Dr. Eva Seidlmayer

Towards machine learning driven quality control between censorship and “Good Research Practice”

EAHIL Workshop // Trondheim

Topic: 02. Power structures in our landscape

June 16, 2023

ZB MED Information Centre for Life Sciences

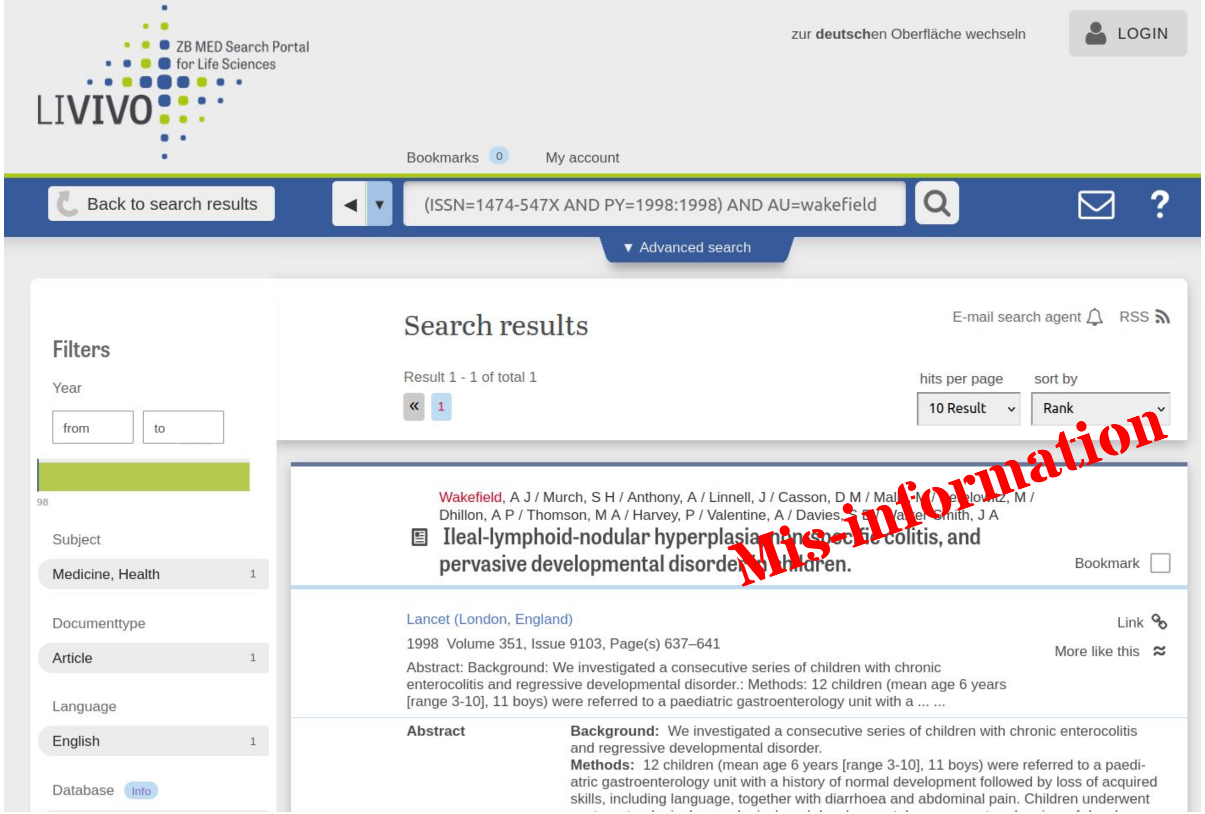
Agenda

- Goal: Enrichment of bibliographic quality information with regard to disinformation
- Machine learning approach and first results
- Further steps

Our goal

Status quo

- LIVIVO discovery system for Life Sciences
- >50 databases (MEDLINE, AGRICOLA, BASE…)
- Information on >80 Mio titles
- Scientific literature portals are affected by mis-information (*Holone 2016*)



Goal: Additional quality information for discovery system metadata

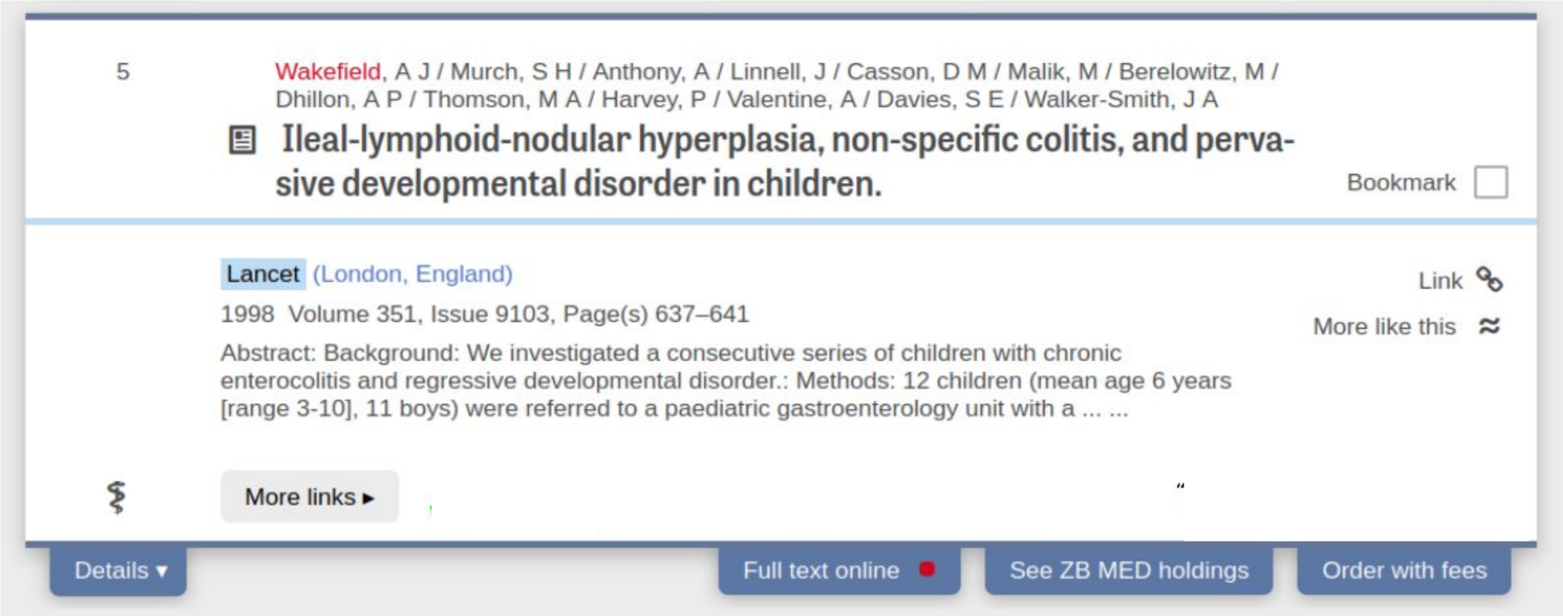
- Mis-information is widespread - also in (Life) Sciences: EU: Homeopathy *(EU 2021);* WHO: One of the ten greatest health hazards worldwide: Vaccination refusal (measles…) *(WHO 2019)*
- Data literacy is better than censorship
- Provision of additional information on:

  1. Metadata Compliance to good scientific practice

  2. Machine Learning: Assignment to machine learning classes

European Commission, Directorate-General for Communications Networks, Content and Technology (2021a). European Commission Guidance on Strengthening the Code of Practice on Disinformation. 52021DC0262 - EN - EUR-Lex (2023-06-13).
WHO (2019). online: Ten threats to global health in 2019 https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 (2023-06-13).

