# Laying the Explorative Groundwork for Document Quality Assessment and Perspective Detection

Lesia Tkacz
Vrije Universiteit Amsterdam
dx9240@gmail.com

Robin Kumar Sharma
Vrije Universiteit Amsterdam
robinkerlaam@gmail.com

Chantal van Son
Vrije Universiteit Amsterdam
c.m.van.son@vu.nl

Davide Ceolin
CWI
davide.ceolin@cwi.nl

## ABSTRACT

In this abstract, we address the problem of information overload that web users face when researching controversial topics on the web. The ultimate goal of our project is to design a tool to aid users in reviewing and learning about the quality and content of large document collections online.

We highlight the challenges as well as the previous and future steps taken to develop a document information quality visualization tool. This lays the explorative groundwork for developing a web browser tool which efficiently informs users about the quality of web documents according to 8 quality dimensions. We also address the challenges of identifying author and source perspectives in documents and collections of controversial debate related text. We use Natural Language Processing (NLP) methods to explore these linguistic features and to extract important textual content with which to contextualize the document information quality. Finally, we motivate our planned crowd-sourced study which is designed to explore the feasibility of automatically evaluating contradictory perspectives across controversial documents. We take the highly publicized vaccination debate as a focal point with which to focus development.

## 1 INTRODUCTION

When searching for information on the web, the typical web user must consume many web documents in order to learn about a controversial topic, as well as the various perspectives through which it can be viewed. Reviewing all the relevant documents on a topic as returned by a web search is an overwhelming task, as there are no measures or indicators describing each document's content. The user's alternative is to only review a handful of documents, but this risks creating an 'information bubble' of documents which promote only one opinion. Thus, reducing the load of information that must be consumed is key to alleviating the user's information overload problem, and can pave the way to helping the user to make informed choices about the information they wish to consume on the web.

Advancements in Information Extraction (IE) and Natural Language Processing (NLP) make it possible to automatically assess the quality of web documents [2], which is understood to be a measure of how valuable a piece of information is with respect to the user's needs. Further, it is possible to extract perspectives from web documents texts [9], such as the beliefs, opinions, stances, and world views of the author and of quoted sources. Building upon studies such as these, our project is able to consider the problem of information overload that web users face.

## 2 BACKGROUND

The challenge of visually expressing information qualities is being attempted by some online platforms.[1,2] These try to raise the user's awareness about the political views expressed in online articles, such as political polarity. However, these existing web implementations are rather course-grained as they focus on pinpointing political leanings, rather than a more fine-grained decomposition of different aspects of information quality as [2] does. Finally, these platforms do not support a range of web document genres, in contrast to [2], which aims to process and measure the quality of web document genres as diverse as online public health advice, news reports, personal blogs, and so forth. Yet automatically identifying perspectives across controversial debate text topics is a serious challenge, as NLP methods which are developed to extract stances from one debate topic are not necessarily transferable to another [7]. We also note that human readers can find it difficult to clearly identify perspectives in such texts as well.

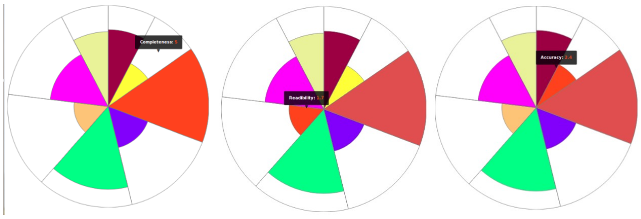


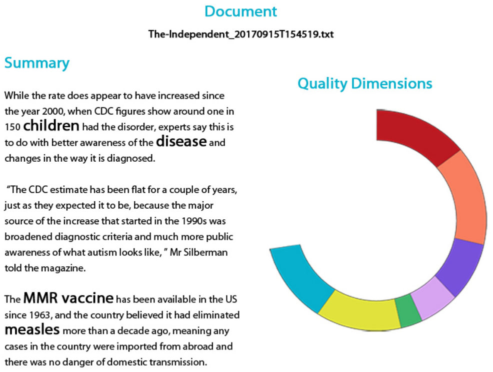
**Figure 1: Current Design**

## 3 PROJECT OUTLINE

We address these drawbacks by first focusing on how to visually present document quality, and text content, to web users. We do this by entering an iterative design process to research visualization methods, which are guided by the goal of generating an informative representation of multiple web documents. This visual representation of 8 quality dimensions is combined and contextualized with summary sentences, which are automatically extracted from the

*Network Institute Academy Assistant Programme, 2017, VU*
.

[1] www.filterbubblan.se
[2] www.readacrosstheaisle.com



Document

The-Independent_20170915T154519.txt

Summary

While the rate does appear to have increased since the year 2000, when CDC figures show around one in 150 **children** had the disorder, experts say this is to do with better awareness of the **disease** and changes in the way it is diagnosed.

"The CDC estimate has been flat for a couple of years, just as they expected it to be, because the major source of the increase that started in the 1990s was broadened diagnostic criteria and much more public awareness of what autism looks like," Mr Silberman told the magazine.

