Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments

Davide Ceolin[1], Julia Noordegraaf[2], and Lora Aroyo[1]

[1] {d.ceolin,lora.aroyo}@vu.nl
VU University Amsterdam
Amsterdam, The Netherlands
[2] j.j.noordegraaf@uva.nl
University of Amsterdam
Amsterdam, Then Netherlands

Abstract. Automatic estimation of the quality of Web documents is a challenging task, especially because the definition of quality heavily depends on the individuals who define it, on the context where it applies, and on the nature of the tasks at hand. Our long-term goal is to allow automatic assessment of Web document quality tailored to specific user requirements and context. This process relies on the possibility to identify document characteristics that indicate their quality. In this paper, we investigate these characteristics as follows: (1) we define features of Web documents that may be indicators of quality; (2) we design a procedure for automatically extracting those features; (3) develop a Web application to present these results to niche users to check the relevance of these features as quality indicators and collect quality assessments; (4) we analyse user's qualitative assessment of Web documents to refine our definition of the features that determine quality, and establish their relevant weight in the overall quality, i.e., in the summarizing score users attribute to a document, determining whether it meets their standards or not. Hence, our contribution is threefold: a Web application for nichesourcing quality assessments; a curated dataset of Web document assessments; and a thorough analysis of the quality assessments collected by means of two case studies involving experts (journalists and media scholars). The dataset obtained is limited in size but highly valuable because of the quality of the experts that provided it. Our analyses show that: (1) it is possible to automate the process of Web document quality estimation to a level of high accuracy; (2) document features shown in isolation are poorly informative to users; and (3) related to the tasks we propose (i.e., choosing Web documents to use as a source for writing an article on the vaccination debate), the most important quality dimensions are accuracy, trustworthiness, and precision.

1 Introduction

Automatically estimating the quality of Web documents is a compelling, yet intricate issue. It is compelling because the huge amount of Web documents

we can access makes their manual evaluation a costly operation. So, to guarantee we access the best documents available on the Web on a given matter, an automated assessment is needed. However, quality is a rather inflated term, that assumes different meanings in different contexts and with different subjects. Quality assessments vary depending on their context (what is the document used for), author (who is judging the document), time (e.g., users may change their assessments about documents as soon they acquire new knowledge), etc. Quality assessments are hard to capture, hence we call them "ineffable".

This paper investigates strategies for capturing such ineffable judgments and assessing their characteristics. In particular, our focus is on the quality assessment of Web documents to be used for professional use (i.e., by journalists and media scholars). Our ultimate goal is to automate the process of document quality assessment, and the contribution of this paper in this direction is threefold. Firstly, we introduce a nichesourcing application for collecting Web document quality assessments (WebQ[3]). Secondly, we present a curated dataset of Web documents (on the topic of vaccinations) enriched with a set of features we extracted, and a set of quality assessments we nichesourced[4]. Thirdly, we describe a thorough set of analyses we performed on these assessments, from which we derive that: (1) given an explicit task at hand, subjects with similar background will provide coherent assessments (i.e., assessments agree with document similarity, measured in terms of shared entities, sentiment, emotions, trustworthiness); (2) users find it difficult to judge document quality based on quantitative features (entities, sentiment, emotions, trustworthiness) extracted from them; however (3) such features are useful to automate the process of quality assessment. The user studies analyzed are based on limited – but highly specialized – judgments, so these findings provide useful insights on how to progress this research.

The rest of the paper is structured as follows. Section 2 introduces related work. Section 3 describes the application we developed for collecting quality assessments, WebQ. Section 4 describes the two case studies we performed, along with the results collected, that are discussed in Section 5. Section 6 concludes.

2 Related Work

The problem of assessing the quality of Web documents and, in general, (Web) data and information, is compelling and has been tackled in many contexts.

The ISO 25010 Model [9] is a standard model for data quality. From this model, we select those data quality dimensions that apply also to Web documents (e.g., precision, accuracy) and ask the users of WebQ to rate Web documents on them. This set of quality dimensions has been extended to include other measures tailored to Web documents, like neutrality and readability.

