Web Credibility: Features Exploration and Credibility Prediction

Alexandra Olteanu*, Stanislav Peshterliev*, Xin Liu, and Karl Aberer

LSIR EPFL
Lausanne, Switzerland

Abstract. The open nature of the World Wide Web makes evaluating webpage credibility challenging for users. In this paper, we aim to automatically assess web credibility by investigating various characteristics of webpages. Specifically, we first identify features from textual content, link structure, webpages design, as well as their social popularity learned from popular social media sites (e.g., Facebook, Twitter). A set of statistical analyses methods are applied to select the most informative features, which are then used to infer webpages credibility by employing supervised learning algorithms. Real dataset-based experiments under two application settings show that we attain an accuracy of 75% for classification, and an improvement of 53% for the mean absolute error (MAE), with respect to the random baseline approach, for regression.

Keywords: web credibility, feature analysis, classification, regression.

1 Introduction

The web is a dynamic environment where users interact with a huge volume of information (e.g., via search engines, social networks). However, due to the openness of the web, anyone can produce any content, a lot of which is published without being rigorously fact-checked. This greatly influences people's daily activities as many users rely on the web as their primary information source to make decisions. For example, earthquakes rumors succeeded to induce panic within the targeted communities[1]. The speed at which the numerous examples of hoaxed web information spread (e.g., kidnapping rumours[2], attack rumours[3]), poses important challenges to users. Users generally lack evidence about the factors that characterize *credibility*, such as the author expertise and trustworthiness in delivering credible information [1]. It is therefore imperative to provide users

* The first two authors have equal contribution to this work.

[1] `http://www.instantriverside.com/2010/04/california-earthquake-rumors-untrue-quake/`

[2] `http://chemgen.wordpress.com/2012/03/31/rumour-mongers-pranksters-and-crying-wolf/`

[3] `http://nepalkhabaronline.com/2012/08/17/attack-rumours-trigger-panic-among-nepalis-in-banglore/`

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 557–568, 2013.

appropriate tools to help them correctly assess the *credibility* of the webpages they visit and avoid being affected by misleading or false information.

Assessing information credibility is not new [2]; yet, while peer reviewing has been ensuring content quality in traditional media (e.g., books or scholarly articles), the web has enabled anyone to publish without restriction. On the one hand, existing approaches to credibility assessment rely on humans judging visualizations of web page aspects [3, 4]. On the other hand, one may leverage the "wisdom of the crowds", i.e., other users' credibility evaluations. However, due to the large, fast-growing volume of information on the web, a significant fraction of pages will receive little or no evaluation [5]. To address this problem, we aim to automate the task by using solely the information available on the web, that is, by investigating the various characteristics of web page content.

Contributions and Paper Organization. In this paper, we show how webpage credibility can be automatically and accurately assessed by employing machine learning algorithms. To do so, we first identify a set of features that are expected to be relevant for web credibility assessment. Then, we comprehensively study webpage characteristics by investigating various information sources in order to building a general-purpose and automated web credibility assessment framework. Specifically, in §3, we explore features falling in two main categories: (1) *content features* which refer to features that can be computed either based on the textual content of the webpages, *text-based features*, or based on the webpage structure, *appearance and metadata features*; and (2) *social features* which include features that reflect the online popularity of a webpage and its link structure. In §4, we show how the selected features are applied to assess webpages credibility as a binary result (i.e., classification) and on a five-point Likert scale (i.e., regression) using supervised learning algorithms. Experiments conducted on a real dataset show that for binary classification, our approach achieves 70% or more precision and recall; for regression, our approach improves the mean absolute error (MAE) by 53%. We conclude in §5.

2 Related Work

Some previous work on web credibility evaluation aims to identify the most relevant features for credibility assessments and make them more prominent to users [3, 4], while others use such features to automate the credibility assessment[6–9]. In this section, we survey work that (1) identifies relevant features and (2) proposes methods for automatic web credibility assessment.

