Credibility Microscope: relating Web page credibility evaluations to their textual content

Wojciech Jaworski*†, Emilia Rejmund* and Adam Wierzbicki*
*Polish-Japanese Institute of Information Technology,
Koszykowa 86, 02-008 Warsaw, Poland (s11444,adamw)@pjwstk.edu.pl
†Institute of Informatics, University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland wjaworski@mimuw.edu.pl

***Abstract*—The popularization of the Internet made it a primary information source for many people. Unfortunately quality of information available on the Internet varies. Therefore, evaluation of credibility of web page content, especially while making important decisions like those concerning health care, medical information, and large purchases, is crucial, but users often lack a necessary knowledge. The main goal of the paper is to study to what extent the textual content of a webpage determines its credibility evaluations. This goal is achieved by an experiment in which we ask respondents to rate webpage credibility as well as credibility and importance of each statement from this site. We formulate a number of hypotheses about the nature of the dependence between webpage and statements credibility and we test these hypotheses on the data obtained from experiment. The evaluation of those hypotheses is essential for design of classifier that will aggregate statement credibility into webpage credibility score.**

***Keywords*—Decision support, Credibility assessment**

I. Introduction

Early works on web content credibility subject [3] have focused on structural and author specific features. Among others, authors investigated the impact of credentials of the site, the presence of advertisements, and Web design[11]; credibility of the source of information [1]; the frequency of emoticons and spelling errors, as well as the length of text [13]; content maturity and readability [12]; number of machine extractable facts in text [8]; time distribution correlation with the authoritative source [4].

Unfortunately, all these features can be easily manipulated by marketing experts in order to increase credibility of the content [2], which, in fact, is not credible. The only feature that cannot be falsified in this way, is a textual content. Several systems employing content based credibility assessment have been proposed: TruthGoogles [10] is a browser plugin that scans a Web page text and identifies phrases that are present in an external authoritative source of factchecked statements.

Hypothes.is is a project of creating an annotation maintainer. Its goal is to collect comments concerning every single sentence on the Web.

WISDOM [6] is an already deployed decision support system. It collects pages related to a given topic and then classifies them according to the category of information providers and determine the sentiment towards examined topic. On this basis, WISDOM provides opinion statistics and example opinions for each class. WISDOM does not attempt to directly judge information credibility, but rather helps users judge credibility by aggregating and organizing Web information.

Juffinger et al. [4] proposed a system for estimating blog credibility on the basis of verified content. In order to assess credibility they make use of both structural information and blog contents.

Finally, Murakami et al. [9] described a vision of Statement Map The system will retrieve documents related to a query from the Web and split them into statements. Then, it will determine whether statements are fact or opinions and will find relations between them.

The main goal of the paper is to study to what extent the textual content of a webpage determines its credibility evaluations. We also want to learn automatic credibility evaluation by constructing statements aggregator. In order to achieve this goals, we model webpage textual content as a bag of statements and we perform an experiment where users rate the credibility of the entire page, and of individual statements.

This way we wish to capture the nature of dependence between statement and webpage rating. Such knowledge will be crucial for future systems that will automatically assess webpage credibility on the basis of page and statement ratings.

We investigate the following hypotheses, which we plan to verify in our experiment:

Hypothesis 1: Credibility of a webpage depends on a credibility of its textual content.

Hypothesis 2: Credibility of a text is a function of credibility of important statements that compose that text.

Our experiment also studies hypotheses that can be important for constructing future approaches that use statements to classify Web page credibility automatically:

Hypothesis 3: Credibility and importance are independent.

Hypothesis 4: Evaluations of statements are useful in a process of classification of web page credibility.

Hypothesis 5: A few important unbelievable statement makes the text non credible.

In the rest of the paper, we present experiment set up and results that let us check the above hypothesis.

