Universität Zürich[UZH]

Department of Computational Linguistics

# Helping patients with AI and deep semantics

## *How transformers can transform text analysis with deep semantics in patient care & decision taking*


*Gerold Schneider*

Department of Computational Linguistics

gschneid@cl.uzh.ch

Professor of Computational Linguistics

Linguistic Research Infrastructure (LiRI) NLP Group

Text Crunching Center (TCC)

Center «Language and Medicine»

Project «Mind the Patient Lab»

DSI Health Community

URPP Digital Religion(s)



**Many thanks to Giovanni Spitale and Nikola Biller-Andorno**


01.11.2023

Universität Zürich[UZH]

Department of Computational Linguistics

# Helping patients with AI and deep semantics

## *How transformers can transform text analysis with deep semantics in patient care & decision taking*

*Gerold Schneider*

We have suggestions how AI methods can be used and would like to collaborate with USZ experts. Our approaches could be used in the following areas:

• Shared decision making (SDM): Independent, largely unbiased advice including aspects beyond pre-conceived schemes are vital for supporting patients in SDM. We would like to offer a chatbot-based application to help patients. While there is a danger that tools developed by industry may contain stakeholder biases, we can offer to program a University-based solution, in which research data, patient perspective and hospital requirements are key. Transformers now reach human levels of semantic depth, their ethical values are constantly monitored, and they have unlimited patience for each aspect in the decision process and furnish facts from all areas of expertise, also beyond medical.

• Quality assurance and patient well-being: Patient well-being is more than a binary diagnosis – decisions, coping strategies, transparency and hospital service are further important aspects. Content analysis of unstructured data can detect more patterns than fixed survey questions, find new shades, nuances or forgotten questions, and also bring up individual journeys. Our approaches can extend surveys and bridge the gap from purely statistical approaches to manual reading, from distant to close reading.

# Contents



Datasets for Case Studies:

# 1. Introduction: From Unstructured Data to Quality Control

**IR view**: Textual data = unstructured data. Abundant, and have the potential to contain more information than e.g. surveys.

> Today we have much more data about people's opinions than any single poll could collect, … Such data typically does not constitute an unbiased sample, is noisy, …
> A more important problem, if less often mentioned, is that *we do not know what questions to ask*. (Galbrun & Miettinen 2016)

**Linguistic viewpoint:** language is highly structured, allowing us to express all nuances.
Structurally highly ambiguous <=> semantically highly redundant.
With enough context, it is easy to predict a missing word in a cloze test → BERT/GPT

**Clinical Medicine** View: Unsupervised, data-driven methods are particularly suitable:

- No clear "gold standard" category to predict. Diagnosis?
- *Befund* vs. *Befinden*: patients' experiences are more than binary sentiment or diagnose
- Help the patient: coping strategies, find people with similar concerns
- Ample amounts of rich texts have not been exploited much yet

# 1. Introduction: Research Questions

**RQ: How does Natural Language Processing (NLP) help for applications in Health & Humanities related to Patient Care, where deep semantics & understanding is needed, like in Shared decision making (SDM)?**

1. How does Machine Learning (ML) compare to human annotation?

   → Document classification, BERT zero-shot, GPT-4

2. What level of semantic detail can we reach?

   → BERT zero-shot, GPT-4

# 1. Introduction: Datasets

- Assessment of Doctors by Patients: Automatically predict if satisfied

| rating | happy | content | comment_nolf |
| --- | --- | --- | --- |
| 2 | no | yes | Ich bin franzose und bin seit ein paar Wochen in muenchen. Ich hatte Zahn Schmerzen und m |
| 6 | no | no | Dieser Arzt ist das unmöglichste was mir in meinem Leben je begegnet ist er ist unfreundlich |
| 1 | yes | yes | Hatte akute Beschwerden am Rücken. Herr Magura war der erste Arzt der sich wirklich Zeit fü |
| 1 | yes | yes | Nachdem ich in der Klinik nur ungenaue Angaben erhalten habe welche Sportarten mein her |
| 1 | yes | yes | Frau Dr. Vetter kenne ich seit vielen Jahren sie hat mir bei vielen Problemen sehr kompetent |
| 1 | yes | yes | Dieser Mann weiss was er tut. |
| 2 | no | yes | Kompetent kümmert sich ordentlich |
| 1 | yes | yes | Durch einen Kopfdurchschuß im zweiten Weltkrieg (1941) wurde meine rechte Gesichtshälfte |
| 1 | yes | yes | Ein toller Arzt der sich immer sehr viel Zeit nimmt dabei sehr freundlich ist und überaus kon |

- Dutch Euthanasia Reports: Automatically predict if justified

> The patient, a man in his eighties, had suffered for 10 years from macular degeneration (which causes cells in the centre of the retina to die) in both eyes, which caused his eyesight to deteriorate. Around the same time, an obstructed blood vessel in the retina caused blindness in his right eye. Six months before his death, his left eye deteriorated so much, despite the start of treatment, that he was no longer able to read …

- DIPEX CMI (Intensive Care): Query specific information

> Ms Anna A. “So I remember the time that passed at first very slowly, an immense tiredness, the impression of not understanding what was happening to me. And then the noise, the light that was very strong. And at the beginning, too, when I was under morphine, well, that was unpleasant for me because I started having hallucinations and I couldn't stand that. …”

# 2.1 Document Classification: Doctor Evaluation

Dataset from data.world: German patients' verdicts on their doctors.
They give rating (1-6) and free comment.
1-2: **content (yes)** 4-6: **not content (no)**.

