Semi-Automatically assessing Web Documents

Ozkan Sener
Vrije Universiteit Amsterdam
ozkansener@gmail.com

Davide Ceolin
Vrije Universiteit Amsterdam
d.ceolin@vu.nl

ABSTRACT

Web users are overloaded with information and Webusers often do not know whether a Web document is useful for them or not by just looking at the URL. This impacts the way Webusers perceives information. We tackled this issue by creating a publicly accessible framework where Web users can gather descriptive and prescriptive insights about the Information and the Information Quality of Web documents. We collected an unique set of features and collect annotations via the crowd. We trained an multiple output regression for the task of predicting the Information Quality of Web documents based on historical data where features and assessments where collected, so that for new document where we automatically collect the features and predict the Information Quality based on the historical patterns that our Machine Learning algorithm learned. While we have specifically focused on assessing the Information Quality of the textual content inside Web documents, the combination of the predicted assessment scores and the textual description's that we provide implies that our findings are likely to be of importance to Webusers by pre-assessing the Information Quality of Web documents where we present the Webusers an overview of this so that the Webusers can be more selectively selecting the Web documents based on their information needs. We might be able to see performance improvement using a larger dataset, which we didn't verified yet. However in order to generalize our finding we believe that IQ of Web documents can be scored, but that difference of preferences and context understanding is required, because Machine's can learn only one truth. Another direction is is to have a consensus policy by aggregating the information quality scores and that the Web users that uses these scores agree on these scores. Search engines can use our framework to remove Web document with an extreme low Information Quality. Our framework is also capable of retrieving documents inherent a given topic of interest to the user, and to present their assessment in a comparable manner.

1 INTRODUCTION

Web user are overwhelmed by the quantity of information that is available on the Web [14]. This makes it difficult for a Web user difficult to receive information of high quality. Web user are also subject to human bias, which can be influenced by varying backgrounds and expertise [11]. Information overload occurs when the brain of the Web user process more information then it can handle. When Web user gets confused they will no longer be able assess the quality of the Web documents [13].

We propose an information retrieval framework that can be used to Semi-Automatically Assess the Quality of Websites (SAATQOW). We ATQOW of a new Web document by collecting historical labeled data (quality assessment scores) and features that are (partially) representing the Quality labels. We then training our model to learn a function that can be used to predict the assessment scores (the labels) for a new instance by automatically collecting the features of the Web document. This function is of the form: $\{\phi\}\chi\{\psi\}$ Where $\{\phi\}$ is the set of features and ψ is the IQ score and χ is the function that our model beliefs can be used to predict ψ.

Compared to previous research: 1) Our framework is publicly available by using open-sourced libraries. 2) We added features that that where never used before like numbers of years of existence of the sources, response time of the source and the numbers of pictures counted inside the Web document 3) Experimented with the crowd to collect annotators to label the Web documents 4) Trained our model via a multiple output regression instead of an singular regression. 5) Our framework is capable of retrieving documents inherent a given topic of interest to the user, and to present their assessment in a comparable manner.

Our framework can be valuable to Webusers by pre-assessing the Information Quality of Web documents where we present the Webusers an short overview of the results of the assessment that we performed so that the Webusers can be more selectively selecting the Web documents based on their information preferences.

This paper continues as follows: Section 2 introduces related work; Section 3 describes the methods adopted; Section 4 presents the results collected, that are analyzed in Section 5. Section 6 concludes the paper.

2 RELATED WORK

Several authors [1, 5, 17, 27] report that there is need for a framework that is able to assess the quality of Websites automatically. In America [2, 15] fake news was spread during political election. Facebook and Google was not able to prevent fake news occurring on their platform. Especially younger Web users [6] are vulnerable for fake news.

2.1 Frameworks

Ceolin et al. [5] created an model that semi automatically automatically estimates the assessment quality of Websites. Their model does this by performing computations that creates features that functions as input for their model by assuming that the features that they use would represent the quality scores of Web sites. Ceolin et al. collected the features sentiment, emotions, trustworthiness scores. The feature sentiment and emotion analysis are collected via the API of Alchemy and allowed Ceolin et al. only to collect the first 50000 characters inside the document. Also the feature trustworthiness that they collected changed their services and doesn't provide an average trustworthiness score anymore. Instead they provide an the lowest and the highest scores of the of the trustworthiness score.

The second component are the assessment scores. Ceolin et al. collected their assessment scores by performing a case study with media experts where they asked the media experts to assess the Information Quality of Websites. The third component is the

Machine Learning algorithm that automatically estimates the IQ scores of Websites.

