



EAHIL Workshop // Trondheim
Topic: 02. Power structures in our landscape

June 16, 2023



thank you very much for introducing me.
thank you very much for staying here for the last – and very interesting– talk of
the confernece

I like to present and discuss with our approach to enriching our data with
additional quality information with special regard to disinformation.

Agenda

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

here you see the agenda for my presentation.

first I will shortly introduce you to the status quo at my institute . ZB MED is a information centre for life science in Cologne/Germany.

and then I will dive deep into the progress of our machine learning approach and present the first results

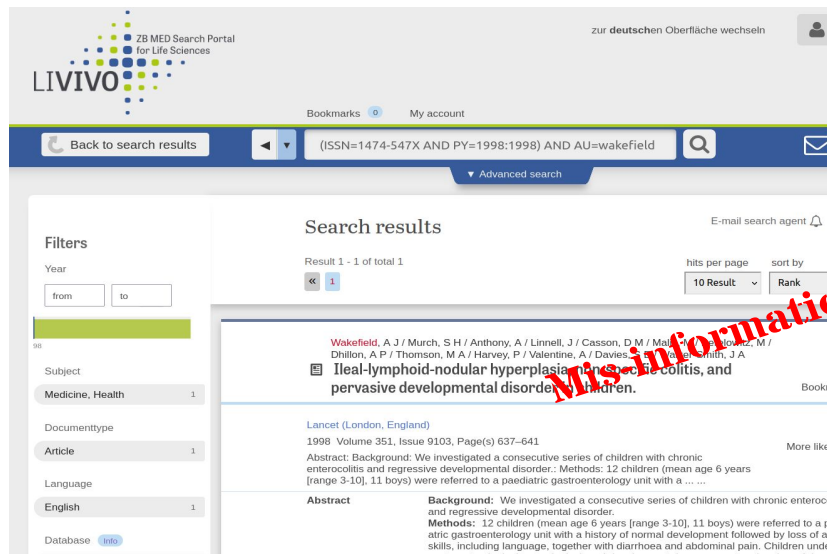
and lastly, I will come to future steps

and here I have some questions for you with regard to the classification and the data set used for the machine learning.

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*)



Here you see the discovery system website of my institut. the discover system is called LIVIVO. it aggregates many key data bases from the life science field and beyond.

LIVIVO takes all the data from those original sources with no filters and without quality control, which results on the one hand in A LOT of data (which is good) but on the other hand this comes at the cost

that also some record of mis / mal/ disinformation which may be included to our metadata

Instead of removing those items from our provided material we opt for labelling the items so users can make an informed decision of whether or not using the item.

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).

And this is exactly what we aim for. Namely to provide additional quality information to our library users

with regard to the increasing threat of disinformation also in science.

By doing this, we like to improve data literacy of our users, instead of patronise them by telling what to read or not.

for providing additional information: we follow a double approach.

first we will give additional information by metadata enrichment

and secondly we work for classifying articles by an machine learning model to three classes.

and here we do some experiments right now.

Goal: Additional quality information for discovery system metadata

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) Link
1998 Volume 351, Issue 9103, Page(s) 637–641 More like this
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

More links ►

Details ▾ Full text online See ZB MED holdings Order with fees

this is the display of one item how it looks

and maybe you know this person andrew wakefield.

//

He was involved in an academic fraud claiming a relation between measles vaccination and autism.

//

The LANCET the journal where the study was published retracted the article later.

Eva Seidlmayer: Towards machine learning driven quality control 06/27/2022 Page 7

//

//

The LANCET the journal where the study was published retracted the article later.

Eva Seidlmayer: Towards machine learning driven quality control 06/27/2022 Page 8

//


//

The LANCET the journal where the study was published retracted the article later.

Goal: Additional quality information for discovery system metadata

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 ▶

“

Details ▼

Full text online 

See ZB MED holdings

Order with fees

so back to the display. right now there is none information on this circumstances regarding this specific study.

Goal: Additional quality information for discovery system metadata

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) Link

1998 Volume 351, Issue 9103, Page(s) 637–641 More like this

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 ...

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

and that is what we strive for instead of how the display with additional information could look like.

So this means this document is **peer reviewed**, it **cites** scientific literature and it is **cited** by **other** papers.

