Assessing the Information Quality of Web documents

Ozkan Sener
Vrije Universiteit Amsterdam
ozkansener@gmail.com

ABSTRACT

Web users who searches for information via a search engine have very less insights about the IQ of Web documents. We trained a multiple output regression for the task of estimating the Information Quality (IQ) of Web documents based on historical data where features and assessments where collected. For retrieving the IQ scores we automatically collect their features and predict the IQ based on the patterns that our algorithm learned. The model for semi-automatically assessing the IQ of Web documents was inspired by the work of Ceolin et al. Compared to their framework our Framework is also capable of retrieving documents inherent a given topic of interest to the user in a comparable manner, we provide descriptive insight about the content of the Web document and we increased the responsivity of the information.

1 INTRODUCTION

Web user are overwhelmed by the quantity of information that is available on the Web [11]. This makes it difficult for a Web user to receive information of high quality. Web users don't know how good the Information Quality (IQ) of a Web document is by just looking at the Uniform Resource Locator (URL). information overload occurs when the brain of the Web user process more information then it can handle [8]. When Web user gets confused they will no longer be able assess the quality of the Web documents [10]. These issues means there is a need for a framework which provides insights about the IQ of Web documents.

We propose an information retrieval framework that can Semi-Automatically Assess the IQ of Websites by collecting historical labelled 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 assessment scores (the labels) for a new instance by automatically collecting the features of the Web document. Compared to previous research: 1) Our framework is publicly available by using open-sourced libraries. 2) We added features like numbers of pictures, numbers of years that a source exist, response time of the source. 3) Experimented with the crowd to collect annotators to label the Web documents 4) Trained our model via a multiple output regression instead of a 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. 6) Our framework is responsive and capable of retrieving multiple Web documents in parallel. 7) We show different interactive methods for presenting the outcome of our analysis.

Our framework can be valuable to Web users by pre-assessing the IQ of Web documents and presenting this output that makes it possible for Web users to decide which Web documents they want to access based on their information preferences.

This paper continues as follows: Section 2 introduces related work; Section 3 describes our research question. Section 4 describes the methods adopted; Section 5 presents the results collected, that are analysed in Section 6. Section 7 concludes the paper.

2 RELATED WORK

Several authors [1, 4, 12, 18] report that there is need for a framework that is able to assess the quality of Websites automatically.

2.1 Frameworks

Ceolin et al. [4] created a model that semi automatically estimates the assessment quality scores of Websites. Their model does this by performing computations that extracts features functioning as input for their model by the assumption that the features represents the quality scores of Web sites. However we observed that this was not the case. The disagreements between Web users are not unambiguous. Selecting the Support Vector Machine algorithm for such a task creates a large amount of variance due to the penalizing property of the algorithm this creates a large amount of unexplainable variance. 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 IQ of Websites. The third component is the Machine Learning algorithm that automatically estimates the IQ scores of Websites.

Ding et al. [5] report that analysis like sentiment and emotion analysis (granular form of sentiment analysis) can be used to discover 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 these bad/good words. Web of Trust (WOT) is a Crowd sourcing platform where Web domains are rated. 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. Ceolin et al. extracted the average trustworthiness score which is not provided retrievable anymore. Instead the API shows the lowest and the highest WOT rating scores where 0 means very untruthful and 100 very trustful. The model that was trained by Ceolin et al. had not the opportunity to learn what each features represents. 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. Extreme values of the features outliers are relevant for pressure points discovery and some algorithms like the weighted least squares algorithm even adds weights to outliers [**?**]. When the interpretation of feature and how the feature influences the assessment scores is known by the Machine Learning Algorithm (MLA), the Machine Learning Algorithm it's believe is more likely to fit with the representation of the Web users. Ceolin et al. performed their assessments with media experts. We hypothesize that annotators with a heterogeneous background is needed in order to

generalize this Framework for all the Web users on the Web. We also extract features like the response time, numbers of pictures and the matureness since our previous work showed that these factors effected the IQ perception of the Web users from the study of Ceolin et al.

