Predicting Webpage Credibility using Linguistic Features

Aleksander Wawer
Institute of Computer Science,
Polish Academy of Sciences
Jana Kazimierza 5
Warsaw, Poland
axw@ipipan.waw.pl

Radoslaw Nielek
Polish-Japanese Institute of
Information Technology
Koszykowa 86
Warsaw, Poland
nielek@pjwstk.edu.pl

Adam Wierzbicki
Polish-Japanese Institute of
Information Technology
Koszykowa 86
Warsaw, Poland
adamw@pjwstk.edu.pl

ABSTRACT

The article focuses on predicting trustworthiness from textual content of webpages. The recent work Olteanu et al. proposes a number of features (linguistic and social) to apply machine learning methods to recognize trust levels. We demonstrate that this approach can be substantially improved in two ways: by applying machine learning methods to vectors computed using psychosocial and psycholinguistic features and in a high-dimensional bag-of-words paradigm of word occurrences. Following [13], we test the methods in two classification settings, as a 2-class and 3-class scenario, and in a regression setting. In the 3-class scenario, the features compiled by [13] achieve weighted precision of 0.63, while the methods proposed in our paper raise it to 0.66 and 0.70. We also examine coefficients of the models in order to discover words associated with low and high trust.

Keywords

credibility prediction, trust, text classification

1. INTRODUCTION

Credibility is an important feature. High credible sources are more valuable. Politicians struggle to appear as credible. Finally, authors and publishers want to make their books and web sites credible. According to the Merriam-Webster dictionary, word credible means "offering reasonable grounds for being believed"ï£¡ (trustworthy is one of synonyms). Credibility, which is partly subjective, should not be confused with truth – usually understood as an objective category. People investigate content to judge whether they should believe in given information or not. The paper focuses on web sites credibility prediction using variety of features, mostly textual and linguistic. In particular, precision of classifiers based on bag-of-words vector spaces and psycholinguistic dimensions has been compared with a state-of-the-art selection of features.

The rest of the paper is organized as follows. In the next Section related works are presented. Section 3 describes web page features used for predicting trust and a brief description of the dataset. The results of classification and regression experiments are presented in Section 4, along with an analysis of the most trustworthy words. Section 5 summarizes the paper and investigates promising future research areas.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579000.

2. RELATED WORK

At the very early days of WWW development, content credibility was not an issue. The number of Internet users was low and mostly limited to academics, big corporations and governments. It was a really small community where (almost) everyone knew each other.

Increasing popularity of the Internet in the mid-90s dismissed this strong control mechanism. Within five years, from 1995 to 2000, the percentage of society having Internet access in developed countries grew from small single-digit numbers to about 40 percent (in Australia or New Zealand it peaked to around 50 percent[7]).

On par with the majority of society issues known from the offline world have appeared on the Internet in their new, modified form (among them content credibility). At that time, the dominant idea of solving these issues was based on developing online ethics and convincing individuals to "internalize norms of behaviour"[8].

Since the early 50s, the concept of credibility has been widely studied by psychologists, media experts and economists. Most publications focus on either persuasive effect of source credibility [18] [6] [15] or media credibility [4] or importance of credibility for economical theories [17]. Most researchers agree that credibility is not a property of an object, person or piece of information but is rather "a perceived quality" [20].

The first scientific paper studying credibility in the context of computing technology was published only 40 year later by Tseng [20]. He proposed four types of credibility: presumed, reputed, surface and experienced. The first two types are based on either stereotypes or third party reports. The last two are derived from individual's own experiences. Some people can argue that the proposed categories are essentially heuristics use to assess credibility and do not define different types of credibility. This view seems to be strengthened by the definition of credibility as believability given by Fogg [3]. According to Fogg et. al [3] credibility can be seen as chance that users believe in given information.

A person that is physically attractive is perceived as more credible [14]. People constantly use some signals to estimate personal credibility and the same type of mechanism exists also for assessing web content credibility. A study on

over 2500 Internet users conducted at the Stanford University revealed 18 areas that people notice when assessing web site credibility [2]. Almost 50% of participants pointed to design and look, one-fourth on information design and information structure. Bias of information and tone of writing are present only in around 10% of comments, what may indicate that the goal of this paper may be hard to reach. On the other hand asking people explicitly about features they use to assess credibility can only reveal heuristics they are aware of.