Goal: Additional quality information for discovery system metadata



Goal: Additional quality information for discovery system metadata

Wikipedia
The Free Encyclopedia

Search Wikipedia

Create account Log in

Andrew Wakefield

31 languages

Contents [hide]

Article Talk

Read View source View history Tools

From Wikipedia, the free encyclopedia

Andrew Jeremy Wakefield (born September 3, 1956)[3][4][a] is a British anti-vaccine activist, former physician, and discredited academic who was struck off the medical register for his involvement in *The Lancet* MMR autism fraud, a 1998 study that falsely claimed a link between the measles, mumps, and rubella (MMR) vaccine and autism. He has subsequently become known for anti-vaccination activism. Publicity around the 1998 study caused a sharp decline in vaccination uptake, leading to a number of outbreaks of measles around the world. He was a surgeon on the liver transplant programme at the Royal Free Hospital in London and became senior lecturer and honorary consultant in experimental gastroenterology at the Royal Free and University College School of Medicine. He resigned from his

Andrew Wakefield

Wakefield at an anti-vaccine rally in Poland, 2019

Born	Andrew Jeremy Wakefield September 3, 1956 (age 66) Eton, Berkshire, England

Goal: Additional quality information for discovery system metadata

Information Centre for Life Sciences
ZB MED

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia

Create account Log in •••

Andrew Wakefield

31 languages

Contents [hide]

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

Andrew Jeremy Wakefield (born September 3, 1956)[3][4][a] is a British anti-vaccine activist, former physician, and discredited academic who was struck off the medical register for his involvement in *The Lancet* MMR autism fraud, a 1998 study that falsely claimed a link between the measles, mumps, and rubella (MMR) vaccine and autism. He has

acti

dec

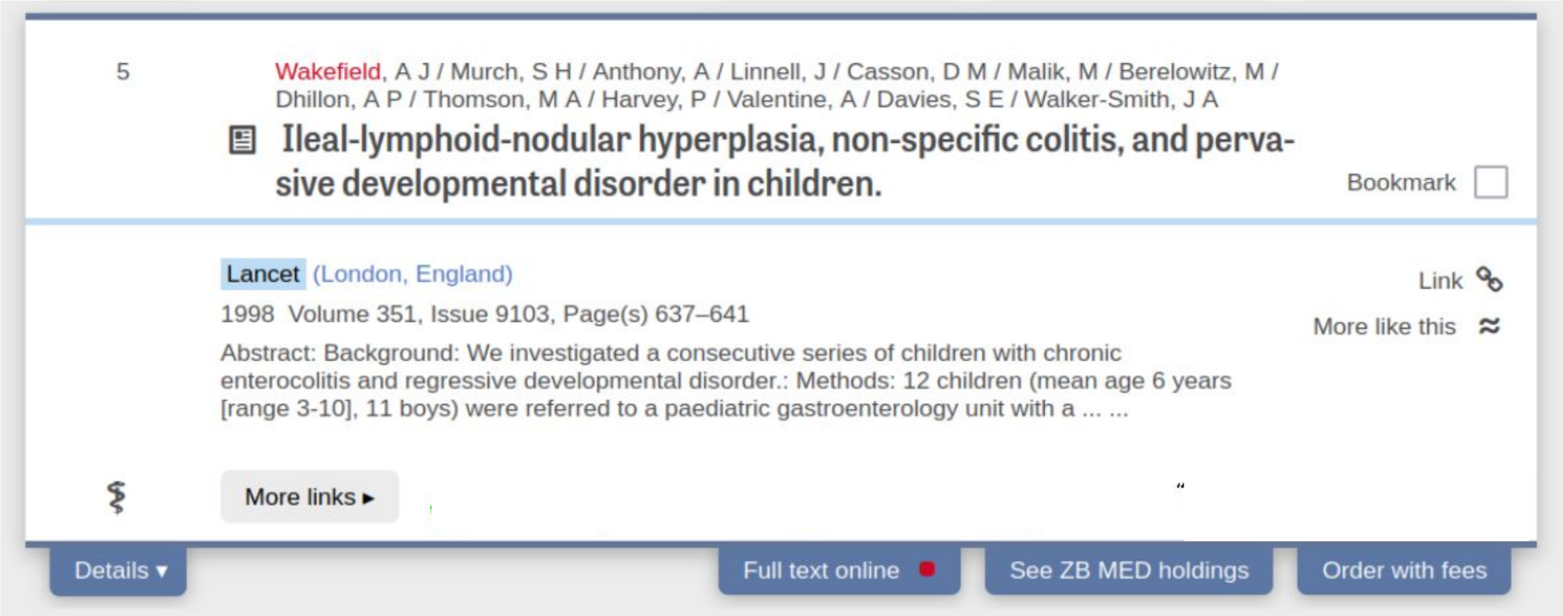
outbreaks of measles around the world. He was a surgeon on the liver transplant programme at the Royal Free Hospital in London and became senior lecturer and honorary consultant in experimental gastroenterology at the Royal Free and University College School of Medicine. He resigned from his