The problem of identifying the documents of higher quality for a given purpose is common in information retrieval. Bharat et al [2] copyrighted a method

[3] The tool is running at `http://webq3.herokuapp.com`, the code is available at `https://github.com/davideceolin/webq`.

[4] The dataset is available at `https://github.com/davideceolin/WebQ-Analyses`.

for clustering online news content based on freshness and quality of content. Clearly, their approach differs from ours as they focus on news, and they aim at clustering documents. However, one of the key features for determining the quality of documents is the (estimated) authoritativeness of the source, both in their and in our approach. Kang and Kim [10] find links between specific quality requirements and user queries. We do not make use of queries: we preselect documents (to guarantee that documents get an even number of assessments) and we predefine the task the users are asked to perform (to allow controlling the definition of quality adopted by users). We still analyze user assessments to derive their specific definition of quality, and might consider analyzing user queries in the future, when we will expand the dataset and tasks at hand.

Following up on the use of specific metadata as markers for quality, Amento et al. [1] use link-based metrics to make quality predictions, showing that these perform as good as content-based ones. In our case, we focus on features we can automatically extract from the documents using AlchemyAPI and WOT. We will consider other features (including link-based ones) in the future.

Regarding the use of niche- or crowdsourcing for collecting information and, in particular, quality assessments, Lee et al. [11] provide a framework tailored to organizations. Zhu et al. [14] propose a method for collaboratively assessing the quality of Web documents that shows some similarity with ours (e.g., we both collect collaborative quality assessments), but the assessments we collect are based on specific tasks, while they rely on contributions via browser plugins. Currently, we focus on niches for collecting quality assessments because the definition of 'quality' is different for different types of users; so, for us, it is necessary to have a controlled user study. In the future, we plan to make use of crowdsourcing, adopting methods for extracting ground truth like CrowdTruth [8].

While this paper proposes a framework that aims at generically identifying markers for quality of Web documents, we evaluate such framework with an emphasis on Digital Humanities applications. Digital Humanities scholars are professionals that are used to critically evaluate the sources they deal with, hence we target this specific class of users to investigate how to extend source criticism practices to cover Web documents as well. Source criticism is the process of evaluating traditional information sources that is common in the (Digital) Humanities. De Jong and Schellers [5] provide an overview of source criticism methods, evaluated in terms of predictive and congruent validity. We will advance such evaluations to identify which Web document features determine their quality. This paper extends the work we presented at the Web Science conference, where we began the exploration of how it is possible to assess the quality of Web documents, especially for the Digital Humanities [4]. In that, we outlined a pipeline for assessing document quality and we provide a preliminary evaluation based on a manual assessment. Here we develop an application for nichesourcing such assessments and we deeply analyze them and their predictability.

Lastly, one aspect that we consider when estimating the quality of Web documents is their provenance. Provenance analysis is used to assess the quality of humanities sources, as Howell and Prevenier mention [7]. In Computer Sci-

ence, Hartig and Zhao [6] use temporal qualities of provenance traces to assess the quality of Web data. More extensively, Zaveri et al. [13] provide a review on quality assessment for Linked Data. We also investigated the assessment of crowdsourced annotations using provenance analysis [3,12].

3 Nichesourcing Web Document Quality Assessments

To collect and analyze judgments about Web documents, we developed the tool WebQ, that aims at understanding three main aspects of Web document quality:

- whether (professional) users are able to estimate the quality of Web documents based on limited sets of features of these documents (e.g., the sentiment of these documents, or the list of entities extracted from them);
- whether assessments are coherent enough over multiple documents and among diverse assessors (i.e., whether assessors assess similar documents in a similar manner; similarity is measured in terms of shared entities, sentiment, emotions, trustworthiness), to allow their automated learning;
- how the overall quality assessments can be explained in terms of specific quality dimensions (precision, accuracy, etc.) when focusing on specific tasks.