Web Credibility Features. Fogg et al. made a quantitative study where users were asked to rate and comment upon the webpages' credibility [10]. They found that users are more influenced by prominent features (e.g., website design, advertising). In [3] the focus is on three categories of webpage features that are difficult for users to assess: *on-page features* – found on the webpage but hard to assess; *off-page features* – information made public by Internet companies/organizations; and *aggregate features* – inferred from information owned by

such companies, but not readily available to users. In [11, 12], the authors determine the quality of users generated content based on a set of factors believed to impact the community preferences. It was observed that the aggregation of several types information sources, e.g., content-based features (e.g., comment informativeness, readability) with user-based features (e.g., category activity, profile views), leads to more accurate results. Other studies focus on tweets credibility perception [8, 13, 14]. For instance, in [13] the non-standard use of grammar and punctuation are found to indicate low credibility perceptions of tweets. By leveraging the works reviewed above, we comprehensively investigate web credibility related features (§3).

Web Credibility Assessment. Recent work proposed to visually augment web pages with features providing evidence about their credibility [3, 4]. Although more meaningful information is provided to users, augmenting web pages brought little benefit due to lack of prominence [3]: showing many features to users is undesirable; each feature would become less prominent due to information clutter. In contrast, we want to automatically infer the webpages' credibility.

Other solutions attempt automated web content credibility assessment [4, 6, 7, 11, 15, 16]. In [15, 16], credibility scores are computed based on Web's link structure. To compute link-based features, at least a local view of the webpage neighborhood in the web graph is needed. From our feature set only pagerank reflects the webpage position in the web graph. Others use weighting schemes to combine different features and predict credibility scores[4, 6, 7]. In [4] the authors showed that Google search enhanced with their ranking system returns more credible webpages than Google search alone. Other similar approaches either lack a thorough evaluation [7], or the obtained scores do not to correlate with human judges [6].

The machine learning based solutions are closer to our work[8, 11, 17]. In [11] the problem of ranking comments from social web is casted as a regression problem. In [8] tweets credibility is automatically assessed. They use a supervised classifier to label the newsworthy tweets as credible or not. Sondhi et. al. employ a supervised learning approach to predict the reliability of health related webpages [17]. In contrast, we aim to assess web credibility without a particular emphasis on a certain type of web content, thus making our approach more generic. Furthermore, we conduct an extensive feature study by exploring various information sources to improve accuracy of machine learning algorithms.

3 Features Exploration

We aim to automatically assess webpages credibility by employing machine learning techniques. To this end, first, we explore useful webpage properties for credibility assessments that can be inferred from readily available information on the webpage or at third party sites. We then define features that characterize each one of these properties. We group these features in two categories: **content features** that are present on the webpage and can be computed using the webpage textual content (e.g., part of speech, punctuation), or the webpage design and meta-information elements (e.g., CSS style definitions, ads); and **social**

Table 1. Features summary (# stands for 'number of', ? for binary variables, and @ for dedicated formulas)

Category	Type	Feature	Description
Content	Text	#exclamations	Number of exclamation marks "!" in the text
		#commas	Number of commas "," in the text
		#dots	Number of dots "." in the text
		#questions	Number of question marks "?" in the text
		#token_count	Text length as the number of words
		?polarity	0 if the page is negative, 1 if the page is positive
		#positive	Number of positive sentences
		#negative	Number of negative sentences
		#subjective	Number of subjective sentences
		#objective	Number of objective sentences
		#spelling_errors	Number of spelling errors
		@text_complexity	Text entropy
		@informativeness	Uniqueness of the page's content relative to other pages
		@smog	Statistical measure of text readability
		category	Web page category, e.g., Entertainment, Business, etc.
		#NN	Number of nouns in the text
		#VB	Number of verbs in the text
		#JJ	Number of adjectives
		#RB	Number of adverbs
		#DT	Number of determiners
	Appearance	#ad_count	Number of ads on the webpage
		#ad_max_size	The area in pixels of the biggest ad
		#ad_body_ratio	Ratio of the area of all ads to the area of the page
		#css_definitions	Number of webpage CSS style definitions
	Meta-Information	domain_type	URL domain type, e.g., .org, .com, .gov
Social	Social Popularity	#fb_share	Number of Facebook shares for a webpage URL
		#fb_like	Number of Facebook likes for a webpage URL
		#fb_comment	Number of Facebook comments for a webpage URL
		#fb_click	Number of Facebook clicks for a webpage URL
		#fb_total	Total Facebook shares, likes, comments and clicks
		#tweets	Number Tweets mentioning a webpage URL
		#bitly_clicks	Number of Bitly short URL clicks for a webpage
		#bitly_referrers	Number of web sites having Bitly short URL for a webpage
		#delicious_bookmarks	Number of Delicious bookmarks for a webpage URL
	General Popularity	@alexa_rank	Alexa rank
	Link structure	#alexa_linksin	Number of web site linkings estimated by Alexa
		@page_rank	Google PageRank

features that may not be available on the webpage, but represent public information available on popular social media platforms (e.g., Facebook, Twitter).