II. Experiment design

We consider a statement as an atomic unit of meaning. Statement is a representation of states of affairs such as: a

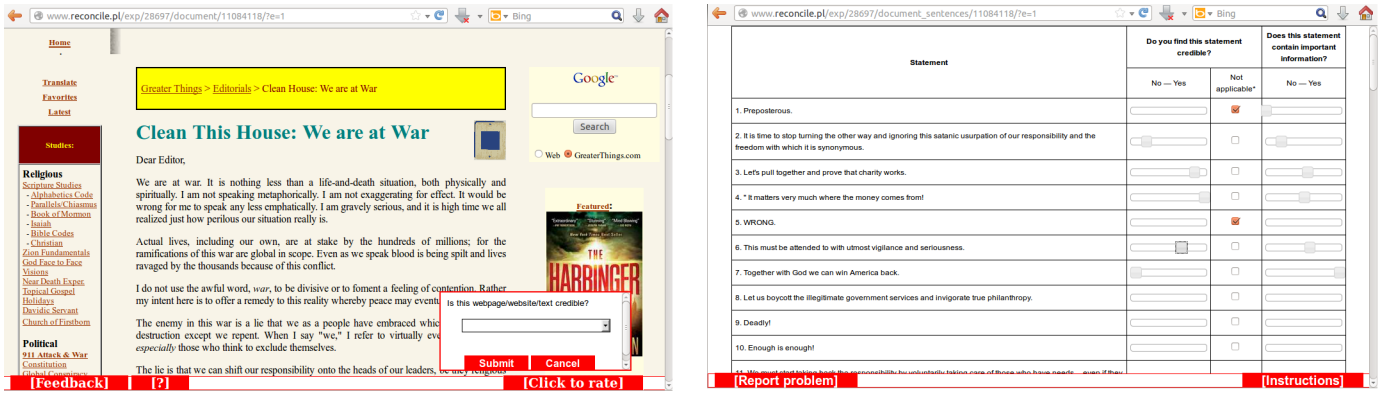

Fig. 1. Experiment: Stages A and C: The whole page is presented to the respondent, Stage B: An excerpt from the text is presented to the respondent

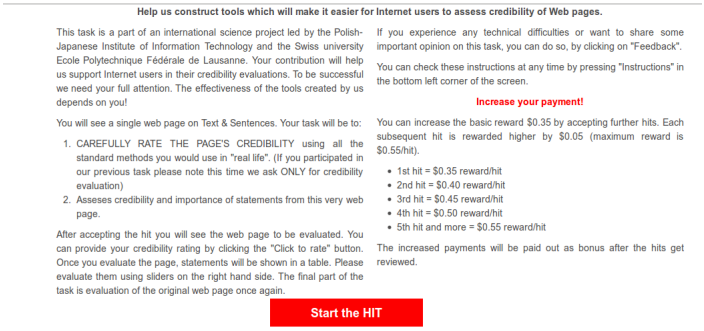

Fig. 2. Experiment instructions

possession of some properties by an object, an occurrence of an event or a certain relation between given objects, existence of a process, etc.

In our approach, we plan to determine text credibility, by means of classifiers, which use attribute values assigned to statements as explanatory variables.

In order to train the classifiers, we need a training set composed of text with manually labeled statements. The minimum unit of text that can be rated by a user is a sentence. However, before showing the text to respondents, we may split it into statements and then present each statement as a sentence.

Since statements are more philosophical abstractions rather than entities explicitly expressed in a text, we will approximate them using NLP techniques. As a first approximation, we follow [9] assuming that statements are equivalent to sentences.

To test our hypotheses and get training data for statement aggregator classifiers we need an experiment in which we obtained a corpus of human rated statements and learned the dependence between credibility of statements and credibility of a text.