Q: Can we predict patient satisfaction based on their comments?

| rating | happy | content | comment_nolf |
|---|---|---|---|
| 2 | no | yes | Ich bin franzose und bin seit ein paar Wochen in muenchen. Ich hatte Zahn Schmerzen und m |
| 6 | no | no | Dieser Arzt ist das unmöglichste was mir in meinem Leben je begegnet ist er ist unfreundlich |
| 1 | yes | yes | Hatte akute Beschwerden am Rücken. Herr Magura war der erste Arzt der sich wirklich Zeit fü |
| 1 | yes | yes | Nachdem ich in der Klinik nur ungenaue Angaben erhalten habe welche Sportarten mein herz |
| 1 | yes | yes | Frau Dr. Vetter kenne ich seit vielen Jahren sie hat mir bei vielen Problemen sehr kompetent |
| 1 | yes | yes | Dieser Mann weiss was er tut. |
| 2 | no | yes | Kompetent kümmert sich ordentlich |
| 1 | yes | yes | Durch einen Kopfdurchschuß im zweiten Weltkrieg (1941) wurde meine rechte Gesichtshälfte |
| 1 | yes | yes | Ein toller Arzt der sich immer sehr viel Zeit nimmt dabei sehr freundlich ist und überaus kom |

With logistic regression, we get about 92% accuracy.

Here: 1-grams, 10-fold cross-validation, 10000 training samples, L2 regularisation

**Train** Name: logit__1grams_1 Feature Selection

Trained Models: logit__1grams

TRAINED_MODEL

Model Evaluation Metrics:

| Metric | Value |
|---|---|
| Accuracy | 0.9182 |
| Kappa | 0.7227 |

Model Confusion Matrix:

| Act \ Pred | no | yes |
|---|---|---|
| no | 1388 | 495 |
| yes | 323 | 7793 |

# 2.2 Document Classification: Euthanasia



Dutch Euthanasia Reviews, scraped from https://english.euthanasiecommissie.nl

Publicly available reports assessing if a physician complied with the strict rules that are laid down by law: the due care criteria. Physicians who fail to observe these statutory requirements could be criminally liable.

Example:

The patient, a man in his eighties, had suffered for 10 years from macular degeneration (which causes cells in the centre of the retina to die) in both eyes, which caused his eyesight to deteriorate. Around the same time, an obstructed blood vessel in the retina caused blindness in his right eye. Six months before his death, his left eye deteriorated so much, despite the start of treatment, that he was no longer able to read, even using aids. In addition to these eyesight problems, he was uncertain when walking, which was aggravated by his near-blindness. In recent years he had become unwell and fallen several times. Because he had become almost totally blind, the patient could no longer read (which was extremely important to him) or pursue his other hobbies. He was suffering from the loss of these activities, which were essential to him. He also suffered from the loss of self-reliance caused by his impaired vision, and the fact that he knew that there was no prospect of improvement whatsoever. The patient, who had always had a wide range of interests and a great intellectual appetite, experienced his suffering as unbearable. The committee found that the physician had plausibly argued that he was reasonably able to conclude that the patient's suffering was unbearable to him and without prospect of improvement, and that it was unlikely that optical aids and possibly surgery would enable him to read again. The other due care criteria were also fulfilled.

# 2.2 Document Classification: Euthanasia

Q: Can we predict the commissions' assessment based on the report?

| Verdict of the committee | Binary assessment |
|---|---|
| due care criteria complied with | y |
| due care criteria not complied with | n |
| not acted in accordance with the due care criteria | n |
| voluntary and well-considered request | y |
| independent assessment | y |
| unbearable suffering without prospect of improvement | y |
| no reasonable alternative | y |
| straightforward case | y |
| exercising due medical care | y |

With logistic regression, we reach 93% accuracy (77% Kappa).
All predicted *n* are correct → Precison = 1
Some *n* are predicted as *y* → Recall = 11/16