Next we provide motivation and a detailed description of the webpage properties that we consider in this work, along with their corresponding features, which are listed in Table 1.

3.1 Content Features

Text-Based Features: *Content categorization.* Users may have different assumptions, levels of involvement, tasks, and interests, which might affect their perception and, thus, interpretation of the web content credibility. Therefore, we start by identifying text-based features that might impact users perception of a webpage credibility, and comply with previous works [3, 11, 12]: (1) Category – users may have different expectations about the writing norms and styles,

and the choice of words and punctuation that a text belonging to a particular category should comform to (e.g., an article in a gossiping magazines is written in a different way than a scientific article). We compute webpages category with AlchemyAPI's text categorization service [18] which classifies webpages in the following categories: arts & entertainment, business, computers & Internet, culture & politics, gaming, health, law & crime, religion, recreation, science & technology, sports, and weather. (2) Part of speech[4] – the number of verbs, nouns, adjectives, and adverbs may give clues about the content type (e.g., a high number of adverbs and adjectives could indicate a descriptive text). (3) The text length may reflect both the author's effort to write it and the reader's effort to evaluate it. To measure it, we do a simple count of the words in a webpage.

Text comprehensibility. Another factor that might affect users' perception of credibility is their level of education. Similar with previous works [11, 12], to capture this factor we compute features that reflect: (1) Text readability – to measure how difficult to understand the text in a webpage is; we use the SMOG score [19], which gives an estimate of the number of years of education a user needs in order to understand the piece of text. (2) Text complexity is given by the text entropy in a webpage w, with n words and f_i the frequency of each word [11]:

$$entropy(w) = \frac{1}{n} \sum_{i=1}^{n} f_i[\log_{10}(n) - \log_{10}(f_i)] \quad (1)$$

(3) Text informativeness measures the singularity of a webpage content with respect to other webpages discussing the same topic[5], and is given by [11]:

$$inform(d) = \sum_{t_i \in d} tf_{i,j} \times idf_i \quad (2)$$

where $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_k}$, with $n_{i,j}$ the number of occurrences of a term t_i in a webpage d; $idf = \log \frac{|W|}{|\{w:t_i \in w\}|+1}$, where the denominator is the number of webpages in which the term t_i appears; W is the set of compared webpages (e.g., the webpages corresponding to a certain topic, i.e., search query), and $|W|$ is the number of webpages in this set.

Non-standard use of grammar and punctuation was found to be a good indicator of low quality content and low credibility perceptions [12, 13]. To evince such characteristics we appraise: (1) the non-standard use of grammar by counting the number of spelling errors in a webpage, and (2) the use of punctuation marks by counting the number of occurrences of a given punctuation mark (e.g., commas, questions marks).

Sentiment Analysis. The content of a webpage may convey its author's personal opinions, beliefs and sentiments, which might be divergent from those of other

[4] To assign part of speech tags we use NLTK's default part of speech tagger (`http://nltk.org/`).

[5] Note that informativeness of a webpage can be measured only when the webpages are semantically clustered, e.g., belong to the same search engine result.

sources of information or of the reader. To capture this we perform sentiment analysis[6] and measure (1) document and sentence level polarity by counting the positive and negative sentences, and (2) sentence level subjectivity by counting the number of objective and subjective sentences.

Appearance and Meta-Information Features. Webpage appearance features, such as webpage design and advertisements, have been found to impact users' perception of credibility [10, 20]. To this end, our feature set includes features that capture (1) the ads prominence in the webpage, e.g., the number of ads on a webpage, the size of the biggest ad, and the ratio between the size of the all ads and the size of the page, and (2) the graphical design of a webpage, which is estimated using the number of CSS style definitions as an approximation of how much effort has been put into the design of a particular webpage.