BUT it is listed as retracted.

which is indicated by the red cross

And on the right hand side, we would like to see some kind of graph showing the similarity to three classes detected by our machine learning model.

Goal: Additional quality information for discovery system metadata

5 **Wakefield, A J / Murch, S H / Anthony, A / Linnell, J / Casson, D M / Malik, M / Berel Dhillon, A P / Thomson, M A / Harvey, P / Valentine, A / Davies, S E / Walker-Smith, I**

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and severe developmental disorder in children.

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 ...

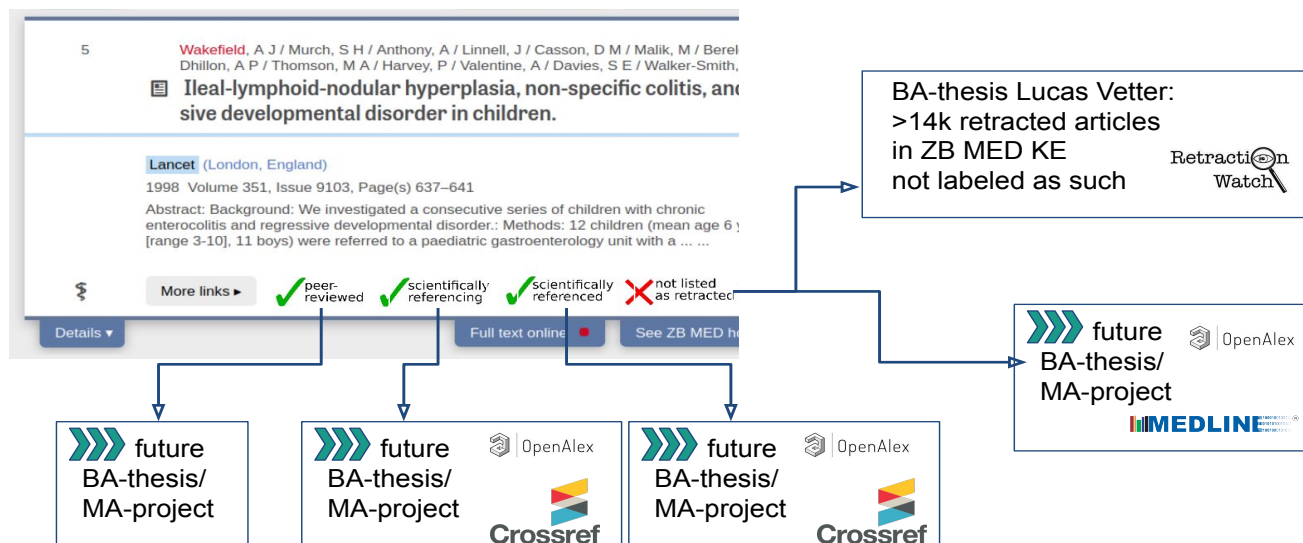
More links ▶

peer-reviewed scientifically referenced scientifically referenced not listed as retracted

Details ▾ Full text online See ZB MED

so again: the four additional information on the left hand side can be provided quite simple by metadata from other data bases:

Goal: Additional quality information for discovery system metadata



These information are conducted by students.

since the project is only running since 6 month right now,
 we had only a first student project on the retraction topic.

And here our student Lucas found out that more than 14k articles provided in LIVIVO are retracted – and not labeled so far.

Lucas used “Retraction watch database”. However, we plan to retrieve the information also from other sources: Pubmed and open Alex also provide meta data on retractions.

also the status whether a text has been reviewed by peers can be found as metadata in other databases. the peer review status is important with regard to the many preprints we provide in our discovery system.

secondly, information, if the text references actively scientific literature and as well if the text itself is scientifically referenced by others in an passive manner. can be checked in other databases.

Machine Learning Approach

Okay. So we come to the main topic I like to talk about. the machine learning approach.
what we are up to is this...

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)
Link

1998 Volume 351, Issue 9103, Page(s) 637–641
More like this

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 ...