Andrew Wakefield

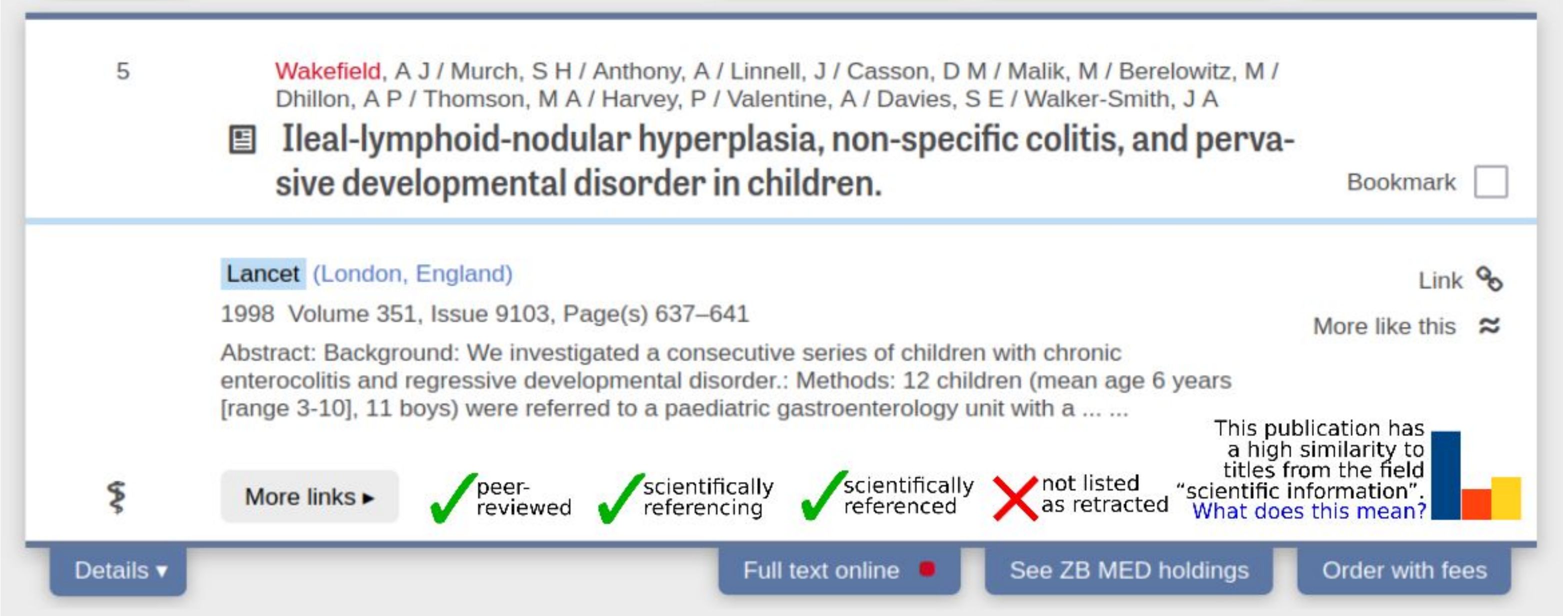
Wakefield at an anti-vaccine rally in Poland, 2019

Born	Andrew Jeremy Wakefield September 3, 1956 (age 66) Eton, Berkshire, England

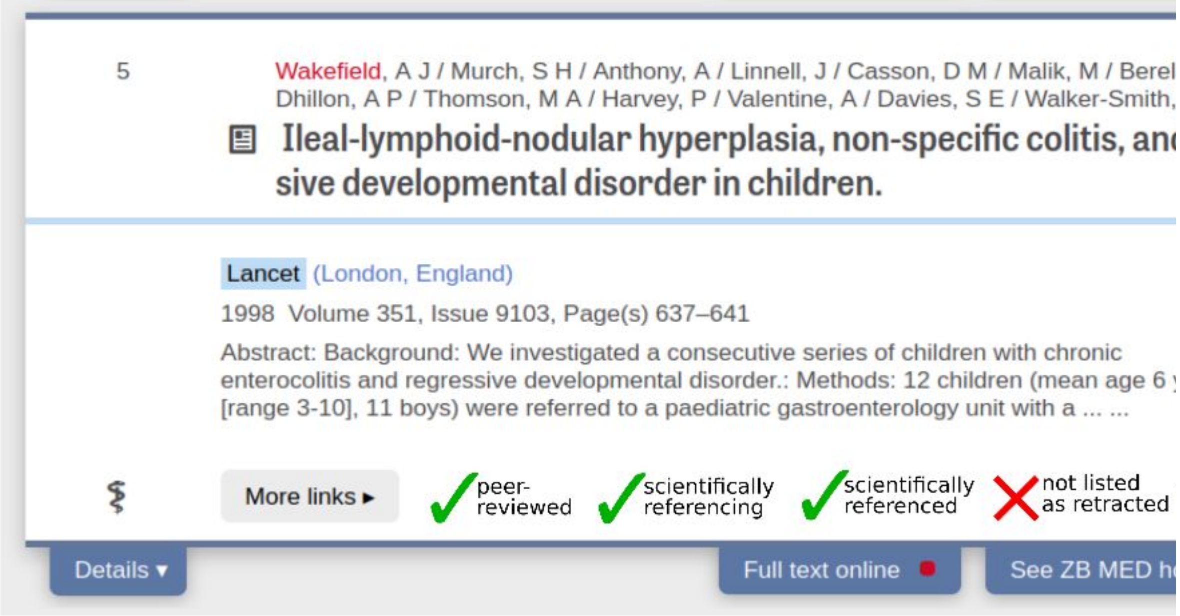
Goal: Additional quality information for discovery system metadata



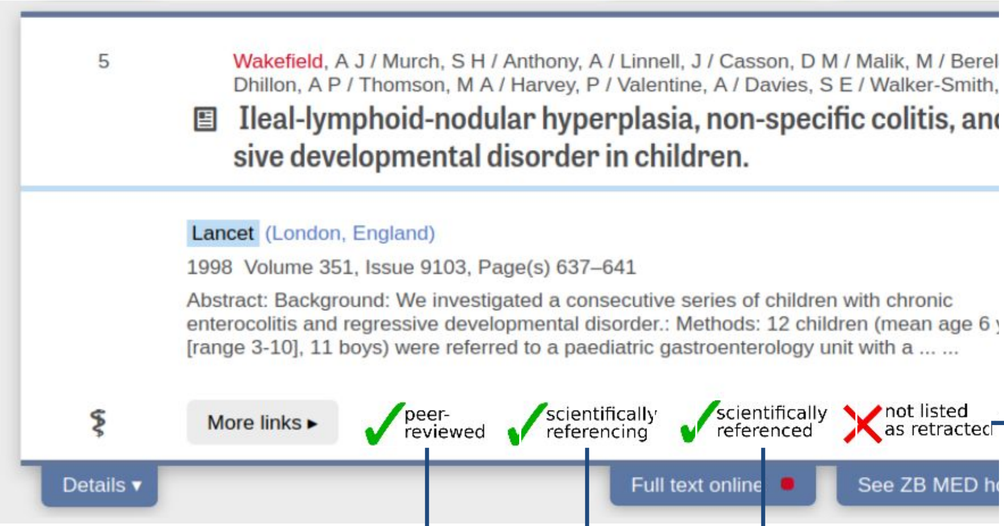
Goal: Additional quality information for discovery system metadata