Inspired by this (and similar) research, the prominence interpretation theory has been proposed [1]. The theory assumes that a user has to notice a particular feature and only then he starts evaluating it. This process is repeated many times by each user for each web site and its efficiency depends strongly on user's motivation and experience. The prominence-interpretation theory is mainly focused on conscious processing and ignores pre-apprehensions and feelings in general. Some features can be difficult to notice by people (e.g. number of question marks or punctuations) but still may be a good approximation of dimensions that are much easier to process manually.

Heuristics and signals used by people to evaluate web site credibility change constantly. New technologies, successful on-line frauds, changes in culture and education shift people's attention from one feature to another and modify rules associated with them. One of the most impressive paradigm changes was caused by Web 2.0 revolution. Virtually all Internet users have started to produce content and quite often publish it in a very structured way enforced by someone else (e.g. user publishing comment under an article does not decide about design, nor surrounding content; for comments but also posts on blogs some limits like length or number of pictures are enforced).

People learn to make the use of web site specific information for assessing credibility. They use position in an Internet search engine (higher position indicates more reliable information [16] or following graphs and presence of shortened URLs for tweets [11]. Popularity is correlated with credibility. Giudice shows that more people visiting web site usually means more credible content [5].

3. FEATURES FOR PREDICTING WEB SITE CREDIBILITY

3.1 Webpage features

Olteanu et al. [13] collected 37 webpage properties, potentially useful for credibility assessment. All features have been divided into two categories: content-related and prominence-related. Both are further divided into text, appearance and meta-information, social popularity, general popularity and link popularity, respectively. The complete list of webpage properties can be found in [13] but to give a quick insight selected examples are presented below:

- **content-related:** number of exclamation marks in the text, polarity, spelling errors, category etc.,
- **prominence-related:** Facebook share, number of tweets mentioning a webpage URL, page rank, Alexa rank, etc.

Some properties are expressed as numbers (usually ranged from zero to infinite), others are binary, and one is discrete (a selection from possible categories). Properties calculation relies heavily on external libraries, scripts and APIs. Webpage category is detected using the Alchemy API[1], all ad related properties are measured using AdBlock scripts[2], text related features are calculated using the NLTK library[3] or simple regular expressions (e.g. for calculating number of exclamations). Some features like PageRank, Facebook share or Alexa rating can be calculated only by invoking an appropriate API delivered by data owners (i.e. Google, Facebook or Alexa).

After feature selection process described in detail in [13] only 22 previously identified webpage properties turn out to be relevant for credibility prediction. Twelve of them are prominence-related and the other ten content-related.

On the one hand, webpage properties selected in [13] cover a broad range of aspects – from presentation through popularity to content. On the other hand, the list of features is not exhaustive. For example, Google uses over 200 webpage properties to rank search results [4]. As is shown in [16] position in search results correlates, although weakly, with webpage credibility judgements. Much stronger positive correlation is observed for popularity [13]. Similar results obtained by authors of this paper are presented on fig. 1.

Although contextual and popularity features correlate with credibility evaluation, it is not a simple cause-effect. This kind of properties are used for approximating webpage credibility. However, credibility in the sense of factual correctness and completeness derives exclusively from text and pictures on the page.

Extensive use of external APIs for features calculation, as in the case of [13], has many advantages but also some drawbacks. First of all, it limits time and costs of implementation and makes research easier to replicate for other scientists. Instead of training new classifiers for webpage categorization, the API provided by Alchemy can be used (and indeed has been used in this research). To calculate some other webpage properties, there exist no reasonable alternatives except for external APIs: calculation of PageRank requires crawling the whole Internet[5], popularity on Facebook or Twitter can be measured only by data owners. APIs provided by these companies are the subject of many limitations, such as the number of requests per day and non-commercial applications. Additionally, vendors may change algorithms that calculate particular features without notification and such changes may have consequences for credibility of classification algorithms. Therefore, solutions based purely on external APIs are difficult to use beyond scientific application and are prone for manipulation. In this paper two exclusively text-based approaches to create features have been tested. Additionally, these approaches can be calculated directly without the use of third-party vendors.