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

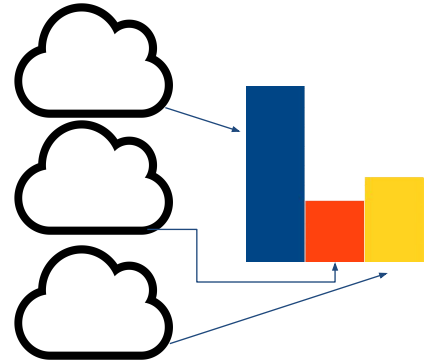
click!



what you see on the right hand side. there is a bar chart indicating the probability of a text belonging to three categories.

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.



Before I will further describe the machine learning process, we need to have a look on the classification we want to support by our machine learning model. and furthermore, we also need to talk shortly about the data set we use for training.

As classification we defined three classes for publications from life science area.:

- 1) scientific information, which is publications from established scientific sources.**
- 2) popular science information, which are easy described scientific topics for common understanding**
- 3) disinformation. which we define as follows:
it is Intentionally spread false information that follows an other purpose than truth. such as profit or a political or religious agenda.**

The success of the machine learning technique rises and falls with the data basis we use for training the model

We know of two data sets for fake news in the life sciences field.

Unfortunately we cannot reuse their data as it is not published properly.

Also they dont use full text, only statements.

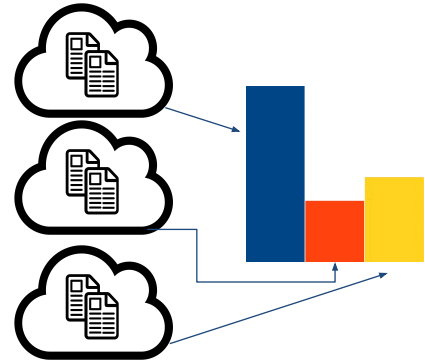
However we can learn from their approach

both data sets use mostly **journalistic** sources and websites. - and this why we feel

encouraged also to use websites as disinformation source for the dataset

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

Eva Seidmayer: Towards machine learning driven quality control

06/27/2022 | Page 18

Before I will further describe the **machine learning process**, we need to have a look on the **classification** we want to support by our machine learning model. and furthermore, we also need to talk shortly about the **data set** we use for training.

As classification we defined three classes for publications from life science area.:

1) scientific information, which is publications from established scientific sources.

2) popular science information, which are easy described scientific topics for common understanding

3) disinformation. which we define as follows:
disinformation is **Intentionally** spread false information that follows an other purpose than truth. such as profit or a political or religious agenda.

The success of the machine learning technique rises and falls with the data basis we use for training the model

We know of two data sets for fake news in the life sciences field.

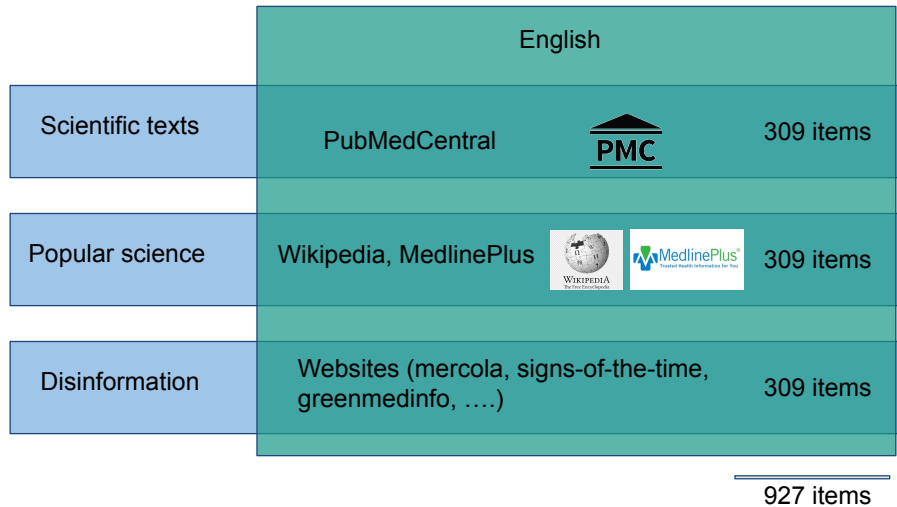
Unfortunately we cannot reuse their data as it is not published properly.

Also they dont use full text, only statements.