The experiment consists of three stages. During the first stage A and last C we show to respondent a webpage on a given topic and ask him for its credibility. Then, in the middle stage B, we show to respondent statements extracted from a text and ask him questions concerning credibility and importance of these statements. The goal of the third stage C is to check whether respondent's opinion about the Web page has changed after he rated single statements. For the example of see Fig. 1. In the stages A and C, we use credibility scale that has either 5 or 101 points depending on the experiment variant. In the stage B, see Fig. 1, we ask rate the credibility of each statement in text in 101 point scale or indicate that credibility measure is not applicable for this statement (titles, exclamations and questions are examples of such statements). We also ask respondent to rate the importance of each statement with respect to the importance of the information that it contains. Finally, we ask whether the Web page contains important information that is not presented in the statement table. Statements in the table appear in the same order as they are presented in the text or in a random order, depending on an experiment variant. We present instruction for users on Fig. 2.

There are three modes of the experiment. In the first and third mode, Webpage is rated in 5 point scale and in second variant in 101 point scale. Moreover, in first and second mode statements are shown in the ordered way and in third mode they are randomly shuffled.

The goal of the second mode is to learn the translation between both scales and the goal of the third mode is to check whether an order of statements affects their ratings. When statements are presented in the same order as on the page they create each other context. Statements in random order simulate the lack of context.

In order to eliminate additional factors, we remove hiperlinks from texts.

We have performed the experiment with 100 pages taken from an alternative medicine domain. Each site has been rated 5 times for each mode. The number of single sentences presented to a user has been limited to 30.

We expect that results will depend on the domain similarity as in previous experiments.[3].

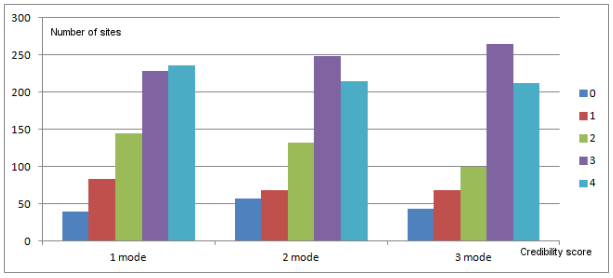

Fig. 3. Sites credibility pre scores histogram

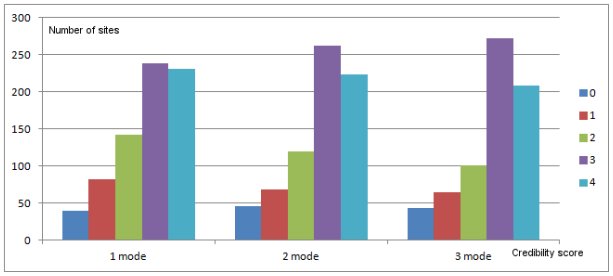

Fig. 4. Sites credibility post scores histogram

III. Analysis & Results

In the first stage of experimental data analysis we focused on site related results to asses evaluator's level of commitment and an influence of the extracted sentence on the evaluation of entire page. Our results contain two marks for every site denoted as pre and post mark that correspond to first and third stage of our experiment. A coefficient correlation of both marks was calculated and result 0,93 shows that this two values are strongly correlated. Further enhancement of analysis was to check correlation for each mode in which experiment was run. Since the same 0-4 scale was used, in first and third mode we expected similar results. The result of our analysis is 0,88 for first mode and 0,89 for third mode. The result for second mode was lower, 0,78, due to experiment set-up, because in this mode was used the 0-100 scale and less accurate slider method of evaluation.

Knowing that pre and post evaluation results are correlated, we set out to evaluate scores evolution. In order to check it, we compared both marks for each page. As the results for slider mode we obtained 7,7% of all answer equal, and for the joined five point scale modes the results was equal with 1% error range and was about 80%. The source of such phenomenon we sought in the accuracy of the clicking in slider. To verify it,we performed an indepth analysis for 101 point scale slider mode. We set the value of equal answer to 80% and changed the error range for two marks to be equal. The resulting accuracy range for 100 point slider was 15 points which constitutes also about 15%.