Ding et al. [7] report that analysis like sentiment and emotion analysis (granular form of sentiment analysis) can be used to discovery opinions and feeling or moods of Web writers. The Sentiment analysis algorithm does this by analyzing the words that are bad or good words. Then the algorithm searches for modifiers that tells something about bad/good words. The other future that is been used by Ceolin et al. is Web of Trust (WOT). WOT is a Crowdsourcing platform which we use to gather ratings of Web users that scored the quality of Web pages. This tell us whether Web users trust the content of the website. The WOT rating are domain based. This limits to see the rate scores for a specific URL. Another limitation of the WOT score is that it is based on crowd sourced ratings. This means that Web users can directly affect the results. This means that the WOT ratings can have other scores in the future. WOT scores are from 0 (very untruthful) till 100 (very trustful). Polarity scores are from -1(very negative) till 1 (very positive). The model that was trained by Ceolin et al. had not the opportunity to learn what each features represents and how this influences the assessment scores. For example: For example the feature sentiment: the feature can have positive and negative values. But the model of Ceolin et al. is only trained for values inside the range: 0 till -0,6. We believe that especially the extreme values of the features outliers carry a lot meaning about the meaning of the features. When the interpretation of feature and how the feature influences the assessment scores is known by the Machine Learning Algorithm (MLA), we will be more accurate at predicting the assessment scores of a Website. Ceolin et al. performed their assessments with media experts. We believe that annotators with a heterogeneous background is needed in order to generalize this Framework for all the Web users on the Web.

Pinto et al. [22] report that there are many Natural Language Processing (NLP) tools. They describe that it is challenging to select one of the many NLP tools. The performance of a NLP tool depends on the type of source of text it is used. APIs like: Calais, Google Natural Language, Havenondemand, Aylien, TheySay PreCeive, Qemotion and Monkey learn are some examples of NLP ApiâĂŹs that provide the same NLP features as IBM Watson API. Ceolin et al. used IBM Watson in order to collect their NLP features. IBM Watson is not able to correctly analyze large Web documents. IBM only analyze the first 50,000 character of these Websites. There are also libraries for programming languages available that can collect these features NLTK, Textblob, tm and Stanford CoreNLP are some examples of open libraries that provide NLP features.

2.1.1 Textual Statistics. Most of the scientific publishers are limiting researchers in the amount pages that a conference report can have in order to be published. The reason for this is that readers of research papers should only present the essentials/the most important/relevant details of their research so that readers easily can understand what the paper is about instead of getting confused (not understandable). Therefore we believe that having features that describes the amount of text on a Web document can be used to assess the quality of the Web. The writing style of a Wikipedia documents can be used to semi-automatically assessing the quality of Wikipedia documents [11, 16]. Dalip et al. [11] report that textual features related to length, structure and style of a Wikipedia document are the most relevant important features in order to assess the quality of a Web document. We believe that Web users are more interested in Web documents that are simple and clear.

Si and Callan [24] developed a model where they used textual statistics in order to predict the readability of the Website. Formulas like the, The Flesch Reading Score and The Dale-Chall can be used to assess the readibility Si and Callan showed that the readability of Web pages are influenced by the writing style. Flesch reading Score (FRS) is very often used by researchers in order to score the readability of Web document about health related information [21]. With the formula bellow we show how the FRS can be calculated:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

The higher the Flesch Reading Score is the easier the Web document is to read. Low Flesch Reading Scores indicates that the Web document is not easy to understand for the average Web user, but for an university graduated person it might be understandable [9].

2.1.2 Source reputation. Alexa Rank ranks websites based on the popularity of the Website over a period of 3 months. It does this by analyzing the number of times a Web domain that has visited by Web users that use the Alexa toolbar over a period of three months. The higher the ranking number is the less often the Web domain has been visited. When a Web domain is scored this means that the Web source has no ranking. Google PageRank (PR) is a metric that is used by Google in order to determine which Web pages will appears first in their search engine based on the importance, reliability and authority of a Web page. However Google hasn't updated page ranks since 2013 and it is not open for public anymore. We believe that Websites that have in general a better reputation/popularity are more likely to have a higher IQ. The reason why we expect this is because users in general understand Web document better if they already can make some inferences on how the web document works technically before carry out there task (reading information).

2.1.3 Response time. Nielsen [20] reports that user-perceived IQ is highly influenced by the response times and the variance of the response times of a Web document. When Websites are taking too long to respond to the Web user request the Web user is less likely to be satisfied about the IQ of a Web document. Web users also think that Websites with a high response time are seen by Web users as incompetent. We believe that the response time also provides information about the maturity of the Web source and that these mature Web sources are more likely to have Web documents with a higher IQ.