Pinto et al. [15] 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 Api's that provide the same NLP features as IBM Watson API. 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 *Features.*

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 most important/relevant details so that readers easily can understand what the paper is about instead of getting confused. Therefore, We hypothesize that having features that describes the amount of text on a Web document can be useful for our representation. Dalip et al. [8] 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 hypothesize that Web users are more interested in Web documents that are simple and clear. A high Flesh Reading Score indicates that it is likely that the content of the Web document is easy to read. A Low Flesh Reading Scores indicates that it's likely that the Web document is not easy to understand for the average Web user, but for a university graduated person it might be understandable [7].

Response time. Nielsen [14] reports that user-perceived IQ is highly influenced by the response time of a Web server to reply on the Web users it's request to access a Web document. When Websites are taking too long to respond to Web user it request it is less likely that the Web user is satisfied about the IQ of a Web document. Web users also think that Websites with a high response time are incompetent. We hypothesize 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.

2.1.2 *Machine Learning.*

Determine which algorithms is the best depends on the usage scenario and the assumptions that we make about the data environment. In our opinion the deep understanding of the extracted pattern and how useful these patterns are depending on the assumptions of the Algorithm Architect. The decision of algorithm selection is subjective, because different people have different views of the same situations. The mathematical properties of a Machine Learning algorithm are a formal view that describes the behavior of Machine Learning algorithms. Algorithm architects are functional experts that are creating functional specification for problem solving by computational learning. Based on this property the architects decide whether they are going to build their own algorithm or use an existing algorithm based on the behavior specification of the algorithm and how satisfy able a existing algorithm if for their problem. Like humans have different learning styles machines also have different styles of learning and in this study we apply algorithms of well known supervised and unsupervised algorithms. Checking the correctness of Algorithm means that we want to assess whether the algorithm shows the desired behaviour. The desired behaviour is in general determined by functional and non functional requirements. 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 algorithms another scenarios where the fit of a model is not perfect, but generalizable needs to be extracted. 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 because the algorithm needs to be able to function in a noisy environment.

We hypothesize that we can provide Web users insight about Web users by using supervised learning and unsupervised learning algorithms. For the Supervised learning task we can similarly to the Framework of Ceolin et al. collect a set of features for the Web document that we assess and also collecting specific annotations for these Web documents from Web users. This means that the structure of Web documents must be known and the goal here then is to estimate an annotation of a Web document by the Belief of our model how the set of features represent the IQ of Web documents.

An alternative method for providing insight about the IQ of Web can be by finding structures within Web documents.

Given a set X of Web documents in the form of X $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ each Web document we compute the dimension vector of each Web document. The numbers of k is determined by us. Before defining the k, in general data scientist run a test with many different values of k and then analyse the outcomes. The larger the numbers of k is the less useless our insight are to the Web users, because we then provide to much information towards the Web users. The hierarchical clustering would be a better solution for this problem, the K-Means is flat cluster type of algorithm which is faster than hierarchical clustering. But the hierarchical clustering provides a layer type of analyse. After the clusters are generated Documents are assigned inside a cluster by performing an optimization learning where the goal is to decrease the variance inside the clusters in general iterations are performed for the task of determine the differences of Web documents and the Variance this caused. A less often applied algorithm due to it's computational requirements are topological types of algorithms. Evolving and adaptive algorithms have a well fondness behaviour which is in mathematical logic used for inductively discovering the semantically connections and organizing this in multiple layers of hierarchy depending on the

input of the algorithm While formal methods will provide more accurate results their computations are more expensive and in general require parallel execution. We hypothesize that clustering would make Information Retrieval via the Web more insightful for Web users by effectively showing grouping Web documents in categories that represents the similarity and dissimilarity between Web documents. A more granular form of similarity and dissimilarity of Web documents can be gathered via distance metrics (cluster algorithms apply distance metrics). Distance metrics are often used for copyright discovery or fraud detection. We hypothesize that we can provide Web users more insights about the content inside the Web documents if we categorize Web documents, assess the similarity between Web documents and estimating the Quality of the Web in terms of added value, but also content and meta information.