Our goal was to learn the recalculation between 101 point to 5 point scales. In order to gain accurate results, we decided to match the probability density function for each scale. A comparison of both scales is presented in table:

5 point scale	101 point scale
0-1	0-4
1-2	5-21
2-3	22-51
3-4	52-80
4-5	81-100

On Fig.3 and Fig.4 is presented pre and post results histogram for sites. Three different modes are included, the values for the second mode has been converted to 5 point scale, using the process described above. The correlation of pre and post marks is clearly visible. Both graphs are very similar for every mode. The second mode graph is a little bit different and we see it as a consequence of accuracy of 101 point slider method, which was used during the evaluation process. To verify independence credibility and importance hypothesis (Hypothesis 3) , we have calculated correlation between credibility and importance evaluation score for all sentences. The results 0,66 shows that this two variables are weakly correlated. As we can observe on scatter plot Fig. 8 the full spectrum of both variables is already covered. The highest density may be observed on the diagonal that corresponds to the positive correlation of coefficient value.

On the Fig. 5 and Fig. 6 we present a histogram of scores for statement credibility and importance. This two graphs are similar, which corresponds to a calculated correlation coefficient value. The spectrum of credibility and importance values are less similar for lower values and more similar for higher ones. We observed an overwhelming amount of the highest range scores 99-100 point.

Comparison of two graphs also shows that there are more pages with zero importance value than pages that were ranked zero credibility.

Our results are biased to the right side of the scale for sites and for statements. This phenomenon was previously observed by other researchers [7]. We encountered such a problem also in previous experiments in our research group [5].

A. Regression based credibility model

In order to prove our hypotheses right, we have constructed the classifier that predicts webpage credibility on the basis of statement credibility and importance. Each web page evaluation resulted with a list of credibility and importance values for each statement. We have transformed this list into a frequency vector in the following way: for a given list $\{cr_i, imp_i\}_i$ we have created a vector $a = \{a_j\}_j$, where the field a_j represents a sum of importance values for statements whose credibility is equal to j.

$$a_j = \sum_i [cr_i = j] imp_i$$

$[cr_i = j]$ is the Iverson bracket i.e.:

$$[P] = \begin{cases} 1 & \text{if } P \text{ is true;} \\ 0 & \text{otherwise.} \end{cases}$$

Since the respondents use 101 scale to rate statement credibility, we do not expect them to select the credibility value precisely. We model this fact by means of fuzzyfication. We

