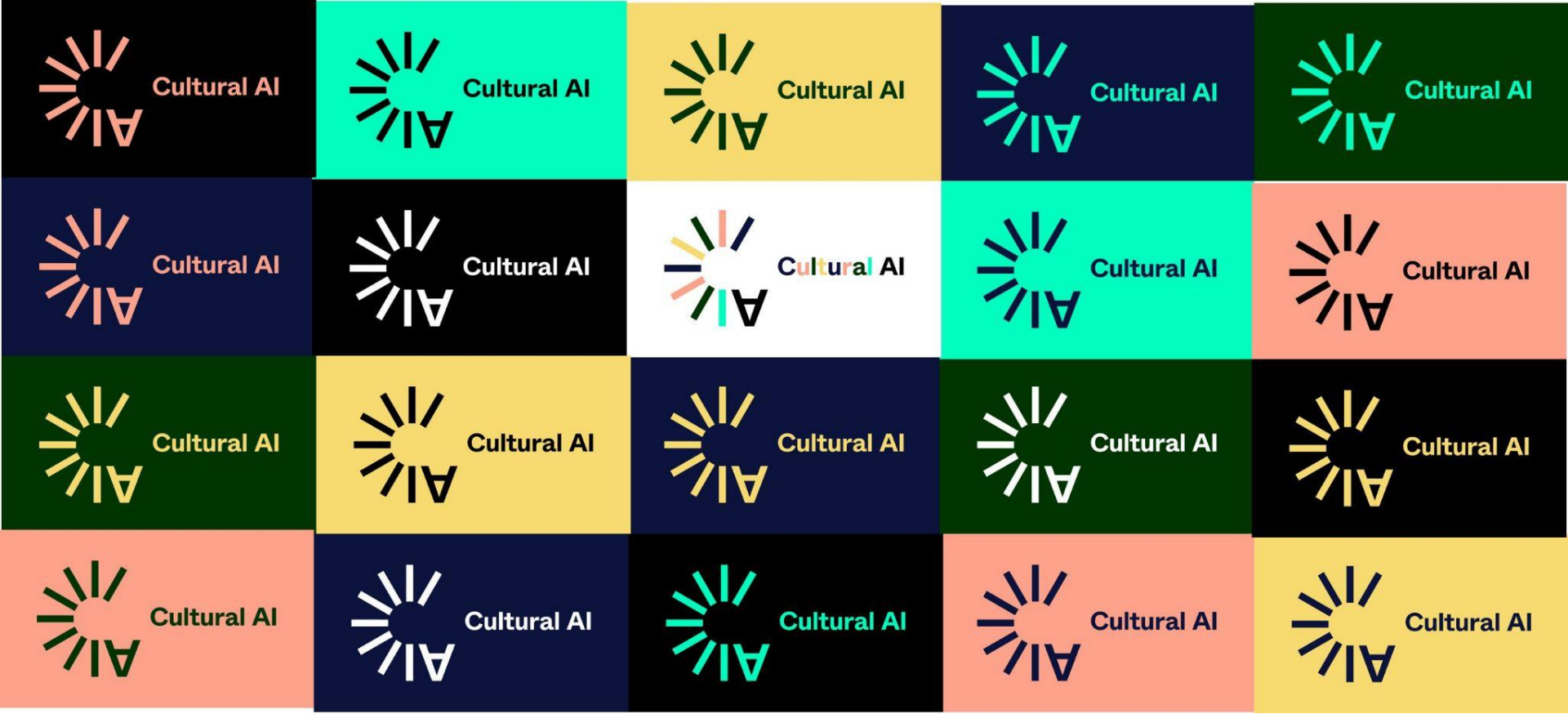


Cultural AI Lab: Engaging AI and Cultural Heritage

Laura Hollink, Marieke van Erp, Martijn Kleppe

CWI Centrum Wiskunde & Informatica

KNAW Humanities Cluster

KB

LIBER 2021 BELGRADE 50th Annual Conference ONLINE 23 -25 JUNE

Cultural AI Lab:
Engaging AI and Cultural Heritage

Laura Hollink
CWI

Marieke van Erp
KNAW HuC

Martijn Kleppe
KB, national library of the Netherlands

Cultural AI
a lab for culturally valued AI

LIBER 2021 BELGRADE 50th Annual Conference ONLINE 23 -25 JUNE

What is Culturally Aware AI ?

Laura Hollink
CWI

Centrum Wiskunde & Informatica

Jacco van Ossenbruggen
Group leader - User Centric Data Science - Vrije Universiteit Amsterdam
jacco.van.ossenbruggen@vu.nl

Marieke van Erp
Research group leader - DHLab - KNAW Humanities Cluster
marieke.van.erp@dh.huc.knaw.nl

Johan Oomen
Head of Research and Heritage Services - Netherlands Institute for Sound and Vision
joomen@beeldengeluid.nl

Martijn Kleppe
Head of the Research Department - KB National Library of the Netherlands
martijn.kleppe@kb.nl

Lotte Wilms
Senior Advisor Digital Scholarship - KB National Library of the Netherlands
lotte.wilms@kb.nl

Victor de Boer
Assistant Professor - User Centric Data Science - Vrije Universiteit Amsterdam
v.de.boer@vu.nl

Ryan Brate
Ph.D. student -
KNAW Humanities Cluster
ryan.brate@dh.huc.knaw.nl

Andrei Nesterov
Ph.D. student -
Centrum voor Wiskunde en Informatica
nesterov@cwi.nl

Saskia Scheltjens
Head of Research Services Department - Rijksmuseum
S.Scheltjens@rijksmuseum.nl

Antal van den Bosch
Director Meertens Instituut -
KNAW Humanities Cluster
antal.van.den.bosch@meertens.knaw.n

Stephan Raaijmakers
Senior Scientist - TNO
stephan.raaijmakers@tno.nl

Laura Hollink
Researcher -
Centrum voor Wiskunde en Informatica l.hollink@cwi.nl

Valentin Vogelmann
Researcher - KNAW Humanities Cluster
valentin.vogelmann@dh.huc.knaw.nl

Cedric Waterschoot
Ph.D. Student -
KNAW Humanities Cluster
cedric.waterschoot@meertens.knaw.nl

Julia Noordegraaf
Professor of Digital Heritage - Department of Media Studies - University of Amsterdam
J.J.Noordegraaf@uva.nl

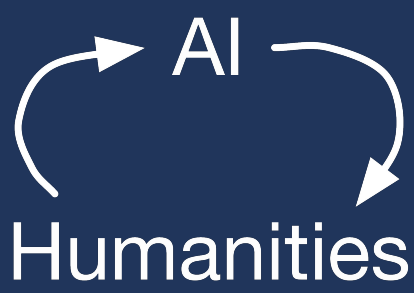
Tobias Blanke
Distinguished University Professor - AI and Humanities -
University of Amsterdam
t.blanke@uva.nl

What is Cultural AI?

