A gentle Introduction to Natural Language Processing

by Chiara Becht
Master Data Science for Life Sciences
23-02-2023

Natural Language Processing in daily business

Translation

DeepL

Multiple tasks

DeepL

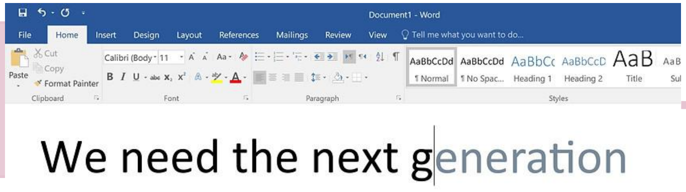
Spam filtering

[1]

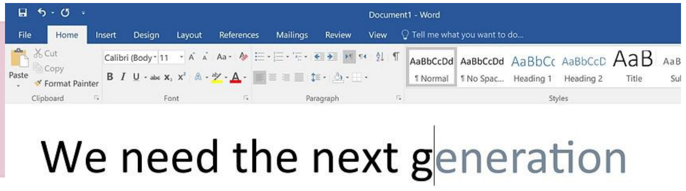
Text classification

[1]

Word prediction



Word prediction



Emotion detection

Customer reviews

Sentiment analysis

Chatbots

Text Generation

„Look closely.
Because the closer you look the less you see“

- now you see me –

How does the magic of NLP comes to life?

- Text vectorization
- Powerful NLP models
- Application for my project

Detect Language

Detect meaningful units

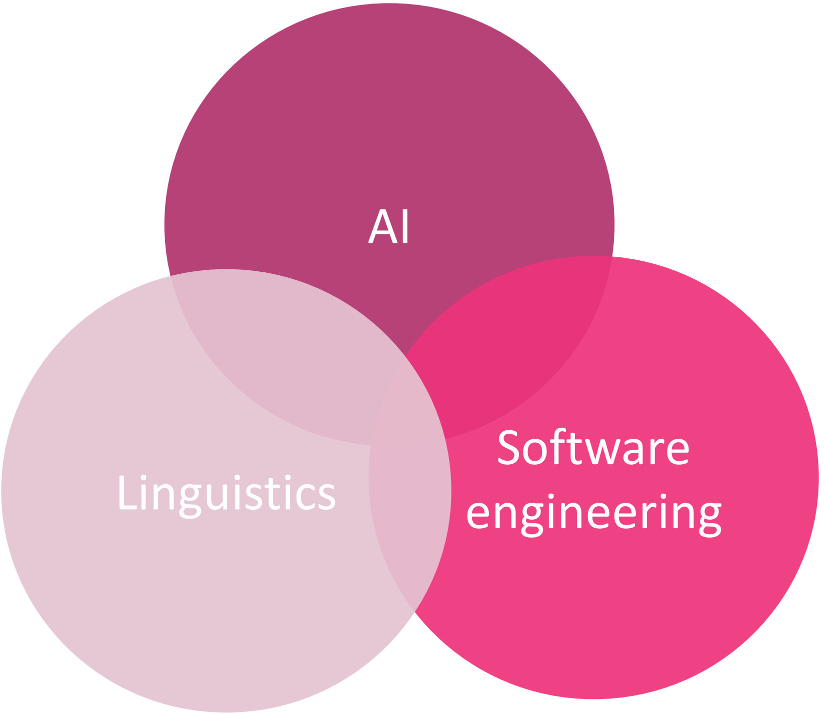
Detect meaning of units

Distinguish between question and answer

Detect language structure and syntax

Abstract meaning of text

Human Brain:
comprehend all at once


AI
Linguistics
Software engineering

Example text

What more could you ask for?

Almost all degree programmes on campus, good bus connections,

smoke-free entire campus and very friendly staff ;)

What more could you ask for?

Almost all degree programmes on campus, good bus connections,

smoke-free entire campus and very friendly staff ;)

Detect Language

Conversational English

What more could you ask for?

Almost all degree programmes on campus, good bus connections,

smoke-free entire campus and very friendly staff ;)

Detect Language

Tokenization

Conversational English

2 sentences, 24 words

What more could you ask for?