Goal: Additional quality information for discovery system metadata



Goal: Additional quality information for discovery system metadata



BA-thesis Lucas Vetter:
>14k retracted articles
in ZB MED KE
not labeled as such

Retraction Watch

future
BA-thesis/
MA-project

OpenAlex

MEDLINE

future
BA-thesis/
MA-project

future
BA-thesis/
MA-project

OpenAlex

Crossref

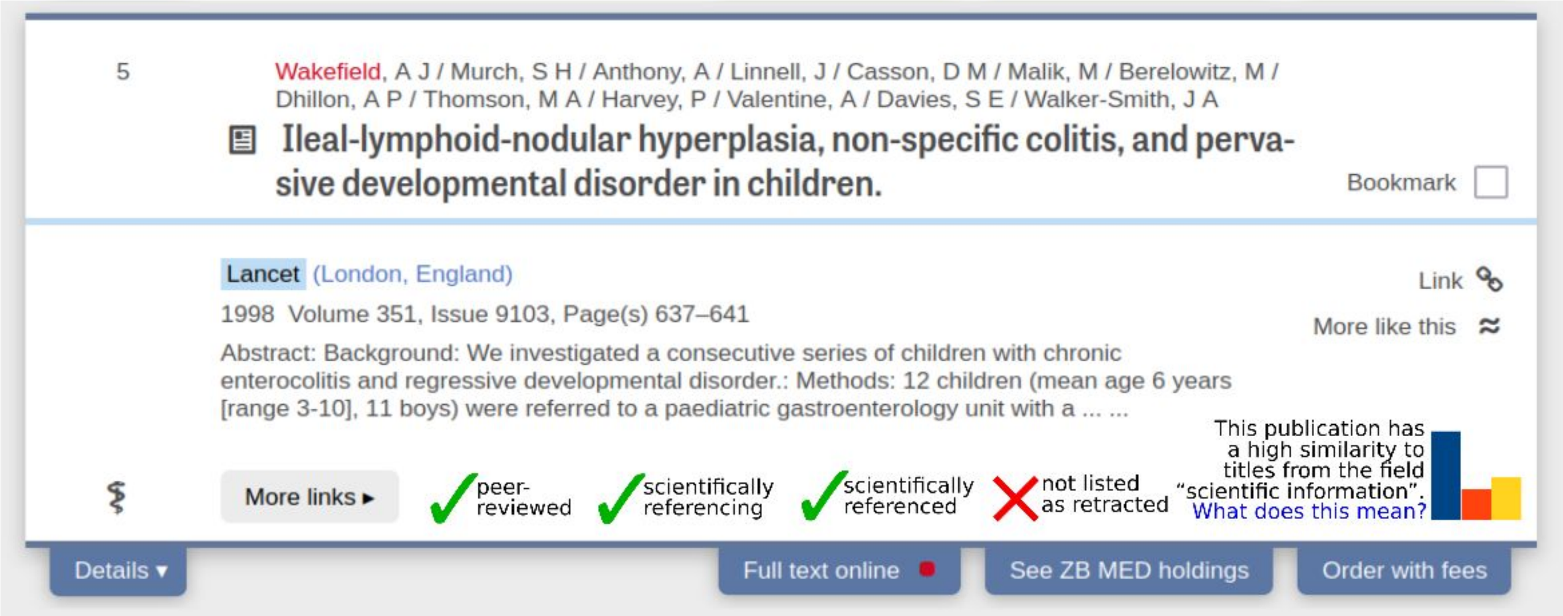
future
BA-thesis/
MA-project

OpenAlex

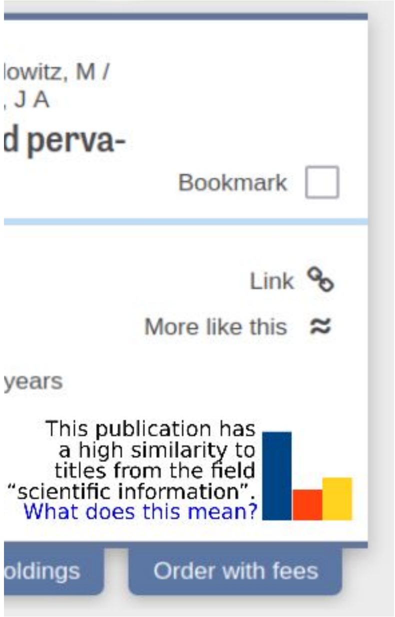
Crossref

Machine Learning Approach

ML-Approach

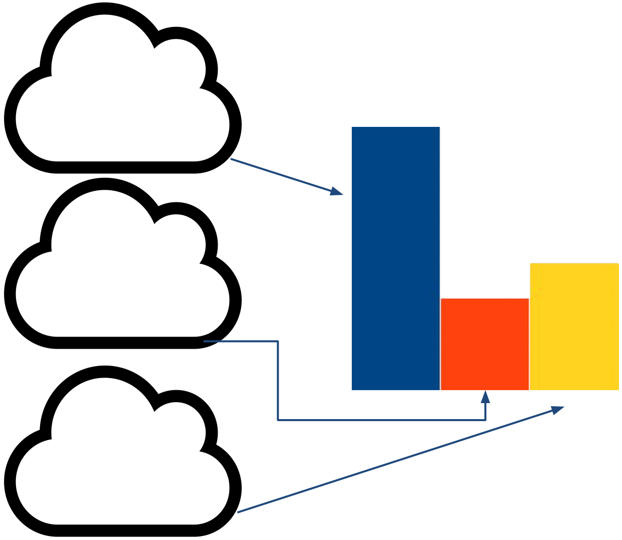

5
Wakefield, A J / Murch, S H / Anthony, A / Linnell, J / Casson, D M / Malik, M / Berelowitz, M / Dhillon, A P / Thomson, M A / Harvey, P / Valentine, A / Davies, S E / Walker-Smith, J A
Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.
Bookmark
Lancet (London, England)
1998 Volume 351, Issue 9103, Page(s) 637–641
Abstract: Background: We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.: Methods: 12 children (mean age 6 years [range 3-10], 11 boys) were referred to a paediatric gastroenterology unit with a ... ...
Link
More like this
More links ▸
peer-reviewed
scientifically referencing
scientifically referenced
not listed as retracted
This publication has a high similarity to titles from the field "scientific information".
What does this mean?
Details ▾
Full text online
See ZB MED holdings
Order with fees

Machine learning approach



ML approach: Classification and dataset for Life Sciences information

- Classification for publications in Life Science
- 3 classes:
  - Scientific texts
  - Popular science
  - Disinformation
- Disinformation: **Intentionally** spread false information that follows an other purpose than truth, such as profit or a political or religious agenda.



ML approach: Classification and dataset for Life Sciences information

- Classification for publications in Life Science
- 3 classes:
  - Scientific texts
  - Popular science
  - Disinformation
- Disinformation: **Intentionally** spread false information that follows an other purpose than truth, such as profit or a political or religious agenda.
- No English data set (full texts) available for Life Science available; PUBHEALTH (Kotonya/Toni 2020), Human Well Being (Singh/Deepak/Anoop 2020)
- Use of journalistic sources
- Compilation of our own data set

Kotonya, Neema und Francesca Toni (2020). Explainable Automated Fact-Checking for Public Health Claims. In: arXiv:2010.09926 [cs]. arXiv: 2010.09926.
Singh, Iknoor, P. Deepak und K. Anoop (2020). On the Coherence of Fake News Articles. In: ECML PKDD 2020 Workshops. Ed. by Irena Koprinska et al. Vol. 1323. Cham, p. 591–607.