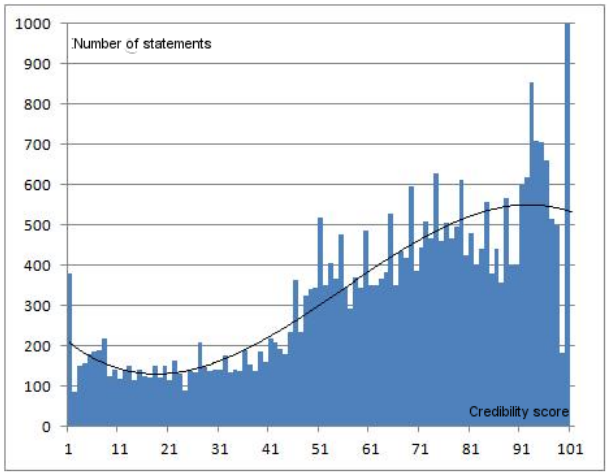

Fig. 5. Statements credibility scores histogram

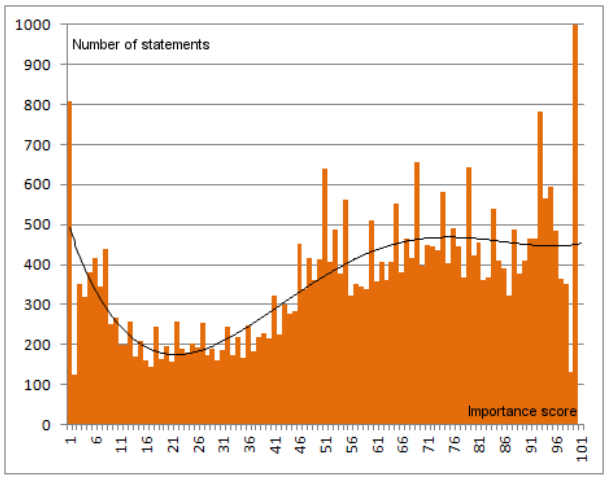

Fig. 6. Statements importance scores histogram

consider each credibility value as an expected value of a Gaussian distribution which predefines standard deviation σ.

$$a_j = \sum_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(cr_i - j)^2}{2\sigma^2}\right) imp_i$$

When $\sigma = 0$ this equation is an equivalent to the previous definition of a. In our analysis we consider σ in the range from 0 to 30.

We have explored linear regression classifiers attributes constructed on the basis of such frequency vectors. We have chosen linear regression because in our experiment, we have ordered decision classes (first and third mode) or numerical decision (second mode). Linear regression is an uniform framework for both decision types.

We considered application of two operations i.e. normalization ($b_i = \frac{a_i}{\sum_j a_j}$) and cumulative sum ($b_i = \sum_{j \le i} a_j$) to the attributes. Finally, we decided to set $\sigma = 10$ and use

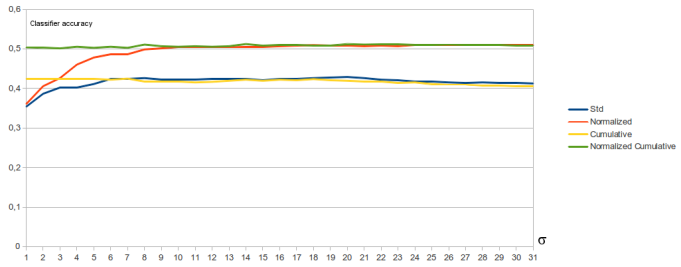

Fig. 7. Dependence of classifier accuracy from σ and preprocessing modes.

normalized attributes.

As an side effect of our preprocessing, neighboring attributes are pretty much lineary dependent. So only few of them contains independent information. We solved this problem restricting the number of attributes with non zero weights to 4 in our model.

Having the parameters set, we performed classification. Table I shows obtained results. Accuracy was calculated by means of 10-fold crossvalidation. A benchmark classifier that predicts the most frequent decision.

B. Neural network credibility model

Because the results obtained using linear regression method did not match our expectations we decided to employ another approach to credibility classification problem. To predict the web page credibility on the statement credibility and important basis we created a classifier feed forward neural network. We decided to use as an input individual statements credibility and importance scores and as an output, we have web page credibility score and five probability classes, corresponding to ranges on the 5 point credibility value scale. We use raw, almost unprocessed experimental data. Since one of the experiment modes used 0-100 scale, data has to be recalculated according to calculated table. Finally since neural network operates in range (-1,1) mandatory normalization has to be applied to the data set. After consideration, we limited the amount of the statements evaluations using in the learning process to 10 which gives us 20 input values for every site evaluation. We construct our network from 4 layers of neurons: input layer that consist of 20 neurons, first hidden layer that counts 23 neurons, second hidden layer consist of 31 neurons and output layer with 6 neurons. The amount of neurons in the hidden layers were estimated empirically based on the parameter sweep to perform lowest error rate. As an activation function tanh was used in all layers except input that used linear activation function. We divided available experiment data into sets,and used 10-fold crossvalidation method to asses classifier accuracy. We trained our network using resilient propagation method (RPROP) in the Encog workbench environment.