Here: 1-grams, 10-fold cross-validation, 70 training samples, L2 regularisation

# 2.2 Document Classification: Euthanasia

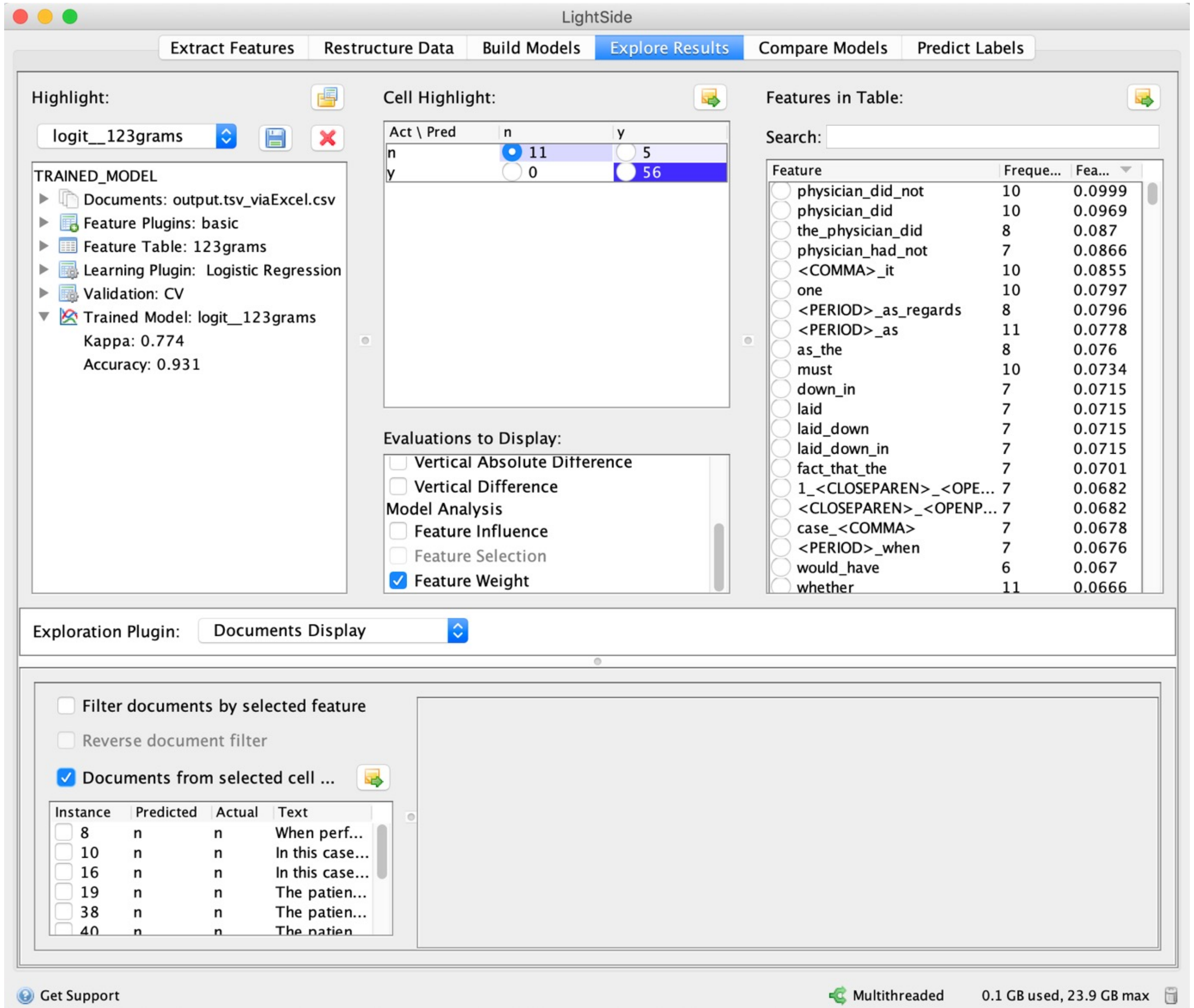
Q: Can we predict the commissions’ assessment based on the report?

Lexical Key Features →

Further Trends:

Adding age:
same classification.
Low age tends for 'n’
(50-60 particularly), high for 'y'

Adding disorders:
same classification.
*Cancer, neurological disorders* tend towards 'n’,
*psychiatric, geriatric, dementia & combination* for 'y'

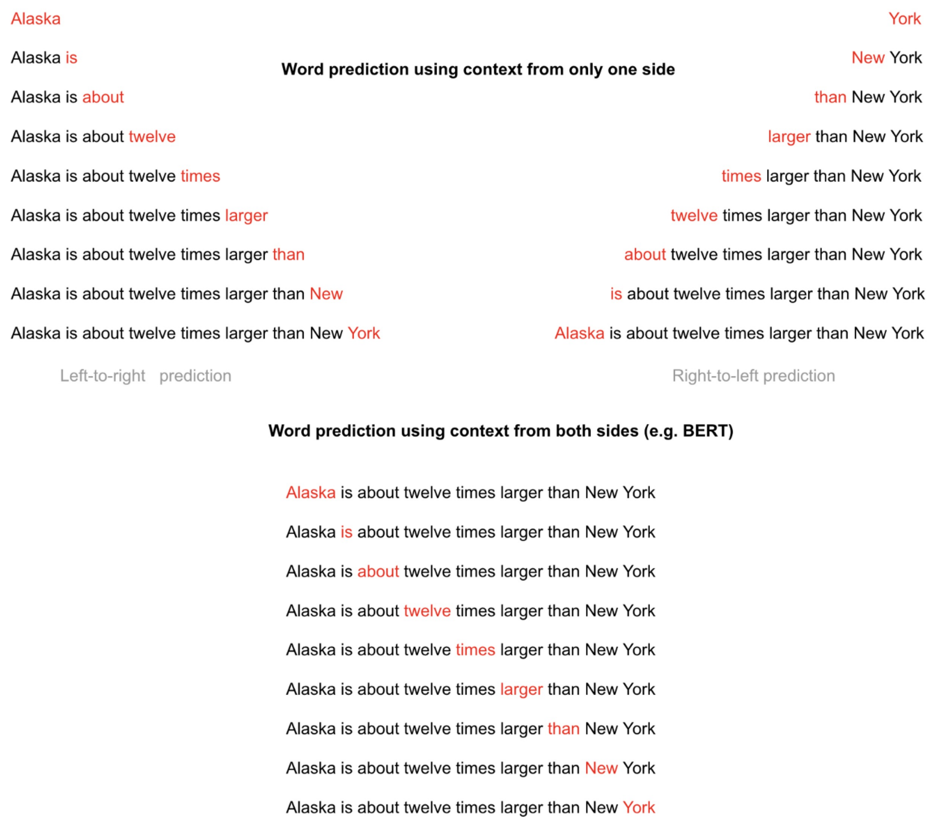


# 3. Transformers
## 3.1 Methods: BERT models

BERT = Bidirectional Encoder Representations from Transformers
predict forward and backward (Bidirectional) simultaneously
and use complex neural networks, so-called transformers.
Good at observing long distances between important events (words)

In BERT masked LM, about 15% of the words are masked (like in a cloze test), a large neural network then tries to best guess them, from both sides.

A *self-supervised* approach, in which a seemingly very remote task is learnt, but with billions of training instances.