2.2 Assessments

There are different standards in the field of information retrieval that describe how the quality of documents can be assessed. The DAMA [19] has created a data quality framework with the dimensions: Accuracy, Consistency, Integrity, Timeliness, Completeness and Validity. Ceolin et al.[3], Wang[25], Held and Lenz [12] provides us the definition of the quality dimensions that we use in this paper: The overall quality score summarizes the opinion about the quality of the website. Accuracy explains how correct the website is. If the accuracy of the Website is high this means that there are not too many wrong (generalized) statements on the website.

Completeness means that the information that is available on the website fulfills our needs. Neutrality means that the author of the website is not pro or con towards the subjects. Relevance is the closeness between the information we need and the information that is provided. Trustworthiness means that we are able to see the protocols and procedures that are used to provide the information that is provided to us. The website is considered readable if we can read the content on the website and that the information on the website is clear to us. Precision means that the website has sufficient detailed information that is required to fulfill our task. In order to compare our framework with the framework of Ceolin et al. we are choosing the same quality dimensions and use the same questions. We believe that crowd-sourcing can be valuable by for us by acquiring a huge amount of occupants in a small amount of time in order to give them the job of assessing the Information Quality of the Web. Crowd sourcing makes it also possible for us to reach Web users all over the world and acquiring their assessment.

Raiber and Kurtland [23] report that the content of the website is not the only key factor that affects the quality scores of our quality dimension attributes. Raiber and Kurtland report that factors like the usability, scalability, originality of the content, personalization of the content, the performance of the Website and the design of the Website are also factors that influence the quality of the website.

2.3 Machine Learning

The choice of choosing an algorithm should be based on the properties of the data set that we want to analyze because different Machine Learning Algorithms (MLA) have different characteristics [18, 26]. Domingos [8] reports that the success of the Machine Learning approach is the combination of highly sophisticated algorithms and large amounts of data of high quality. When assessment scores are not reliable our model will be less accurate at predicting the assessment scores.

Deciding which type of algorithms we apply on our model can be done based on functional and non-functional criteria [10]. Parametric Machine Learning Algorithms are more reliable when the quality of the data is not of high quality compared to non-Parametric algorithms. Parametric algorithms are also easier to understand and do not require complex computation (saves time and computation performance). Parametric algorithms When the sample size of the dataset is small a parametric algorithm is preferred over non-parametric dataset and in scenarios where the fit of a model to the data is not perfect. However these parametric algorithms can not learn the complex patterns from a dataset. For example non-parametric algorithms are capable of interpreting the interactions between our set of features and how these set of features and their interactions represent our IQ scores . However in uncertain scenarios or unseen learning the general patterns of a Machine Learning algorithm is much more important.

Research Question

In this paper we answer the question: "Can we Semi-Automatically Assess the Quality of the Web with the use Machine Learning?"

3 METHODS

The research goal is to develop a holistic understanding of how the Information Quality of Web documents can be semi-automatically asses by using an fixed set of features.

We collect the features for the Web documents from the study of Ceolin et al. that are accessible and where the content of the Web document is in English.

In table 1 we provide an overview of the features that we collect and in table 2 we provide an overview of the labels that we collect.

Table 1: Overview of the features that we collect

Features	Description
Response	The time it took to load the Web document in seconds.
Lowwot (the lowest score) and Highwot(the highest score)	Website reputations received from crowd users. The users answered the question How much do you trust this site (the domain)?
pictures	The numbers of pictures that a Web document contains
polarity	The polarity inside the text. -1 = negative and 1 = positive
subjectivity	The amount of subjectivity inside the text. 0.0 is very objective and 1.0 is very subjective
yearsarchive	The year date that the first archive has been made from this domain.
wordcount	The numbers of words that the Web document contains
Flesch Score	Level of difficulty to understand the text based on the average readability skills of the average reader: Higher scores = easier to read, Lower scores = more difficult to read (range from 0 till 100, rarely the formula doesn't work and you get an score bellow 0 or above 100)

Via the crowd we collect our labels. We did this by asking the same question as Ceolin et al. did in their research for the same set of Web documents. Each Web document is annotated by 15 different annotators.

We perform a test of association with the test of Kendall coefficient in order to test if there is a significance relation between the features that we collected and how our crowdworkers scored these URLs.

Via a multi label regression we trained our framework to learn the function χ. In order to test the quality of our algorithm we train our Machine Learning Algorithm with the labels that we collected via the crowd and test it on the dataset of Ceolin et al.