However we can learn from their approach

both data sets use mostly journalistic sources and websites. - and this why we feel encouraged also to use websites as disinformation source for the dataset

Machine learning approach: First explorative dataset



here you see how the data set is compiled right now. Currently we only have an first dataset in English for exploration .

it consists of 3 times x 309 items. which is a set of 927 items in total – in this early experimental stage.

Also an equivalent German data set is supposed to be compiled later.

One risk is that the ML model learns formal characteristics of a text genre and mistakes them for content characteristics.

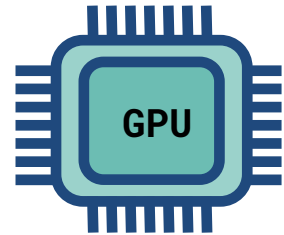
so for example an article or a website have other formal structures than monographs.

This is why we need a careful data pre-processing and data cleaning in the beginning. But we also need to include different data sources.

And this is something which is really need to be improved in the next steps.

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



and this is the computing capacity of our Virtual machine at de.NBI - German Network for Bioinformatics Infrastructure

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.

As machine learning model, we use a pretrained Bidirectional Encoder Representations from Transformers short: BERT model.

BERT is pretrained: this allows us, to only finetune the model with a comparable small data set.

Otherwise we would need ten thousands of text items in order to build a model from scratch

we do supervised learning with a split ratio of 80-20 percentage

BUT: there is a crucial limitation of BERT. BERT cannot process texts with more than 512 tokens.

when a text is longer than 512 tokens the memory consumption will just cut of the rest.

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



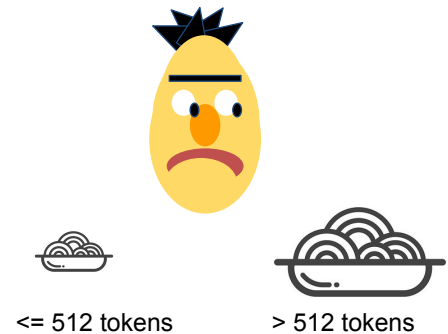
≤ 512 tokens

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



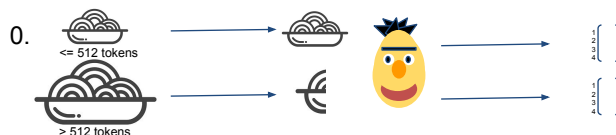
Token: small unit ~ roughly a word.

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

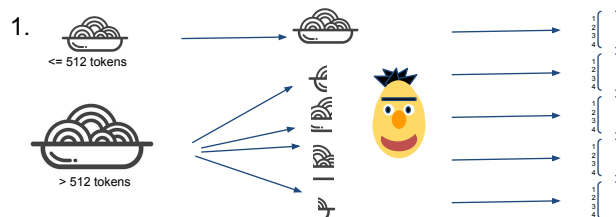
0. Original BERT:

- only first 512 tokens discard remainings



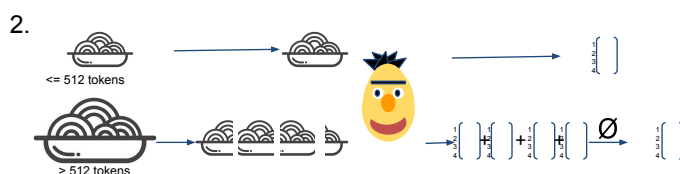
1. Cutting documents in parts

- disadvantage: parts will be taken as whole documents



2. Sliding window

- overlapping parts
- mean of windows represent the document



first you see the original bert model. if you feed it a long document it will just process the first 512 tokens.

there are two ways we thought of dealing with the problem in order to use full texts.

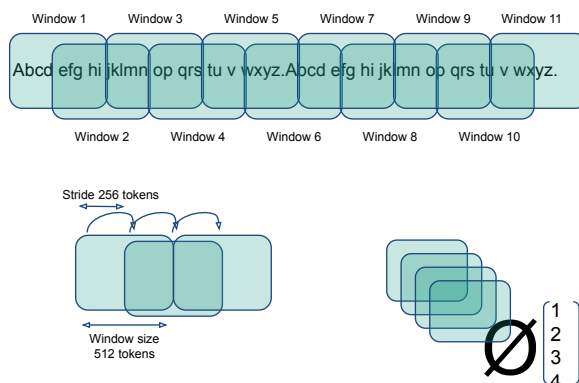
first, just cutting the documents into parts and take all the parts as single documents. this is really misleading for our task, as for example a bibliography which is an essential part of a scientific text, would be considered as a single document in this solution.