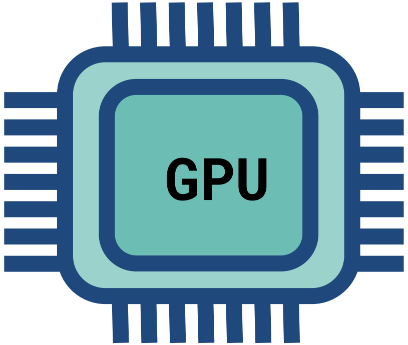
Machine learning approach: First explorative dataset

	English	
Scientific texts	PubMedCentral	309 items
Popular science	Wikipedia, MedlinePlus	309 items
Disinformation	Websites (mercola, signs-of-the-time, greenmedinfo, ….)	309 items

927 items

Machine learning approach: Computing Capacity

- de.NBI (German Network for Bioinformatics Infrastructure)
- Virtual Machine parameter:
  - Flavour: de.NBI GPU medium
  - total core: 14
  - total RAM: 64GB
  - total GPUs: 1
  - Storage Limit 500 GB



ML approach: BERT set up

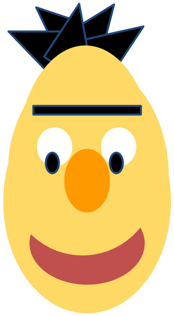
- Bidirectional Encoder Representations from Transformers: BERT-base-uncased, BioBert (Devlin et al. 2019)
- Finetuning
- 3 categories
- Supervised learning
- Split ratio: 80% Training, 20% Validation
- Limitation: BERT cannot deal with long texts: maximum position embedding is 512 tokens in BERT

Token: small unit ~ roughly a word.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

ML approach: BERT set up

- Bidirectional Encoder Representations from Transformers: BERT-base-uncased, BioBert (Devlin et al. 2019)
- Finetuning
- 3 categories
- Supervised learning
- Split ratio: 80% Training, 20% Validation
- Limitation: BERT cannot deal with long texts: maximum position embedding is 512 tokens in BERT



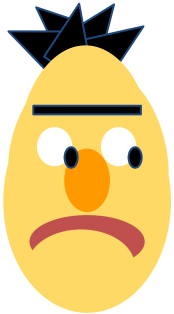
<= 512 tokens

Token: small unit ~ roughly a word.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

ML approach: BERT set up

- Bidirectional Encoder Representations from Transformers: BERT-base-uncased, BioBert (Devlin et al. 2019)
- Finetuning
- 3 categories
- Supervised learning
- Split ratio: 80% Training, 20% Validation
- Limitation: BERT cannot deal with long texts: maximum position embedding is 512 tokens in BERT



<= 512 tokens

> 512 tokens

Token: small unit ~ roughly a word.

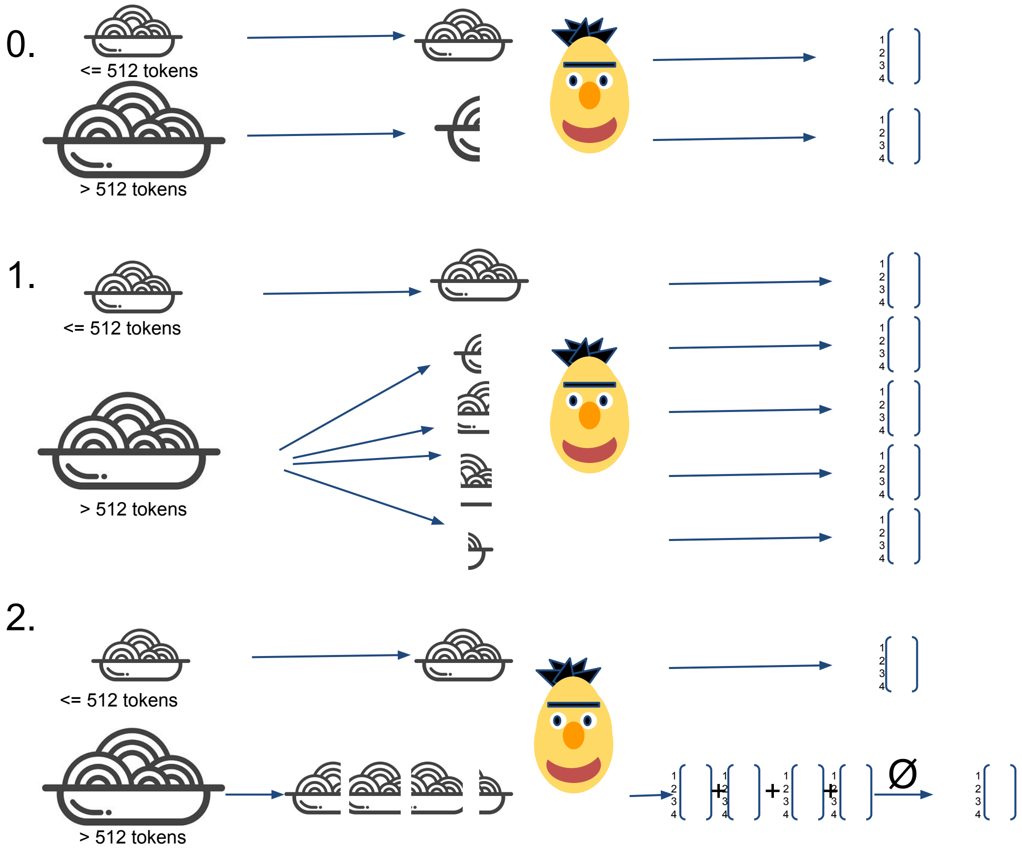
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

BERT model: Modification for long texts

0. Original BERT:
   - only first 512 tokens discard remainings

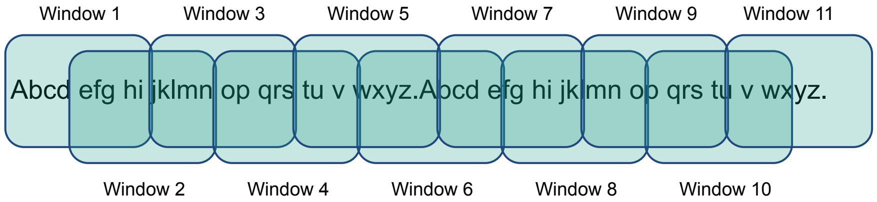
1. Cutting documents in parts
   - disadvantage: parts will be taken as whole documents

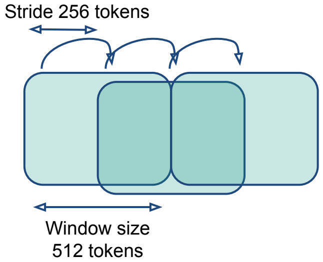
2. Sliding window
   - overlapping parts
   - mean of windows represent the document