"Cultural AI is the study, design and developmentof AI systems that are implicitly or explicitly awareof the subtle and subjective complexity of human culture."

Bias
Ethics
Cultural differences
Perspectives

AI for humanities, and humanities for AI



“Cultural AI is as much about using AI for understanding human culture as it is about using knowledge and expertise from the humanities to analyze and improve AI technology.”

Research Topics

- Context & Connections

  How can we automatically contextualise collection objects and link them to each other and across collections & information sources?

- Trust & Polyphony/Polyvocality

  How can we make other voices in data sets explicit?

CWI
Centrum Wiskunde & Informatica

DHLAB

Research Topics (ctd)

- Change & Variation
  How can we make AI tools deal with differences and evolution across time and space?

- Exploration & Interaction
  How can we make contextualised and polyvocal data insightful to users?

ICAI

Innovation Center for
Artificial Intelligence

The National Innovation Center for Artificial Intelligence (ICAI) has the mission to keep the Netherlands at the forefront of knowledge and talent development in AI.

Creating and nurturing a national AI knowledge and talent ecosystem.

More info: icai.ai

Layers of bias

- Bias in the data
- Bias introduced by tool creators
- Technological bias

We do not want to erase bias, we want to make it visible!

Algorithmic transparency

KB newspaper archive (100M articles, 6 months server logs) study:

- Reran 1M real user queries through the (black box) search engine
- Counted how often each document appears in top 10 (100, 1000)
- Analysed correlation technical document features with retrieval counts

Findings:

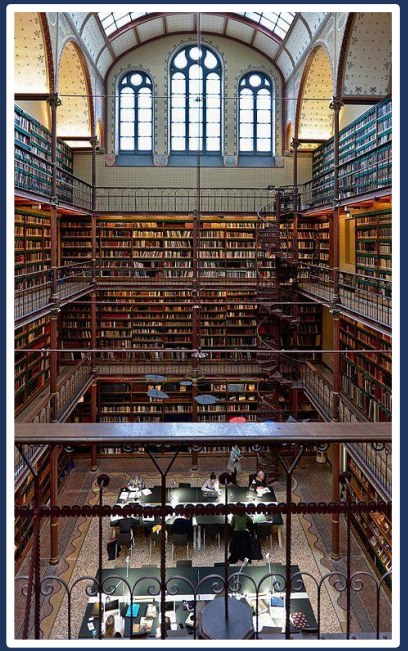
- 96% of the articles **never** make it into the top 10 (76% never in top 100)
- Engine discriminates against very short and very long documents
- Best scoring articles contain long list of names (local elections, swimming diplomas, ...)

CWI
Centrum Wiskunde & Informatica

DHLAB

Lab organisation and mission

- PhD students and researchers spend time at academic AND heritage partners
- Close collaboration through “data sprints”, monthly reading club, joint conference/workshop visits
- Core partners, Associate partners, Affiliate partners.
- **Innovation Center** for **Artificial Intelligence**



Current projects in the Cultural AI Lab

- BETTER-Mods (funded by NWO) 2 PhD students
- Culturally Aware AI (funded by NWO) 2 PhD students
- SABIO (funded by NDE) 1 researcher
- RE-FRAME (funded by Sound and Vision) 1 PhD student

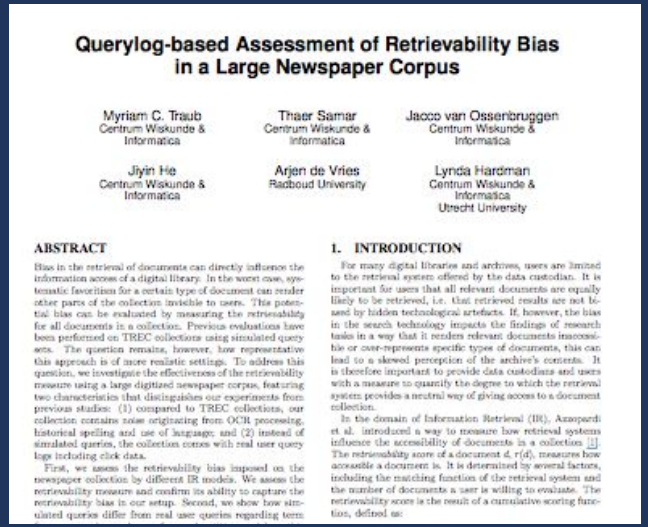
Upcoming:

- Library AI Principles (funded by National Library) 1 PhD student
- Researcher in Residence (funded by National Library) 1 postdoc
- Transparent pipelines (funded by NWO/NLeSc) 1 postdoc
- Responsible AI in public media (funded by NWO) 1 PhD student

Algorithmic transparency

KB newspaper archive (100M articles, 6 months server logs) study:

- Reran 1M real user queries through the (black box) search engine
- Counted how often each document appears in top 10 (100, 1000)
- Analysed correlation technical document features with retrieval counts



Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus

Myriam C. Traub, Centrum Wiskunde & Informatica
Thaer Samar, Centrum Wiskunde & Informatica
Jacco van Ossenbruggen, Centrum Wiskunde & Informatica
Jiyin He, Centrum Wiskunde & Informatica
Arjen de Vries, Radboud University
Lynda Hardman, Centrum Wiskunde & Informatica, Utrecht University

ABSTRACT

Bias in the retrieval of documents can directly influence the information access of a digital library. In the worst case, systematic favoritism for a certain type of document can render other parts of the collection invisible to users. This potential bias can be evaluated by measuring the retrievability for all documents in a collection. Previous evaluations have been performed on TREC collections using simulated query sets. The question remains, however, how representative this approach is of more realistic settings. To address this question, we investigate the effectiveness of the retrievability measure using a large digitized newspaper corpus, featuring two characteristics that distinguishes our experiments from previous studies: (1) compared to TREC collections, our collection contains noise originating from OCR processing, historical spelling and use of language, and (2) instead of simulated queries, the collection comes with real user query logs including click data.

First, we assess the retrievability bias imposed on the newspaper collection by different IR models. We assess the retrievability measure and confirm its ability to capture the retrievability bias in our setup. Second, we show how simulated queries differ from real user queries regarding term

1. INTRODUCTION

For many digital libraries and archives, users are limited to the retrieval system offered by the data custodian. It is important for users that all relevant documents are equally likely to be retrieved, i.e. that retrieved results are not biased by hidden technological artefacts. If, however, the bias in the search technology impacts the findings of research tasks in a way that it renders relevant documents inaccessible or over-represents specific types of documents, this can lead to a skewed perception of the archive's contents. It is therefore important to provide data custodians and users with a measure to quantify the degree to which the retrieval system provides a neutral way of giving access to a document collection.