Source:
https://qphs.fs.quoracdn.net/main-qimg-7009ad77a7fc4dae106dbd4b0d5f8286

Universität Zürich[UZH]

# 3.1 DIPEx CMI: deep semantics

Department of Computational Linguistics

Predicting the 6 basic mental states (fine-grained sentiment detection)


- Sad
- Happy
- Afraid
- Angry
- Surprised
- Disgusted

Plus, domain-specific:

- Confused
- Ill

Detect mental states to offer personal help to the patient

# Example with DIPEx CMI

Extracting specific info from texts:
- Find out if patients have specific complaints

Text used here:

Ms Anna A. was very tired during her stay in the intensive care unit. She was particularly disturbed by the noise and the light, but she also received a lot of care from the nursing staff. “So I remember the time that passed at first very slowly, an immense tiredness, the impression of not understanding what was happening to me. And then the noise, the light that was very strong. And at the beginning, too, when I was under morphine, well, that was unpleasant for me because I started having hallucinations and I couldn't stand that. And then the fact of lying in bed all the time and then (almost) not being able to move, that was difficult.   I: Do you have a more specific, practical example where you felt that way?   E:Yes, already the fact that I had to be washed, for example, completely, or, that's it, I couldn't get up. So I was in bed and the nurses had to wash me. That was a bit difficult; well, they did it very well, it was admirable and they really did everything to not make me feel indebted or whatever; well, very well, but for me it was difficult. And the other thing that was also difficult was the heat and the thirst. So I couldn't drink, I was thirsty, that was difficult. And the heat, I was always hot.”

# 3.2 BERT NLI with DIPEx CMI: deep semantics

```
You felt that you were in very good and safe hands
Score for Sad Class :  0.009092215448617935
Score for Happy Class :  0.5244683027267456
Score for Afraid Class :  0.0027242163822055
Score for Angry Class :  0.0143034690991044O4
Score for Surprised Class :  0.09013856947422028
Score for Disgusted Class :  0.0018479727441444993
Score for Ill Class :  0.003864304395392537
Score for Confused Class :  0.06891942769289017

I had a tonsil operation in the hospital, it was the last time I was in
the hospital and it went really, really badly
Score for Sad Class :  0.9595155119895935
Score for Happy Class :  0.00019919435726478696
Score for Afraid Class :  0.25847625732421875
Score for Angry Class :  0.15531715750694275
Score for Surprised Class :  0.5397513508796692
Score for Disgusted Class :  0.7282556295394897
Score for Ill Class :  0.8842177391052246
Score for Confused Class :  0.5764621496200562
```

# 3.2 BERT NLI with DIPEx CMI: deep semantics

```
Especially the high-pitched sounds
Score for Sad Class :  0.043813709169626236
Score for Happy Class :  0.08523856848478317
Score for Afraid Class :  0.1930558979511261
Score for Angry Class :  0.20833750069141388
Score for Surprised Class :  0.7467048168182373
Score for Disgusted Class :  0.33189794421195984
Score for Ill Class :  0.03922383114695549
Score for Confused Class :  0.3544618487358093

So for 72 hours I was intubated
Score for Sad Class : 0.2933671772480011
Score for Happy Class : 0.009790988638997078
Score for Afraid Class : 0.1792670339345932
Score for Angry Class : 0.054960042238235474
Score for Surprised Class : 0.6244845390319824
Score for Disgusted Class : 0.09834091365337372
Score for Ill Class : 0.8878562450408936
Score for Confused Class : 0.6522014737129211
```

# BERT model:

# QAsper

Hugging Face Search models, datasets, users...

z-uo / **bert-qasper** like 0

Question Answering · PyTorch · Transformers · z-uo/qasper-squad · en · bert · AutoTrain

**Model card** · Files and versions · Community

### bert-base for QA with qasper

Train from bert-base-uncased.

How to use by python code:

```
from transformers import AutoModelForQuestionAnswering, AutoTokenizer, pipel
ine

# Load model with pipeline
model_name = "z-uo/bert-qasper"
nlp = pipeline('question-answering', model=model_name, tokenizer=model_name)

# Get predictions
QA_input = {
    'question': 'what they propose?',
    'context': "In this paper, we provide an innovative contribution in the
research domain dedicated to crop mapping by exploiting the of Sentinel-2 sa
```

# Example with DIPEx CMI

Extracting *specific* info from texts:

- Find out if patients have specific complaints

```
Did the patient feel thirsty? {'score': 0.08876413106918335, 'start': 1146,
'end': 1177, 'answer': "I couldn't drink, I was thirsty"}

What disturbed the patient? {'score': 0.6534686088562012, 'start': 108,
'end': 127, 'answer': 'noise and the light'}

Did the patient sleep well? {'score': 0.022327510640025l4, 'start': 0, 'end':
25, 'answer': 'Ms Anna A. was very tired'}
```

Extracting *specific* info from texts:

- Mine the scientific literature for medical info

```
What are the dangers of XXX operation?
When is XXX recommended?
What are frequent side effects of XXX?
```

# 3.1 BERT NL Inference

How deep is their semantic general knowledge? Let’s test.
https://huggingface.co/svalabs/gbert-large-zeroshot-nli

Trained on Wikipedia, comprehensive web dumps (OSCAR, machine-translated).
Can it predict patient satisfaction?