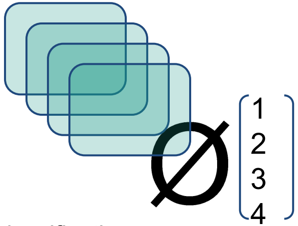


BERT model: Modification for long texts

- Sliding Window approach (Pappagari et al. 2019, Wang et al. 2018)
- Implemented to forward-function (starting line 1533) from Huggingface's BERT implementation (Transformers 2023)
- Window stride: 256 tokens
  Window size: 512 tokens
- Mean of multiple trained window vectors is taken for optimizer update

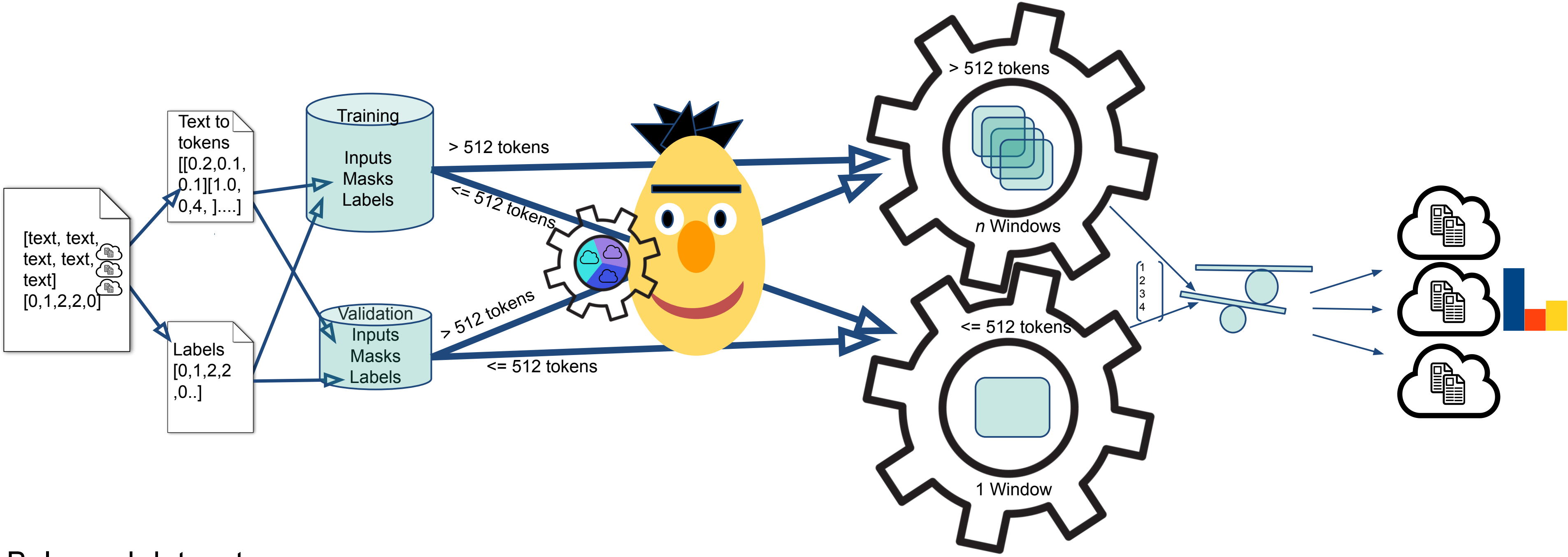






Raghavendra Pappagari, Piotr Żelasko, Jesus Villalba, Yishay Carmiel, and Najim Dehak (2019). Hierarchical Transformers for long document classification, https://arxiv.org/pdf/1910.10781.pdf.
Wei Wang, Ming Yan, and Chen Wu (2018.: Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714, Melbourne, Australia. Association for Computational Linguistics, doi: doi.org/10.18653/v1/P18-1158.
Transformers (2023): Code on Huggingface, https://github.com/huggingface/transformers/blob/v4.28.1/src/transformers/models/bert/modeling_bert.py#L1533

BERT model: Workflow



Balanced dataset (3 categories): list of texts, list of labels

Split ratio: 80:20

Model Initiation (BERT-base-uncased/BioBERT with 3 categories)

Training *n* epochs with *n* sliding windows and Update of optimizer

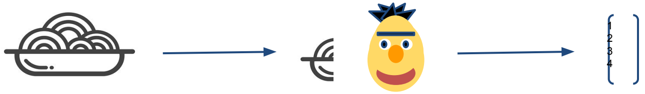
Statistical predicted class distribution

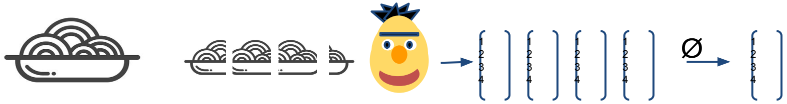
First results from machine learning: 512 tokens versus full texts

Bert-base-uncased, learning rate 3e-5

512 tokens	F1-score
2 epochs	0.9750
3 epochs	0.9765
4 epochs	0.9851
5 epochs	0.9851
6 epochs	0.9837

10k tokens	F1-score
2 epochs	0.6166
3 epochs	0.7254
4 epochs	0.9653
5 epochs	0.8254
6 epochs	0.9344





Precision: fraction of correct instances among the retrieved instances
Recall: fraction of complete group of relevant instances that had been retrieved
F1-score: F1-score= 2 * (precision * recall) / (precision + recall); highest possible value of an F-score is 1.0, indicating perfect precision and recall

