3.2.1 Website credibility evaluation

Microsoft Dataset (Schwarz and Morris, 2011) consists of thousands of URLs and their credibility ratings (five-point Likert Scale[8]), ranging from 1 ("very non-credible") to 5 ("very credible"). In this study, participants were asked to rate the websites as credible following the definition: "*A credible webpage is one whose information one can accept as the truth without needing to look elsewhere*". Studies by (Olteanu et al., 2013; Wawer et al., 2014) use this dataset for evaluation. **Content Credibility Corpus (C3)**[9] is the most recent and the largest credibility dataset currently publicly available for research (Kakol et al., 2017). It contains 15.750 evaluations of 5.543 URLs from 2.041 participants with some additional information about website characteristics and basic demographic features of users. Among many metadata information existing in the dataset, in this work we are only interested in the URLs and their respective five-point Likert scale, so that we obtain the same information available in the Microsoft dataset.

3.2.2 Fact-checking influence

In order to verify the impact of our web credibility model in a real use-case scenario, we ran a fact-checking framework to verify a set of input claims. Then we collected the sources (URLs) containing proofs to support a given claim. We used this as a dataset to evaluate our web credibility model.

The primary objective is to verify whether our model is able, on average, to assign *lower* scores to the websites that contain *proofs* supporting *claims* which are labeled as *false* in the FactBench dataset (i.e., the source is providing false information, thus should have a lower credibility score). Similarly, we expect that websites that support *positive* claims are assigned with higher scores (i.e., the source is supporting an accurate claim, thus should have a higher credibility score).

The (gold standard) input claims were obtained from the **FactBench** dataset[10], a multilingual benchmark for the evaluation of fact validation algorithms. It contains a set of RDF[11] models (10 different relations), where each model contains a singular fact expressed as a *subject-predicate-object* triple. The data was automatically extracted from DBpedia and Freebase KBs, and manually curated in order to generate true and false examples.

The website list extraction was carried out by DeFacto (Gerber et al., 2015), a fact-checking framework designed for RDF KBs. DeFacto returns a set of websites as pieces of evidence to support its prediction (true or false) for a given input claim.

3.3 Final Features

We implemented a set of **Content-based** features (Section 3.1) adding more lexical and textual based features. **Social-based** features were not considered due to financial costs associated with paid APIs. The final set of features for each website w is defined as follows:

1. *Web Archive*: the temporal information w.r.t. cache and freshness. Δ_b and Δ_e correspond to the temporal differences of the first and last 2 updates, respectively. Δ_a represents the age of w and finally Δ_u represents the temporal difference for

[2] authors applied ANOVA test confirming this finding
[3] https://developers.facebook.com/
[4] https://adblockplus.org/
[5] www.alchemyapi.com
[6] excepting for heuristic computations, calculation of PageRank requires crawling the whole Internet
[7] https://www.alexa.com/siteinfo
[8] https://en.wikipedia.org/wiki/Likert_scale
[9] also known as Reconcile Corpus
[10] https://github.com/DeFacto/FactBench
[11] https://www.w3.org/RDF/

the last update to today. γ is a penalization factor when the information is obtained from the *domain* of w (w_d) instead w.

$$f_{arc}(w) = \left(\left[\frac{1}{\log(\Delta_b \times \Delta_e)} + \log(\Delta_a) + \frac{1}{\Delta_u}\right]\right) \times \gamma$$

2. *Domain*: refers to the (encoded) domain w (e.g. org)

3. *Authority*: searches for authoritative keywords within the page HTML content w_c (e.g., contact email, business address, etc..)

4. *Outbound Links*: searches the number of different outbound links in $w \wedge w_d \in d$, i.e., $\sum_{n=1}^{P} \phi(w_c)$ where P is the number of web-based protocols.

5. *Text Category*: returns a vector containing the probabilities P for each pre-trained category c of w w.r.t. the sentences of the website w_s and page title w_t: $\sum_{s=1}^{w_s} \gamma(s) \frown \gamma(w_t)$. We trained a set of binary multinomial Naive Bayes (NB) classifiers, one per class, as follows: *business*, *entertainment*, *politics*, *religion*, *sports* and *tech*.