We use the coefficient of determination R^2 to determine the amount of variance in the Information Quality dimensions scores variables that are predictable from our set of features. In our context we use it to measure how close or how well our predicted values (notation as $\hat{Y}_i$) fits compared to the the actual value (notation as

Table 2: Overview of the labels that we collect

Labels	Question
complete	how complete is the information in this document
accuracy	how correct is the information in this document
precise	how precise is the information in this document as opposed to vague
readable	how readable is the article
relevant	how relevant is the document
trustworthy	how trustworthy is the source
overall	in overall how good is the is the information quality of this article
neutral	is the document neutral with respect to the topic addressed or does it show a clear stance eg pro against

Y_i) with labels which it has never seen before. Bellow we show how the R^2 (the coefficient of determination) score is computed.

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2 \tag{1}$$

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2 \tag{2}$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \tag{3}$$

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{4}$$

In order to the determine to what extent our model approaches the actual values we perform three different studies.

In the first study we train our model by training our model with the data that we collected via the crowd and where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL. We test it (the test set Y_i is the set of actual value) on the dataset of the Media experts to predict their Information Quality scores. The predicted scores are $\hat{Y}_i$ here.

In the second study we train our model by training our model with the data that we collected via the crowd and where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL. We test it (the test set Y_i is the set of actual value) on the dataset that we collected via the crowd (without aggregating the Information Quality scores) to predict their Information Quality scores. The predicted scores are $\hat{Y}_i$ here.

In the third study we train our model by training our model with the Media experts dataset where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL. We test it (the test set Y_i is the set of actual value) on the dataset of the Media experts (without aggregating the Information Quality scores) to predict their Information Quality scores. The predicted scores are $\hat{Y}_i$ here.

3.1 Method limitations

We believe that the design of the Web documents and that media type of content and ads inside Webdocuments have an influence on the Information quality of Web documents. In this research we mainly focus on the textual content of these Websites.

The quality of the features trustworthiness that we collected via the API has a changing behavior. This implies that tomorrow we can have other scores. Our Natural Language Processing (NLP) features are an prediction. This implies that there can be an difference between the predicted scores and the actual NLP features scores. The feature response time is not very reliably collected, because the response time of a Webserver depends on many factors that we did not controlled (for example the geographically distance between the Webusers and the Webservers). These type of issues are important because the when there is a difference between the true underlying quantitative parameters of the features and the estimated or collected scores of our features this will influences the feature weights (the decision that our algorithm took during the training process of determining the importance of features) that our machine learning algorithms assigns to our set of features.

Different Web users have a different perception of the Information Quality of an Web document. Also the stance of the Web users towards a topic that is mentioned can change over time.

4 RESULTS

We managed to collect the features of 39 Web documents from the study of Ceolin et al.[4] about vaccination. In table 3 we show the distribution of the set of features that we collected.

Each URL has been assed by 15 different annotators. We then computed the average Information Quality scores for each Web document. In table 4 we present our correlation matrix for the crowd dataset and in table 5 we do this for the media experts dataset.

4.1 Model fit

In table 6 you see the outputs of all the Information Quality dimension scores their R^2 score for the tree studies that we mentioned in method section 3 and we show the overall R^2 score of our model by computing the average of the outputs of all the Information Quality dimensions by summing up their R^2 dividing them by numbers of Information Quality dimensions.

5 DISCUSSION

To our knowledge this is the first study where there is a platform created that provides insights about the Information Quality of Web documents. Our framework does this at high performance by distributing the collection of several data sources (data sources that are needed for processing the set of features) over a pool of threads. Our framework is also capable of retrieving documents inherent a given topic of interest to the user, and to present their assessment in a comparable manner. We can do this at large scale by assessing the Information Quality score of these Web documents in parallel.

Unfortunately the set of features that we collected did not fully utilized the range of the possible distribution range and therefore this limits our Framework to fully understand the meaning of the feature, because in Machine Learning it is sometimes the special

Table 3: Distribution of the features

fleschscore	highwot	lowwot	pictures	polarity	reponsetime	subjectivity	wordcount	yeararchive
Min. : 42.90	Min. : 2.00	Min. : 0.00	Min. : 1.00	Min. :-0.01284	Min. :0.004049	Min. :0.0000	Min. : 10.0	Min. :1996
1st Qu.: 56.30	1st Qu.:56.00	1st Qu.:13.00	1st Qu.: 8.00	1st Qu.: 0.05374	1st Qu.:0.006046	1st Qu.:0.4458	1st Qu.: 558.5	1st Qu.:1998
Median : 61.60	Median :85.00	Median :32.00	Median :13.00	Median : 0.08996	Median :0.014754	Median :0.4629	Median : 804.0	Median :2005
Mean : 62.81	Mean :72.18	Mean :34.54	Mean :17.95	Mean : 0.08508	Mean :0.200786	Mean :0.4562	Mean :1154.9	Mean :2004
3rd Qu.: 67.31	3rd Qu.:92.00	3rd Qu.:57.50	3rd Qu.:25.50	3rd Qu.: 0.10373	3rd Qu.:0.307840	3rd Qu.:0.4943	3rd Qu.:1465.5	3rd Qu.:2009
Max. :108.70	Max. :94.00	Max. :68.00	Max. :95.00	Max. : 0.20000	Max. :1.274372	Max. :0.5767	Max. :4295.0	Max. :2014