Random shuffling of statements results in 2% to 3% accuracy decreases in linear regression classifier. Surprisingly, the neural network classifier accuracy increasing with random shuffling, but the values are within 5% range. This leads us to conclusion that context consideration doesn't affect

Mode	Attributes	Stage	Benchmark	Accuracy	Model
First	Normalized	1st	0.322404	0.485080	$3.5 - 34.7a_0 - 51.7a_{24} + 14.2a_{84} + 12.7a_{96}$
First	Norm Cumul	1st	0.322404	0.495983	$3.39 - 2.39a_{36} + 0.763a_{84} - 6.87a_{88} + 6.38a_{92}$
First	Normalized	3rd	0.325136	0.533950	$3.65 - 49.6a_4 - 73.4a_{36} + 27.9a_{44} + 20.6a_{92}$
First	Norm Cumul	3rd	0.325136	0.525787	$3.36 - 20.2a_0 - 1.96a_{52} - 4.5a_{88} + 5.13a_{92}$
Second	Normalized	1st		0.475000	$38.2 - 985a_{16} + 593a_{64} + 2820a_{96} - 2270a_{100}$
Second	Norm Cumul	1st		0.470833	$50.6 - 97.5a_{52} + 58a_{60} - 75.8a_{84} + 77.2a_{92}$
Second	Normalized	3rd		0.583333	$10.4 + 592a_{52} + 914a_{68} + 1070a_{84} + 714a_{96}$
Second	Norm Cumul	3rd		0.598611	$51.9 - 43.6a_{40} - 22.3a_{56} - 123a_{88} + 138a_{92}$
Third	Normalized	1st	0.385174	0.453325	$3.73 - 48.5a_8 - 26.2a_{28} - 16a_{48} + 21.4a_{88}$
Third	Norm Cumul	1st	0.385174	0.467924	$3.96 - 3.04a_8 - 1.18a_{52} - 1.86a_{80} + 1.55a_{92}$
Third	Normalized	3rd	0.395348	0.514578	$3.66 - 50.7a_8 - 11.4a_{24} - 26.7a_{36} + 22.7a_{88}$
Third	Norm Cumul	3rd	0.395348	0.514557	$4.27 - 0.976a_{16} - 1.55a_{32} - 2.1a_{76} + 1.15a_{88}$

TABLE I. INFERRED MODELS.

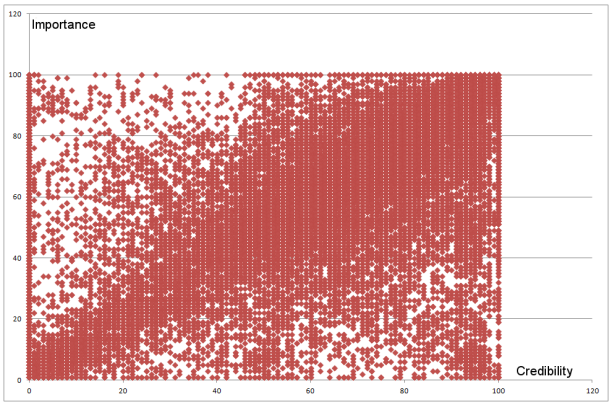

Fig. 8. Importance as a credibility function for statements

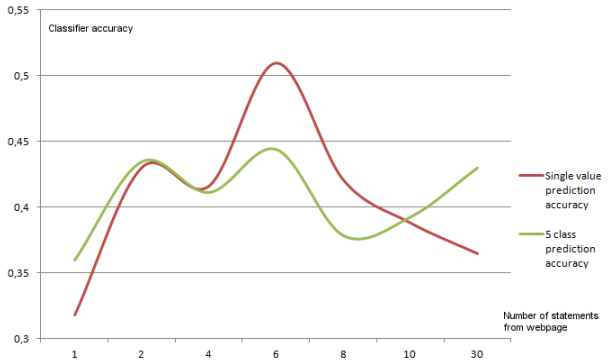

Fig. 10. Neural network accuracy for different number of statements, 30 means all statements from web page were used

automatically predicted credibility value of web page. In both modes, it is easier to predict page credibility measured after statements ratings. Linear regression classifiers perform about 4% better. Thus decision was made that for simplicity we will strive to predict only the second site evaluation score.