3.2 Dataset

To evaluate the performance of our approach, we use a dataset built by Microsoft for a study that analyzes if famil-

[1] www.alchemyapi.com
[2] www.adblockplus.org
[3] www.nltk.org
[4] http://www.searchenginejournal.com/infographic-googles-200-ranking-factors/64316/
[5] There are some heuristics which make it possible to estimate PageRank with less calculation

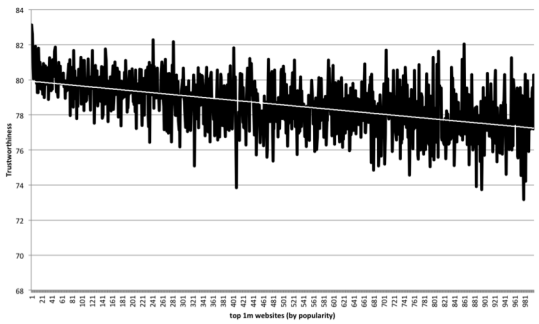

Figure 1: Average trustworthiness of websites on MyWOT in relation to their position in Alexa ranking (ordered from the most popular)

iarizing users with some web page features which are hard to estimate for them (e.g. web popularity) can help them to better assess webpage credibility [16]. 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[6]. All these webpages are rated on a five-point Likert scale (where 1 means "very non-credible" and 5 "very credible").

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

Proposed extension of the existing approaches is based on two types of features. The first one, covered in Section 3.3 is based on existing dictionaries, built upon various theoretical views of categorization of textual contents. The second is an opposite one, based on inferring dictionaries directly from texts labelled with trustworthiness, in a supervised fashion. In both approaches, predictions are based on models trained on word lists, but whereas in the first case the lists are predefined and input vectors represent aggregations over existing lists of words, in the second case each word may become a feature and thus obtain a score reflecting its relationship to the predicted category.

Despite numerous attempts, we could not replicate the exact results reported by [13]. Even though the dataset is the same one as in our experiments, the numbers were different. We suspect that this fact may be explained by likely differences in the implementations of evaluation and parameter estimation procedures. Therefore, we computed the results for features compiled in [13] using the same evaluation procedures for every other feature set under evaluation.

3.3 General Inquirer

The General Inquirer (GI) [19][7] is one of the most well-known content analysis tools. It consists of an application and an associated dictionary.

The dictionary contains 11,767 word senses[8] mapped to 183 categories. The notion of category is central to content analysis. As [10] puts it, a category is a group of content – in this case word senses – that shares a commonality (possesses a shared feature or attribute). GI categories are linked with multiple psycholinguistic and psychosocial categories. The list of GI categories includes for example topic-based ones (politics, economy, religion), several emotion-related categories such as pleasure, pain, feelings or arousal. Two of the categories, Positive and Negative, represent evaluative dimension (sentiment). Category membership is binary: words either belong to a category or not.

For text processing, the GI application uses a dictionary-backed lemmatizer and word sense disambiguation rules by [9]. For each document in the dataset, the GI application produces a vector of 183 numbers, which represent counts of each category in every analysed text.

Obviously, there are other linguistic resources and approaches available that might be used to compute word categories. For example, one might apply a word clustering algorithm or word grouping based on WordNet. However, we believe that the General Inquirer is a resource especially relevant to trustworthiness measurement due to its psychosocial characteristics, especially multiple psychological traits of language.

4. RESULTS

4.1 Regression

In the regression setting, we compare the results using several well-known error (and goodness of fit) measures: the R^2, root mean-square error ($RMSE$), mean absolute error (MAE) and explained variance ($ExplVar$). In each case, the scores represent average values obtained in a 10-fold cross-validation. 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). While the authors use SVM and ERT's (Extremely Randomized Trees) variants for regression, we compare the datasets using simple linear models such as ridge and linear regression. We believe simpler methods are more stable, especially compared to the ERT which tends to produce estimators of varied quality on the same data, and thus more appropriate for comparisons of feature sets.

We preceded the computing of regression models with an optional feature selection according to a percentile of the highest scoring features (as to their F scores). We tested the choice of 20, 50, 80 and 100 percentiles (thus, no selection) of features and found that in every case, the best performing regression models used only top 20th percentile of features. The optimum percentile was selected in a cross-validation setting of 10 folds.