Membership. The domain type of a webpage might suggest that the page belongs or not to a set of trusted webpages (e.g, a .gov domain type might indicate that its content is approved by a governmental institution, a .edu domain type could belong to a educational institution).

3.2 Social Features

The popularity of a webpage can be a good indicator of its credibility [3]. To this end, we consider a webpage popularity on several social media platforms, and from internet traffic monitors. (1) We gathered information about webpages' social popularity from social networks sites (i.e., how many times a webpage was liked, shared, clicked, commented on Facebook, and tweeted on Twitter), from the URL shortener service Bitly[7] (i.e., the number of clicks and referrers of the shortened URL), and from the social bookmarking service Delicious[8] (i.e., the number of bookmarks for a webpage). (2) However, these social media metrics measure rather the influence, and not necessary user browsing behaviors. To capture this, we gathered information about the general popularity of webpages from the analytical service provided by Alexa[10].

Finally, the incoming links to a webpage can be seen as endorsements for this webpage and, thus, they could be a good indicator for a webpage reliability. This property is encompassed by features such as (1) Google PageRank and (2) Alexa rank[9].

4 Automatic Assessment of Web Credibility

In this section we show how, given a set of webpages, we can automatically predict the (level of) credibility of each webpage. Two application settings are

[6] For sentiment analysis we train three NLTK Naive Bayesian classifiers with movie reviews data from `http://www.cs.cornell.edu/people/pabo/movie-review-data/`. Polarity is classified with 90% accuracy, and subjectivity with 80% accuracy.

[7] `http://bitly.com/`

[8] `http://delicious.com/`

[9] `http://www.alexa.com/`

considered: (1) assessing if a webpage is credible or not, case in which we cast the credibility assessment problem as a binary classification problem, and (2) assessing a webpage level of credibility on a five-point Likert scale, for which we approach the credibility assessment problem as a regression problem.

4.1 Dataset

To evaluate the performance of our approach, we use a dataset built by Microsoft for a study that analyzes if showing users a set of features can help them to better assess the webpages credibility [3]. This dataset consists of 1000 URLs (along with their credibility ratings) that point to webpages falling in five categories that exhibit both credible and non-credible web content. All these webpages are rated on a five-point Likert scale (where 1 means "very non-credible" and 5 "very credible"). The evaluation has been done according to the following definition: *"A credible webpage is one whose information one can accept as the truth without needing to look elsewhere" [3].*

Since the values of some features are time-dependent (e.g., Alexa rank, social popularity), we remove from this dataset staled URLs (i.e., URLs that point to invalid web locations, or for which the content of the webpage changed since the dataset has been built). As a result, we are left with 883 URLs for which we compute the features detailed in Table 1. For these URLs, the rating distribution is: 1 - 4% (33), 2 - 15% (131), 3 - 22% (193), 4 - 42% (373), and 5 - 17% (153). For binary classification we label as credible only the webpages that received a credibility rating of 4 or 5, ending up with 60% (526) credible webpages and 40% (357) non-credible ones. We assume that a rating of 3 corresponds to a borderline/ambigous credibility valuation, and, thus, prefer to train the classifiers to label such webpages as non-credible.

4.2 Feature Selection

Before employing supervised learning algorithms to assess the webpages' credibility, we first filter out the features that are statistically proved to be irrelevant to the credibility prediction task. To do so, we apply three statistical tests, each of which tests a different *null hypothesis*, to select the discriminative features.

We first use the *Spearman* ρ test to check if the null hypothesis that there is no correlation between the feature values and the credibility ratings can be rejected. In particular, the link-based and some use-of-punctuation based features (i.e., the number of question and exclamation marks) were found to exhibit high correlation with the credibility ratings. In total, the Spearman ρ test rejected 15 features (out of which 14 are content-based features) for which the test was not statistically significant (p-value > 0.01).