3.1 Document Features and Document Quality Dimensions

We characterize documents by means of features we automatically extract about them. In Section 4 we analyze the existence of correlations between these automatically extracted features and the nichesourced features of quality.

Document Features These are a series of attributes we automatically extract by means of Web APIs. These features aim at identifying commonalities among documents, opening up for the possibility of predicting their qualities (provided that features and qualities correlate). These features are:

Entities, Sentiment, Emotions We use AlchemyAPI[5] to extract all the entities mentioned in the documents, along with an assessment of their relevance to the document. Also, AlchemyAPI provides us with a quantification of the sentiment expressed by the document (positive or negative, and its strength), and its emotions (joy, fear, sadness, disgust and anger, and their strength).

Trustworthiness In this case, we use the Web Of Trust API[6] to obtain crowdsourced trustworthiness assessments about the source publishing the article.

Document Quality Dimensions These are a series of abstractions of the documents qualifying the information therein contained. We ask the users to assess the documents based on each quality dimension reported as follows:

[5] `http://www.alchemyapi.com`
[6] `http://www.mywot.com`

Overall Quality provides an overall indication of the quality of a document. It summarizes the other quality dimensions in a single value representing the suitability of the document for a given task, in a given context.

Accuracy quantifies the level of the truthfulness of the document information.

Precision determines whether the document information is precise or vague.

Completeness determines whether the document information is complete.

Neutrality determines whether a particular stance (e.g., pro or anti a given topic) is represented in the document.

Readability quantifies whether the document reads well.

Trustworthiness quantifies the perceived level of trustworthiness of the information in the document. Note that the Web Of Trust score refers to the source, while this quality refers to the specific document evaluated.

3.2 Structure of WebQ

Below we describe the structure of WebQ, illustrated in Figure 1.

Architecture The application is developed based on the Flask Python library[7]. As backend storage for Web document assessments, we use MongoDB[8].

Annotations We use AnnotatorJs[9] to allow users to indicate which specific parts of a document mark particular qualities of the whole document. AnnotatorJs is a javascript library run on the client side that records the document annotations by sending HTTP messages to a storage server. We adapted to this purpose the Annotation Store[10], which relies on ElasticSearch[11].

HTTP Proxy We developed an HTTP proxy to provide the users with the Web documents to be annotated within WebQ. This proxy allows the system to present the documents within our application and allows users to annotate them by enabling AnnotatorJs. In this manner, the users see the exact same document they would see on the Web, but they are able to annotate it, remaining in the context of our application. This proxy is tailored to the documents in our dataset and renders them at their best. In particular, it addresses the following issues:

- **replace relative paths with absolute ones** in image, CSS and link addresses, so the page can refer to the absolute addresses of the accessory files;
- **correctly detect and utilize charsets** to properly render the documents;
- **forward the browser headers** because some websites allow being accessed only via (some) browsers, and not being scraped. The proxy accesses them programmatically on behalf of a browser.

In the future, we will extend our dataset, so we will extend further this proxy.

[7] `http://flask.pocoo.org/`
[8] `http://mongodb.com`
[9] `http://annotatorjs.org`
[10] `https://github.com/openannotation/annotator-store`
[11] `https://www.elastic.co/`

Randomizer WebQ is designed for collecting Web document quality assessments via one or more user studies. In such a scenario, users access the application more or less simultaneously. We assign to each user a random sequence of documents to assess (we set the length of such sequence to six), but we also guarantee that the dataset is uniformly assessed: documents should get approximatively

$$n_{ass} = |dataset| \text{ div } |users|$$

assessments, where $|dataset|$ is the cardinality of the document dataset (50), div is the integer division and $|users|$ is the cardinality of the set of users. Offline, we generate n_{ass} random permutations of documents. We split them in consecutive sequences of six documents, uniquely assigned to users when they register.