Features	Method	R^2	$RMSE$	MAE	$EVar$
[13]	Ridge	0.055	0.911	0.761	0.154
	Linear	0.055	0.079	0.762	0.155
[13] and GI	Ridge	0.074	0.900	0.759	0.164
	Linear	0.079	0.882	0.751	0.147

Table 1: [13] and GI features (regression)

[6]Dataset can be downloaded from `http://research.microsoft.com/en-us/projects/credibility/`

[7]`http://www.wjh.harvard.edu/~inquirer/`

[8]Most of the dictionary words have a single sense, though.

The results indicate that the introduction of General Inquirer features increases prediction performance. The combined feature space of [13] and GI performs better in according to three metrices: it achieves higher R^2 values and lower $RMSE$ and MAE scores. Only in the case of explained variance $EVar$, linear models explain less variance in the combined feature space than in the features of [13].

4.2 Classification

4.2.1 Three classes

In this experiment, we divide the data according to their trustworthiness so that sites scored lower than 2 are now not trusted (label *not* – 164 cases), value of 3 is labeled as medium trust (label *med* – 191 cases), and sites scored higher than 4 are marked as of high trust (label *high* – 524 cases).

We use the above data to compare different features. For the evaluation, we use scores obtained from 10-fold cross-validation and parameter selection in a grid search scenario. We present three different metrics: precision, recall and F-measure, averaged between 10 folds. Also, we present the data in two general settings, first by providing class-level statistics (for **not**, **med** and **high**) and as total averages, computed as three metrics:

- 'micro': Calculate metrics globally by counting the total true positives, false negatives and false positives.
- 'macro': Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.
- 'weighted': Calculate metrics for each label, and find their average, weighted by support (the number of true instances for each label). This alters ï£¡macroï£¡ to account for label imbalance; it can result in an F-score that is not between precision and recall.

Both micro and weighted metrics take into account class imbalance, therefore are more suitable for datasets such as this.

In all of the comparisons below we use the logistic regression classification algorithm and select features using ANOVA scores. However, as the percentage of selected features is optimized in parameter selection procedure based on cross-validation, it is also possible that the best performance might be obtained using only a subset of features.

Table 2 presents the results for the dataset described in [13]. The best performing parameter combination was the one with 20th percentile of features with the highest scores, l1 penalty and C value of 100.

	Precision	Recall	F
not	0.498	0.579	0.528
med	0.366	0.304	0.329
high	0.762	0.767	0.764
micro	0.63	0.63	0.63
macro	0.542	0.55	0.54
weighted	0.631	0.63	0.627

Table 2: [13] features (3-class)

Table 3 presents the results for the dataset described in [13] extended with 183 variables from the General Inquirer (GI) [19]. The best performing parameter combination was the one with 50th percentile of features with the highest scores, l1 penalty and C value of 10.

Class	Precision	Recall	F
not	0.508	0.608	0.545
med	0.401	0.347	0.366
high	0.795	0.782	0.787
micro	0.654	0.654	0.654
macro	0.568	0.579	0.566
weighted	0.66	0.654	0.652

Table 3: [13] and GI features (3-class)

Comparing tables 2 and 3 demonstrates that the General Inquirer psycholinguistic features improve the performance of features compiled in [13].

Table 4 presents the results obtained using bag-of-words method in the 3-class scenario.

	Precision	Recall	F
not	0.65	0.601	0.62
med	0.483	0.349	0.397
high	0.776	0.882	0.823
micro	0.716	0.716	0.716
macro	0.636	0.611	0.613
weighted	0.702	0.716	0.699

Table 4: Learning from bag-of-words lexical space (3-class)

4.2.2 Two classes

Now, we divide the data so that sites scored lower than or equal to 3 are now marked for low trust (label *low* – 355 cases), and sites scored higher or equal to 4 are marked as of high trust (label *high* – 524 cases).

As before, we use the above data to compare feature sets, present scores obtained from 10-fold cross-validation and search optimum set of parameters in a grid search scenario. Because the data is binary, we present only precision, recall and F-measure for each class, averaged between 10 folds.

Table 5 presents the results for the dataset described in [13]. The best performing parameter combination was the one with 20th percentile of features with the highest scores, l1 penalty and C value of 100.