Table 4: Correlation Matrix crowd dataset

	complete	accuracy	precise	readable	relevant	trustworthy	overall	neutral	fleschscore	highwot	lowwot	pictures	polarity	reponsetime	subjectivity	wordcount	yeararchive
complete	1.00	0.54	0.77	0.40	0.63	0.67	0.57	0.31	-0.14	0.11	0.12	-0.02	0.12	-0.01	0.27	0.28	-0.07
accuracy	0.54	1.00	0.66	0.53	0.55	0.81	0.60	0.45	-0.23	0.20	0.08	0.06	0.16	-0.01	0.15	0.14	-0.29
precise	0.77	0.66	1.00	0.41	0.68	0.72	0.55	0.33	-0.15	0.21	0.25	0.01	0.02	0.06	0.14	0.14	-0.24
readable	0.40	0.53	0.41	1.00	0.37	0.55	0.51	0.24	-0.17	0.03	-0.07	0.22	0.18	0.23	0.34	-0.07	-0.07
relevant	0.63	0.55	0.68	0.37	1.00	0.62	0.42	0.36	-0.40	0.13	-0.07	-0.15	0.09	-0.12	0.09	0.09	-0.20
trustworthy	0.67	0.81	0.72	0.55	0.62	1.00	0.54	0.34	-0.14	0.13	-0.02	0.00	0.16	-0.08	0.15	0.04	-0.21
overall	0.57	0.60	0.55	0.51	0.42	0.54	1.00	0.45	-0.12	0.39	0.33	-0.15	0.05	-0.18	0.27	-0.12	-0.52
neutral	0.31	0.45	0.33	0.24	0.36	0.34	0.45	1.00	-0.30	0.17	0.12	-0.20	0.02	-0.29	0.05	-0.14	-0.41
fleschscore	-0.14	-0.23	-0.15	-0.17	-0.40	-0.14	-0.12	-0.30	1.00	0.16	0.40	0.17	-0.10	0.29	-0.33	-0.01	-0.13
highwot	0.11	0.20	0.21	0.03	0.13	0.13	0.39	0.17	0.16	1.00	0.63	-0.32	0.03	-0.41	-0.14	-0.35	-0.55
lowwot	0.12	0.08	0.25	-0.07	-0.07	-0.02	0.33	0.12	0.40	0.63	1.00	-0.28	-0.12	-0.12	-0.14	0.03	-0.65
pictures	-0.02	0.06	0.01	0.22	-0.15	0.00	-0.15	-0.20	0.17	-0.32	-0.28	1.00	-0.11	0.42	0.25	0.22	0.19
polarity	0.12	0.16	0.02	0.18	0.09	0.16	0.05	0.02	-0.10	0.03	-0.12	-0.11	1.00	-0.10	0.29	-0.04	0.06
reponsetime	-0.01	-0.01	0.06	0.23	-0.12	-0.08	-0.18	-0.29	0.29	-0.41	-0.12	0.42	-0.10	1.00	0.10	0.43	0.29
subjectivity	0.27	0.15	0.14	0.34	0.09	0.15	0.27	0.05	-0.33	-0.14	-0.14	0.25	0.29	0.10	1.00	0.18	-0.00
wordcount	0.28	0.14	0.14	-0.07	0.09	0.04	-0.12	-0.14	-0.01	-0.35	0.03	0.22	-0.04	0.43	0.18	1.00	0.26
yeararchive	-0.07	-0.29	-0.24	-0.07	-0.20	-0.21	-0.52	-0.41	-0.13	-0.55	-0.65	0.19	0.06	0.29	-0.00	0.26	1.00