In the domain of Information Retrieval (IR), Azzopardi et al. introduced a way to measure how retrieval systems influence the accessibility of documents in a collection [1]. The retrievability score of a document d, r(d), measures how accessible a document is. It is determined by several factors, including the matching function of the retrieval system and the number of documents a user is willing to evaluate. The retrievability score is the result of a cumulative scoring function, defined as:

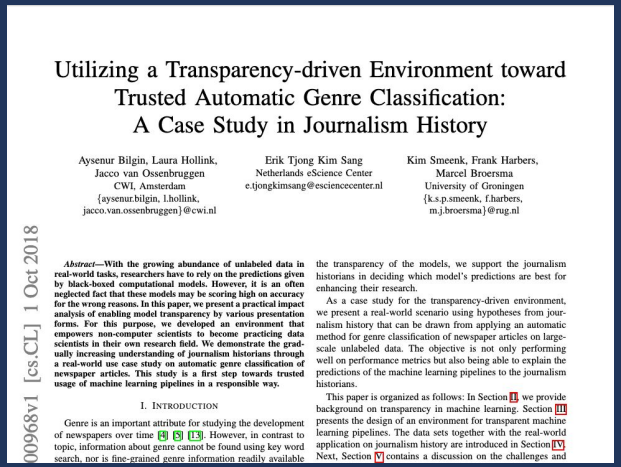
Findings:

- 96% of the articles never make it into the top 10 (76% never in top 100)
- Engine discriminates against very short and very long documents
- Best scoring articles contain long list of names (local elections, swimming diplomas, ...)

NEWSGAC project: Transparent Machine Learning Pipelines

Can we support humanities scholars using AI with transparency at every step in an AI pipeline?

- from data selection and data preparation, to choosing an algorithm selection, and inspection of the results



Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History

Aysenur Bilgin, Laura Hollink, Jacco van Ossenbruggen
CWI, Amsterdam
{aysenur.bilgin, l.hollink, jacco.van.ossenbruggen}@cwi.nl

Erik Tjong Kim Sang
Netherlands eScience Center
e.tjongkimsang@esciencecenter.nl

Kim Smeenk, Frank Harbers, Marcel Broersma
University of Groningen
{k.s.p.smeenk, f.harbers, m.j.broersma}@rug.nl

***Abstract*—With the growing abundance of unlabeled data in real-world tasks, researchers have to rely on the predictions given by black-boxed computational models. However, it is an often neglected fact that these models may be scoring high on accuracy for the wrong reasons. In this paper, we present a practical impact analysis of enabling model transparency by various presentation forms. For this purpose, we developed an environment that empowers non-computer scientists to become practicing data scientists in their own research field. We demonstrate the gradually increasing understanding of journalism historians through a real-world use case study on automatic genre classification of newspaper articles. This study is a first step towards trusted usage of machine learning pipelines in a responsible way.**

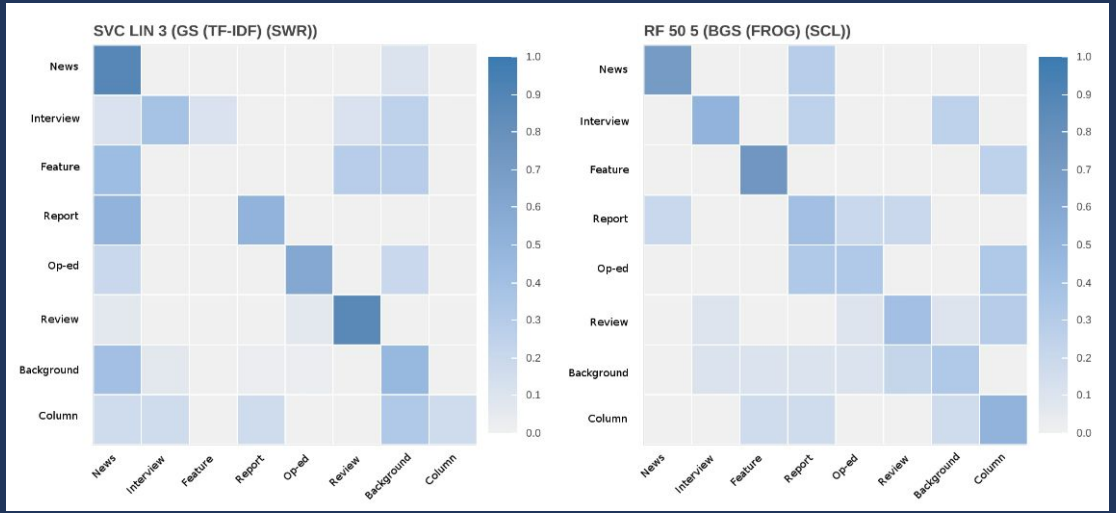
I. Introduction

Genre is an important attribute for studying the development of newspapers over time [4], [5] [13]. However, in contrast to topic, information about genre cannot be found using key word search, nor is fine-grained genre information readily available

the transparency of the models, we support the journalism historians in deciding which model's predictions are best for enhancing their research.

As a case study for the transparency-driven environment, we present a real-world scenario using hypotheses from journalism history that can be drawn from applying an automatic method for genre classification of newspaper articles on large-scale unlabeled data. The objective is not only performing well on performance metrics but also being able to explain the predictions of the machine learning pipelines to the journalism historians.

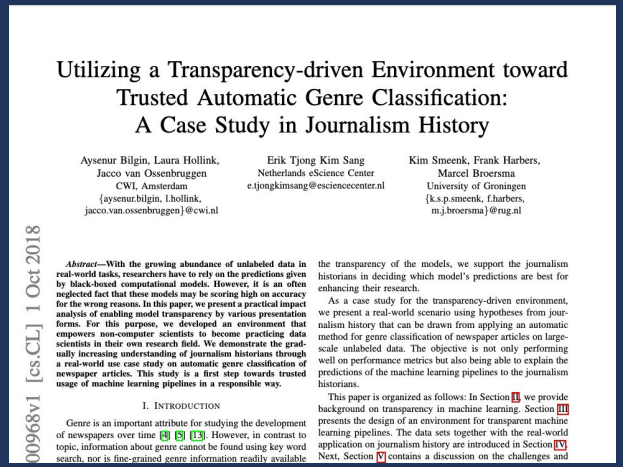
This paper is organized as follows: In Section II, we provide background on transparency in machine learning. Section III presents the design of an environment for transparent machine learning pipelines. The data sets together with the real-world application on journalism history are introduced in Section IV. Next, Section V contains a discussion on the challenges and



NEWSGAC project: Transparent Machine Learning Pipelines

Can we support humanities scholars using AI with transparency at every step in an AI pipeline?