	Precision	Recall	F
low	0.624	0.734	0.67
high	0.798	0.698	0.74

Table 5: [13] features

Table 6 presents the results for the dataset described in [13] extended with 183 variables from the General Inquirer (GI) The best performing parameter combination was the one with 20th percentile of features with the highest scores, l1 penalty and C value of 10.

Finally, we try another well-known method: machine learning from word occurrences. The approach, also known as

	Precision	Recall	F
low	0.665	0.768	0.712
high	0.825	0.738	0.778

Table 6: [13] and GI features (2-class)

bag-of-words or unigram-based learning, disregards word order and grammar. We evaluated a set of parameters applicable to this method including TF-IDF weighting. The optimum-performing combination was the one without TF-IDF weights and models trained using 100% of features selected by ANOVA scores, l2 penalty and C value of 100. Table 7 presents the results obtained using bag-of-words method in the 2-class scenario:

	Precision	Recall	F
low	0.757	0.737	0.740
high	0.824	0.835	0.827

Table 7: Learning from bag-of-words lexical space (2-class)

4.3 Words of High and Low Trust

Using logistic regression over vector space models, one may examine coefficients associated to words in order to find out words that invoke high trust and identify those that are associated with untrustworthiness and low trust. We divided the data set into 524 pages with ratings over 3 and 355 pages with ratings below 4. Then we trained a logistic regression classifier on the dataset. To produce weights for all possible words (unigrams) we did not apply feature selection, thus the full list includes 72280 items. We removed proper names, numbers and inflected variants of words that already appear on the list. Tables 8 and 9 presents the results, narrowed to the first 20 and followed by another 10, manually selected from the top 100.

We should note that the annotated sample of pages is rather small and any interpretations or generalizations should involve a degree of caution, as the results presented may be a consequence of over-representation of a certain kind of web pages or content types.

The lists hint at possible variation in the apriori trustworthiness of different topics. In other words, some topics may be generally more trustworthy than others. These considerations, confirmed recently in [12], point to the possibility that bag-of-word representation could indeed be aligned with topic-based classification. However, the verification of this hypothesis falls out of the immediate scope of this paper and may need more robust data sets with topic labels assignment.

Topic-wise, the examination might be carried by focusing on nouns. High trust appears to be associated with *retirement*-related content, *energy*, *research* (*publications*, *reprints*) and government content (*department*, *fed*, *gov*). Quite unsurprisingly, the list included words such as *clean*, *safety*, *security*, *ethics*, *green*. Selected healthcare-related words appear on the list (*symptoms*, *clinic*).

The list of words with negative scores, associated with low trust, contains multiple financial items. The items are very specific though, as most of the words are associated with borrowing money (*refinancing*, *debt*, *bank* and *loans*); the only words, related with investments, are *invest* itself and *gold*. Words of low trust include typical phrases associated with user generated content (*blog*, *forums*, *posted*, *facebook*), *questions* and *answers*. Health-related words include *diabetes*.

5. CONCLUSIONS

The article is a follow-up to the former work on trustworthiness of web pages by [13]. We demonstrate that the results can be significantly improved in two ways. First, by applying machine learning methods on vectors obtained from the General Inquirer (a well-known tool and dictionary for content analysis, vocabulary lists related to key psychosocial and psycholinguistic theories). Second, by using machine learning applied in a supervised fashion, in a bag-of-words approach (learning word occurrences). The second method is highly-dimensional (over 70 thousands features) but at the same the most accurate.

In a 3-class scenario, the dataset of [13] achieves average weighted precision of 0.63, the introduction of the General Inquirer features raises it to 0.66. The application of supervised learning on unigrams raises the weighted precision further to reach 0.70. Similar improvements can be observed also in the 2-class scenario. Also in the case of regression, proposed improvements outperform the features compiled by [13].

We also examine coefficients of the supervised models in order to discover words associated with low and high trust. We observe that people are less likely to trust user-generated contents and financial services and operations related to borrowing money. Words of high trust are associated with the government and include vocabulary linked to safety.

