OCR quality assessment: Beyond ground truth
KB, National Library of the Netherlands

About KB, National Library of the Netherlands
The KB, national Library of the Netherlands collects everything that is published in the Netherlands and about the Netherlands. The KB collections include books, newspapers, and periodicals, but also websites and e-books. With 450 fte the KB is one of the largest cultural heritage institutions in the Netherlands. The Research department (20 fte) focuses on topics ranging from providing insights into the development of the public library sector to applying AI techniques to make the collections accessible to a broad international audience. The KB is an active player in the field of Artificial Intelligence: it is co-founder of the Cultural AI Lab and the working group Culture&Media of the Netherlands AI Coalition.

Our challenge: OCR quality assessment beyond ground truth
Heritage organizations such as libraries and archives host a wealth of written resources of great cultural relevance to researchers and the public. Thanks to considerable investments in the digitization of historical resources, digital images of an unprecedented number of documents are now available. At the current moment, the KB collection already includes more than one hundred million digitized books, newspapers, and periodicals.

While this is a step in the right direction when it comes to making historical resources more accessible, it falls short where searchability is concerned. Only the availability of digital texts can dramatically improve access to digitized resources. This is why Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are so central to the strategy of libraries and archives. Recent advances in machine learning have indeed allowed to automatically extract texts from historical resources at scale, yet a question always remains open: is the OCR quality good enough? This question is critical as it has been shown that the quality of OCR impacts the usability of the resulting texts [4,6].

The systematic assessment of OCR ideally requires high-quality ground truth, manually checked by human experts [2]. This is a laborious and costly endeavor, which has so far slowed down digitization efforts that, invariably, include OCR as a key component of their workflow. We propose a challenge centered on devising, developing, testing and comparing approaches to OCR quality assessment not requiring ground truth.

This timely topic has only recently started to be explored in the literature, with promising results. Possible approaches include using lexicon-based methods such as dictionary lookup and pre-trained language models [5], combining OCR assessment metrics [1], or applying metrics devised in domains of research, such as reading comprehension [3]. Time is ripe for making a step towards a systematic comparison of such automated OCR quality assessment methods not requiring ground truth. This task will be based on the use of ground truth provided by the KB to measure how well non-ground truth methods compare against established, ground truth ones. The data proposed for the challenge will also allow to compare results over language, source typology and time, and more. This challenge is centered around ICT competences, and will be of interest to computer scientists, linguists, and digital humanists.

Bibliography
5. Ströbel, Phillip Benjamin, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel,


**Team Leaders**

**Academic Team Leaders**

- Dr. Giovanni Colavizza, University of Amsterdam

KB, national Library of the Netherlands

- Mirjam Cuper, Data Scientist
Explainable Fall Risk Prediction for Elderly People
VeiligheidNL

About VeiligheidNL
VeiligheidNL is the Dutch knowledge center for injury prevention. For more than 35 years, we have been committed to making the lives of millions of people safer by encouraging safe behaviour in a safe environment. VeiligheidNL is a Public Benefit Organisation (ANBI, Algemeen nut beogende instelling). Around 60 employees work for VeiligheidNL. The yearly budget is around 7.5 million euros per year. Most of our funding is provided by the Ministry of Health, Welfare and Sport (VWS). Other sources of funding include the Ministry of Infrastructure and Water Management, ZonMw and local authorities. Core activities are research, professional development, consulting and education.

Our challenge:
Elderly fall prevention is one of the six core programs of VeiligheidNL. Every 5 minutes, a person over 65 ends up in a Dutch Emergency Room after a fall accident. And this is only 10% of the total number of falls of the elderly. The consequences of a fall are enormous, both in terms of personal suffering and social care costs. A fall accident has a major impact on the self-reliance of the elderly, the ability to live at home longer and the quality of life. The direct medical care costs of fall accidents have already exceeded 1 billion euros in 2020. One of the goals of VeiligheidNL is to reduce the number of fall injuries in the elderly by developing effective interventions and disseminating knowledge and expertise (https://www.veiligheid.nl/kennisaanbod/valpreventie-bij-ouderen).
Fall prevention has been included in the 2021-2025 coalition agreement. This makes substantial resources available that contribute to the structural implementation of effective fall prevention for the elderly. Municipalities are given the task of providing fall prevention programs for their residents aged 65 and over. This requires investments in detection of elderly people at a high risk of falling, and the development and execution of fall prevention programs.

In this case study we would like to investigate if we can predict the fall risk of elderly people based on information that is collected in the primary health process, such as emergency room transcripts and electronic health records, and studies such as the Longitudinal Aging Study Amsterdam (LASA, https://lasa-vu.nl/).