In case of second mode, we have introduced the following loss function accuracy measure:

$$L(y,d) = \begin{cases} 0 & \text{if } |y-d| < 15 \\ 1 & \text{if } |y-d| \geq 15 \end{cases}$$

i.e. we assume that if the difference between classifier output and the decision is less than 15, there is no error, otherwise, there an error occurs. As we stated above, 15 is the accuracy range for 100 point slider. The results for credibility measured before statements ratings are comparable to those for first mode, while the results for credibility measured after statements ratings are 5% better.

Thus, we have proved that credibility of webpage depends on the credibility of its textual content (Hypothesis 1) and that evaluations of statements are useful to classify web page credibility (Hypothesis 4).

The truth of the Hypothesis 5 depends on the strict definition of an important unbelievable statement. If we assume that statement is important when its importance is greater than 50 and unbelievable when its credibility is lower then 20, the Hypothesis will be true with the probability 41% and 44% in the 1st and 3rd stage respectively. Empty fields in tables Fig. 9 indicate the lack of examples satisfying given criteria.

To validate this hypothesis by neural network we change the number of input neurons and we trained the network to see how limitations of input values cripples neural network accuracy. We change the number of statements from 10 to 1 count every other. As shown on the Fig. 10 results from classification by neural network also supported this hypothesis. Results shows that predicting credibility of the site basing only on small part of it's textual content is possible and gives results comparable to those with all statements from site used Fig.10 Hypothesis 2 seems to be strong: our experiment shows that credibility of the webpage depends on the credibility and importance of what composes its text. However, this is probabilistic rather than functional dependence and not important statements also affects this dependence.

IV. CONCLUSIONS

The attempt to understand the credibility of Web content is an objective of social computer science. We described experiment that shows us how users see credibility, and how the credibility of the whole site is connected with the credibility of statements, treated as an atomic parts of the page.

Two completely different approaches, for web page credibility evaluation based on textual content, were presented in this paper. Linear regression approach required extensive data preprocessing, while neutral network approach used almost raw data from the experiment. Despite of completely different

		Importance										
		0	10	20	30	40	50	60	70	80	90	100
C	0											
r	10	0.25	0.31	0.42	0.50	0.61	0.56	0.64				
e	20	0.18	0.22	0.30	0.24	0.37	0.41	0.44	0.38			
d	30	0.13	0.14	0.18	0.17	0.21	0.32	0.35	0.26	0.29	0.27	
i	40	0.14	0.14	0.14	0.13	0.17	0.22	0.28	0.18	0.23	0.19	
b	50	0.08	0.11	0.11	0.11	0.13	0.15	0.22	0.23	0.20	0.24	
i	60	0.08	0.13	0.12	0.11	0.12	0.13	0.14	0.17	0.21	0.18	
l	70	0.03	0.09	0.10	0.09	0.10	0.12	0.10	0.11	0.14	0.11	0.18
i	80	0.06	0.13	0.12	0.10	0.12	0.13	0.12	0.10	0.12	0.06	0.27
t	90	0.04	0.19	0.17	0.11	0.17	0.21	0.19	0.13	0.14	0.08	0.17
y	100			0.46	0.40	0.40	0.39	0.26	0.16	0.17	0.09	0.10