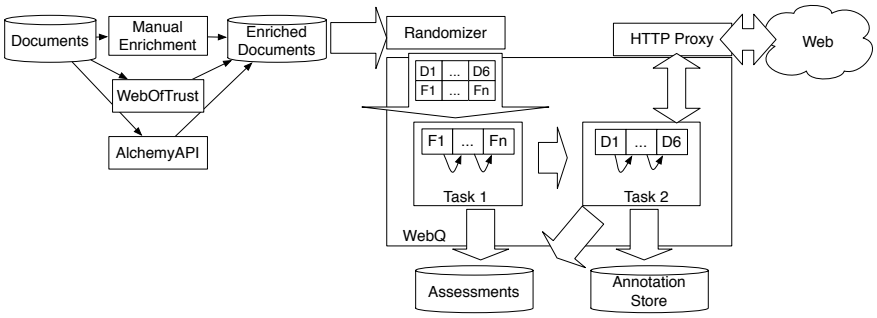

Fig. 1. Overview of the WebQ application. The document set is enriched by using AlchemyApi, Web of Trust, and manually. A random selection of six documents is presented to the users for the first task: identifying the highest quality documents on the basis of the value of one feature. After all the features (sentiment, etc.) have been evaluated, users assess each of the six documents assigned (task 2). Documents are rendered through an HTTP proxy, to allow annotating them within the app.

3.3 Tasks Description

In WebQ we ask the users to perform two tasks. The first task aims at exploring whether single document features could be used as quality indicators. The second task aims at collecting assessments about the documents presented. The two tasks are described as follows, first in general terms, and then, in section 4, as adapted according to a specific scenario for the two case studies.

Task 1 Task 1 is structured as follows:

1. We assign to each user a set of six documents from our overall dataset.
2. We identify six classes of potentially useful features about the documents, namely: the document's sentiment and emotions, its trustworthiness, its title, its source and the list of entities we extract from it.

3. We show the values for each of these features to the user. First, we present the user with the lists of entities extracted from the six documents, then we present the user with the sentiment and the emotions detected in each document, and so on, each feature at a time. Users do not know the documents, they only know the values of the features we present. Every time we present features we shuffle the document order, and we change document identifiers.
4. We ask the user to select which documents among these six she will use as a source for her article, based on the information displayed.
5. Lastly, we ask the user to make the selection again, on the basis of all the features presented together.

Task 2 We ask them to assess the quality of each article in depth. Based on the same selection of six articles the user was assigned to in task 1, she:

1. Reads the article
2. Assesses the overall quality of the article, as well as the following quality dimensions: accuracy, precision, completeness, readability, neutrality, trustworthiness. Assessments are indicated in a 1 to 5 Likert scale.
3. Highlights in the article the words or sentences that motivate her assessments, tagging each selection with the name of the corresponding quality dimension and indicating if it represents a positive or negative observation.
4. Revises their quality assessments (step 2.) if she wishes so.

4 Case Studies

In this section, we describe the two case studies we run. Both case studies are based on the same set of documents, which we describe as follows.

4.1 Dataset and Scenario

The dataset we base our experiments on is composed by Web documents about the vaccination debate triggered by the measles outbreak that happened at Disneyland, California, in 2015[12]. This dataset contains 50 documents, diversified in terms of: **stance** (some are pro vaccinations, some anti, some neutral) and **type of source** (e.g., we include: official reports, editorial articles, blog posts).

The scenario we hypothesize is that users have to write an article about the vaccination debate triggered by such measles outbreak. We propose diverse types of Web documents to the users, and we ask to select those they would use as a source for their article (i.e., those they consider of a higher quality). Thus, we consider selection a marker of relatively high quality.

4.2 Case Study 1 - Journalism Students

Experimental Setup The first case study involved a class of 20 last-year journalism students from the University of Amsterdam. The students performed both tasks of WebQ in a time frame that lasted between 45 and 60 minutes.

[12] The dataset is available at `https://goo.gl/cLDTtS`

Results We present here a series of analyses on the results collected.

Document Assessments Collected We collected 104 complete assessments about the diverse quality dimensions of the documents and 238 annotations.