Table 5: Correlation Matrix Media experts dataset

	complete	accuracy	precise	readable	relevant	trustworthy	overall	neutral	fleschscore	highwot	lowwot	pictures	polarity	reponsetime	subjectivity	wordcount	yeararchive
complete	1.00	0.81	0.63	0.23	0.53	0.66	0.64	0.57	-0.10	0.15	0.34	-0.09	-0.11	0.05	0.17	0.32	-0.11
accuracy	0.81	1.00	0.68	0.31	0.64	0.88	0.82	0.68	-0.01	0.29	0.36	-0.18	0.05	-0.16	0.06	-0.00	-0.23
precise	0.63	0.68	1.00	0.28	0.73	0.63	0.70	0.57	-0.17	0.13	0.21	-0.27	0.03	-0.12	0.13	0.14	-0.08
readable	0.23	0.31	0.28	1.00	0.38	0.19	0.20	0.16	-0.10	-0.21	0.04	0.07	-0.02	0.22	0.43	-0.09	-0.10
relevant	0.53	0.64	0.73	0.38	1.00	0.55	0.57	0.60	0.09	0.16	0.25	-0.13	0.11	-0.10	0.15	0.12	-0.06
trustworthy	0.66	0.88	0.63	0.19	0.55	1.00	0.81	0.74	-0.07	0.36	0.30	-0.32	0.12	-0.33	-0.07	-0.25	-0.19
overall	0.64	0.82	0.70	0.20	0.57	0.81	1.00	0.73	-0.03	0.49	0.35	-0.10	0.11	-0.20	0.01	-0.12	-0.32
neutral	0.57	0.68	0.57	0.16	0.60	0.74	0.73	1.00	-0.02	0.35	0.35	-0.36	0.23	-0.25	0.01	-0.14	-0.24
fleschscore	-0.10	-0.01	-0.17	-0.10	0.09	-0.07	-0.03	-0.02	1.00	0.17	0.40	0.17	0.28	0.30	-0.13	-0.02	-0.13
highwot	0.15	0.29	0.13	-0.21	0.16	0.36	0.49	0.35	0.17	1.00	0.63	-0.31	0.10	-0.43	-0.12	-0.35	-0.54
lowwot	0.34	0.36	0.21	0.04	0.25	0.30	0.35	0.35	0.40	0.63	1.00	-0.27	-0.00	-0.13	-0.08	0.02	-0.65
pictures	-0.09	-0.18	-0.27	0.07	-0.13	-0.32	-0.10	-0.36	0.17	-0.31	-0.27	1.00	-0.20	0.42	0.28	0.22	0.19
polarity	-0.11	0.05	0.03	-0.02	0.11	0.12	0.11	0.23	0.28	0.10	-0.00	-0.20	1.00	-0.09	-0.16	-0.16	0.04
reponsetime	0.05	-0.16	-0.12	0.22	-0.10	-0.33	-0.20	-0.25	0.30	-0.43	-0.13	0.42	-0.09	1.00	0.05	0.38	0.30
subjectivity	0.17	0.06	0.13	0.43	0.15	-0.07	0.01	0.01	-0.13	-0.12	-0.08	0.28	-0.16	0.05	1.00	0.14	-0.02
wordcount	0.32	-0.00	0.14	-0.09	0.12	-0.25	-0.12	-0.14	-0.02	-0.35	0.02	0.22	-0.16	0.38	0.14	1.00	0.26
yeararchive	-0.11	-0.23	-0.08	-0.10	-0.06	-0.19	-0.32	-0.24	-0.13	-0.54	-0.65	0.19	0.04	0.30	-0.02	0.26	1.00

Table 6: Test results of the three studies

TestNo	Average	complete	accuracy	precise	readable	relevant	trustworthy	overall	neutral
1	-0.17	-0.37	-0.14	-0.09	-0.13	-0.02	-0.07	-0.09	-0.42
2	0.05	0.05	0.04	0.06	0.02	0.05	0.06	0.06	0.05
3	0.31	0.27	0.37	0.33	0.26	0.25	0.38	0.3	0.32

values (e.g. extreme values or combination of values) that have an actual meaning.

Using a Random Forrest regression multi model algorithm our framework learned several functions in order to predict multiple quality scores at the same time by using meta-analysis that aggregates the Information that we can learn from our set features in combination with the set of labels. This will lead to a higher statistical power and a more robust point estimate compared to the framework of Ceolin et al. [4]. We trained our framework with

the Random Forrest regression because it is often referred to be resilient in dealing with skewness labels. We do believe that our set of features can be used in order to predict the Quality of a Web document if there is exactly one Truth about the Information Quality of a Web document.

We observed that some Information Quality dimensions have an strong relation with other Information Quality dimensions. This maybe means that there is some overlap with the semantically meaning of the Information Quality dimension. We also observed that our set of features in general have a weak relationship with the Information Quality scores, but that there are some Quality dimensions where there is moderate relationship strength between the features and the Information Quality scores. We also observed that different type of Webusers reacted different to the effect of the features on the Information Quality scores.