6. *Text Category - LexRank*: reduces the noisy of w_b by classifying only top N sentences generated by applying LexRank (Erkan and Radev, 2004) over w_b ($S\prime = \Gamma(w_b, N)$), which is a graph-based text summarizing technique: $\sum_{s\prime=1}^{S\prime} \gamma(s\prime) \frown \gamma(w_t)$.

7. *Text Category - LSA*: similarly, we apply Latent Semantic Analysis (LSA) (Steinberger and Jeek, 2004) to detect semantically important sentences in w_b ($S\prime = \Omega(w_b, N)$): $\sum_{s\prime=1}^{S\prime} \gamma(s\prime) \frown \gamma(w_t)$.

8. *Readability Metrics*: returns a vector resulting of the concatenation of several R readability metrics (Si and Callan, 2001)

9. *SPAM*: detects whether the w_b or w_t are classified as spam: $\psi(w_b) \frown \psi(w_t)$

10. *Social Tags*: returns the frequency of social tags in w_b: $\bigcup_{i=1}^{R} \varphi(i, w_b)$

11. *OpenSources*: returns the open-source classification (x) for a given website:

$$x = \begin{cases} 1, & \text{if } w \in \mathcal{O} \\ 0, & \text{if } w \notin \mathcal{O} \end{cases}$$

12. *PageRankCC*: PageRank information computed through the CommonCrawl[12] Corpus

13. *General Inquirer* (Stone and Hunt, 1963): a 182-lenght vector containing several lexicons

14. *Vader Lexicon*: lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments

15. *HTML2Seq*: we introduce the concept of *bag-of-tags*, where similarly to *bag-of-words*[13] we group the HTML tag occurrences in each web site. We additionally explore this concept along with a sequence problem, i.e. we encode the tags and evaluate this considering a window size (offset) from the header of the page.

4 Experiments

Previous research proposes two **application settings** w.r.t. the classification itself, as follows: (A.1) casting the credibility problem as a classification problem and (A.2) evaluating the credibility on a five-point Likert scale (regression). In the classification scenario, the models are evaluated both w.r.t. the 2-classes as well as 3-classes. In the 2-classes scenario, websites ranging from 1 to 3 are labeled as "low" whereas 4 and 5 are labeled as "high" (credibility). Analogously, in the 3-classes scenario, websites labeled as 1 and 2 are converted to "low", 3 remains as "medium" while 4 and 5 are grouped into the "high" class.

We first explore the impact of the *bag-of-tags* strategy. We encode and convert the tags into a sequence of tags, similar to a sequence of sentences (looking for opening and closing tags, e.g., <a>and). Therefore, we perform document classification over the resulting vectors. Figures 1a to 1d show results of this strategy for both 2 and 3-classes scenarios. The x-axis is the log scale of the paddings (i.e., the offset of HTML tags we retrieved from w, ranging from 25 to 10.000). The charts reveal an interesting pattern in both gold-standard datasets (Microsoft Dataset and C3 Corpus): the first tags are the most relevant to predict the credibility class. Although this strategy does not achieve state of the art performance[14], it presents reasonable performance by just inspecting website metadata: F1-measures = 0.690 and 0.571 for the 2-classes and 3-classes settings, respectively. However, it is worth mentioning that the main advantage of this approach lies in the fact that it is language agnostic (while current research

[12] http://commoncrawl.org/

[13] https://en.wikipedia.org/wiki/Bag-of-words_model

[14] F1 measures = 0.745 (2-classes) and 0.652 (3-classes).