From comment (free text), predict the probability of four derivable labels:

zufrieden, unzufrieden, glücklich, enttäuscht

```
[9]: sequence = "Hatte akute Beschwerden am Rücken. Herr Magura war der erste Arzt der sich wirklich Zei
     labels = ["zufrieden", "glücklich", "unzufrieden", "enttäuscht"]
     hypothesis_template = "In diesem Satz ist der Patient {}."

[10]: zershot_pipeline(sequence, labels, hypothesis_template=hypothesis_template)

[10]: {'sequence': 'Hatte akute Beschwerden am Rücken. Herr Magura war der erste Arzt der sich wirklich Z
     eit für einen Therapieplan genommen hat um nachhaltig meine Schmerzen zu beseitigen',
      'labels': ['zufrieden', 'glücklich', 'enttäuscht', 'unzufrieden'],
      'scores': [0.5109395980834961,
       0.20857857167720795,
       0.1795770525932312,
       0.10090477764606476]}
```

jlab_server Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 31 BART_German.ipynb

# 3.1 BERT NL Inference

zufrieden, unzufrieden, glücklich (→ zufrieden), enttäuscht (→ unzufrieden)

Evaluation: Accuracy = 93.7%

Better than our (simple) Document classification method, where we had 92%

The generic knowledge of the model is surprisingly domain-independent

BERT Models, ***without any specific training***, reach similar performance to trained document Classification.

| F | Manual | Zero-shot | True Pos | False Pos | Accuracy |
|---|---|---|---|---|---|
| 6840 | 1 | zufrieden | 6840 | | |
| 480 | 2 | zufrieden | 480 | | |
| 59 | 3 | zufrieden | | 59 | |
| 41 | 4 | zufrieden | | 41 | |
| 27 | 5 | zufrieden | | 27 | |
| 11 | 6 | zufrieden | | 11 | 0.9815 |
| 55 | 1 | unzufrieden | | 55 | |
| 32 | 2 | unzufrieden | | 32 | |
| 49 | 3 | unzufrieden | 49 | | |
| 95 | 4 | unzufrieden | 95 | | |
| 125 | 5 | unzufrieden | 125 | | |
| 97 | 6 | unzufrieden | 97 | | 0.8079 |
| 245 | 1 | enttäuscht | | 245 | |
| 142 | 2 | enttäuscht | | 142 | |
| 160 | 3 | enttäuscht | 160 | | |
| 298 | 4 | enttäuscht | 298 | | |
| 493 | 5 | enttäuscht | 493 | | |
| 410 | 6 | enttäuscht | 410 | | 0.7786 |
| 304 | 1 | glücklich | 304 | | |
| 18 | 2 | glücklich | 18 | | |
| 6 | 3 | glücklich | | 6 | |
| 3 | 4 | glücklich | | 3 | |
| 6 | 5 | glücklich | | 6 | |
| 3 | 6 | glücklich | | 3 | 0.9471 |
| ∑ | | | 9369 | 630 | **0.9370** |

## Doc 20 from Euthanasia, misclassified by bag-of-words document classification:

The patient, a man in his fifties, was diagnosed with Parkinson's disease four years before his death. He was treated with medication and because he was having difficulty coping with the disease, he received psychotherapy and other treatments at various stages of his illness. He twice underwent deep brain stimulation (electrodes implanted in the brain send electrical impulses to suppress specific symptoms); the second procedure took place around five months before his death. None of this achieved the desired result. After the last treatment, the patient's symptoms worsened. This caused tension and feelings of anxiety and helplessness. The patient experienced his suffering as without prospect of improvement and asked his physician for euthanasia. At the physician's request, the patient was seen by a psychiatrist who found him to be decisionally competent. In the psychiatrist's opinion, there was a psychological aspect, in addition to the Parkinson's disease, that had not yet been treated sufficiently. The psychiatrist recommended a trial course of medication for depression. The patient stopped taking the medication after a few days, as he felt it aggravated his symptoms. He did not want any more psychotherapy to alleviate the symptoms of Parkinson's disease, which could no longer be treated and were increasing. The attending neurologist found that the patient had a mild form of Parkinson's disease, in which the tremor (shaking movements in the limbs) was largely determined by emotional factors. He was also of the opinion that the man's fear of the future was the dominant factor. The neurologist thought that adequate treatment of this fear and the underlying mood disorder was the appropriate course of action. The neurologist was unable to support the patient's request for euthanasia on the grounds of the severity of his Parkinson's disease. He also considered that, as he was unable to support the request from a neurological point of view and the patient wished no further psychiatric treatment, it was impossible to properly assess whether the man's suffering was without prospect of improvement. The notifying physician was satisfied that this suffering was unbearable to the patient and with no prospect of improvement according to prevailing medical opinion; he performed euthanasia. The committee had questions about the absence of a reasonable alternative. The physician was therefore first asked to give a written explanation, later followed by an oral one. The physician was of the opinion that, given the patient's medical history, personality and life history, they had nothing more to offer him. When asked by the committee whether he was satisfied that if it had been possible to treat the stress suffered by the patient, the symptoms of Parkinson's disease would have become milder and therefore the tremors would also lessen, the physician replied that he was not satisfied that that was the case. The committee referred to the psychiatrist's assessment (that the psychological component had been treated insufficiently) and the neurologist's assessment (that it was a mild form of Parkinson's disease in which treatable psychological factors played a role) and pointed out that the process had taken very little time (the physician had talked with the man twice in eight days). The committee noted that if the process is short it attaches great importance to intensive communication, not just between the physician and the patient, but also between the physician and other persons involved. In such a case the physician must do everything that is reasonably possible to obtain all the information that may be relevant. The committee was of the opinion that the physician should not have disregarded the neurologist's advice and the psychiatrist's opinion without further enquiry. He should have consulted with them or with another specialist who was an expert in the field. Particularly in view of the speed at which the process was conducted and the fact that the physician had only spoken twice with the patient, the physician should have used such consultation to assess his own views against those of the specialists. The committee therefore found that the physician had not plausibly argued that he was reasonably able to conclude that the patient was suffering unbearably without prospect of improvement or that there were no reasonable alternatives that could alleviate his suffering.