2.2 information Dimensions

There are different standards in the field of information retrieval that describe how the quality of documents can be assessed. Raider and Kurland [16] 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. In this study we mainly focus on the textual content of the Web documents and originality of the textual content.

The DAMA [13] has created a data quality framework with the dimensions: Accuracy, Consistency, Integrity, Timeliness, Completeness and Validity. Ceolin et al.[2], Wang[17], Held and Lenz [9] 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. Because we are reusing the IQ Assessment scores from the study of Ceolin et al. we do not use the IQ dimensions of Raider and Kurland [16]. We hypothesize 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 IQ of the Web. Crowd sourcing makes it also possible for us to reach Web users all over the world and acquiring their assessment.

2.2.1 Crowd-sourcing. Via the crowd we can acquire a huge amount of different occupants in a small amount of time in order to give them the job of assessing the IQ of the Web. Crowd sourcing makes it also possible for us to reach Web users all over the world and acquiring their assessment.

3 RESEARCH QUESTION

In this work, we answer the following question: "How can we semi-automatically assess the IQ of Web documents?" Concerning this issue, we address two sub research questions:

Can we (partially) represent how Web users perceive the IQ of Web documents? Here we like to see whether we can extract features that are (partially) representing the IQ of Web document.

Can we provide Web users insights about the information inside Web documents? with the use of Unsupervised Machine Learning? Unsupervised Machine Learning provide Web users with relevant insights about the information of Web documents.

4 METHODS

4.1 Extraction, Transformation, Load (ETL)

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.

Via the crowd we collected more assessment scores. 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.

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.

4.2 Evaluation

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 crowd workers scored these URLs.

We use the coefficient of determination R^2 to determine the amount of variance in the IQ dimensions' scores variables that is predictable from our set of features [6]. 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 Y_i) with labels of annotators which it has never seen before. Bellow we show how the R^2 (the coefficient of determination) score is computed.

Table 1: Overview of the features that we collect

Features	Description
latency	The time it took to load the Web document in seconds.
reputation High (the highest score) and Reputation Low (the lowest 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
subjective	The amount of subjectivity inside the text. 0.0 is very objective and 1.0 is very subjective
maturity	The year date that the first archive has been made from this domain.
word count	The numbers of words that the Web document contains
readability	Flesh Kincaid 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 a score bellow 0 or above 100)

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 IQ of this article
neutral	is the document neutral with respect to the topic addressed or does it show a clear stance e.g. pro against

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

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

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

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

In order to the determine to what extent our model approaches the actual values we perform four 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 IQ scores by computing the grouped average IQ 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 IQ 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 IQ scores by computing the grouped average IQ 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 IQ scores) to predict their IQ 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 IQ scores by computing the grouped average IQ 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 IQ scores) to predict their IQ scores. The predicted scores are $\hat{Y}_i$ here.

5 RESULTS

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

In our user study on the crowd we assessed each URL by 15 different annotators. We then computed the average IQ 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.

5.1 Model fit

For a simple explanation of our IQ score we trained a standard linear regression because we can easily present on understandable Web with the use of visualization to Web users as a powerful communication tools. However, the amount of variance that we observed and the lack of learning capabilities of our model caused to much variance for us to have trust in linear models. For example, sometimes our model predicted IQ scores very often to high. We also don't have insights how the noise inside is caused due the features and how it effects the IQ scores. However it provides insight to Web users when their is a strong disagreements of how features influences the IQ scores because we think that it might be more natural for Web users to interpret the output of these models by using the computation that our model performed in a visual so that the Web users can use the insights that they gathered via our framework to understand how our model influenced their preferences by checking whether behaving of our model is reasonable with their preferences and easily adjusting their preferences with an simple extra form where the Web users can indicate their preference since