This is why we discard this solution possibility.

Instead, we opt for a second solution: the sliding window approach. here we create overlapping parts of the document, run BERT on it and take the mean of the all the vectors to update the Model optimizer.

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 Zelasko, 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

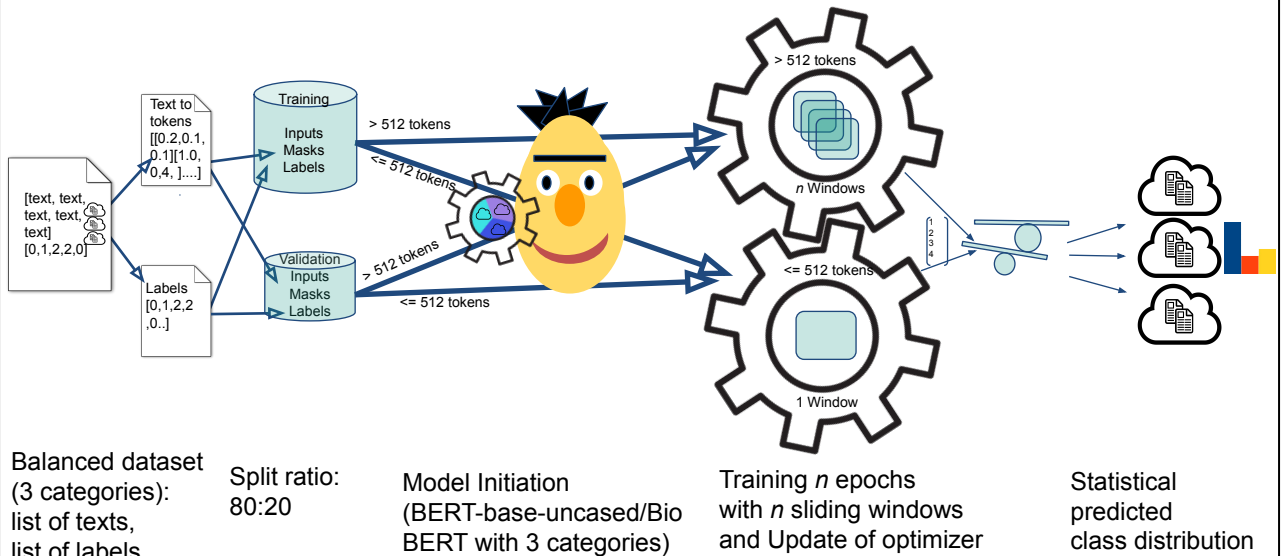
we implemented the sliding window to the forward function of the original code.

The window size is 512 tokens and we use a stride of half of a window size in order to make the windows overlap.

BERT calculates a vector on every window.

after each loop the vectors are summed up and the mean is taken in the end. so there is only one vector per document.

BERT model: Workflow



Here again, you see the workflow in total
so on the left side there is the split of the dataset in train-data and validation data
then, the model set up
and the sliding window technique is applied if necessary.
in the end we get statistical probabilities.

And this is what we like to provide as information on the display in the discovery
system for our users.

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: $F1\text{-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$; highest possible value of an F-score is 1.0, indicating perfect precision and recall

here you see the some first results

it's a comparison between the "normal" BERT approach, which cuts off after 512 tokens, and the sliding window BERT

Basically: the F1-score indicates how good model performs. as closer to 1 the better.

to me, surprisingly, the results of the "normal" BERT is even better, than the one, that takes care of the whole document.