Microsoft Dataset (Gradient Boosting, $K = 25$)			
Class	Precision	Recall	F1
low	0.851	0.588	0.695
high	0.752	0.924	0.829
weighted	0.794	0.781	0.772
micro	0.781	0.781	0.781
macro	0.801	0.756	0.762
C3 Corpus (AdaBoost, $K = 75$)			
Class	Precision	Recall	F1
low	0.558	0.355	0.434
high	0.732	0.862	0.792
weighted	0.675	0.695	0.674
micro	0.695	0.695	0.695
macro	0.645	0.609	0.613

Table 1: Text+HTML2Seq features (2-class): best classifier performance

Microsoft Dataset (Gradient Boosting, $K = 75$)			
Class	Precision	Recall	F1
low	0.567	0.447	0.500
medium	0.467	0.237	0.315
high	0.714	0.916	0.803
weighted	0.626	0.662	0.626
micro	0.662	0.662	0.662
macro	0.583	0.534	0.539
C3 Corpus (AdaBoost, $K = 100$)			
Class	Precision	Recall	F1
low	0.143	0.031	0.051
medium	0.410	0.177	0.247
high	0.701	0.916	0.794
weighted	0.583	0.660	0.598
micro	0.660	0.660	0.660
macro	0.418	0.375	0.364

Table 2: Text+HTML2Seq features (3-class): best classifier performance

focuses on English) as well as less susceptible to overfitting.

We then evaluate the performance of the textual features (Section 3.3) isolated. Results for the 2-classes scenario are presented as follows: Figure 2a highlights the best models performance using textual features only. While this as a single feature does not outperform the lexical features, when we combine the *bag-of-tags* approach (predictions of probabilities for each class) we boost the performance (F1 from 0.738 to 0.772) and outperform state of the art (0.745), as shown in Figure 2b. Tables 1 to 3 shows detailed results for both datasets (2-classes, 3-classes and 5-classes configurations, respectively). For 5-class regression, we found that the *best pad = 100* for the Microsoft dataset and *best pad = 175* for the C3 Corpus. We preceded the computing of both classification and regression models with feature selection according to a percentile of the highest scoring features (*SelectKBest*). We tested the choice of 3, 5, 10, 25, 50 75 and K=100 percentiles (thus, no selection) of features and did not find a unique K value for every case. It is worth noticing that in general it is easy to detect high credible sources (F1 for "high" class around 0.80 in all experiments and both datasets) but recall of "low" credible sources is still an issue.

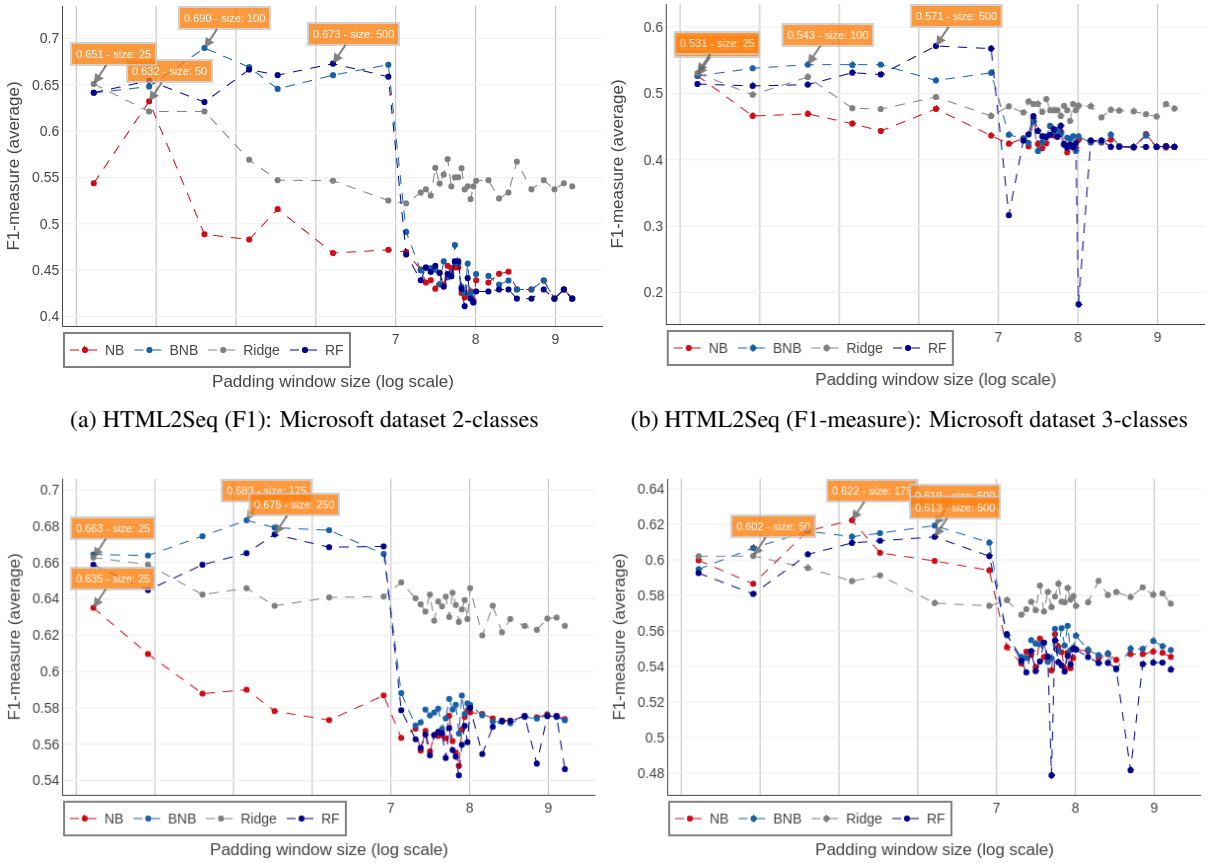
Table 4 shows statistics on the data generated by the fact-checking algorithm. For 1500 claims, it collected pieces of evidence for over 27.000 websites. Table 5 depicts the impact of the credibility model in the fact-checking context. We collected a small subset of 186 URLs from the FactBench dataset and manually annotated[15] the credibility for each URL (following the Likert scale). The model corrected labeled around 80% of the URLs associated with a positive claim and, more importantly, 70% of non-credible websites linked to false claims were correctly identified. This helps to minimize the number of non-credible information providers that contain information that supports a *false* claim.