- from data selection and data preparation, to choosing an algorithm selection, and inspection of the results



Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History

Aysenur Bilgin, Laura Hollink, Jacco van Ossenbruggen
CWI, Amsterdam
{aysenur.bilgin, l.hollink, jacco.van.ossenbruggen}@cwi.nl

Erik Tjong Kim Sang
Netherlands eScience Center
e.tjongkimsang@esciencecenter.nl

Kim Smeenk, Frank Harbers, Marcel Broersma
University of Groningen
{k.s.p.smeenk, f.harbers, m.j.broersma}@rug.nl

***Abstract*—With the growing abundance of unlabeled data in real-world tasks, researchers have to rely on the predictions given by black-boxed computational models. However, it is an often neglected fact that these models may be scoring high on accuracy for the wrong reasons. In this paper, we present a practical impact analysis of enabling model transparency by various presentation forms. For this purpose, we developed an environment that empowers non-computer scientists to become practicing data scientists in their own research field. We demonstrate the gradually increasing understanding of journalism historians through a real-world use case study on automatic genre classification of newspaper articles. This study is a first step towards trusted usage of machine learning pipelines in a responsible way.**

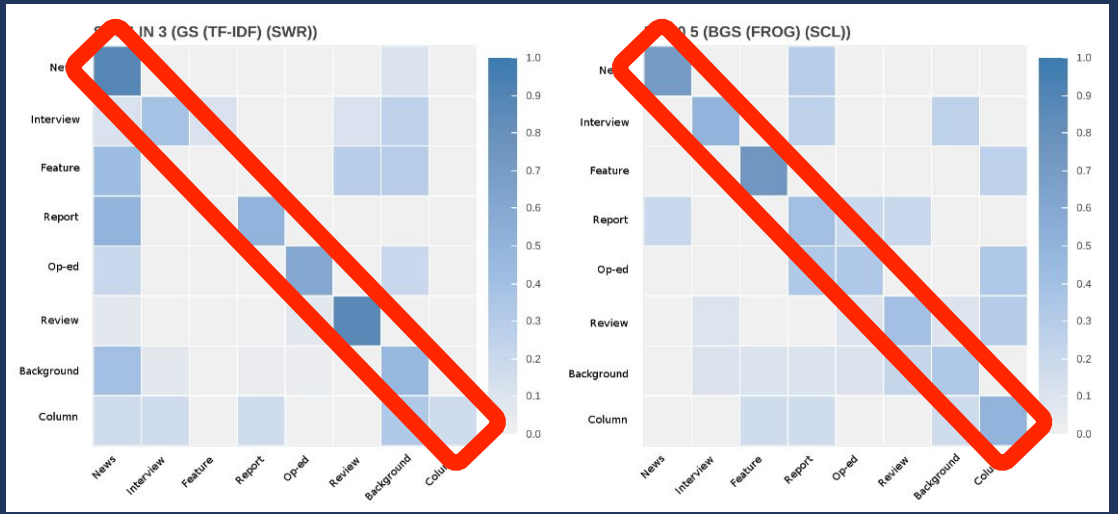
I. Introduction

Genre is an important attribute for studying the development of newspapers over time [4], [5] [13]. However, in contrast to topic, information about genre cannot be found using key word search, nor is fine-grained genre information readily available

the transparency of the models, we support the journalism historians in deciding which model's predictions are best for enhancing their research.

As a case study for the transparency-driven environment, we present a real-world scenario using hypotheses from journalism history that can be drawn from applying an automatic method for genre classification of newspaper articles on large-scale unlabeled data. The objective is not only performing well on performance metrics but also being able to explain the predictions of the machine learning pipelines to the journalism historians.

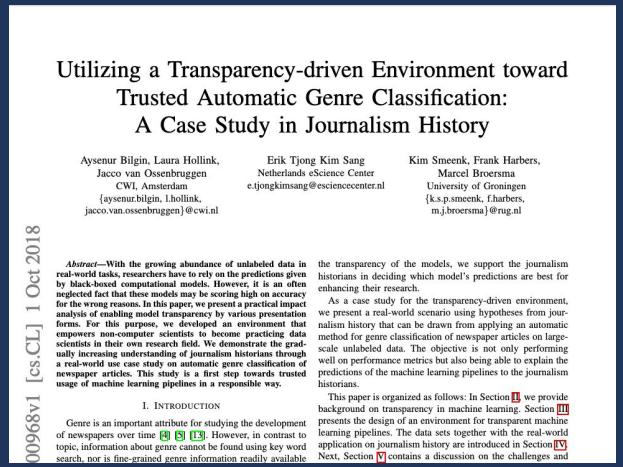
This paper is organized as follows: In Section II, we provide background on transparency in machine learning. Section III presents the design of an environment for transparent machine learning pipelines. The data sets together with the real-world application on journalism history are introduced in Section IV. Next, Section V contains a discussion on the challenges and



NEWSGAC project: Transparent Machine Learning Pipelines

Can we support humanities scholars using AI with transparency at every step in an AI pipeline?

- from data selection and data preparation, to choosing an algorithm selection, and inspection of the results



Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History

Aysenur Bilgin, Laura Hollink, Jacco van Ossenbruggen
CWI, Amsterdam
{aysenur.bilgin, l.hollink, jacco.van.ossenbruggen}@cwi.nl

Erik Tjong Kim Sang
Netherlands eScience Center
e.tjongkimsang@esciencecenter.nl

Kim Smeenk, Frank Harbers, Marcel Broersma
University of Groningen
{k.s.p.smeenk, f.harbers, m.j.broersma}@rug.nl

***Abstract*—With the growing abundance of unlabeled data in real-world tasks, researchers have to rely on the predictions given by black-boxed computational models. However, it is an often neglected fact that these models may be scoring high on accuracy for the wrong reasons. In this paper, we present a practical impact analysis of enabling model transparency by various presentation forms. For this purpose, we developed an environment that empowers non-computer scientists to become practicing data scientists in their own research field. We demonstrate the gradually increasing understanding of journalism historians through a real-world use case study on automatic genre classification of newspaper articles. This study is a first step towards trusted usage of machine learning pipelines in a responsible way.**

I. Introduction

Genre is an important attribute for studying the development of newspapers over time [4], [5] [13]. However, in contrast to topic, information about genre cannot be found using key word search, nor is fine-grained genre information readily available

the transparency of the models, we support the journalism historians in deciding which model's predictions are best for enhancing their research.

As a case study for the transparency-driven environment, we present a real-world scenario using hypotheses from journalism history that can be drawn from applying an automatic method for genre classification of newspaper articles on large-scale unlabeled data. The objective is not only performing well on performance metrics but also being able to explain the predictions of the machine learning pipelines to the journalism historians.