Then, we apply the *Chi2* test to see whether the occurrence of a feature is independent of the occurrence of a class (for this test we split the webpages into two classes, credible and non-credible, as explained in Section 4.1). Social features, e.g., Facebook features and Alexa rank, are particularly good indicators of credible webpages (the higher the feature value, the more credible the

Table 2. Best features selected using statistical tests

Category	Count	Features
Content	10	#exclamations, #questions, ?polarity, #negative, #subjective, @informativeness, @smog, #css_definitions, #RB, domain_type
Social	12	#fb_share, #fb_like, #fb_comment, #fb_click, #fb_total, #tweets, #bitly_clicks, #bitly_referrers, #delicious_bookmarks, @alexa_rank, #alexa_linksin, @page_rank

Table 3. Results for the classification of webpages according to their credibility

Class	FP rate	Precision	Recall	F1 score
Random Baseline, Accuracy: 0.49				
NC	0.50	0.40	0.47	0.43
C	0.53	0.57	0.50	0.53
W avg	0.52	0.50	0.49	0.49
Selected features, Accuracy: 0.75				
NC	0.16	0.74	0.63	0.68
C	0.37	0.76	0.84	0.80
W avg	0.28	0.75	0.75	0.75

(a) Cross-validation

Class	FP rate	Precision	Recall	F1 score
Random Baseline, Accuracy: 0.53				
NC	0.47	0.39	0.54	0.45
C	0.46	0.67	0.53	0.59
W avg	0.46	0.57	0.53	0.54
Selected features, Accuracy: 0.76				
NC	0.16	0.68	0.62	0.65
C	0.38	0.80	0.84	0.82
W avg	0.30	0.76	0.76	0.76

(b) Test set

webpage is). Overall, this test rejected only 6 features, for which the test was not statistically significant.

One way ANOVA test is applied to analyze whether the feature values exhibit different means across classes (credible vs. non-credible webpages). Only few features (9 out of 35 features) were found to exhibit a significant difference between classes. For instance, pagerank and Alexa rank showed the highest discriminative capacity among these features: higher values are more related to credible webpage. Other features for which this test is statistically significant include the informativeness (i.e., informative content relates with credible webpages), how attentive the webpage design is, its social popularity, and text features (i.e., the number of adverbs, text polarity and the number of question marks).

Barring the categorical features (i.e., domain type, webpage category, webpage polarity) for which Spearman ρ test cannot be applied, we select the features that proved to be statistically significant in at least two of the tests (ensuring this way the features selection reliability). From the categorical features we keep all those significant in at least one test (e.g., *Chi2* test). We, thus, ended up with 22 features, listed in Table 2.

4.3 Credibility Prediction

For evaluation, we used *scikit-learn*[10] toolkit to train supervised classification and regression models. We experimented with several learning schemes such

[10] `www.scikit-learn.org`

Table 4. Results for regression analysis

R^2	**RMSE**	**MAE**	**Spearman** ρ
Random Baseline			
-1.82	1.82	1.46	0.06 (p-value> 0.01)
Selected features			
0.35	0.87	0.69	0.59 (p-value< 0.01)

(a) Cross-validation

R^2	**RMSE**	**MAE**	**Spearman** ρ
Random Baseline			
-2.86	1.77	1.45	0.01 (p-value> 0.01)
Selected features			
0.26	0.78	0.61	0.52 (p-value< 0.01)

(b) Test set

as support vector machine (SVM), decision trees, extremely randomized trees (ERT), and naive bayes for classification; and SVM and ERT's variants for regression. The obtained results were similar, but ERT[11] achieved slightly better results. Tables 3 and 4 summarize the results obtained with ERT for both classification and regression, respectively. To obtain the optimal model for ERT, we used grid search to determine the best parameter combination[12]. For all our experiments, the dataset has been split into 80% (706 webpages) learning set (used for 3-fold cross-validation) and 20% (177 webpages) test set (used to test how well the learning models generalize to new data).

Evaluation Metrics. To evaluate the classification performance we use popular metrics such as accuracy, precision, recall, and F1-score. We also look at the false positives (FP) rate, which is particularly important for the pages labeled as credible (i.e., indicating that the non-credible webpages are predicted to be credible, which is undesirable). To evaluate the performance of the regression models we use root mean squared error (RMSE) and mean absolute error (MAE) to compare between different models, and the coefficient of determination (R^2) and the rank correlation (Spearman ρ) to determine to what extent our predicted values can explain the real ratings of the webpages.