Almost all degree programmes on campus, good bus connections,

smoke-free entire campus and very friendly staff ;)

Detect Language	Tokenization	Text Vectorization
Conversational English	2 sentences, 24 words	Numerical representation: • Words • Sentences • paragraph

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Detect Language	Tokenization	Text Vectorization
Conversational English	2 sentences, 24 words	Numerical representation: • Words • Sentences • paragraph

Basic Model building blocks

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Named Entity Recognition (NER)

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Named Entity Recognition (NER)

Part of speech (POS) tagging

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Named Entity Recognition (NER)

Dependency tagging

Part of speech (POS) tagging

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Named Entity Recognition (NER)

Dependency tagging

Part of speech (POS) tagging

Stemming / Lemmatization

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

Named Entity Recognition (NER)

Dependency tagging

Part of speech (POS) tagging

Stemming / Lemmatization

Sequence Tagging Tasks

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

DON'T STUDY HERE!

Awful way of teaching. Horrible teachers. Do not waste your money on this hilarious institution.

NLP Downstream Tasks

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

DON'T STUDY HERE!

Awful way of teaching. Horrible teachers. Do not waste your money on this hilarious institution.

Text Classification

E.g. Sentiment analysis

NLP Downstream Tasks

What more could you ask for?

Almost all degree programmes on campus, good bus connections, smoke-free entire campus and very friendly staff ;)

DON'T STUDY HERE!

Awful way of teaching. Horrible teachers. Do not waste your money on this hilarious institution.

Text Classification

E.g. Sentiment analysis

Question Answering

E.g. should I study at the Hanze?

NLP Downstream Tasks

How does the magic of NLP comes to life?

- **Text vectorization**
- Powerful NLP models
- Application for my project

Ronaldo, Messi, Dicaprio

How can we define this words numerically?

[2]

Ronaldo, Messi, Dicaprio

	isRonaldo	isMessi	isDicaprio
Ronaldo	1	0	0
Messi	0	1	0
Dicaprio	0	0	1

One hot encoding

Ronaldo, Messi, Dicaprio

	isRonaldo	isMessi	isDicaprio
Ronaldo	1	0	0
Messi	0	1	0
Dicaprio	0	0	1

	isFootballer	isActor
Ronaldo	1	0
Messi	1	0
Dicaprio	0	1

Embedding

[2]

Ronaldo, Messi, Dicaprio

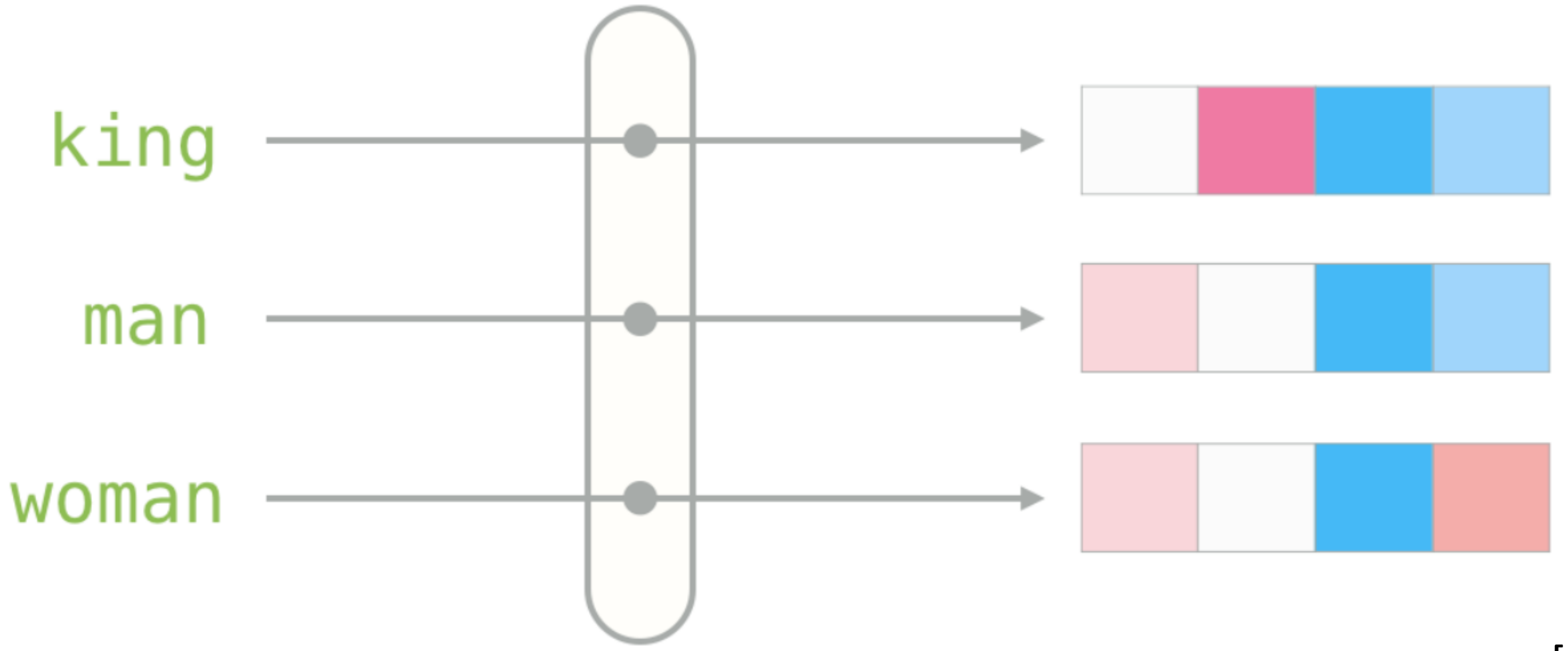
	isRonaldo	isMessi	isDicaprio
Ronaldo	1	0	0
Messi	0	1	0
Dicaprio	0	0	1