Comparison of the two document assessments in task 2 We ask users to assess the documents twice: when they first read the documents, and after having highlighted the motivations for their assessments. These two assessments show no significant difference using a Wilcoxon Signed-rank test at 95% confidence.

Document Assessments Predictability The first analysis we perform regards the predictability of Web documents assessments. Only two or three assessments are provided per document, but if users assess the documents coherently enough (i.e., following similar policies), and if the features we extracted (entities, sentiment, emotions, trustworthiness) are considered by the users' policies, then we might be able to automatically learn such predictions. Table 1 shows the results of such predictions using the Support Vector Classification algorithm.

Table 1. Accuracy of 10-fold cross-validation using Support Vector Classification with different number of features, and predicting either 5 classes (as in the 1-5 Likert scale used in WebQ) or 2 classes (i.e., high- and low-quality documents). We calculated the performance for all possible combinations of the four classes of features. For each cardinality of such combination (1,2,3,4) we show the best performing combination.

Features used	**SVC 5 classes**	**SVC 2 classes**
trustworthiness	48%	75%
sentiment, trustworthiness	46%	78%
sentiment, emotions, trustworthiness	38%	72%
sentiment, emotions, trustworthiness, entities	39%	72%

Correlation between quality dimensions and overall quality Table 2 shows the results for each quality dimension.

Table 2. Correlation between each quality dimension and the overall quality score attributed to the documents.

Quality dimension	**Correlation with Overall Quality**
Accuracy	0.89
Completeness	0.69
Neutrality	0.46
Relevance	0.63
Trustworthiness	0.80
Readability	0.67
Precision	0.77

Correlation between document selection (task 1) and document assessments (task2) In task 1 we ask the users to select documents they think are of high quality based on diverse document features. If many users select a document, we derive that it has a high probability to be of high quality. Since each document has been proposed to only either two or three users, we compute such probability using a smoothing factor that allows accounting for the uncertainty due to the small samples observed (see Equation (1)). Smoothing allows treating differently documents that have been proposed two or three times: if a document has never been selected when it has been proposed two times, its probability to be of high quality is 0.25; if it has been proposed three times, 0.2. This allows us comparing probabilities based on different amount of evidence in an unbiased manner. The resulting probability is equivalent to the expected value of a Beta probability distribution with a non-informative prior: we add 1 and 2 to the numerator and denominator exactly because we do not know a priori if a given document is of high or low quality (hence its probability of being of high quality is 50%).

$$P = \frac{\#selection + 1}{\#samples + 2} \tag{1}$$

In task 2, users assess these same documents. Table 3 shows the correlation between the probability from task 1 and the overall quality score from task 2. Entities, sentiment, and title show a poor correlation, close to zero: probabilities from these features (task 1) are not correlated with assessments from task 2. Trustworthiness, sources and all show a slightly higher but still weak correlation: between 20% and 30% of the times, their probabilities agree with assessments.

Table 3. Correlation (Spearman) between the probability of documents to be selected in task 1 and their overall quality assessment from task 2.

Feature shown (task 1)	**Correlation with Overall Quality (task 2)**
Entities	-0.07
Sentiment	0.09
Trustworthiness	0.20
Sources	0.29
Title	-0.07
All	0.20

User Evaluation We asked the users to complete a questionnaire about their experience[13]. The quantitative results of the 13 respondents (52% of the total) are reported in Table 4, which shows the percentage of users that indicated a feature or quality as important. Moreover, the majority (∼70%) of users gave a low score (1 or 2 on a 1-5 Likert scale) to the whole experience, to its easiness, and to the fact that the experiment resembles their process when writing an article.

[13] The questionnaire is available at: `http://goo.gl/forms/2pIjjpIp0PtyPxd72`

Users agree on the importance of most of the features and qualities we identified, but they negatively assess the experience they had. We use such information to improve the experiment design in the next case study, as we explain below.

Table 4. Results of the user evaluation questionnaire.