Screening tools have been developed to analyse the fall risk of an individual elderly person such as the Fall risk 65+ (https://interventies.loketgezondleven.nl/leefstijlinterventies/interventieszoeken/1401588). However, these tools require substantial effort from medical professionals to fill the detailed analysis out, while there is already a lack of available medical professionals. Therefore we would like to use machine learning, and in particular text mining to make predictions about fall risk.

One of the major activities in such a study is to select the predictors, only predictors that are likely to contribute to the prediction, and preferably predictors that have proven to be a risk factor in earlier research will be taken into account. In this case study we will not select the predictors manually, but take a data driven approach and use machine learning to train an optimal model to predict fall risk, taking into account all the available information, i.e. open text (answers) as well as categorical data. So by using machine learning we aim to have a more efficient way to develop and execute fall risk prediction tools. The feasibility of this approach is also confirmed in a case study by Ye et al [2020] (https://doi.org/10.1016/j.ijmedinf.2020.104105) on English EHR Data, but they did not use text mining on the notes in the record.

An important requirement of the models to be developed is that they are able to explain why the model has come to the conclusion of a high fall risk, i.e. which aspects are the risk factors. These risk factors can then be used to determine the most efficient intervention. A part of the case study could also consist of developing an interface in which the results of the explainable models are presented to the medical professionals.

**Team Leaders**

Academic Team Leaders

- Dr. Maya Sappelli, HAN University of Applied Sciences

VeiligheidNL

- Dr. Rianne Kaptein
Brand Value Classification in Text
Textmetrics

About Textmetrics
Textmetrics offers an augmented writing assistant that gives writers support and feedback during their writing process. Most of our users are recruiters writing job applications. Our editor works with a number of modules, addressing different contextual aspects of text. This includes various discrimination checks but also feedback to the tone of voice and compliance to brand identity.

Our challenge:
For the brand identity module we are looking for a way to classify if a sentence contains a company's brand identity. Marketing departments of a company usually have a set of brand values they want their company to present. Popular values are Loyalty, Honesty, Trust, Ingenuity, Accountability, Simplicity, Respect etc. As these are broad terms, companies usually have their own definition of how they want to communicate these values and which emphasis they want to put on which value.

Our idea would be to have a solution that allows us to input a certain set of sample sentences that align with one of the brand values. This set would be used to train a classifier that eventually gives feedback if a sentence is addressing a brand value or not. This does not have to be an exclusive score. One sentence may contain several brand values. Either combined or in different sub clauses of the sentence.

This approach should be easily extendable so that it can be trained easily for new customers who might have a slightly different definition of a similar brand value or a completely different, not yet seen brand value. You can think of this somewhat like a sentiment classifier, but instead of sentiments it would predict brand values.

The challenge in this assignment is the relatively small amount of positive-only training data. There are many more ways the same brand identity can be described so you cannot just assume that all sentences that are not in the set of sample sentences do not fit the brand identity. Further, sentences can express more than one brand value.

In order to use the model in production, we need a good architecture for the training procedure that allows us to extend the model for a new customer easily. We therefore do not only need the model for prediction, but we also need a well designed training procedure that can be used to add labels to the model without impacting the performance of already existing labels.

Team Leaders

Academic Team Leaders

• TBD

Textmetrics

• Kyrill Poelmans
Gender Bias Detection in Job Ads
Textmetrics

About Textmetrics
Textmetrics offers an augmented writing assistant that gives writers support and feedback during their writing process. Most of our users are recruiters writing job applications. Our editor works with a number of modules, addressing different contextual aspects of text. This includes various discrimination checks but also feedback to the tone of voice and compliance to brand identity.

Our challenge:
One of our discrimination modules checks for gender bias in job advertisements. Fields which are dominated by a certain gender, job advertisements contain gendered wording. Especially for women, this can be a reason to not apply to a position they are perfectly qualified for. For this task, we would like you to come up with a way to predict gender bias in job advertisement. Our current approach involves word lists that have shown to be either female or male connoted. However, at this moment we are only taking single words into consideration, because there are no scientific resources with lists of short expressions etc. Furthermore, the word list is only disambiguated by POS tagging. We hope that the gender bias which was shown to be encoded in the vector space of large language models can be utilized to detect gender biased words and expressions. Translating social psychological concepts into algorithms is challenging. There is no ground truth and no labeled data. There are pointers as to which concepts and words contain gender bias, you can use them as anchor points in your approach. Additionally, the final product should be explainable in a way that the bias can be "reduced" to sets of words, phrases or sentences in the job advertisement. On the one hand this allows to give actionable advice. On the other hand, this adds the necessary transparency to the model, which is especially needed when it comes to discrimination topics.