The **MMR vaccine** has been available in the US since 1963, and the country believed it had eliminated **measles** more than a decade ago, meaning any cases in the country were imported from abroad and there was no danger of domestic transmission.

**Figure 2: Second Design**

documents in order to give the user an impression of the document's content and debate stance. Our ultimate design goal is currently progressing, with plans to reach a stage where the information quality visualization tool it is incorporated into browsers as a web information assessment tool that is accessible to the average user. Our second focus is on exploring how best to harness NLP methods and resources for approaching the difficult task of automatically identifying perspectives in controversial debate texts, with an eye to incorporating successes into the web browser tool.

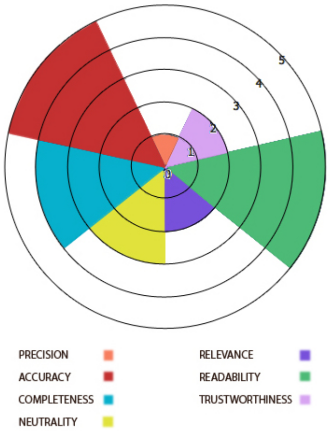


**Figure 3: Third Design**

## 4 TOOL CHALLENGES AND DESIGN

Our information quality visualization tool has gone through several design iterations (*Fig. 2* and *Fig. 3*). In each iteration, we address the challenges of how to graphically represent numerical data and textual content without implying false relationships, as well as how to compare multiple web documents [8]. In its fourth and current design iteration (*Fig. 1*), the now responsive visualization graphically models the output of a quality assessment tool [2] which automatically evaluates the text of a web page according to 8 quality dimensions: Accuracy, Completeness, Neutrality, Relevance, Readability, Trustworthiness, Precision, and Overall Quality. Each of these are scored on a scale from 1 to 5. Additionally, we draw on methods from information retrieval and NLP to help contextualize the quality scores with textual content (seen in *Fig. 2*). Automatic text summarization based on TextRank [5] is used to extract the top most relevant sentences from the documents being visualized, and Term Frequency - Inverse Document Frequency (TF-IDF) is employed to extract the most relevant keywords from the single document, as well as from the document collection.

The next steps in the development of this information visualization tool portion of our project is to perform demonstrations, as well as a user testing study.

## 5 NLP CHALLENGES AND STUDY DESIGN

Identifying the overall stance (*pro/con/neutral*) towards topics within a controversial debate as expressed in a document is the first step towards identifying a document's perspective [7]. Next, we aim to find contradictory or non-contradictory relationship between information. For example, one document may contain sentences suggesting a causal link between vaccines and autism, whereas another document may state that there is no evidence to support this claim. Detecting contradictions between sentence pairs is reminiscent of the established NLP task of Recognizing Textual Entailment (RTE) [1]. However, it has been shown that the performance of existing RTE systems has been overestimated due to statistical irregularities in the datasets [3, 4, 6]. In addition, our own investigations in processing, extracting, grouping, and comparing texts from our vaccination corpus[3] suggest that actual, nuanced debate texts, rather than clear-cut constructed sentence pairs, can pose classification difficulties to humans. This is an important point to keep in mind, as the optimal NLP results should typically match human performance. In order to gauge the difficulty of the task, we are implementing a crowd-sourced study to ascertain whether non-expert human judges are able to annotate both stance and contradictions in the vaccination corpus, and if they can reach a high level of agreement with each other. The results, then, could indicate if it is realistic to expect automatic RTE methods to also be able to perform the task with success. This can pave the way for further judging as to whether or not evaluating or fine-tuning existing RTE resources may advance automatic perspective detection in controversial debate texts.

## 6 CONCLUSION

Our web document information quality visualization tool will soon reach its user testing stage, and is progressing in parallel to the detailed design and execution of the crowd-sourced stance annotation study. These two portions of our project are moving together towards an interdisciplinary full paper contributing to research in NLP, as well as Computer and Web Science.

## REFERENCES

[1] G. Bowman, S R.and Angeli, C. Potts, and C D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[2] D. Ceolin, J. Noordegraaf, and L. Aroyo. 2016. Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments. In *Proceedings of the European Knowledge Acquisition Workshop (EKAW).* 83–97.

[3] Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th*

[3] www.vaccinationcorpus.wordpress.com

*Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia, 650–655.

[4] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of NAACL-HLT*. New Orleans, Louisiana, 107–112.

[5] R. Mihalcea and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[6] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*. New Orleans, Louisiana, 180–191.

[7] S. Somasundaran and J. Wiebe. 2010. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. 116–124.

[8] L. Tkacz, R K. Sharma, D. Ceolin, and C. van Son. 2018. Visualizing Information Quality and Perspective on the Web. In *WebSci'18 Main Conference Poster Session Pre-Proceedings*. 23–24.

[9] C. van Son, T. Caselli, A. Fokkens, I. Maks, R. Morante, L. Aroyo, and P. Vossen. 2016. GRaSP: A Multilayered Annotation Scheme for Perspectives. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 1177–1184.