Table 3: Distribution of the features

fleschscore	highwot	lowwot	pictures	polarity	latency	subjectivity	word count	year archive
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	word count	year archive
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
word count	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
year archive	-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 expert's dataset

	complete	accuracy	precise	readable	relevant	trustworthy	overall	neutral	fleschscore	highwot	lowwot	pictures	polarity	reponsetime	subjectivity	word count	year archive
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
year archive	-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

the formula of a Regression model is:

$$\begin{aligned}\text{IQ score} = \alpha &+ \beta_1 \text{Pictures} \\ &+ \beta_2 \text{Maturity} \\ &+ \beta_3 \text{Polarity} \\ &+ \beta_4 \text{Subjectivity} \\ &+ \beta_5 \text{LowWot} \\ &+ \beta_6 \text{HighWot} \\ &+ \beta_7 \text{Words} \\ &+ \beta_8 \text{Readability} + \epsilon\end{aligned}$$

This would mean if we could assume linearity that the each estimated effects of β_i could have been controlled by individual Web users and this might even. Looking at the mathematical behaviour of this model we see that this model would not be capable of learning the complex patterns of gathering the complex insight of IQ and the assumptions of linearity does not hold in our model.

In table 6 you see the outputs of all the IQ dimension scores their R^2 score for the tree studies that we mentioned in method section 4 and we show the overall R^2 score of our model by computing the average of the outputs of all the IQ dimensions by summing up their R^2 dividing them by numbers of IQ dimensions.

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

6 DISCUSSION

To our knowledge this is the first study where Web intelligence is applied for the task of information Retrieval via analysis of Web documents and providing this information in a multidimensional format for increasing the flexibility of selecting factors that are relevant for the task of information election. Our framework does this by distributing the collection of several data sources (data sources that are needed for processing the set of features) over a pool of threads. This increases the responsivity by simultaneously scheduling the execution of multiple http request for the extraction of multiple data sources for a single Web documents. 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 perform these assessments these over a pool of independent workers that are executed in parallel.

The functionality of Machine Learning dependent on it score performance (for example the R^2), the amount relative diversity ($R_A(x) = \frac{(Ax,x)}{(x,x)}$ relative diversity is total) and learning characteristics of the base learners. In order to tell what aspects or insight of IQ can be represented by the different configuration of our model and whether the representation dependent on the Web users, the Web documents and the quality of representatives of the content of the Web documents by the set of Features. Because the set of features that we collected did not fully utilized the possible distribution range and therefore this limits our Framework to fully understand the meaning of the features, because in Machine Learning it is sometimes the special values (e.g. extreme values or combination of values) that have an actual meaning. For example, the Natural Language Processing libraries that we used was trained on a movie reviews corpus and this might not be accurately re-presentable for the meaning of the features polarity and subjectivity. However, on the Web there are annotated datasets of news articles, tweets that can be used to train larger model that extracts the features of sentiments.

When linearity would hold (or can be controlled) for Web users and Web users want to adjust the behaviour of model to indicate how they think that the features influences the IQ scores or even a more complex adjustment would be in the formula $E(y) = (B(x|\alpha))$ adjusting for each individual Web user α the weight that has to be given to the feature if the feature is more important to the Web user and how the feature is important by changing the estimated effect B of our model. We could also decide to improve our parameters of the random forest that is capable of learning complexes machine learning algorithm of understanding individual preferences of Web users, because We hypothesize that the quality of our model depends on the effect of unobserved latent variables. More specific: to build a customized framework for a given Web user. We could be identifying the preferences of the past Web users who are similar to the present Web users and using their data to train a model that fits better with the preferences of the present Web users due to some similarity of Web users and we know that similarity metrics can be used for discovering the latent variables so it might even be used as a feature generator for adjusting the model of the Web users.

It was not our goal to achieve the highest scores, because it is knowing that algorithm exist that can easily explain the training data because it was stricter trained to learn the patterns from the training data. Instead we designed to train our model learn generalizable patterns by limiting the depth of the random forest algorithm. Strategic application of learning algorithms requires the capability of adaptability of correct functioning of a Machine Learning algorithm in an uncontrollable environment. Learning algorithms that focused on learning deep patterns are less likely to have generalizable properties. They are often also computational expensive and query map types of algorithms are understandable and showed some generational without the application of regulation and other optimizing strategies.