	isFootballer	isActor
Ronaldo	1	0
Messi	1	0
Dicaprio	0	1

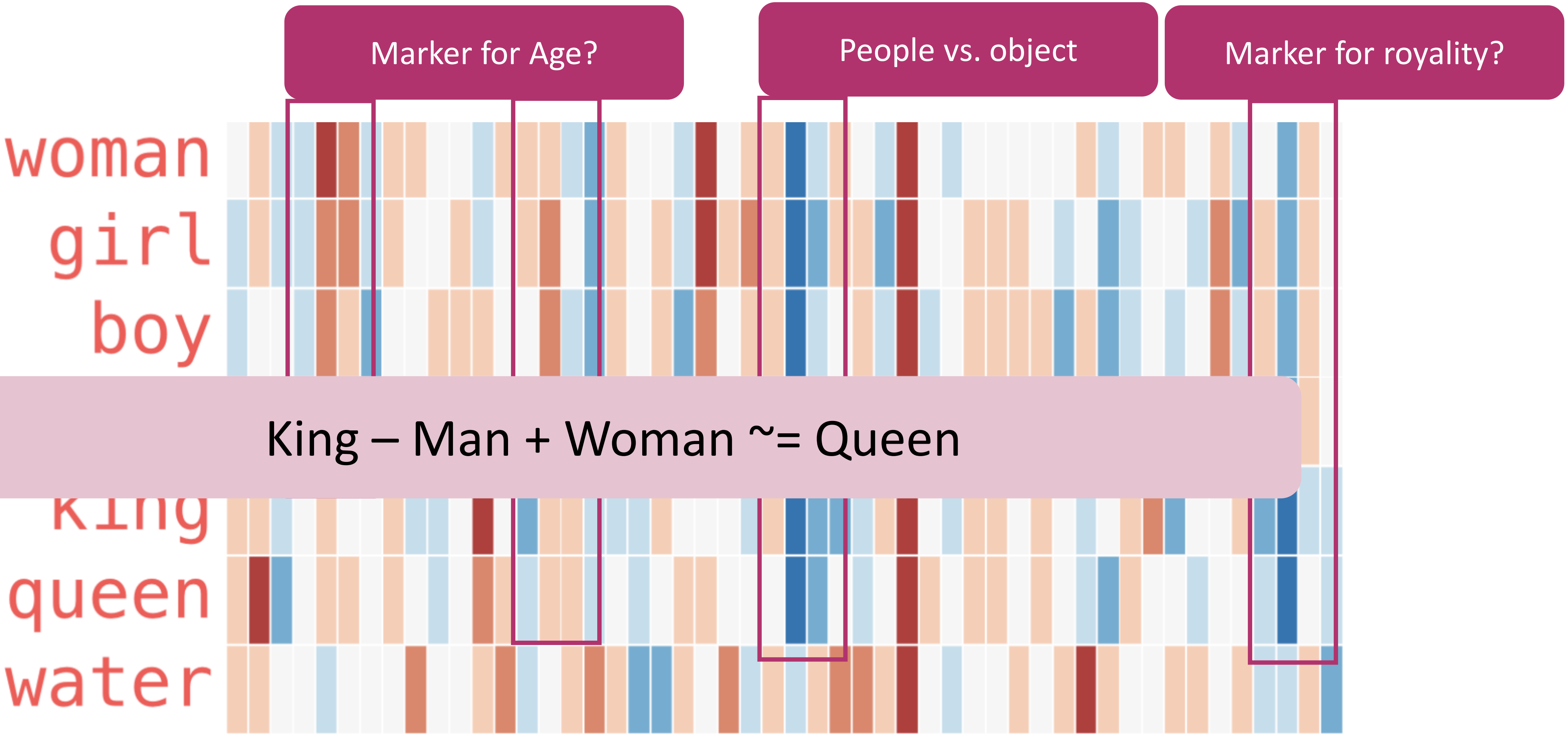
	isFootballer	isActor	Popularity	Gender	Height
Ronaldo	1	0	...	...	...
Messi	1	0	...	...	...
Dicaprio	0	1	...	...	...