Feature	**Users choosing it**	**Quality**	**Users choosing it**
Sentiment	0%	Accuracy	30.8%
Entitites	23.1%	Completeness	23.1%
Emotions	0%	Neutrality	15.4%
Source	76.9%	Precision	30.8%
Title	46.2%	Trustworthiness	**69.2%**
Trustworthiness	**100%**	Relevance	38.5%

Quality Definition and Qualitative Analysis of Annotations and Remarks Lastly, from a qualitative evaluation of the annotations and of the remarks collected, we derive that users assume that *the documents of higher quality are those showing the following qualities: high trustworthiness, high accuracy, and high precision.*

4.3 Case Study 2 - Media Scholars

Experimental Setup This case study involves 20 media scholars (RMA and PhD students as well as senior scholars) attending the Research School for Media Studies (RMeS) summer school in Utrecht (27 May 2016). Based on the user evaluation of case study 1, we add a walk-through session to guide the users in the application, and we improve the task descriptions and the user experience (e.g., landing pages). The users had about 45 minutes at their disposal.

Results We present the results obtained and their analyses.

Document Assessments Collected In this experiment we collected 47 complete assessments about the documents in our dataset and 89 annotations.

Comparison of the two document assessments in task 2 We observe no significance difference between the two series of assessments, for any quality dimension.

Document Assessments Predictability Like with the previous case study, we use 10-fold cross-validation to test the predictability performance of Support Vector Classifier on the overall quality assessment. Results are reported in Table 5.

Correlation between quality dimensions and overall quality Table 6 shows the results for each quality dimension.

Table 5. Accuracy of the prediction of the overall quality assessments case study 2. We show the best performing combination of features per set cardinality (1,2,3,4).

Features used	**SVC 5 classes**	**SVC 2 classes**
trustworthiness	63%	89%
sentiment, trustworthiness	53%	86%
sentiment, entities, trustworthiness	34%	85%
sentiment, entities, trustworthiness, emotions	34%	85%

Table 6. Correlation between each quality dimension and the overall quality score.

Quality dimension	**Correlation with Overall Quality**
Accuracy	0.89
Completeness	0.69
Neutrality	0.45
Relevance	0.64
Trustworthiness	0.78
Readability	0.66
Precision	0.76

Correlation between document selection (task 1) and document assessments (task2) We computed the probability of documents to be of high quality based on the number of selections collected in task 1 (see Equation (1)). Table 7 shows the correlation between such probability and the overall quality from task 2. Again, the probabilities show a weak correlation with the quality assessments.

Table 7. Correlation (Spearman) between the probability of documents to be selected in task 1 and their overall quality assessment from task 2.

Feature shown (task 1)	**Correlation with Overall Quality (task 2)**
Entities	0.38
Sentiment	0.19
Trustworthiness	0.21
Sources	0.25
Title	0.15
All	0.24

User Evaluation The results of the user evaluation questionnaire[14] are reported in Table 8. To these quantitative results, we add the fact that users indicate accuracy and also indicators from social media (e.g., discussion on the topic, likes) as possible quality markers and that the majority of the users (75%-100%) rate the experience and its easiness fairly (2-3 in a 1-5 scale). Users disagree

[14] The questionnaire is available at: `http://goo.gl/forms/ZwvaqDidGeC8FCXm1`.

on whether or not this resembles the process of writing an article. Only four participants responded to the questionnaire.

Table 8. Results of the user evaluation questionnaire.

Feature	**Users choosing it**	**Quality**	**Users choosing it**
Sentiment	0%	Accuracy	25%
Entitites	0%	Completeness	0%
Emotions	0%	Neutrality	25%
Source	100%	Precision	0%
Title	50%	Trustworthiness	50%
Trustworthiness	100%	Relevance	25%

Quality Definition and Qualitative Analysis of Annotations and Remarks From a qualitative evaluation of the annotations and of the remarks collected, we can derive that users assume that *the documents of higher quality are those showing the following qualities: high trustworthiness, high accuracy, and high precision.*

4.4 Comparison between Case Study 1 and 2

We compare the results obtained in case study 1 and 2. We use a Wilcoxon signed-rank test to compare the performance obtained by support vector machines (Tables 1 and 5). We observe no significant difference neither with 2 nor with 5 classes. Also comparing the correlations between the quality dimensions and the overall quality (Tables 2 and 6), we observe no significant difference. Neither the results of Tables 3 and 7, i.e., the correlation between probabilities of a document to be selected and its quality, show any significant difference between task 1 and 2. The second user questionnaire has been completed only by a very limited number of users. A Wilcoxon signed-rank test and a χ^2 test agree that the results from the two case studies are not significantly different but, in this case, the sample sizes are so small that we can hardly rely on these results.