First results from machine learning: BERT base versus BioBert

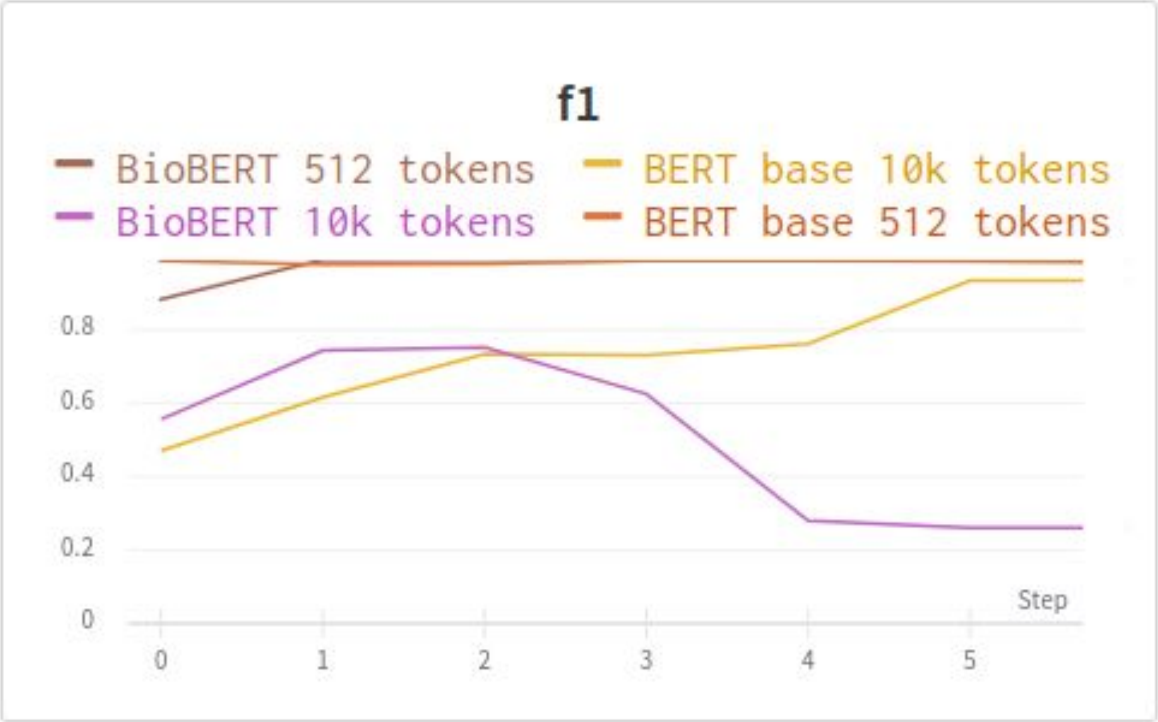
10k tokens, learning rate 3e-5

Bert base	F1-score
2 epochs	0.6166
3 epochs	0.7254
4 epochs	0.9653
5 epochs	0.8254
6 epochs	0.9344

BioBert	F1-score
2 epochs	0.7443
3 epochs	0.7518
4 epochs	0.6250
5 epochs	0.2797
6 epochs	0.2611

Precision: fraction of correct instances among the retrieved instances
Recall: fraction of complete group of relevant instances that had been retrieved
F1-score: F1-score= 2 * (precision * recall) / (precision + recall); highest possible value of an F-score is 1.0, indicating perfect precision and recall

First results from machine learning



Precision: fraction of correct instances among the retrieved instances
Recall: fraction of complete group of relevant instances that had been retrieved
F1-score: F1-score= 2 * (precision * recall) / (precision + recall); highest possible value of an F-score is 1.0, indicating perfect precision and recall

First results from machine learning: A closer look on categories

Bert-base-uncased, 10k tokens, learning rate 3e-5

6 epochs	precision	recall	F1-score
class scientific	1.00	0.97	0.98
class popular science	1.00	0.87	0.93
class disinformation	0.92	0.84	0.88

Precision: fraction of correct instances among the retrieved instances
Recall: fraction of complete group of relevant instances that had been retrieved
F1-score: F1-score= 2 * (precision * recall) / (precision + recall); highest possible value of an F-score is 1.0, indicating perfect precision and recall

First results from machine learning: Test classification

ZB MED Information Centre for Life Sciences

- Probabilities for unknown text
- PlosOne not source of data set yet
- BERT-base, 10k tokens, 4 epochs
- Estimated probabilities:
  - Scientific texts: 0.4197
  - Popular science: 0.1947
  - Disinformation: 0.4711


PLOS
Publish with PLOS | Research Communities | Open Science | Resources | About PLOS

PLOS BLOGS
EveryONE

About This Blog | About PLOS ONE | Browse all PLOS Blogs

World Malaria Day – A community effort to achieve ZERO

April 25, 2023 / Johannes Stortz / PLOS ONE Listicle

Image credit: Pixabay License

While tremendous progress has been made in fighting malaria, the disease still poses a significant threat to global human health. Especially in hard-to-reach remote and rural areas, fighting malaria remains a challenge. Therefore, this year's WHO World Malaria Day emphasizes the need for innovative strategies and measurements to combat malaria in the Western Pacific Region with the overall goal of eliminating the burden of malaria worldwide.

To emphasise the efforts made by the research community to achieve zero malaria, we are highlighting publications in *PLOS ONE* that strengthen our understanding of the disease and develop innovative strategies for controlling and eradicating malaria.

First results from machine learning: Test classification

- Probabilities for unknown text
- PlosOne not source of data set yet
- BERT-base, 10k tokens, 4 epochs
- Estimated probabilities:
  - Scientific texts: 0.4197
  - Popular science: 0.1947
  - Disinformation: 0.4711

PLOS

Publish with PLOS · Research Communities · Open Science · Resources · About PLOS

PLOS BLOGS

EveryONE

About This Blog · About PLOS ONE · Browse all PLOS Blogs

World Malaria Day – A community effort to achieve ZERO

April 25, 2023 / Johannes Stortz / PLOS ONE Listicle

Image credit: Pixabay License

While tremendous progress has been made in fighting malaria, the disease still poses a significant threat to global human health. Especially in hard-to-reach remote and rural areas, fighting malaria remains a challenge. Therefore, this year's WHO World Malaria Day emphasizes the need for innovative strategies and measurements to combat malaria in the Western Pacific Region with the overall goal of eliminating the burden of malaria worldwide.

To emphasise the efforts made by the research community to achieve zero malaria, we are highlighting publications in *PLOS ONE* that strengthen our understanding of the disease and develop innovative strategies for controlling and eradicating malaria.

Results and further steps

Results and further steps

- Summary of results:
  - Workflow is running properly.
  - First experiments seem to be promising - 512 tokens versus full text (> 512 tokens) need more exploration
  - Training data need to be improved especially with more and more diverse data sources
- Further steps:
  - Additional workflow for detection of bot created content
- Question for you:
  - “Disinformation” class - rewording to “not-scientific information”?
  - Just two categories “scientific text” and “non scientific text”?

References:

Chartered Institute of Library Information Professionals (CILIP) (2018). Definitions of Information Literacy 2018, https://infolit.org.uk/ILdefinitionCILIP2018.pdf (2022-07-16).

Crossref: online: https://www.crossref.org/ (2023-05-17).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Commission, Directorate-General for Communications Networks, Content and Technology (2021a). European Commission Guidance on Strengthening the Code of Practice on Disinformation. 52021DC0262 - EN - EUR-Lex (2023-06-13).

Holone, Harald (2016). The filter bubble and its effect on online personal health information. In: Croatian Medical Journal 57.3. doi: 10.3325/cmj.2016.57.298, p. 298–301.

Kotonya, Neema und Francesca Toni (Okt. 2020). Explainable Automated Fact-Checking for Public Health Claims. In: arXiv:2010.09926 [cs]. arXiv: 2010.09926.

LIVIVO, online: www.livivo.de (2023-06-13).

Open Alex, online: https://openalex.org/ (2023-06-13).

Oshikawa, Ray, Jing Qian und William Yang Wang (2020). A Survey on Natural Language Processing for Fake News Detection. In: arXiv:1811.00770 [cs]. arXiv: 1811.00770.