This paper is organized as follows: In Section II, we provide background on transparency in machine learning. Section III presents the design of an environment for transparent machine learning pipelines. The data sets together with the real-world application on journalism history are introduced in Section IV. Next, Section V contains a discussion on the challenges and

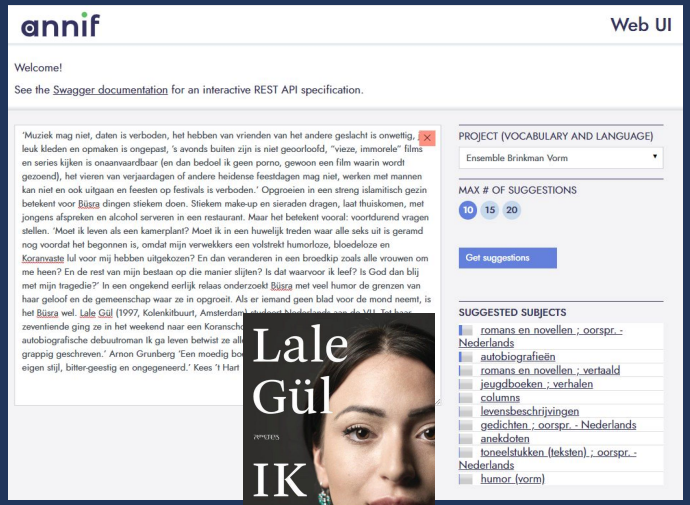
AI in the Library

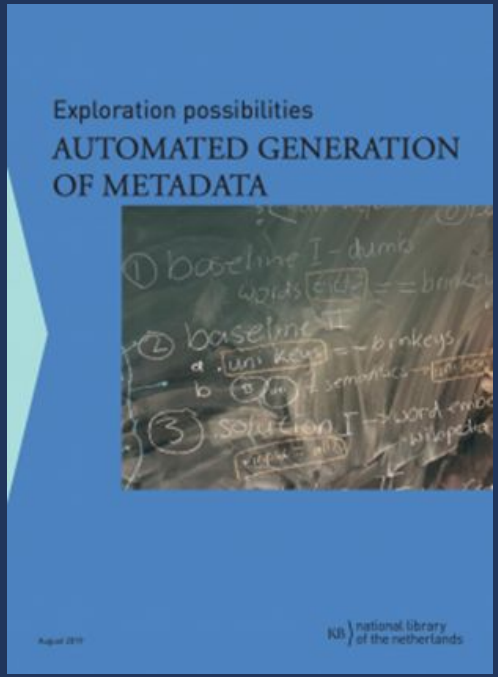
Martijn Kleppe
KB, national library of the Netherlands

KB nationale bibliotheek

AI in the Library

- AI to improve our internal processes





https://www.kb.nl/en/news/2019/kb-explores-artificial-intelligence-to-generate-metadata
https://zenodo.org/record/3375192
https://lab.kb.nl/tool/assisted-keyword-assignment-using-annif

AI in the Library

- AI to improve our internal processes
- AI to improve our services to our clients

https://www.onlinebibliotheek.nl/catalogus/424478382/een-vorm-van-verraad-lieneke-dijkzeul

AI in the Library

- AI to improve our internal processes
- AI to improve our services to our clients
- Libraries as platform to discuss AI & ethics

public spaces
internet for the common good

https://publicspaces.net/manifesto/
https://publicspaces.net/2021/04/30/verantwoorde-inzet-van-ai-is-meer-dan-regelgeving/
https://www.bibliotheeknetwerk.nl/digitaal-burgerschap

AI in the Library

- AI to improve our internal processes
- AI to improve our services to our clients
- Libraries as platform to discuss AI & ethics
- Libraries as means to contribute to development of AI

Cultural AI

"It is as much about using AI for understanding human culture as it is about using knowledge and expertise from the humanities ***to analyse and improve AI technology****."*

Cultural AI Case Studies
SABIO & ConConCor

Marieke van Erp
KNAW DH Lab

SABIO* - The SociAl BIas Observatory

Valentin Vogelmann
Researcher
KNAW HuC
DHLab

SABIO* - The SociAl BIas Observatory

dutch digital heritage network

AFRIKA MUSEUM

MUSEUM VOOR VOLKENKUNDE

WERELD MUSEUM

TROPEN MUSEUM

SUDOX

DHLAB

Modest, Wayne & Lelijveld, Robin (editors) 2018. *Words Matter*, Work in Progress I. National Museum of World Cultures

Cultural AI

Bias ~~Detection~~ Navigation

https://hdl.handle.net/20.500.11840/525228

Bias depends on: perspective, goals, context

Finding and seeing bias is **not** a problem to be solved (in the ML sense)

Navigating bias is a social action

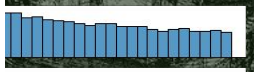
The process of navigation itself is informative about bias, both societal and individual

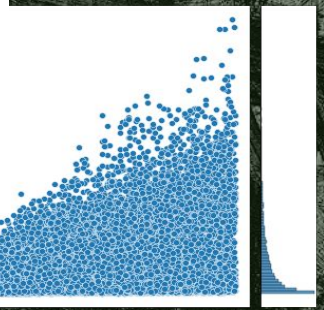
from the collection of the *Museum van Wereldculturen*, https://hdl.handle.net/20.500.11840/525228

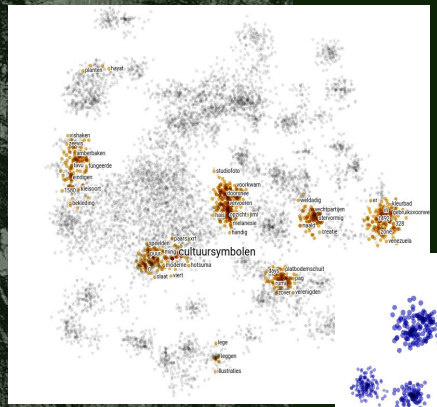
Cultural AI

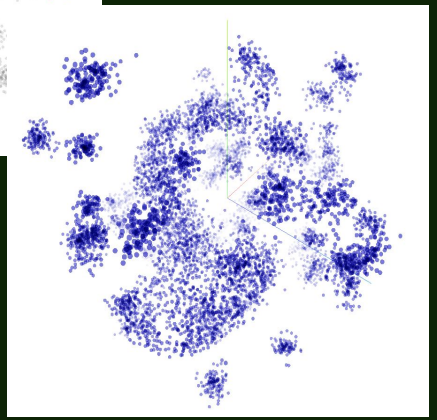
Avoiding the *Black Box*

w1	w2	PMI	f(w1)	f(w2)	f(w1w2)
nuoc	con	16.2251	6	7	6
naaldbomen	ceders	16.2251	6	7	6
dreiging	heerst	16.2251	7	6	6
rolle	berlijn	16.2251	5	7	5
gardens	xochimilco	16.2251	7	7	7
ibn	salam	16.2251	6	7	6
ernesto	castillo	16.2251	7	6	6
ditzelfde	collectienummer	16.2251	7	6	6
medemens	decennia	16.2251	4	7	4
chiang	mai	16.2251	5	7	5
ingekerfde	vierhoek	16.2251	7	4	4
jurfepin	litteken	16.2251	7	4	4
limburg	stirum	16.2251	7	7	7
goudmijn	placer	16.2251	7	4	4
spring	flowering	16.2251	7	6	6
pijnbomen	naaldbomen	16.2251	7	6	6
fantasielandschappen	voorstelden	16.2251	6	7	6
dasavatara	kaartspel	16.2251	7	7	7
floating	gardens	16.2251	7	7	7
harpoenspits	srepang	16.2251	7	7	7
oerang	oetang	16.2251	7	7	7
bersiap	period	16.2251	5	7	5
anatomical	identification	16.2251	6	7	6
con	nguoi	16.2251	7	6	6
srepang	masik	16.2251	7	7	7
expeditiefotografie	antropometrisch	16.2251	6	7	6
trin	chen	16.2251	7	7	7
pottenbakster	margarita	16.1845	6	6	5
padalarang	krawang	16.1845	6	6	5
extreme	care	16.1845	6	6	5