4.1 Discussion

Reproducibility is still one of the cornerstones of science and scientific projects (Baker, 2016). In the following, we list some relevant issues encountered while performing our experiments:

Experimental results: this gap is also observed w.r.t. results reported by (Olteanu et al., 2013), which is acknowledged by (Wawer et al., 2014), despite numerous attempts to replicate experiments. Authors (Wawer et al., 2014) believe this is

[15] By four human annotators. In the event of a tie we exclude the URL from the final dataset.



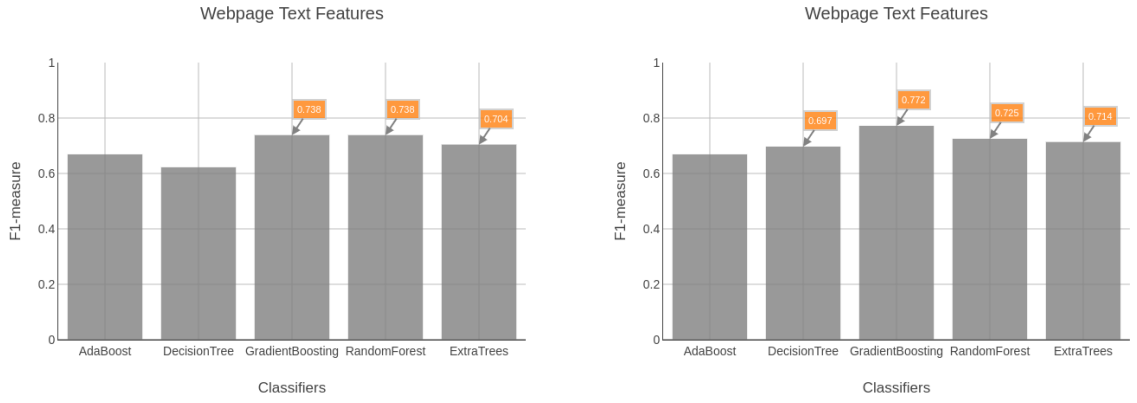
(a) HTML2Seq (F1): Microsoft dataset 2-classes

(b) HTML2Seq (F1-measure): Microsoft dataset 3-classes

(c) HTML2Seq (F1): C3 Corpus 2-classes

(d) HTML2Seq (F1): C3 Corpus 3-classes

Figure 1: HTML2Seq (F1-measure) over different padding sizes.