Many interesting research questions have arisen during preparation of the paper. Does the focus on single subject (e.g. medicine, investment etc.) or webpage type (blogs, news portals, e-commerce) will substantially influence precision and recall? How efficient are the same linguistic features for websites written in languages other than English? Popularization of feature-based credibility evaluation will eventually ignite cat-and-mouse play between researchers and people interested in manipulating such algorithms. Therefore, immunize algorithms against their attacks will be, both, interesting and challenging.

6. ACKNOWLEDGMENTS

This work was partly supported by the grant âĂİReconcile: Robust Online Credibility Evaluation of Web ContentâĂİ from Switzerland through the Swiss Contribution to the enlarged European Union. Adam Wierzbicki has been supported by Polish National Science Centre grant 2012/05/B/ST6/03364.

7. REFERENCES

[1] B. J. Fogg. Prominence-interpretation theory: explaining how people assess credibility online, 2003.

[2] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants, 2003.

[3] B. J. Fogg and H. Tseng. The elements of computer credibility, 1999.

Top 20						Manually selected 10	
Words	**Weights**		**Words**	**Weights**		**Words**	**Weights**
review	3,95		calculator	3,94		company	2,43
guide	3,5		retirement	3,42		symptoms	2,37
energy	3,34		clean	3,19		security	2,32
research	3,02		genetic	2,93		ethics	2,32
range	2,91		reprints	2,9		green	2,31
provided	2,76		department	2,73		amortization	2,26
parents	2,69		safety	2,68		efficient	2,22
alerts	2,68		ride	2,51		clinic	2,12
publications	2,5		notice	2,5		fed	1,97
included	2,48		additional	2,48		gov	1,95

Table 8: Top words of high trust

Top 20						Manually selected 10	
Words	**Weights**		**Words**	**Weights**		**Words**	**Weights**
invest	-4,74		tea	-4,55		posted	-2,91
posts	-4,39		debt	-4,35		forums	-2,8
article	-3,99		gold	-3,8		doityourself	-2,47
court	-3,72		reply	-3,65		refinancing	-2,4
question	-3,64		panels	-3,54		toxic	-2,4
answers	-3,54		best	-3,46		diabetes	-2,31
return	-3,41		wise	-3,41		bank	-2,17
party	-3,2		blog	-3,18		wikipedia	-2,16
great	-3,12		good	-3,08		loans	-2,05
birthplace	-2,96		years	-2,95		facebook	-2,01

Table 9: Top words of low trust

[4] C. Gaziano and K. McGrath. Measuring the concept of credibility. *Journalism Quarterly*, 63(3):451–462, 1986.

[5] K. D. Giudice. Crowdsourcing credibility: the impact of audience feedback on web page credibility, 2010.

[6] C. I. Hovland and W. Weiss. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15(4):635–650, 1951.

[7] W. D. Indicators. Internet users (per 100 people) | data | table, 2012.

[8] D. G. Johnson. Ethics online. *Commun. ACM*, 40(1):60–65, 1997.

[9] E. Kelly and P. Stone. *Computer Recognition of English Word Senses*. North-Holland, Amsterdam, 1975.

[10] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 1980.

[11] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions, 2012.

[12] R. Nielek, A. Wawer, and A. Wierzbicki. Temporal, cultural and thematic aspects of web credibility. In *The proceedings of Social Informatics - 5th International Conference*, Lecture Notes in Computer Science, pages 419–428. Springer, 2013.

[13] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer. Web credibility: features exploration and credibility prediction. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 557–568, Berlin, Heidelberg, 2013. Springer-Verlag.

[14] G. L. Patzer. Source credibility as a function of communicator physical attractiveness. *Journal of Business Research*, 11(2):229–241, 1983.

[15] C. Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2):243–281, 2004.

[16] J. Schwarz and M. Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1245–1254, New York, NY, USA, 2011. ACM.

[17] J. Sobel. A theory of credibility. *Review of Economic Studies*, 52(4):557–573, 1985.

[18] B. Sternthal, R. Dholakia, and C. Leavitt. The persuasive effect of source credibility: Tests of cognitive response. *Journal of Consumer Research*, 4(4):252–260, 1978.

[19] P. J. Stone, D. C. Dunphy, D. M. Ogilvie, and M. S. Smith. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[20] S. Tseng and B. J. Fogg. Credibility and computing technology. *Commun. ACM*, 42(5):39–44, 1999.