## Experiments with BERT NL Inference as Q&A

```
[13]: QA_input = {'question': 'Did the patient suffer?',
                 'context' : "The patient, a man in his fifties, was diagnosed with Parkinson’s disease fo
      }
      res = nlp(QA_input)
      print(res)
```

```
{'score': 0.1983867734670639, 'start': 643, 'end': 715, 'answer': 'The patient experienced his suffe
ring as without prospect of improvement'}
```

```
[17]: QA_input = {'question': 'Does the committee agree that the patient suffered?',
                 'context' : "The patient, a man in his fifties, was diagnosed with Parkinson’s disease fo
      }
      res = nlp(QA_input)
      print(res)
```

```
{'score': 0.023891253396868706, 'start': 643, 'end': 715, 'answer': 'The patient experienced his suf
fering as without prospect of improvement'}
```

```
[18]: QA_input = {'question': 'Does the committee agree that euthanasia was justified?',
                 'context' : "The patient, a man in his fifties, was diagnosed with Parkinson’s disease fo
      }
      res = nlp(QA_input)
      print(res)
```

```
{'score': 0.004644709639251232, 'start': 1737, 'end': 1796, 'answer': 'The neurologist was unable to
support the patient’s request'}
```

Relatively good understanding.

# 3.1 BERT NL Inference

**What does BERT know about Euthanasia?** (BART large MLNI model)

```
[105]: input_text="euthanasia is";
       candidate_labels = ["assisted suicide","death","murder","the end of life",
                           "an illness","a patient","a famous painter","a Greek godess",
                           "an ethical dilemma","an ethical question","a political debate",
                           "connected to suffering","a consequence of suffering",
                           "a consequence only permitted for extreme suffering","a reason for suffering"]
```

```
[106]: # multi_label=True will return confidence score for both labels independently
       model_dict = classifier(input_text, candidate_labels, multi_label=True)
       # Zip results to dict
       result_dict = dict(zip(model_dict.get('labels'), model_dict.get('scores')))
```

```
[107]: result_dict
```

```
[107]: {'assisted suicide': 0.9878346920013428,
        'the end of life': 0.9851932525634766,
        'connected to suffering': 0.9771237373352051,
        'death': 0.9526866674423218,
        'an ethical dilemma': 0.9075977802276611,
        'a consequence of suffering': 0.8390344977378845,
        'an ethical question': 0.8190814256668091,
        'a consequence only permitted for extreme suffering': 0.6314218044281006,
        'a reason for suffering': 0.3726317286491394,
        'a patient': 0.11474824696779251,
        'a political debate': 0.06330907344818115,
        'murder': 0.03578542172908783,
        'an illness': 0.033096276223659515,
        'a Greek godess': 0.0007784005138091743,
        'a famous painter': 0.00019802608585450798}
```

Relatively good understanding. Classification?
We used the prompt: ‘The background text X from a discussion on Euthanasia is "…". Given X, does the committee agree that euthanasia was justified? Please answer just with yes or no.’ Its accuracy is 63/72 = 88%. Kappa = 45%

# 3.2 ChatGPT (with GPT-4)

**Experiments with ChatGPT:**

Q: The background text X from a discussion on Euthanasia is "…". Given X, does everyone agree that euthanasia was justified?

The background text X indicates there were disagreements between the medical professionals involved in this case about whether euthanasia was justified.

The attending physician believed euthanasia was warranted, citing the patient's unbearable suffering and the lack of prospects for improvement. However, both the psychiatrist and the neurologist involved felt there were aspects of the patient's condition, specifically psychological and emotional factors, which had not been adequately addressed and might have improved the patient's quality of life. They both saw room for more treatment, making them less supportive of the decision for euthanasia.

Furthermore, the committee overseeing this case expressed concern over the speed of the process, the lack of intensive communication, and the absence of consultation with relevant specialists, and did not seem to find the physician's argument for euthanasia convincing.

So, based on the information given, it is clear that there is not a consensus that euthanasia was justified in this particular case.

# 3.2 ChatGPT (with GPT-4)

**What is the performance of ChatGPT on the binary decision task?**

E.g. Doc 6

Careful formulation of prompts is essential

Q: The background text X6 from a discussion on Euthanasia is "…". Given X6, does **the committee** agree that euthanasia was justified? Please answer just with yes or no.

Although the second independent physician disagreed with this assessment, and the procedure was eventually performed, the initial disagreement suggests that not everyone involved in the process agreed that euthanasia was justified in this case. Therefore, while it's clear that a decision was reached, the text does not suggest that this decision was universally agreed upon.

Furthermore, societal and ethical views on euthanasia are varied and contentious, and different people and societies may interpret the circumstances and ethical considerations of this case in different ways. So even outside of the specific people involved in this case, there would likely not be universal agreement that euthanasia was justified.

Given X6, does the committee agree that euthanasia was justified? Please answer with just yes or no.

Yes

# 3.2 ChatGPT (with GPT-4)

**What is the performance of ChatGPT on the binary decision task?**

Q: The background text X from a discussion on Euthanasia is "…". Given X, does **the committee** agree that euthanasia was justified? Please answer just with yes or no.

With this prompt, **ALL** classifications (72 cases) are correct (P=R=1).
Also detailed justifications can be requested.

E.g. Doc ID 10

thiopental. According to the committee, it could not be completely ruled out that the patient was in an insufficiently deep coma and that for that reason she might have perceived the effects of the muscle relaxant. The committee could therefore only conclude that the physician did not exercise due medical care when terminating the patient's life."
Given X10, does the committee agree that euthanasia was justified? Please answer with just yes or no.