We observed that in our first study where we trained our model with the data that we collected via the crowd and where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL and tested it on the dataset of the Media experts to predict their Information Quality scores that the model that we trained predicted the Information Quality scores really poorly compared to the actual values.

In our second study where we trained our model with the data that we collected via the crowd and where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL and tested it om the test set crowd dataset (without aggregation) and observed that the this trained model fits better in this study compared to the results of the first study but still the model fits poorly with actual values.

In our third study where we trained our model with the Media expert dataset and where we aggregated the Information Quality scores by computing the grouped average Information Quality score for each URL and tested it on the the Media expert dataset (without aggregation) and observed that the this trained model has the best fits compared to the results of the first two studies this model has some fits with the dataset.

We believe that the errors can be explained by looking the impact of outliers, and the order of how the annotators assed the Web documents. The lack of consensus between annotators has also a negative impact on the fit of our Model. There can be several reason for this: the subjective nature of the annotators or the definitions of the IQ criteria can be the cause human bias. It can be the case that our set of features over or underestimated the actual values of the features. We also believe that differences of methodology, IQ criteria, the capability of assessing the IQ criteria, demo graphical factors, and the content and context in which the IQ scores was scored could have influence on our results. Therefore we would like to see whether our tool can be more valuable if we collected reviews of trained critical reviewers.

We believe that our framework should not behave statically but also dynamically. Ideally we would like to receive real-time feedback so that the execution of our framework will provide a representative assessment scores that is personalized. In order to validate this we would like to perform an experiment where we gather feedback from Web users about our predictions about the Information quality scores that we present.

In this study only English language sites were evaluated, and therefore the findings may not be generalization to those websites written in other languages. Our framework only works our framework was able to collect a complete set of features.

6 CONCLUSION

This paper answered the question: "Can we Semi-Automatically Assess the Quality of the Web with the use Machine Learning?" We tested the model: $\{\phi\}\chi\{\psi\}$ where $\{\phi\}$ is named the set of features and ψ is named the IQ assessment scores (the labels) and χ is the function that our model learned in order to Semi-Automatically Assess the Quality of the Web documents (SAATQOTW).

Our framework:

- is the first publicly available framework where Webusers can Semi-Automatically Assess the Quality of the Web documents
- has an unique set of features
- could be personalized when users of the tool provides us some insights about their information quality perception by providing feedback on the Information Quality predictions provided by us.
- is capable of retrieving documents inherent a given topic of interest to the user, and to present their assessment in a comparable manner.

While we have specifically focused on assessing the Information Quality of the textual content inside Web documents, the combination of the predicted assessment scores and the textual description's that we provide implies that our findings are likely to be of importance to Webusers by pre-assessing the Information Quality of Web documents where we present the Webusers an overview of this so that the Webusers can be more selectively selecting the Web documents based on their information needs. We might be able to see performance improvement using a larger dataset, which we didn't verified yet.

We believe that IQ of Web documents can be scored, but in order to generalize the findings of our framework difference of preferences and context understanding is required, because Machine's can learn only one truth. We could also take the perception of one well trained experts in the field of assessing the Information Quality of Web documents and train our model based on the perception of the expert and assume that the expert knows what is good for everybody else. An example where this is applied for example is in the field of Spam filters. Another direction is to have an consensus policy by aggregating the information quality scores and that the Web users that uses these scores agree on these scores or for example by being very exclusively at predicting the Information Quality of a Website if and only of more then 95% of the annotators would agree on this. We also believe that our framework can be useful for search engines in order to filter Web document of extreme low quality.

In terms of feature research we particularly suggest:

- Our architectural design of our framework can be adapted for information workers. On the Web there is so much knowledge and Web users have very less insights whether Information on the web is relevance for them or not. By combining Information sources together and finding dependencies by

between analyzing different Web data sources and performing multidimensional dimensional analysis and mapping this with Information Question of Information Worker we believe that we can make these information workers aware of situations that are relevant for them by reasoning. This would encourage decision makers to take evidence based decisions by analyzing historical data where decisions where taken and how these decisions effected the critical performance indicators and the the perspectives of the audience and it's stakeholders towards the decision.