Ideally we would like to be able to determine gendered words / phrases and suggest changes to the users of our platform. Given the challenging and open nature of this project, we do not want to define the final project too rigidly but give you the freedom to decide what fits. Your final product could be an extended list of words and phrases that you found while exploring the vector space or an approach to classify/quantify the gender bias in a text while keeping the explainability in mind.

Team Leaders

Academic Team Leaders

- TBD

Textmetrics

- Kyrill Poelmans
Automatic Trailer Storyline Generation
RTL Nederland

About RTL Nederland
RTL Nederland is the largest commercial broadcaster in the Netherlands, with a yearly turnover of around €350m. RTL has the mission to tell unmissable stories that touch heart and mind. With a team of 10 data scientists, RTL works on challenges such as personalization, forecasting and automatic content generation.

Our challenge:
Every day, RTL produces many hours of video content that aims to touch the heart and mind of viewers. With advanced artificial intelligence technologies, we believe that we can support our creative people, by supporting the creative process of promotional content creation. For an earlier case at ICT with Industry 2022, we picked up the challenge of automatically selecting scenes that could be used in trailers for TV programs. Our goal was to generate from a full length video a short teaser that increases interest in viewing the content, without spoiling the experience. While we were able to successfully rank scenes by their “trailerness”, we noticed we were missing key storytelling capabilities for making a compelling trailer. This is what we would like to study in our case for ICT with Industry 2023. For our trailerness model, we created a dataset of episodes from our daily soap opera GTST with shot-level annotations of recurring video content. Each GTST episode starts with about a minute of recaps from previous episodes and ends with a short preview of the next episode. This recurring content makes for ideal training material in a weakly-supervised learning task for what content is interesting to include in promotional material. We know from directors that good storytelling in trailers is key, as it compels attention. A better understanding of the narrative in video content would unlock an important aspect of the automatic generation of promotional video content. We found in our previous case that just selecting “trailerworthy” scenes is not enough, because a compelling trailer tells a story and understands the storyline. We are seeking to close the gap between a good human-made trailer and automatically selected content. Our goal is to understand storylines better and to generate trailers that summarize future content more coherently. Combining the applied expertise of RTL and the Hogeschool Utrecht, will make for exactly the right atmosphere for tackling this challenge.

Team Leaders

Academic Team Leaders

- Dr. Stefan Leijnen, Hogeschool Utrecht
- Floor Schukking, Hogeschool Utrecht

RTL Nederland

- Dr. Daan Odijk
- Dr. Hendrik Vincent Koops
Modeling and Model Based Testing of 3D games
Axini

About Axini

One of the challenges in gaming, and virtual racing in particular, is testing. A big and important part of testing is manual testing done by a human, often professional, driver. There are also automated tests and it is a big challenge to keep them running. To thoroughly test a game we need thousands of test-scripts and this is a big challenge to maintain. Wouldn’t it be great to have automated testing without the need to program test-scripts and test-data. Look no further, that technique exists and is called model-based-testing, also known as no-code scriptless testing. Axini is a leader in the development and application of model-based testing with their Axini Modeling Platform. Axini is wondering how model-based testing can be applied to gaming like serious virtual racing.

Our challenge:

Are you aware that a serious game like a Formula 1 simulator is tested by (expensive) Formula 1 drivers in order to ensure that the simulator correlates with reality. But even the testing of less serious 3D games like first person shooters costs millions of euros. On top of that, most of the testing is manual. We’d like to get this one step further: model based testing of 3D games. How cool would it be if we could model and model-based test professional Formula 1 simulators?

![Image of Formula 1 car](https://youtu.be/JgZGZcb3Zyg)

Today racing games are developed using all kinds of configuration parameters to simulate the behavior of the car on a track. Final testing is always performed by a (professional) human test driver. The test driver’s feedback is used to tweak the track and environmental conditions to get as close as possible to the reality of a real racing car. A first step to a pure virtual racing environment is the replacement of a test driver by a model. How can we transform the professional test diver into a model to execute the primary testing and reduce the testing effort of humans by 80%.

During the workshop we would like to investigate all aspects and physical feedback parameters of a professional test driver and create a model out of it for testing purposes at first. Next steps are to take the physical aspects of the track, the environmental conditions and complex interactions between driver and environment into consideration.

As we learned in the last ICT with industry, maybe professional Formula 1 simulators is one step too far for current theory and practice, but modeling and testing of 3D games in general is well within reach. In this ICT with Industry we want to come up with the next level of test automation for 3D games.

Team Leaders

Academic Team Leaders

- TBA

Axini

- Dr. Machiel van der Bijl