Giovanni Battista Piranesi 1745. *Carceri. Folder 7.* Kupferstichkabinett Dresden

Cultural AI

PMI Engine

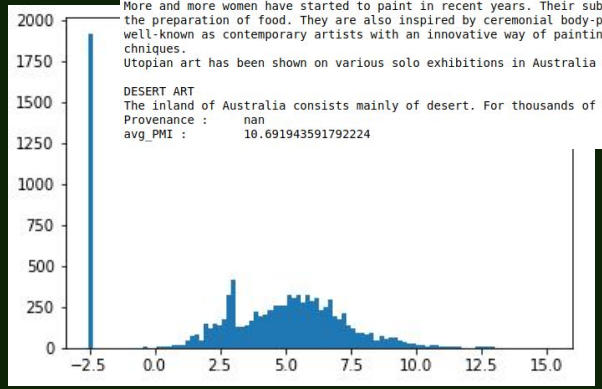
$$PMI('bias','social') = \log \frac{P('bias'|'social')}{P('bias')} = \log \frac{P('social','bias')}{P('social') * P('bias')}$$

	pair	PMI
0	een mensen	-1.353165
1	met mensen	-0.455868
2	mensen de	-0.211570
3	mensen </s>	0.357199
4	mensfiguren </s>	0.487596
...	...	...
389	thila bovenmenselijke	17.563169
390	royle namens	17.563169
391	menstruatieperiode voorbij	18.148131
392	mensenhoofden juwelensnoer	18.148131
393	mensenhanden geschapen	19.148131

			Title	Description	avg_PMI
			NaN	[VO] mensengezicht met haar en baard	3.137183
			NaN	[DE] dansend mensfiguur	3.515280
			NaN	Afbeelding van een grote groep negroïde mensen...	4.123839
			N DIERFIGUREN	Schuitvormige bak met rechthoekig uitgehold de...	4.452035
			ERK PIJPENKOP	De pijpenkop is vervaardigd van terracotta. Op...	4.493420
...	...		...	...	...
6801	125226	BRONZEN BEELD VAN DE DANSENDE SHIVA		Shiva Nataraja, de dansende Shiva.\r\n\r\nMet ...	8.714545
727	905210	' ABORIGINAL CEREMONIAL 'THE BORA TREE'		De Kerry & Co studio \r\nDe Kerry & Co studio ...	8.774003
3287	754671	NAKNOK MASKER		Naxnox mask; Tsimshian, probably Nishga; ca. 1...	9.608883
5446	1046747	zonder titel		The main theme in the paintings from Utopia is...	10.691944
8326	1101335	NaN		[VO] godheid & dharmapala & drag-gsed & yi-dam...	11.877365

```
ObjectID :        1046747
DepartmentID :    9
ClassificationID :        154
ObjectName :      schildering
Title :  zonder titel
Description :     The main theme in the paintings from Utopia is the strong bond that the Aborigines have with thei
r country. Men paint the " Songlines":  the roads that the mythological ancestors travelled when they gave the wor
ld her shape. Traditionally you will find in these paintings concentric circles and connecting lines so that the a
rtworks look like topographical maps.
Most of the paintings of Utopia consist of dots and circles. The men were the first to transfer to  the acrylic-ca
nvas medium. Their work is highly symbolic: representing ancestral history and nomadic routes. The paintings often
look like maps, seen from the sky. Only a person with considerable knowledge of the depicted locations and the dre
amings related to them can understand the symbols properly. A circle may represent a campsite, waterhole or fire.
An  arc represent  a person, whether man or woman will be explained by the symbols next to it.
More and more women have started to paint in recent years. Their subjects often have to do with the search for and
the preparation of food. They are also inspired by ceremonial body-paint designs.Many female artists are becoming
well-known as contemporary artists with an innovative way of painting. They experiment with colours, shapes and te
chniques.
Utopian art has been shown on various solo exhibitions in Australia as well as in Asia, Europe and the U.S.A.

DESERT ART
The inland of Australia consists mainly of desert. For thousands of years the Aborigines have wandered
Provenance :     nan
avg_PMI :         10.691943591792224
```



SABIO * | Wereldculturen

MY ARCHIVE 5 | BROWSE | ABOUT

QUERY

RESULTS 3,011 / 13,020,122 (0.02%, 15.6log%)

* INDICATOR SCORES

OBJECT METADATA

Keyword

Koloniaal

Start date: 1878/05/13

End date: 1966/01/08

Location: Indonesia

Object type:

INDICATOR

Engine: Word pairs | INFO

Minimal score: 35%

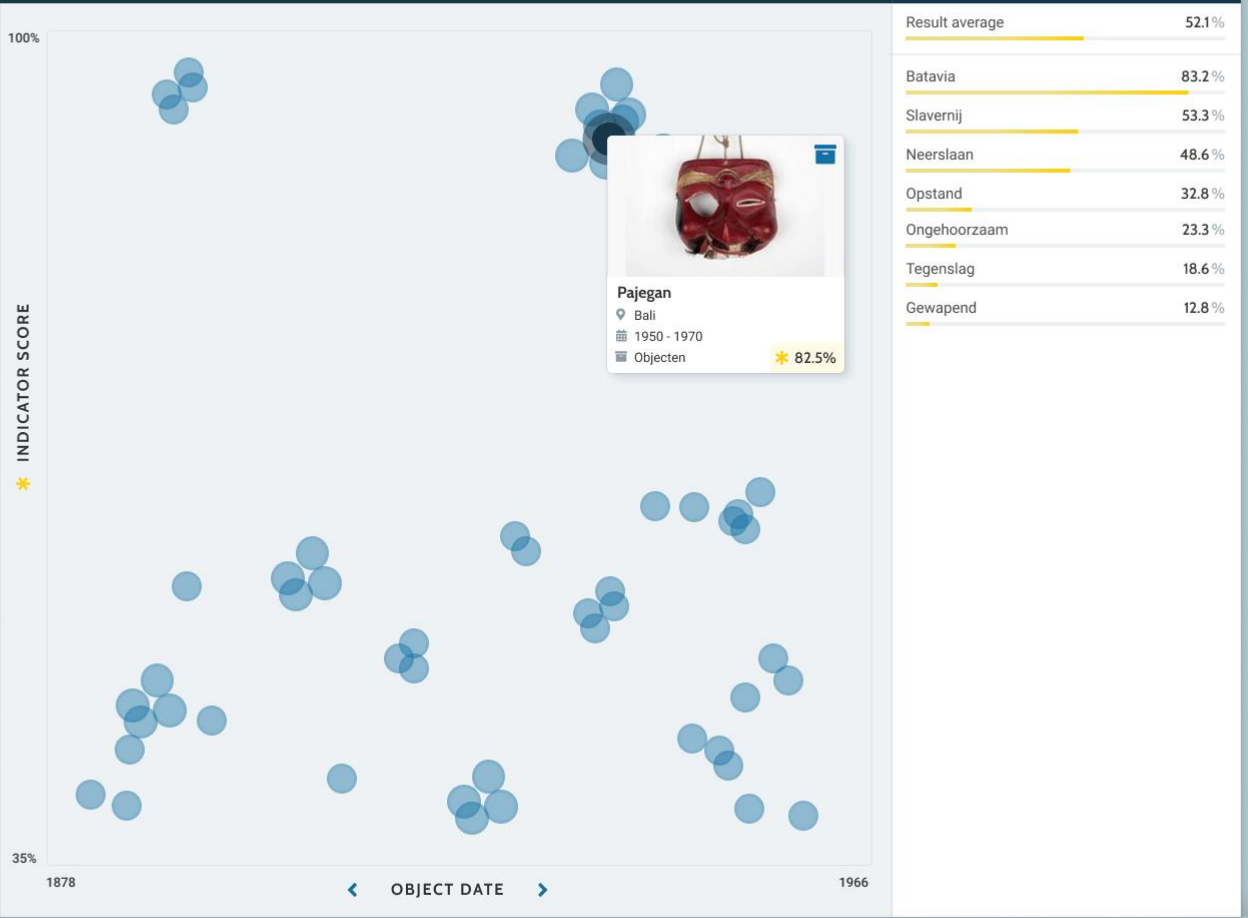
Maximal score: 100%

Scale: Linear

VOCABULARY

Terms | PRESETS

bediende, slaaf, Batavia, Mumbai, inboorling



* INDICATOR SCORES	
Result average	52.1 %
Batavia	83.2 %
Slavernij	53.3 %
Neerslaan	48.6 %
Opstand	32.8 %
Ongehoorzaam	23.3 %
Tegenslag	18.6 %
Gewapend	12.8 %

SABIO * | Wereldculturen

MY ARCHIVE 5 | BROWSE | ABOUT

QUERY | RESULTS 3,011 / 13,020,122 (0.02%, 15.6log%) | OBJECT DETAILS

V2

OBJECT METADATA

Keyword

Koloniaal

Start date: 1878/05/13

End date: 1966/01/08

Location: Indonesia

Object type:

INDICATOR

Engine: Word pairs | INFO

Minimal score: 35%

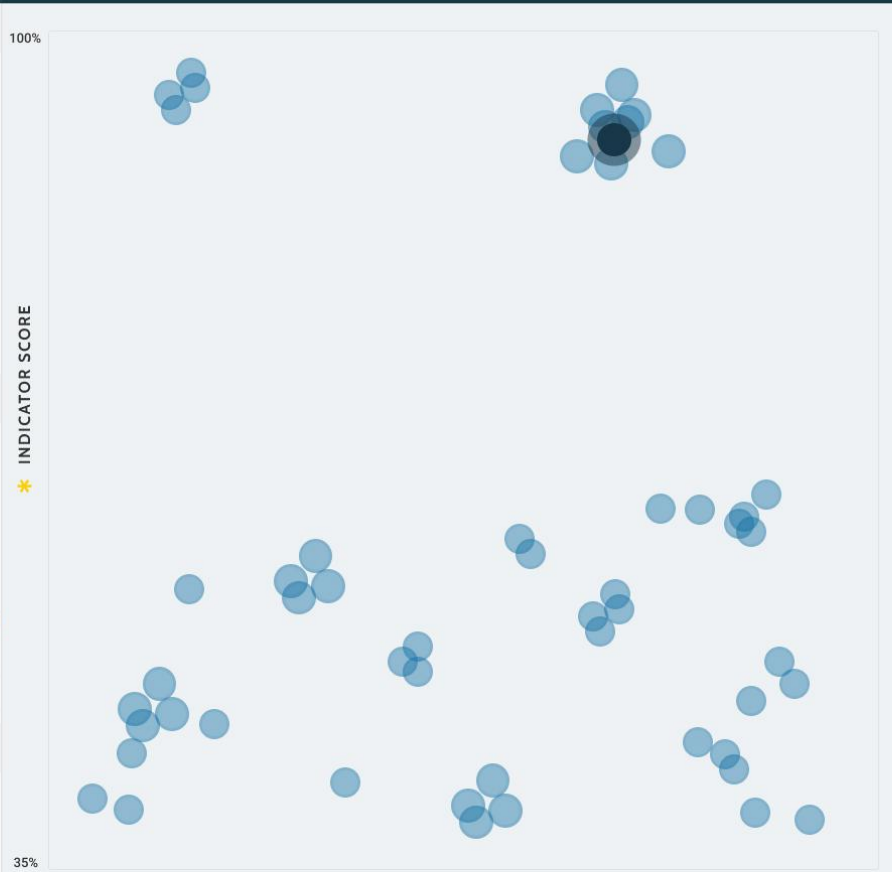
Maximal score: 100%

Scale: Linear

VOCABULARY

Terms | PRESETS

bediende, slaaf, Batavia, Mumbai, inboorling



Pajegan 82.5%

Bali | Objecten

1950 - 1970 | 7082-S-2295-1

Halfmasker, ziektemasker. Het stelt een gewone man voor, met het linkeroog toegeknepen, het rechter oog groot rond en grote wenkbrauwen. Hij heeft een grote haakneus met grote gaten, hazenlip aan de rechterkant, vier tanden en bolle wangen. De rechterwang is zwart en de voorhoofdsknobbel is versierd met bogen.

VIEW SOURCE

INDICATOR SCORES

Batavia	83.2%
Slavernij	53.3%
Weerstand	48.6%
Ongehoorzaam	12.8%

COMMENTS

Add comment...

JOSHUA 2021-08-05

This is biased because X and Y.

MICHELLE 2021-06-05

For me this is not biased, because it A and of course B.

V2

ConConCor: The Contentious terms in Context Corpus

EuropeanaTech Challenge for Europeana AI/ML Datasets

Andrei Nesterov
PhD Student

Laura Hollink
Research group leader
Centrum Wiskunde & Informatica
Human-Centered Data Analytics group

Valentin Vogelmann
Researcher

Ryan Brate
PhD student
KNAW Humanities Cluster
DHLab

Marieke van Erp
Research group leader

ConConCor: Background

Funding for:

- the creation of an annotated dataset of ‘contentious’ terms in Dutch Newspapers in Europeana;

- to bootstrap and evaluate (semi-)automatic methods for detecting such terms in cultural heritage collections.