5 Discussion

Our long-term goal is to allow automatic assessment of Web document quality tailored to specific user requirements and context. Such a process relies on the possibility to identify document features that indicate quality (if these exist). In this paper, we perform two case studies that shed a light on how professionals evaluated Web documents. Here we discuss the results presented in Section 4 by means of a series of statements that emerge from the analysis of the results. Even though the sets of assessments are small, they are large enough to support the statistical test run in Section 4. Only the tests run to compare the evaluation test are based on a very small dataset, and thus are less conclusive.

User assessments are stable and coherent. In both case studies, we observe that the first and the second document assessments are not significantly different. Moreover, in both cases, we can use Support Vector Classifier to automatically learn and predict the quality of documents. This means that, even if users assess different documents (the same document has been assessed by three users at most), assessments are coherent enough to be learned. The features we identified (entities, sentiment, emotions, trustworthiness) correlate with these judgments enough to allow using them as features for prediction, at least in this case.

User assessments are highly related to the task at hand. The extremely high similarity between the results in Tables 2 and 6 shows that, when assessing the quality of documents, the task at hand is the most important factor. Here the users were asked to pretend they were writing an article about the vaccination debate. So, they focused on identifying the most accurate and trustworthy documents. Neutrality is the least significant quality of these documents because, to represent the whole spectrum of the debate, users have to consider also the least neutral documents, provided that they are accurate enough. Different tasks can imply different quality requirements. This facilitates the definition of future user studies that will provide assessments that are mergeable to the existing dataset (provided, for instance, that they show no statistically significant difference between the existing ones, or that this difference is manageable). So, we will scale up our current approach: even though different case studies will have to be based on limited groups of (diverse) users, their contributions will be used to incrementally build a larger set of document assessments. To guarantee that assessments are handled and merged properly, keeping track of their provenance will be crucial. In this light, although in some cases we observe that by considering only a subset of features we obtain a better performance (up to +6% in some cases), we still prefer to consider all the features we collected so far. In fact, we do not know if, by extending the set of documents considered (or by diversifying the tasks at hand), some of the features could gain or loose importance, and it may be extremely difficult (if not impossible) to know when this would happen.

Features in isolation are hardly meaningful (but the user experience plays a role here). Showing entities, sentiment, and emotions, trustworthiness, title and source (especially in isolation) is hardly useful to users to decide if a document is of high quality or not (see Tables 3 and 7). The fact that these features are profitably used to learn the quality assessments of the documents using SVC means that they are good markers of quality (e.g., the fact that a given document expresses an extremely positive sentiment or show specific entities is correlated with its quality). Nevertheless, users are hardly able to determine the document quality on the basis of a quantification of such features. What is true is that in the second case study, although the performance is still pretty low, the results are slightly better than those of the first use case. This might be due to the different user background (more senior level scholars in case study 2), as to the fact that we improved the setup of the WebQ application and explained the logic behind it better in the introduction and walk through.

The application setup should take the user experience into consideration. We aim at collecting annotations from users, so we need to balance a couple of trade-offs between the application requirements and user-based constraints. First, our target users have a professional background that is not necessarily Information or Computer Science. So, even if the application is able to capture all the necessary information, the way its functionality is presented to the user and the way she is guided plays an important role. In fact, after having better explained the logic of the setup of the application we observed (both via the questionnaire and via a post-study discussion) an improvement in the perception of the experience from case study 1 to 2. Second, our goal is to collect as many assessments as possible, but we must take into account that the user attention decreases over time. So, in a situation like case study 2, we need to either extend the duration of the experiment or to reduce the number of documents assessed by each user (e.g., to preserve a uniform number of assessments per document).