(a) Textual Features.

(b) Textual+HTML2Seq (best padding) Features.

Figure 2: Evaluating distinct classifiers in the 2-classes setting (Microsoft dataset): increasing almost +3% (from 0.745 to 0.772) on average F1 (Gradient Boosting). Feature selection performed with ANOVA *SelectKBest* method, K=0.25.

Microsoft Dataset					
model	K	R^2	RMSE	MAE	EVar
SVR	3	0.232	0.861	0.691	0.238
Ridge	3	0.268	0.841	0.683	0.269
C3 Corpus					
model	K	R^2	RMSE	MAE	EVar
SVR	25	0.096	0.939	0.739	0.102
Ridge	25	0.133	0.920	0.750	0.134

Table 3: Text+HTML2Seq: regression measures (5-class). Selecting top K lexical features.

FactBench (Credibility Model)		
label	**claims**	**sites**
true	750	14.638
false	750	13.186
-	1500	27.824

Table 4: FactBench: Web sites collected from claims.

due to the lack of parameters and hyperparameters explicitly cited in the previous research (Olteanu et al., 2013).

Microsoft dataset: presents inconsistencies. Although all the web pages are cached (in theory) in order to guarantee a deterministic environment, the dataset - in its original form[16] - has a number of problems, as follows: (a) web pages not physically cached (b) URL not matching (dataset links *versus* cached files) (c) Invalid file format (e.g., PDF). Even though these issues have also been previously identified by related research (Olteanu et al., 2013) it is not clear what the URLs for the final dataset (i.e., the support) are nor where this new version is available.

Contradictions: w.r.t. the divergence of the importance of visual features have drawn our attention (Dong et al., 2015) and (Fogg, 2003; Shah et al., 2015) which corroborate to the need of more methods to solve the web credibility problem, in practice. The main hypothesis that supports this contradiction relies on the fact that feature-based credibility evaluation eventually ignites cat-and-mouse play between scientists and people interested in manipulating the models. In this case, *reinforcement learning* methods pose as a good alternative for adaptation.

FactBench (Sample - Human Annotation)				
label	**claims**	**sites**	**non-cred**	**cred**
true	5	96	57	39
false	5	80	48	32
-	10	186	105	71
FactBench (Sample - Credibility Model)				
label	**non-cred**	**%**	**cred**	**%**
true	40	0.81	31	0.79
false	34	0.70	24	0.75

Table 5: FactBench Dataset: analyzing the performance of the credibility model in the fact-checking task.

Proposed features: The acknowledgement made by authors in (Wawer et al., 2014) that "*solutions based purely on external APIs are difficult to use beyond scientific application and are prone for manipulation*" confirming the need to exclude **social features** from research of (Olteanu et al., 2013) contradicts itself. In the course of experiments, authors mention the usage of all features proposed by (Olteanu et al., 2013): "*Table 1 presents regression results for the dataset described in [13] in its original version (37 features) and extended with 183 variables from the General Inquirer (to 221 features)*".

Therefore, due to the number of relevant issues presented w.r.t. reproducibility and contradiction of arguments, the comparison to recent research becomes more difficult. In this work, we solved the technical issues in the Microsoft dataset and released a new fixed version[17]. Also, since we need to perform evaluations in a deterministic environment, we cached and released the websites for the C3 corpus. After scraping, 2.977 URLs were used (out of 5.543). Others were left due to processing errors (e.g., 404). The algorithms and its hyperparameters and further relevant metadata are available through the MEX Interchange Format (Esteves et al., 2015). By doing this, we provide a computational environment to perform safer comparisons, being engaged in recent discussions about mechanisms to measure and enhance

[16]The original dataset can be downloaded from http://research.microsoft.com/en-us/projects/credibility/

[17]more information at the project website: https://github.com/DeFacto/WebCredibility

the reproducibility of scientific projects (Wilkinson et al., 2016).