No

why?

In case X10, the committee did not agree that euthanasia was justified because the physician did not adhere to the established euthanasia procedures and guidelines.
Specifically, the physician administered a lower dosage of the coma-inducing substance than recommended and did not adequately check the depth of the patient's coma before administering the muscle relaxant. This could have resulted in the patient being in an insufficiently deep coma, meaning they might have been able to perceive the effects of the muscle relaxant.
Therefore, the committee concluded that the physician did not exercise due medical care when terminating the patient's life, which is a requirement for euthanasia to be considered justified and legally compliant.

# Comparison of Methods

| Method | Pros | Cons |
|---|---|---|
| Document Classification | Fast, large-scale<br>Overview of word features | Annotation needed<br>Limited semantic detail |
| Conceptual Maps | Good bird's eye overview<br>Fields between Classes<br>No annotation needed | Hard to evaluate<br>Limited semantic detail |
| BERT models | Good semantic detail<br>Still easy to run locally | Blackbox<br>Below human level |
| GPT-3.5 / 4 | Excellent semantic detail<br>Human-level performance<br>Can drive Chatbots / Art. Adviser | Blackbox<br>Hard to run locally, slow<br>May hallucinate |

# Conclusions & Outlook

**RQ: How does Natural Language Processing (NLP) help for applications in Health & Humanities such as SDM?**

1. How does Machine Learning (ML) compare to human annotation?

   Good enough for many tasks. Document classification is used for many.
   Zero-shot BERT models attain similar performance with no training.
   GPT-4 reaches human performance.

2. What level of semantic detail can we reach?

   BERT zero-shot lag only little behind humans.
   GPT-4 reaches human performance on most tasks

# Outlook: How NLP/AI approaches can help

*AI Algorithms*

**Predictive Modeling:** Offers forecasts of treatment outcomes based on historical data.

**Data Mining:** Extracts hidden patterns in patient history, enabling more individualized care.

*Combined NLP/AI Features*

**Interactive Empathy-Driven Virtual Assistants:** Engages patients in pre-consultation discussions to gather preliminary information. Utilize sentiment analysis and machine learning to gauge emotional states and provide clinicians with insights into patient feelings, enhancing patient-centered dialogue.

**Qualitative analysis:** Analyzes patient stories and feedback through NLP to extract qualitative data, enriching quantitative medical data for a more comprehensive view.

**Automated Pre-Consultation Summaries:** Use speech-to-text to transcribe patient interviews and apply AI to distill key points, providing a quick summary for physicians to review before consultations.

**Real-Time Explanation Generators:** Employ machine learning models to analyze complex medical data and use NLP to generate easily understandable explanations for cognitively diverse patients during the consultation.

# Outlook: Shared decision making

**Shared Decision Making (SDM) is a collaborative approach where clinicians and patients work together to make healthcare decisions.**

*Key Components*

**Information Exchange:** Comprehensive and unbiased sharing of medical knowledge and options.

**Negotiation and Deliberation:** Balanced discussion on preferences, values, and needs of all stakeholders.

**Consensus:** Arriving at a mutual decision that is agreeable to both the healthcare provider and the patient.

**Autonomy:** Respects patients' freedom to make choices about their own healthcare.

**Informed Consent:** Promotes fully informed choices by sharing complete information on risks, benefits, and alternatives.

**Patient-Centered Care:** Aligns medical decisions with patients’ values, preferences, and needs.

# Outlook: How NLP/AI approaches can help

**Personalized Information:** NLP/AI can curate personalized medical information, making it easier for patients to understand options.

**Predictive Analytics:** NLP/AI can simulate outcomes, providing tangible data for deliberation.

**Neutral Mediation:** NLP/AI tools can offer unbiased information to balance the medical decision-making process.

**Sentiment Analysis & Deep Semantics:** Understands patient feedback and concerns, aiding in a more empathetic approach.

**Decision Support:** NLP/AI algorithms can assist in identifying the best treatment paths, saving time for meaningful dialogue.

**Language Translation:** Breaks down language barriers between physicians and non-English speaking patients.

# 3.2 ChatGPT (with GPT-4)

**What is the performance of ChatGPT on the binary decision task?**

E.g. Doc 6 (classified correctly by Document classification)

Q: The background text X6 from a discussion on Euthanasia is "…". Given X6, does everyone agree that euthanasia was justified? Please answer just with yes or no.

Performing euthanasia when her husband could no longer care for her properly was therefore in line with her advance directive. Together, the physician and the patient could be satisfied that there was no reasonable alternative in the patient's situation."
Given X6, does everyone agree that euthanasia was justified? Please answer with just yes or no.

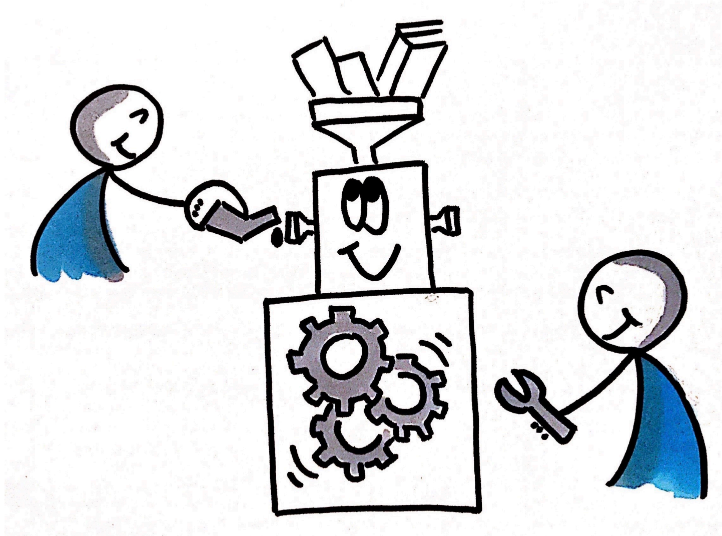
No

Why?