6 Conclusion

Automatically assessing the quality of Web documents is crucial to benefit from the vast amount of online information. In this paper, we present WebQ, a Web application to nichesource quality assessments. We also describe two datasets of Web documents, enriched with assessments resulting from two case studies involving journalists and media scholars. WebQ provides the necessary functionalities (i.e., rating and annotating documents) to collect such assessments, and the user evaluations collected allowed fine tuning it. Our last contribution is a set of thorough analyses on the resulting dataset. Through such analyses, we showed that if we assign a clearly defined task to users with a similar background we can obtain uniform document quality assessments. These can be automatically estimated (in our case, using SVC) but, given their tight relation to the context, their provenance needs to be precisely tracked to allow their future reuse. Also, by decomposing overall quality assessments into quality dimensions, we can identify which quality definition (expressed in terms of quality dimensions) is adopted by users. For the task performed (selecting documents to be used as a source for an article on the vaccination debate), the most important dimensions are accuracy, precision, and trustworthiness. We show that the results collected in the two case studies are assimilable: this allows creating a uniform collection of document assessments. Lastly, the user experience in such application matters, and while it is a delicate balance, small changes lead to improvements.

We plan to extend our application in several directions. We will consider other typologies of users and extend the tasks evaluated. Clearly, we intend to extend also the dataset of documents considered, and to incorporate additional features in our models, including link- and network-based features (e.g., based on document interlinking) and social media-based features (e.g., the number of likes a given article received on social media sites, or the number of followers a given blog has). Besides nichesourcing, we will also make use of crowdsourcing, to reach out more contributors. However, such step will require particular attention

to assimilate expert and laymen assessments. Lastly, as a consequence of such extension, we will have to consider methods for scaling up our prediction models.

Acknowledgements This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities. We thank the students of the UvA journalism course and the RMeS summer school participants for participating our user studies.

References

1. Amento, B., Terveen, L., Hill, W.: Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents. In: SIGIR. pp. 296–303. ACM (2000)
2. Bharat, K., Curtiss, M., Schmitt, M.: Method and apparatus for clustering news online content based on content freshness and quality of content source (2016), `https://www.google.com/patents/US9361369`, uS Patent 9,361,369
3. Ceolin, D., Groth, P., Maccatrozzo, V., Fokkink, W., van Hage, W.R., Nottamkandath, A.: Combining user reputation and provenance analysis for trust assessment. Journal of Data and Information Quality 7(1-2), 6:1–6:28 (Jan 2016)
4. Ceolin, D., Noordegraaf, J., Aroyo, L., van Son, C.: Towards Web Documents Quality Assessment for Digital Humanities Scholars. In: WebSci. pp. 315–317. ACM (2016)
5. De Jong, M., Schellens, P.: Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods? (2000)
6. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: SWPM (2009)
7. Howell, M., Prevenier, W.: From Reliable Sources: An Introduction to Historical Methods. Cornell University Press (2001)
8. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.J.: Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In: ISWC. pp. 486–504. Springer International Publishing (2014)
9. International Organization for Standardization: ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model. Tech. rep., ISO (2008)
10. Kang, I.H., Kim, G.: Query type classification for web document retrieval. In: SIGIR. pp. 64–71. ACM (2003)
11. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: Aimq: A methodology for information quality assessment. Inf. Manage. 40(2), 133–146 (2002)
12. Nottamkandath, A., Oosterman, J., Ceolin, D., de Vries, G.K.D., Fokkink, W.: Predicting quality of crowdsourced annotations using graph kernels. In: IFIPTM. pp. 134–148. Springer International Publishing
13. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web Journal (2015), `http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey`
14. Zhu, H., Ma, Y., Su, G.: Collaboratively assessing information quality on the web. In: ICIS sigIQ Workshop (2011)