Table 3 shows that, overall, the classification accuracy is around 75%, significantly higher than of a random predictor. However, the prediction yields different performance for each class. While the F1 score is high for credible webpages ('C' in the table), indicating a good tradeoff between precision and recall, for non-credible webpages ('NC' in the table) the F1 score is much lower, in particular, due to a lower recall. Additionally, the false positives (FP) rates show that while only a few credible webpages are labeled as non-credible, roughly 2x more non-credible webpages are classified as credible. This along with the low recall for non-credible webpages raises a concern that the classifier is too optimistic and assesses non-credible pages as credible. The results obtained on the test set are similar with those obtained with cross-validation, showing that the predictive model is generalizing well to new data.

[11] In this learning scheme, the trees were built using the CART learning algorithm.

[12] We tried different values for number of trees, maximum depth, and minimum sample split. The best combination found is 50, 15 and 1 for classification, and 50, 10 and 3 for regression. Other adjustable parameters in scikit-learn are set to default.

Table 5. Classification results when different features types are used

Class	FP rate	Precision	Recall	F1 score
Content features, Accuracy: 0.73				
NC	0.16	0.72	0.56	0.63
C	0.44	0.73	0.84	0.78
W avg	0.32	0.73	0.73	0.72
Social features, Accuracy: 0.68				
NC	0.22	0.64	0.55	0.59
C	0.45	0.71	0.78	0.74
W avg	0.35	0.68	0.68	0.68

(a) Cross-validation

Class	FP rate	Precision	Recall	F1 score
Content features, Accuracy: 0.72				
NC	0.13	0.66	0.46	0.54
C	0.54	0.74	0.87	0.80
W avg	0.39	0.71	0.72	0.71
Social features, Accuracy: 0.65				
NC	0.25	0.52	0.48	0.50
C	0.52	0.72	0.75	0.74
W avg	0.43	0.65	0.65	0.65

(b) Test set

Table 6. Results for regression analysis

R^2	RMSE	MAE	Spearman ρ
Content Features			
0.24	0.95	0.76	0.50 (p-value< 0.01)
Social Features			
0.26	0.93	0.74	0.55 (p-value< 0.01)

(a) Cross-validation

R^2	RMSE	MAE	Spearman ρ
Content features			
0.26	0.77	0.63	0.49 (p-value < 0.01)
Social Features			
0.07	0.87	0.70	0.43 (p-value < 0.01)

(b) Test set

Misclassified Webpages. The high FP rates for non-credible webpages motivated us to analyze the feature values for the misclassified webpages. We noticed that webpages with short textual content, but high social popularity, were predicted as credible, while in the dataset they were labeled as non-credible. This could be explained by the definition used to rate the webpages in the Microsoft dataset (Section 4.1), which might account credible, but poor in information, webpages as non-credible.

For regression, table 4 shows that slightly more than 30% of the ratings variation is explained by the variation in our predicted values, compared with the random model which does not even follow the trend of the ratings. Both RMSE and MAE show a significant improvement w.r.t. the random model. Additionally, the Spearman ρ indicates a statistically significant, positive and moderate monotonic dependence between the predicted and the true credibility ratings.

4.4 Prediction Performance for Different Feature Types

We now study how good predictors either the social features, or the content features are. For classification (Table 5), the content features generally yield better performance over all metrics. In fact, the results obtained with the content features are close to those obtained with all features (Table 3), the few percent dropped in accuracy being caused by the poorer detection of non-credible webpages. Nevertheless, the non-credible webpages are in general harder to predict.

Table 7. Top 5 best features among all the experiments

All features	
Classification	Regression
@page_rank	@page_rank
#css_definitions	@smog
@alexa_rank	#css_definitions
@smog	?polarity
#alexa_linksin	#alexa_linksin

Content features	
Classification	Regression
#questions	@smog
@smog	domain_type
@informativeness	#questions
#css_definitions	@informativeness
domain_type	#exclamations

Social features	
Classification	Regression
@page_rank	@page_rank
@alexa_rank	@alexa_rank
#alexa_linksin	#alexa_linksin
#tweets	#delicious_bookmark
#fb_total	#tweets

In contrast, for regression (Table 6), using the social features leads to slightly better results: smaller error rates and higher correlations. However, while the model built with content features generalize well, the one build with social features overfits the data. This could be explained by the higher number of features that the regression model needs to account, and the more noisy features (while high values for the social popularity based features relate more with credible webpage, their absence is not necessary a good indicator for lack of credibility).