[2]

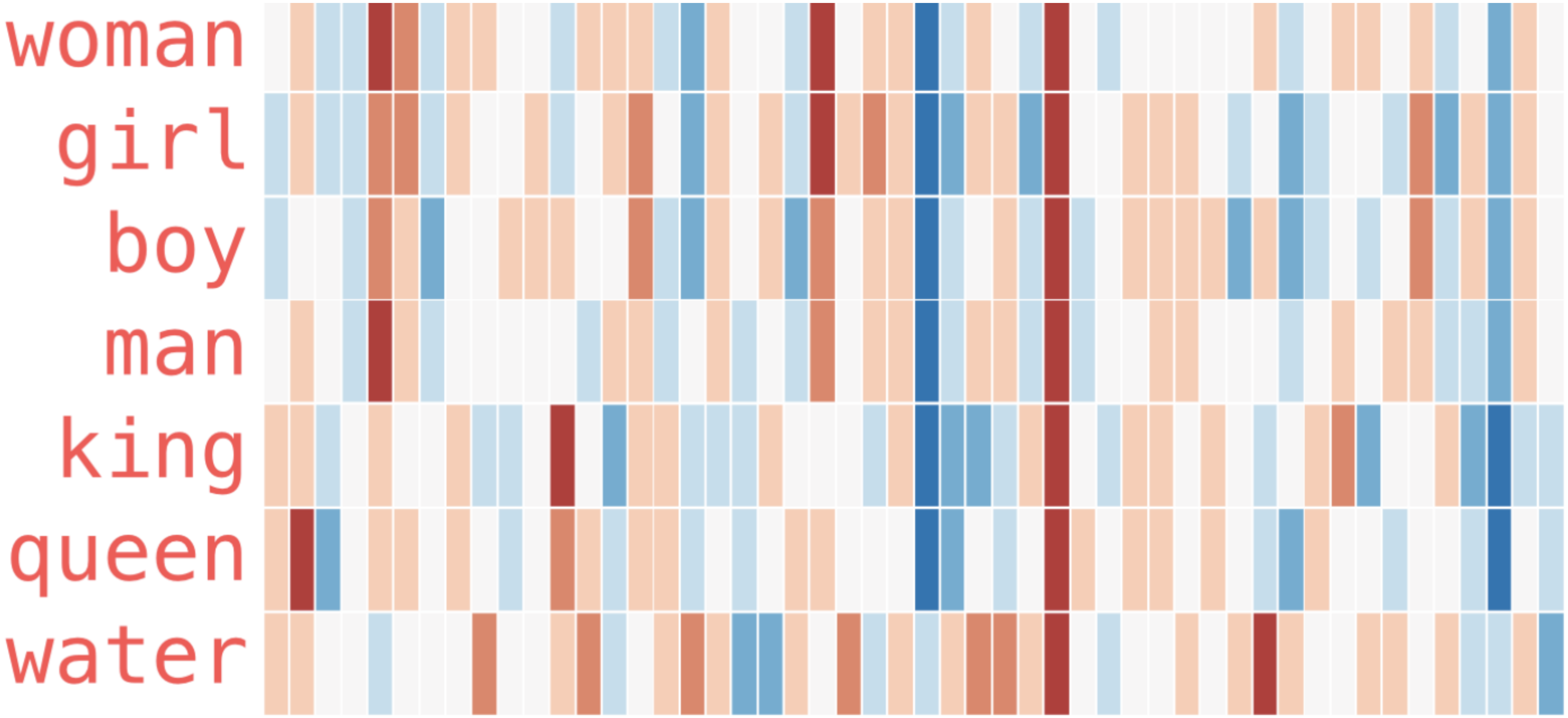
Can a Neural Network do this for us?