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

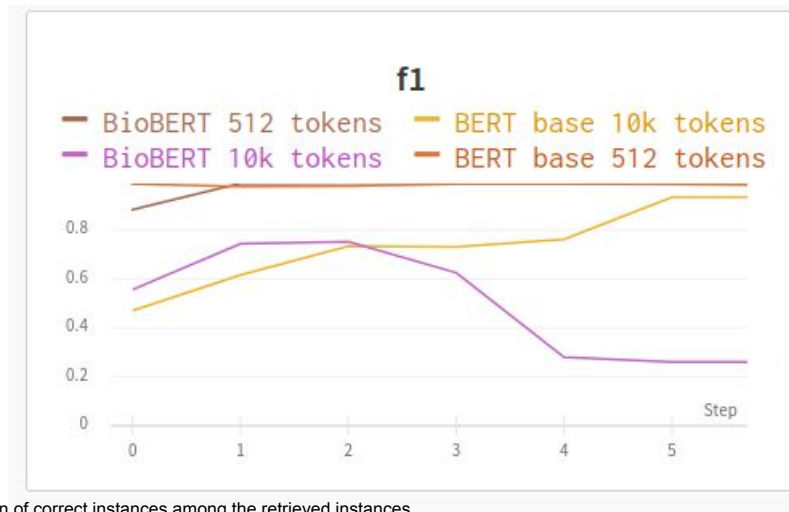
Recall: fraction of complete group of relevant instances that had been retrieved

F1-score: $F1\text{-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$; highest possible value of an F-score is 1.0, indicating perfect precision and recall

second is a comparison between two models from the BERT-family. the basic BERT and the Bio Bert. BioBERT is trained on texts from the Biological field. Therefore, we expected it to be more sufficient for our task. which deals with life science.

Also, here, the Bert base model performed really good. And even better than the bioBert model.

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: $F1\text{-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$; highest possible value of an F-score is 1.0, indicating perfect precision and recall

Here we see the results of the F1-score again in a graph regarding both models: bio bert and BERT base

The original BERT approach with only the first 512tokens outperforms our sliding window approach in both cases.

Here we need more experiments, more epochs, additional models or a better understanding of this behaviour.

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: $F1\text{-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$; highest possible value of an F-score is 1.0, indicating perfect precision and recall

When we take a closer look on the single classes, classified. we see a quite good result for the class probabilities.

The class disinformation differs. It has a slightly worse result.

the precision is 0.92 - zero point nine- two)- means, about 90 percentage of the items classified as disinformation are actually, labeled as disinformation.

The second parameter, recall, tell us, than only about 80 percentatge of all the items labeled as disinformation had been identified.

From this we can see, there is still improvement for better result. and for this we need better training data. especially for the disinformation class.

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**



We also tried, to test our fine-tuned BERT model on a text which was unknown to BERT. so which was not part of the training data or test data.

And we decided for a PLOS One article.

Also the Plos One data base is not part of the data set

the result shows a quite high probabilities for class disinformation.

we guess, the reason is, that the text is taken from a website.

also many items for the disinformation class are taken from websites

and so the reason or this classification might be that the BERT model learned the formal characteristics of websites rather than the language characteristics.

so what we learn again here is, we need more data for training from different sources.

— also to include websites to the scientific text class - for example

```
without softmax tensor([0.4197, 0.1947, 0.4711]), grad_fn=<SigmoidBackward0>
with softmax tensor([0.3897, 0.1303, 0.4800]), grad_fn=<SoftmaxBackward0>
```


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



those estimated probability values for the classes can be used
to create the introduced bar chart in the display of the discovery system.

and with this I come to my last slides.

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”?

What can we learn from this.

The workflow is running properly. And also the first results appear to be promising the issue on the full text and the sliding window approach needs to be better understand by additional experiments - more epochs.

We definitely, need to include more data.

more distinct data to the classes. more data from different data sources for all categories in order to avoid misleading classification.

so, what is next:

I am currently in discussion with my colleagues about the topic of detection of bot generated texts. so maybe we will also include this. but this is still in the air.

An I also have a question to you:

I have a lot of discussion on the term “disinformation” - maybe “disinformation” is to broad? or does it tends to be a political judgement?

Maybe an different wording would be better serve the approach, as just and simple “not scientific information”?

I f you have an opinion please tell me or write me an email:

and, now I thank for the attention

and I am looking forward for your questions.