5 Conclusion

In this work, we discuss existing alternatives, gaps and current challenges to tackle the problem of web credibility. More specifically, we focused on automated models to compute a credibility factor for a given website. This research follows the former studies presented by (Olteanu et al., 2013; Wawer et al., 2014) and presents several contributions. First, we propose different features to avoid the financial cost imposed by external APIs in order to access website credibility indicators. This issue has become even more relevant in the light of the challenges that have emerged after the shutdown of Google PageRank, for instance. To bridge this gap, we have proposed the concept of bag-of-tags. Similar to (Wawer et al., 2014), we conduct experiments in a highly-dimensional feature space, but also considering web page metadata, which outperforms state of the art results in the 2-classes and 5-classes settings. Second, we identified and fixed several problems on a gold standard dataset for web credibility (Microsoft), as well as indexed several web pages for the C3 Corpus. Finally, we evaluate the impact of the model in a real fact-checking use-case. We show that the proposed model can help in belittling and supporting different websites that contain evidence of true and false claims, which helps the very challenging fact verification task. As future work, we plan to explore deep learning methods over the HTML2Seq module.

Acknowledgments

This research was partially supported by an EU H2020 grant provided for the WDAqua project (GA no. 642795) and by DAAD under the project "International promovieren in Deutschland für alle" (IPID4all).

References

Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, and Jay F Nunamaker Jr. 2010. Detecting fake websites: the contribution of statistical learning theory. *Mis Quarterly*, pages 435–461.

Sadia Afroz and Rachel Greenstadt. 2011. Phishzoo: Detecting phishing websites by looking at them. *2012 IEEE Sixth International Conference on Semantic Computing*, 00:368–375.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Diego Esteves, Diego Moussallem, Ciro Baron Neto, Tommaso Soru, Ricardo Usbeck, Markus Ackermann, and Jens Lehmann. 2015. Mex vocabulary: A lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems*, SEMANTICS '15, pages 169–176, New York, NY, USA. ACM.

Diego Esteves, Anisa Rula, Aniketh Janardhan Reddy, and Jens Lehmann. 2018. Toward veracity assessment in rdf knowledge bases: An exploratory analysis. *Journal of Data and Information Quality (JDIQ)*, 9(3):16.

BJ Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, et al. 2001. What makes web sites credible?: a report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68. ACM.

BJ Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87. ACM.

Brian J Fogg. 2003. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723. ACM.

Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15. ACM.

Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, and René Speck. 2015. Defacto - temporal and multilingual deep fact validation. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Katherine Del Giudice. 2010. Crowdsourcing credibility: The impact of audience feedback on web page credibility. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47*, page 59. American Society for Information Science.

Alexander Haas and Julian Unkel. 2017. Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology*, 36(12):1285–1298.

Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. 2017. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 53(5):1043–1061.

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth finding on the deep web: Is the problem solved? *Proc. VLDB Endow.*, 6(2):97–108.

Xin Liu, Radoslaw Nielek, Paulina Adamska, Adam Wierzbicki, and Karl Aberer. 2015. Towards a highly effective and robust web credibility evaluation system. *Decision Support Systems*, 79:99–108.

Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. 2007. *Trustworthiness Analysis of Web Search Results*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Netcraft. 2016. Netcraft survey (2016). http://www.webcitation.org/6lhJlHtez. Accessed: 2017-10-01.

Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web credibility: features exploration and credibility prediction. In *European conference on information retrieval*, pages 557–568. Springer.

Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM.

Asad Ali Shah, Sri Devi Ravana, Suraya Hamid, and Maizatul Akmar Ismail. 2015. Web credibility assessment: affecting factors and assessment techniques. *Information Research*, 20(1):20–1.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA. ACM.

Himani Singal and Shruti Kohli. 2016. Trust necessitated through metrics: Estimating the trustworthiness of websites. *Procedia Computer Science*, 85:133–140.

Joel Sobel. 1985. A theory of credibility. *The Review of Economic Studies*, 52(4):557–573.

Josef Steinberger and Karel Jeek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *In Proc. ISIM 04*, pages 93–100.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *CoRR*, abs/1806.07687.

Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. 2014. Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd international conference on world wide web*, pages 1135–1140. ACM.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2017. Detection and resolution of rumours in social media: A survey. *CoRR*, abs/1704.00656.