Raghavendra Pappagari, Piotr Zelasko, Jesus Villalba, Yishay Carmiel, and Najim Dehak (2019.: Hierarchical Transformers for long document classification, https://arxiv.org/pdf/1910.10781.pdf.

Retraction Watch, online: https://retractionwatch.com/ (2023-06-13).

Singh, Iknoor, P. Deepak und K. Anoop (2020). On the Coherence of Fake News Articles. In: ECML PKDD 2020 Workshops. Ed. by Irena Koprinska et al. Vol. 1323. Cham, p. 591–607.

Transformers (2023). Code on Huggingface: https://github.com/huggingface/transformers/blob/v4.28.1/src/transformers/models/bert/modeling_bert.py#L1533

Wei Wang, Ming Yan, and Chen Wu (2018). Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714, Melbourne, Australia. Association for Computational Linguistics, doi: doi.org/10.18653/v1/P18-1158.

WHO (2019). online: Ten threats to global health in 2019 https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 (2023-06-13).

ZB MED, online: www.zbmed.de (2023-05-17).

Thanks…

to:
Prof. Dr. Konrad Förstner, ZB MED, Cologne/Germany
Dr. Lukas Galke, Max Planck Institute for Psycholinguistics, Nijmegen/Netherlands
Dr. des. Lisa Kühnel, ZB MED, Bonn/Germany
My unit *Data Science and Services*

Eva Seidlmayer, Dr. phil., M.LIS
Data Sciences and Services, Research Fellow
ORCID: 0000-0001-7258-0532
Twitter: @kivilih
Mastodon: @eta_kivilih

ZB MED - Information Centre for Life Sciences
Gleueler Straße 60
50931 Cologne
Germany

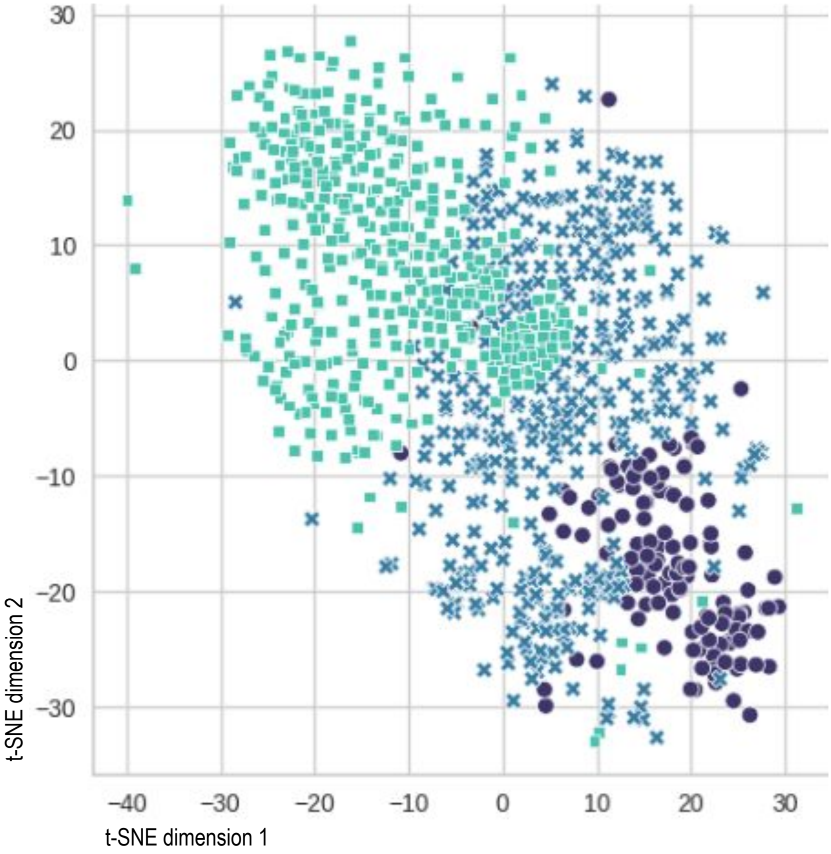
seidlmayer@zbmed.de
http://www.zbmed.de/

INFORMATION. KNOWLEDGE. LIFE.

ML-approach

Unsupervised clustering of the German test data set with Doc2Vec (t-SNE projection)

- Specialized texts
- Popular-science texts
- Mis-informative texts



Project goal

ML-Approach: Results

Result representation

- Display of all estimated probabilities not normalized over all classes

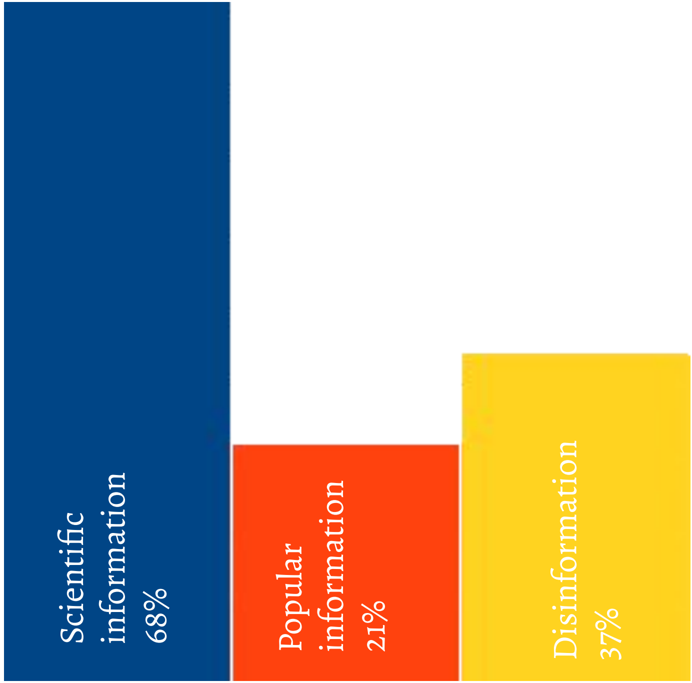
  -> multi-label classification

  vs. single-label classification

  neutral: low percentage values

- Explainability of the workflow *(EU 2021b)*

This publication has a high probability to titles from the field "scientific information".
What does this mean?



ML-Approach: Data base schema

id	category
1	scientific
2	popular science
3	disinformation

text-id	id	text
10.3390/jcm11071855	1	Predictive Markers for Immune Checkpoint Inhibitors in Non-Small Cell Lung Cancer….
10.3390/jcm11071964	1	Predictors Associated with Adverse Pregnancy Outcomes in a Cohort of Women with Systematic Lupus Ery….
Biodiversity	2	Biodiversity or biological diversity is the variety and variability….
Ecosystem	2	An ecosystem (or ecological system) consists of all the organisms…
sott-13	3	Let's consider the claim that *Covid*-19 vaccines can alter our DNA….