Using a Random Forrest regression multi model algorithm our framework learned several functions in order to predict multiple quality scores at the same time 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. [3]. We decided to train our framework with the Random Forrest regression because it is often referred to be resilient in dealing with scenes labels. The Support Vector Machine algorithm has the property of learning by punishing the learned pattern when this pattern does not fit and We hypothesize that this is not reliable for analyzing data in the wild which in general is noisy.

We observed that some IQ dimensions have a strong relation with other IQ dimensions. This maybe means that there is some overlap with the semantically meaning of the IQ dimension. We also observed that our set of features in general have a weak relationship with the IQ scores, but that there are some quality dimensions where there is moderate relationship strength between the features and the IQ scores. We also observed that different type of Web users reacted differently to the effect of the features on the IQ 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 IQ scores by computing the grouped average IQ score for each URL and tested it on the dataset of the Media experts to predict their IQ scores that the model that we trained predicted the IQ 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 IQ scores by computing the grouped average IQ score for each URL and tested it om the test set crowd dataset (without aggregation) and observed that 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 IQ scores by computing the grouped average IQ score for each URL and tested it on the the Media expert dataset (without aggregation) and observed that 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 hypothesize that the errors can be explained by looking the impact of outliers, and the order of how the annotators annotated the Web documents which we could not control. 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 hypothesize 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 representation that is likely to be relevant for Web users by presenting the recommendation that are likely to interesting to Web users. In order to validate this, we would like to perform experiments where we gather explicit or implicit feedback from Web users about the usefulness of the insights that we provide about the IQ scores and the information Insights that we present.

Compared to search engines and other methods of information retrieval this strategy makes effective information Retrieval possible. While many Data Scientist agree that one of the critical performance indicators of Knowledge or information Engineering is to have a high quality Extraction, Transformation, Load strategy this is not possible given the current usage scenario of the Web. Unknown security concerns on the Web might decrease for example malicious infecting Websites that profiled Web users and even create threats in society.

In this study only English language sites were evaluated, and therefore the findings may not be generalization to those websites written in other languages. The default parameter of the package Sklearn that we used was not investigated. We where not able to perform a more solid types of experiment studies so this decreased control of the annotators and their effect on the assessments. We hypothesize that the design of the Web documents and that media type of content and ads inside Web documents have an influence on the IQ 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 behaviour. This implies that tomorrow we can have other scores. Our Natural Language Processing (NLP) features are a prediction. This implies that there can be a 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 Web users 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.

7 CONCLUSION

This paper set out to answer the question: How can we semi-automatically assess the IQ of Web documents?. Through semi-automatically assessing the IQ of Web documents, we have found that our set of features provides some insights about the IQ of Web documents. These findings are important contributions to Web users and search engines for gathering insights about the IQ scores of Web documents.

Our framework:

- Has a unique combination set of features and some of these features like the years of archive, numbers of pictures inside the Web document.
- is capable of retrieving documents inherent a given topic of interest to the user, and to present their assessment in a comparable manner.
- Shows different ways of presenting the analysis

While we have specifically focused on assessing the IQ 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 Web users by pre-assessing the IQ of Web documents where we present the Web users an overview of this so that the Web users can be more selectively selecting the Web documents based on their information needs.

8 ACKNOWLEDGEMENTS

This research learned me that there are enormous parameter spaces inside textual document which causes a Huge amount of variance. On the Web there is a huge amount of Data and we are curious whether we can apply information Retrieval tactics for knowledge extraction by inferentially reasoning. I want to thank my supervisor Davide Ceolin for introducing me to topics like Natural Language Processing, Crowd-sourcing and the feedback he provided.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] Dougherty, E. R., Kim, S., and Chen, Y. Coefficient of determination in nonlinear signal processing. *Signal Process. 80,* 10 (Oct. 2000), 2219–2235.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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).

[12] 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.

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

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

[15] 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).

[16] 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.

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

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