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 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 Koprinka et al. Vol. 1323. Cham, p. 591–604. Springer, 2020.
- 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

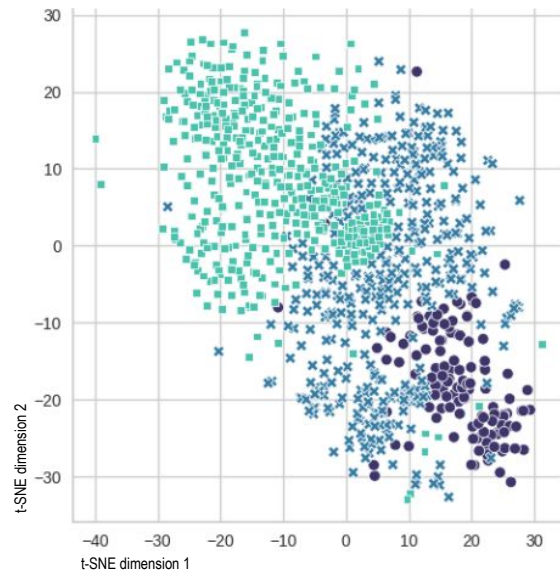
seidlmayer@zbmed.de
<http://www.zbmed.de/>

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



To prove the feasibility of the approach, we created an exploratory dataset, which contains nearly 1 k German documents.

We used doc2vec to test unsupervised classification of the data. With gensim the texts were preprocessed and tokens were created from single words.

We then used gensim to create a doc2vec vector for each document with a length of 40 dimensions and trained a model in 60 epochs. The labels of each text type were added to each data point afterwards..

The resulting matrix for the document corpus shows a fairly clear delineation of text types in the t-SNE projection.

However, it **also** shows the risk of a wrong classification, since characteristics are not fully distinct

so we really need to take care of the data set.

Project goal

The screenshot shows the PubMed.gov search results for the query "wakefield children". The top navigation bar includes the NIH logo, "National Library of Medicine", and "National Center for Biotechnology Information". A search bar contains the text "wakefield children" with a "Search" button. Below the search bar, there are links for "Advanced" and "User Guide". The search results section shows a "Retracted article" warning in a red box. The article title is "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children". The authors listed are A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, and J A Walker-Smith. The article is from Lancet, 1998 Feb 28;351(9103):637-41. The PMID is 9500320 and the DOI is 10.1016/s0140-6736(97)11096-0. There is an "Erratum in" section below the main article, which is a retraction of an interpretation from Lancet, 2004 Mar 6;363(9411):750. The PMID is 15016483 and it states "No abstract available". On the right side, there are links for "FULL TEXT LINKS" (THE LANCET), "ACTIONS" (Cite, Collections), "SHARE" (Twitter, Facebook, LinkedIn), "PAGE NAVIGATION" (Retraction notice), and a list of related items (Title & authors, Erratum in, Retraction in, Expression of concern).

and that is what we strive for
or how the display could look like.

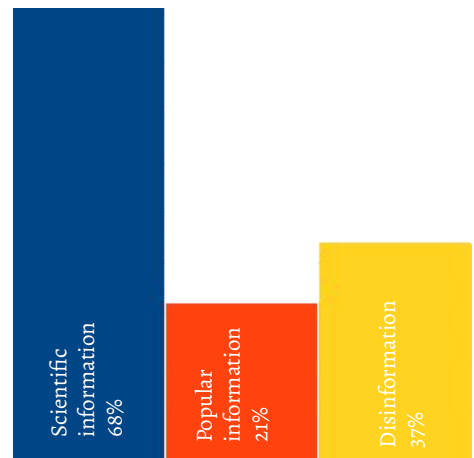
Next to the feature **we already have seen**, is a bar chart indicating similarity to the classes.

By providing a transparent labelling of different aspects of publications, the public is empowered to make informed decisions while still being presented with all the information. No information is excluded or censored, allowing the inclusion of, e.g., dissenting positions, borderline cases and new currents.

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?](#)



Before I will further describe the machine learning process as well as the classification.

let me just shortly

We ensure that we get three values output for three classes, that indicate the similarity to each class.

We think, it's also important, to explain to the user how the model works. This means that for each estimated publication also an explanation is displayed, why exactly this example got a certain assignment. Especially in the context of disinformation, it is important that the ML models are comprehensible for the users.

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....

here you see the data base scheme