Based on the text provided, it's evident that the situation was complex and heavily debated. The approval of euthanasia, in this case, was primarily based on the patient's voluntary and well-considered request, her unbearable suffering, and the absence of reasonable alternatives. However, the decision was not unanimous. Notably, the first independent physician did not initially believe that the due care criteria for euthanasia had been met because the patient did not express a current wish for euthanasia during his visit.
Although the second independent physician disagreed with this assessment, and the procedure was eventually performed, the initial disagreement suggests that not everyone involved in the process agreed that euthanasia was justified in this case. Therefore, while it's clear that a decision was reached, the text does not suggest that this decision was universally agreed upon.
Furthermore, societal and ethical views on euthanasia are varied and contentious, and different people and societies may interpret the circumstances and ethical considerations of this case in different ways. So even outside of the specific people involved in this case, there would likely not be universal agreement that euthanasia was justified.

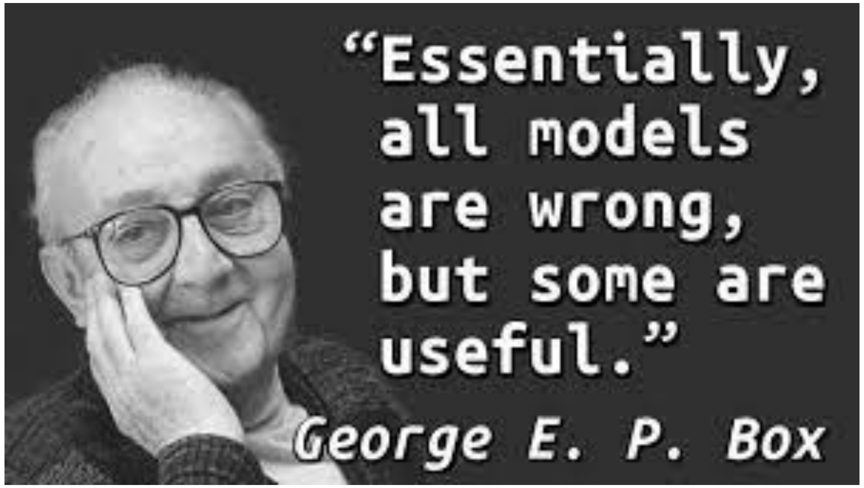
# Conclusions

- Motivated the use of data-driven/unsupervised methods on unstructured=textual data
- The challenge of evaluation is hard
- Detecting new and confirming known patterns
- Structured and unstructured data can mutually benefit
- Multidimensional, open spaces
- The boundaries between supervised and unsupervised are fading (distance learning, meta in maps, few-shot)
- Unsupervised: from niche to majority
- Shown several case studies:
    - Document classification (History, Politics)
    - Distributional Semantics (MS)
    - Topic Modelling (Medical History)
    - Conceptual Maps (History, DIPEx)
    - Zero-shot multi-label classification (DIPEx)
- Outlook: Chatbots. Evaluation.



# What is a model?

Factors in Interaction
Make Predictions
Blackbox Models
Parsimony & Interpretability
Ceiling Effects
Sparse Data
Data Loss




The actual quote is from *Science and Statistics*, George E. P. Box, *Journal of the American Statistical Association*, December 1976, pp. 791-799.

## 2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

# Key References

Ananiadou, Sophia, Kell, Douglas B., and Tsujii, Jun-ichi. 2006. "Text mining and its potential applications in systems biology". *Trends in Biotechnology*, 24, 12, 571 - 579.

Baroni, Marco and Lenci, Alessandro. 2010. "Distributional Memory: A general framework for corpus-based semantics". *Computational Linguistics*, 36, 4, 673-721.

Fitzmaurice, Susan, Robinson, Justyna A., Alexander, Marc, Hine, Iona C., Mehl, Seth, and Dallachy, Fraser. 2017. "Linguistic {DNA}: Investigating Conceptual Change in Early Modern English Discourse". *Studia Neophilologica*, 1--18.

Grimmer, Justin and Stewart, Brandon. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis*, 21, 3, 267-297.

Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva P., Veres, Adrian, Gray, Matthew K., , Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A., and Aiden, Erez Lieberman. 2010. "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science*, 331, 6014, 176-182.

Moretti, Franco. 2013. *Distant Reading*. Verso, London.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral Thesis, University of Stockholm.

Schneider, Gerold (in press). “Medical topics and style from 1500 to 2018”. In Turo Hiltunen and Irma Taavitsainen (eds.) *Corpus pragmatic studies on the history of medical discourse*. Amsterdam: Benjamins. [ watch presentation at https://tube.switch.ch/videos/951fe35c ]

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. OUP, Oxford.

Taavitsainen, Irma and Gerold Schneider (2019). "Scholastic argumentation in Early English medical writing and its afterlife: new corpus evidence." In Carla Suhr, Terttu Nevalainen & Irma Taavitsainen, eds. *From data to evidence in English language research*. Language and Computers, Volume 83. Leiden: Brill. 191-221. [ https://brill.com/view/title/54063 ]

Tognini-Bonelli, Elena. 2001. Corpus Linguistics at Work. John Benjams, Amsterdam .