		Importance										
		0	10	20	30	40	50	60	70	80	90	100
C	0											
r	10	0.28	0.28	0.39	0.43	0.48	0.44	0.45				
e	20	0.21	0.27	0.35	0.29	0.37	0.44	0.44	0.31			
d	30	0.15	0.17	0.20	0.19	0.22	0.31	0.33	0.26	0.21	0.27	
i	40	0.08	0.10	0.11	0.12	0.17	0.22	0.25	0.18	0.19	0.19	
b	50	0.04	0.07	0.10	0.07	0.11	0.15	0.24	0.23	0.20	0.24	
i	60	0.06	0.10	0.13	0.10	0.12	0.16	0.16	0.18	0.19	0.18	
l	70	0.03	0.08	0.11	0.09	0.10	0.13	0.13	0.13	0.14	0.11	0.18
i	80	0.04	0.13	0.14	0.13	0.15	0.17	0.13	0.10	0.14	0.09	0.27
t	90	0.04	0.19	0.20	0.14	0.19	0.23	0.19	0.13	0.13	0.07	0.17
y	100			0.54	0.50	0.47	0.48	0.28	0.18	0.17	0.09	0.10

Fig. 9. Probability that single important unbelievable statement makes text non-credible in the 1st and 3rd stage of an experiment

evaluation methods employed, results for those classifiers are quite similar, therefore we concluded that this is best accuracy to be achieved on data available.

The evaluation of hypotheses stated in this paper is essential for future design and functionalities of web page credibility tool that is being developed by our team. Our goal is to create a universal framework for credibility evaluation on the basis of the meaning of the text contained on the webpage. Our experiment studies hypotheses that can be important for constructing future approaches that use statements to classify Web page credibility automatically. When we grasp the relationship between a text and statements that compose it, we will be able to develop our system to automatic site credibility assessment. As a one of the features of our system we consider a Web browser extension that highlights sentences in a Web pages and presents their credibility calculated on the basis of ratings obtained from other users.

Furthermore, evaluations of statement credibility will be reusable if algorithms for the aggregation of statement credibility into Web page credibility can be designed.

Acknowledgements: This work was financially supported by the European Community from the European Social Fund within the *INTERKADRA* project.

It was also 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] A. Amin, J. Zhang, H. Cramer, L. Hardman, and V. Evers. The effects of source credibility ratings in a cultural heritage information aggregator. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 35–42. ACM, 2009.

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

[3] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, et al. 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, 2001.

[4] A. Juffinger, M. Granitzer, and E. Lex. Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 51–58. ACM, 2009.

[5] M. Kąkol, M. Jankowski-Lorek, K. Abramczuk, A. Wierzbicki, and M. Catasta. On the subjectivity and bias of web content credibility evaluations. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1131–1136. International World Wide Web Conferences Steering Committee, 2013.

[6] T. Kawada, S. Akamine, D. Kawahara, Y. Kato, Y. I. Leon-Suematsu, K. Inui, S. Kurohashi, and Y. Kidawara. Web information analysis for open-domain decision support: system design and user evaluation. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, pages 13–18. ACM, 2011.

[7] V. Kostakos. Is the crowd's wisdom biased? a quantitative asessment of three online communities. *CoRR*, abs/0909.0237, 2009.

[8] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. Measuring the quality of web content using factual information. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 7–10. ACM, 2012.

[9] K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui, and Y. Matumoto. Statement map: assisting information credibility analysis by visualizing arguments. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 43–50. ACM, 2009.

[10] D. Schultz. Truth goggles: automatic incorporation of context and primary source for a critical media experience, 2012.

[11] M. Wassmer and C. M. Eastman. Automatic evaluation of credibility on the web. *Proceedings of the American Society for Information Science and Technology*, 42(1):n–a, 2005.

[12] N. Weber, K. Schoefegger, J. Bimrose, T. Ley, S. Lindstaedt, A. Brown, and S.-A. Barnes. Knowledge maturing in the semantic mediawiki: A design study in career guidance. *Learning in the Synergy of Multiple Disciplines*, pages 700–705, 2009.

[13] W. Weerkamp and M. De Rijke. Credibility improves topical blog post retrieval. In *ACL-08: HLT*, pages 923–93. Association for Computational Linguistics (ACL), 2008.