- We would like to validate the coding of the Web document according to the standards of the World Wide Web Consortium (W3C) and classify the output of the validator as an document of high quality or low quality based on the labels that we collected for the Web document.
- To place our framework in it is usage context. By this we mean that we want to understand (with the use of explicit or implicit feedback) how differences of the topics and the differences in the perception of the Information Quality can be assessed so that our framework can dynamically estimate the Information Quality of a Web document, given that we know the stance of the reader towards the content of the Web document and given that we know how the reader of the Web document perceive the information Quality of a Web document.
- The crowd provides us demographically information the annotators that can be useful for understanding differences of preferences and dealing with these demographical factors.
- Investigating the effects of media (video's and pictures) inside Web document and the Design of the Website can provide insights that affect the IQ of a Web document. We also believe that feature learning can help us to automatically discover the representations needed these type of media inside Web documents in order to deal with it's effect on the Online Web documents.

7 ACKNOWLEDGMENTS

This work was supported by the Qupid Project.

REFERENCES

[1] Aladwani, A. M., and Palvia, P. C. Developing and validating an instrument for measuring user-perceived web quality. 467 – 476.

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research, January 2017.

[3] Ceolin, D., Noordegraaf, J., and Aroyo, L. Capturing the ineffable: Collecting, analysing, and automating web document quality assessments. In *20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024* (New York, NY, USA, 2016), EKAW 2016, Springer-Verlag New York, Inc., pp. 83–97.

[4] Ceolin, D., Noordegraaf, J., and Aroyo, L. Web data quality assessment.

[5] Ceolin, D., Noordegraaf, J., Aroyo, L., and van Son, C. Towards web documents quality assessment for digital humanities scholars. In *Proceedings of the 8th ACM Conference on Web Science* (New York, NY, USA, 2016), WebSci '16, ACM, pp. 315–317.

[6] Clark, L., and Marchi, R. *Young People and the Future of News: Social Media and the Rise of Connective Journalism.* Communication, Society and Politics. Cambridge University Press, 2017.

[7] Ding, X., Liu, B., and Zhang, L. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2009), KDD '09, ACM, pp. 1125–1134.

[8] Domingos, P. A few useful things to know about machine learning. *Commun. ACM 55*, 10 (Oct. 2012), 78–87.

[9] Fitzsimmons, P., Michael, B., Hulley, J., and Scott, G. A readability assessment of online parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh 40*, 4 (December 2010), 292âĂŤ296.

[10] Graczyk, M., Lasota, T., Telec, Z., and Trawiski, B. *Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 111–120.

[11] Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY, USA, 2009), JCDL '09, ACM, pp. 295–304.

[12] Held, J., and Lenz, R. Towards measuring test data quality. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops* (New York, NY, USA, 2012), EDBT-ICDT '12, ACM, pp. 233–238.

[13] Ho, J., and Tang, R. Towards an optimal resolution to information overload: An infomediary approach. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work* (New York, NY, USA, 2001), GROUP '01, ACM, pp. 91–96.

[14] Kapyla, T., Niemi, I., and Lehtola, A. Towards an accessible web by applying push technology. In *Fourth ERCIM Workshop on "User Interfaces for All"* (Stockholm, Sweden, 1998).

[15] Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 591–600.

[16] Lipka, N., and Stein, B. Identifying featured articles in wikipedia: Writing style matters. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 1147–1148.

[17] Loiacono, E. T., Watson, R. T., and Goodhue, D. L. Webqual: A measure of website quality. *Marketing theory and applications 13*, 3 (2002), 432–438.

[18] Mitchell, T. M. *Machine Learning.* McGraw-Hill, 1997.

[19] Nicola Askham, Ulrich Landbeck, J. S. The six primary dimensions for data quality assessment, defining data quality dimensions. DAMA UK Working Group.

[20] Nielsen, J. *Designing Web Usability: The Practice of Simplicity.* New Riders Publishing, Thousand Oaks, CA, USA, 1999.

[21] Paasche-Orlow, M. K., Taylor, H. A., and Brancati, F. L. Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine 348*, 8 (2003), 721–726. PMID: 12594317.

[22] Pinto, A. M., Oliveira, H. G., and Alves, A. O. Comparing the performance of different nlp toolkits in formal and social media text. In *SLATE* (2016).

[23] Raiber, F., and Kurland, O. Using document-quality measures to predict web-search effectiveness. In *Proceedings of the 35th European Conference on Advances in Information Retrieval* (Berlin, Heidelberg, 2013), ECIR'13, Springer-Verlag, pp. 134–145.

[24] Si, L., and Callan, J. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (New York, NY, USA, 2001), CIKM '01, ACM, pp. 574–576.

[25] Wand, Y., and Wang, R. Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM 39*, 11 (Nov. 1996), 86–95.

[26] Witten, I. HFrank, E. *Data mining.* Morgan Kaufmann, 2000.

[27] Zhung, Y., and Mecer, R. A machine learning approach for rating the quality of depression treatment web pages.