ConConCor: Data Collection

- 83 words of both ‘contentious’ and ‘alternative’ type collated from words matter, and used to sample the Europeana/ KB collection (these 83 words match against 3.4M articles);

- For each of the 83 words, across 6 decades 1890-1941, 200 random samples are taken, collecting the metadata and OCR for approx 70K articles;

- Unigram & Bigram probabilities were calculated for the 70K articles, and for each word, decade and (1 of 6 newspaper circulations), 5-sentence extracts were sampled, weighted by ‘typicality'. Distilling to approx 3,000 extracts for annotation & presented to annotators.

ConConCor: The Task

For each **boldfaced term**

given the sentence it occurs in

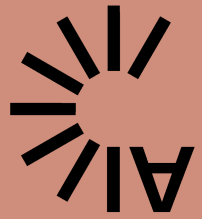
the sentence before it

and the sentence after it

Omstreden	Niet omstreden	Weet ik niet
☐	☐	☐

UIT DE ZUIDZEE. Koppensnellers van het eiland Villa Lavella in hun opgesmukte oorlogskano.

Would you say it is contentious?



ConConCor: Stats

~3,000 extracts over 63 unique google forms sheets, each sheet passed to ~7 participants

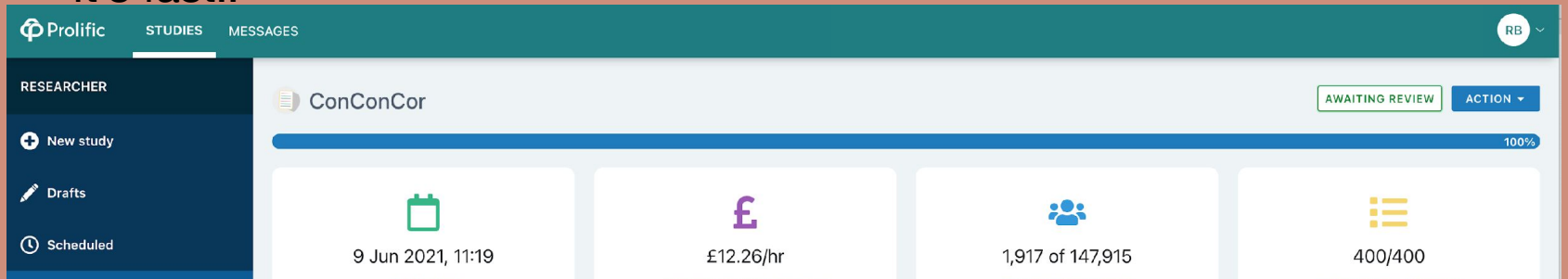
2 rounds:

- 10 participants from KNAW HuC ☞ used to finetune the instructions
- 399 participants on Prolific crowdsourcing platform

ConConCor: First observations

Prolific demographic information

- Far-flung nationalities: Korea, Pakistan, Belarus, India, Iran, Armenia, Turkey, Hungary, Greece, Lebanon
- youngest = 1 (apparently), oldest = 75
- It’s fast!!



ConConCor: Data Analysis (work in progress)

Inter-annotator agreement

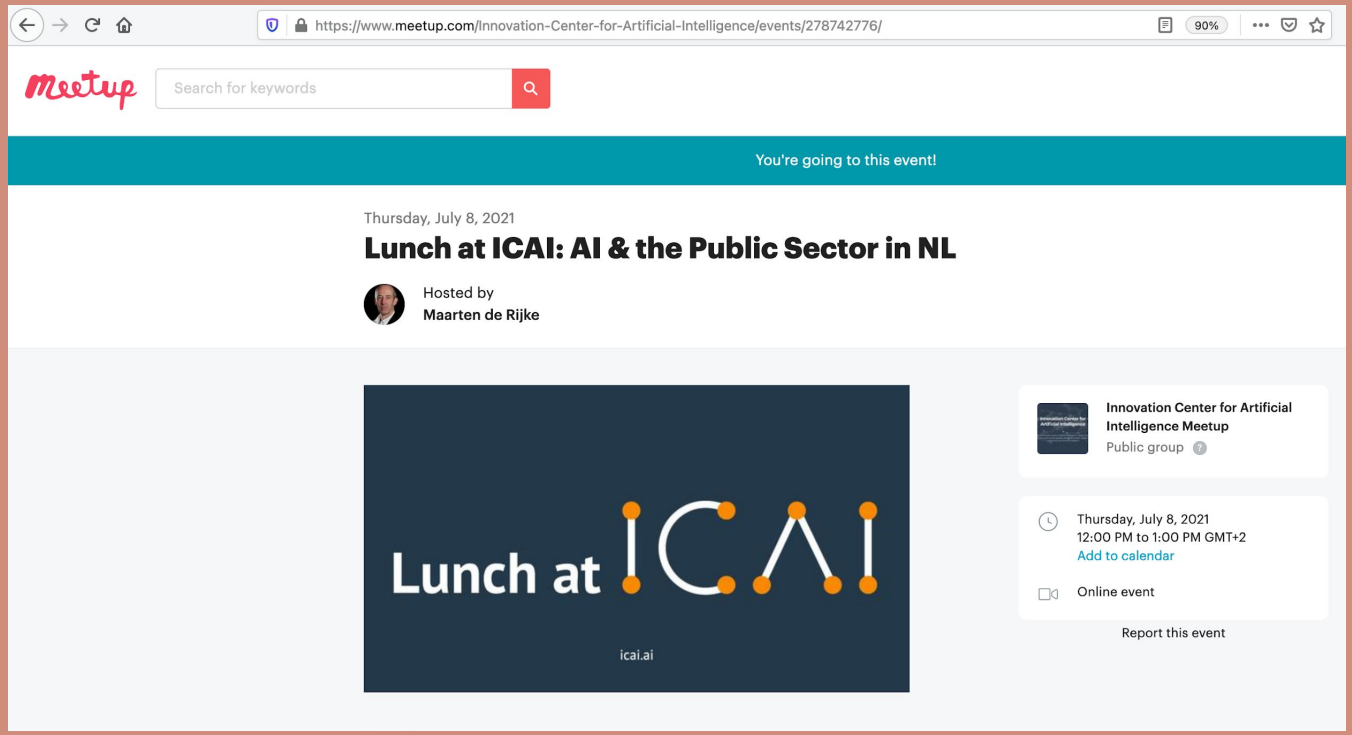
Most agreed on terms

Analysing contexts

ConConCor: Want to know more? Join the ICAI Lunch 8 July!

Andrei Nesterov
CWI

Jacco van Ossenbruggen
Vrije Universiteit Amsterdam



https://www.meetup.com/Innovation-Center-for-Artificial-Intelligence/events/278742776/

Wrap-up

What we're doing at the Cultural AI Lab

- Collaborate across 8 Dutch research and cultural heritage institutions
- Interdisciplinary teams
- Data & User-driven
- True 'All together now':

> "It is as much about **using AI** for understanding human culture as it is about using **knowledge and expertise from the humanities** to analyse and improve AI technology."

https://cultural-ai.nl

cultural_ai