Best Features. Table 7 shows the top best features for both classification and regression tasks, which are ranked based on their relative distances from the root of trees in ERT model. Link-based features (e.g., @pagerank and #alexa_linksin) appear consistently close to the root when the social features are used. In contrast, Alexa rank appears to be a good feature for the classification task only. On the other hand, the number of CSS definitions and SMOG score are good features when the content features are included. When only the content features are used, the number of question marks, the domain type and the text informativeness are relevant for both classification and regression tasks. When relying only on the social features, the social popularity-based features are placed very close to the root (in particular, the number of tweets). Additionally, Alexa rank becomes relevant for the regression task as well.

5 Conclusions

Assessing web credibility is important for avoiding being misled by inappropriate sources of information. However, it also exposes important challenges due to the open nature and the large scale of the web. To provide end users support, in this paper, we present a general-purpose fully automated web credibility assessment framework, based on supervised learning algorithms. To this end, we first did an extensive survey of the relevant literature, which resulted in a super-set of 37 webpage features that encompass numerous attributes involved in web credibility assessment and deemed relevant by the prior work in different contexts. We leverage these features, which are gathered from various sources: textual content, webpage design, link structure, social popularity, etc. We presented how to refine them into a relevant subset by applying statistical tests, on which a supervised learning algorithm yields good results on a real dataset: for classification we obtained an accuracy of 75%, while for regression we obtained an improvement of roughly 53% for both MAE and RMSE over the random baseline approach.

Acknowledgments. We thank Stefan Bucur and Tri Kurniawan Wijaya for their valuable feedback on the earlier versions of this work. This work was partially supported by the grant *Reconcile: Robust Online Credibility Evaluation of Web Content* from Switzerland through the Swiss Contribution to the enlarged European Union.

References

1. Fogg, B.J., Tseng, H.: The elements of computer credibility. In: Proc. of CHI (1999)
2. Burbules, N.: Paradoxes of the web: The ethical dimensions of credibility, vol. 49. University of Illinois Library School, Urbana (2001)
3. Schwarz, J., Morris, M.: Augmenting web pages and search results to support credibility assessment. In: Proc. of CHI (2011)
4. Yamamoto, Y., Tanaka, K.: Enhancing credibility judgment of web search results. In: Proc. of CHI (2011)
5. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering (2005)
6. Aggarwal, S., Van Oostendorp, H.: An attempt to automate the process of source evaluation. In: Proc. of ACE (2011)
7. Rubin, V., Liddy, E.: Assessing credibility of weblogs. In: Proc. of CAAW (2006)
8. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proc. of WWW (2011)
9. Weerkamp, W., de Rijke, M.: Credibility improves topical blog post retrieval. In: Proc. of ACL (2008)
10. Fogg, B.J., Soohoo, C., Danielson, D.R., Marable, L., Stanford, J., Tauber, E.R.: How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In: Proc. of DUX (2003)
11. Hsu, C.F., Khabiri, E., Caverlee, J.: Ranking comments on the social web. In: Proc. of CSE (2009)
12. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proc. of WSDM (2008)
13. Morris, M., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing?: understanding microblog credibility perceptions. In: Proc. of CSCW (2012)
14. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on twitter. In: Proc. of SIAM (2012)
15. Caverlee, J., Liu, L.: Countering web spam with credibility-based link analysis. In: Proc. of PODC (2007)
16. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proc. of VLDB (2004)
17. Sondhi, P., Vydiswaran, V.G.V., Zhai, C.: Reliability Prediction of Webpages in the Medical Domain. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 219–231. Springer, Heidelberg (2012)
18. LLC, O.: AlchemyApi, `http://www.alchemyapi.com/` (retrieved on September 2012)
19. Mc Laughlin, G.: Smog grading-a new readability formula. Journal of reading (1969)
20. Fogg, B.J.: Prominence-interpretation theory: explaining how people assess credibility online. In: Proc. of CHI (2003)