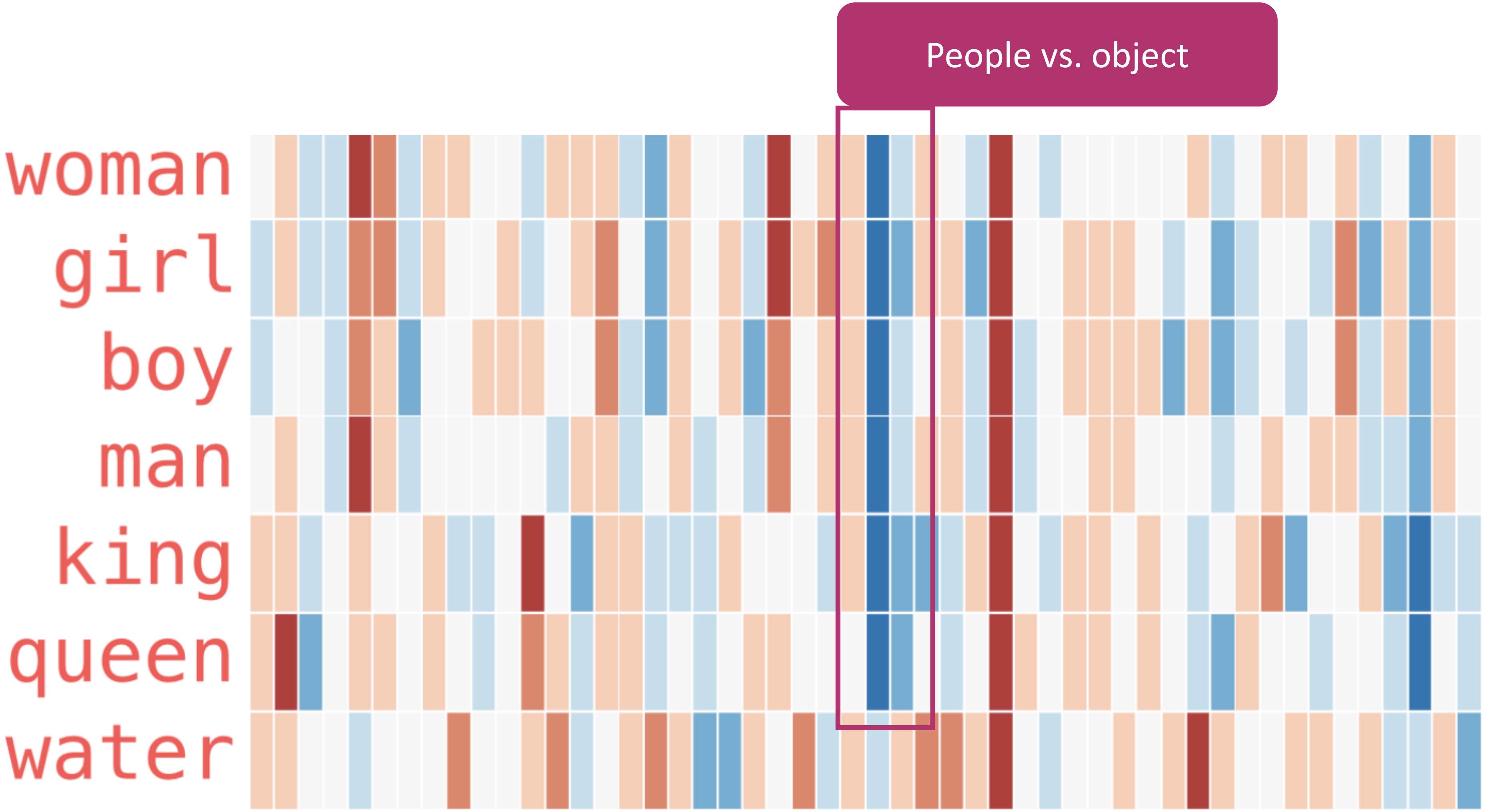
[3]


Marker for Age?
People vs. object
Marker for royality?
woman
girl
boy
King – Man + Woman ~= Queen
king
queen
water

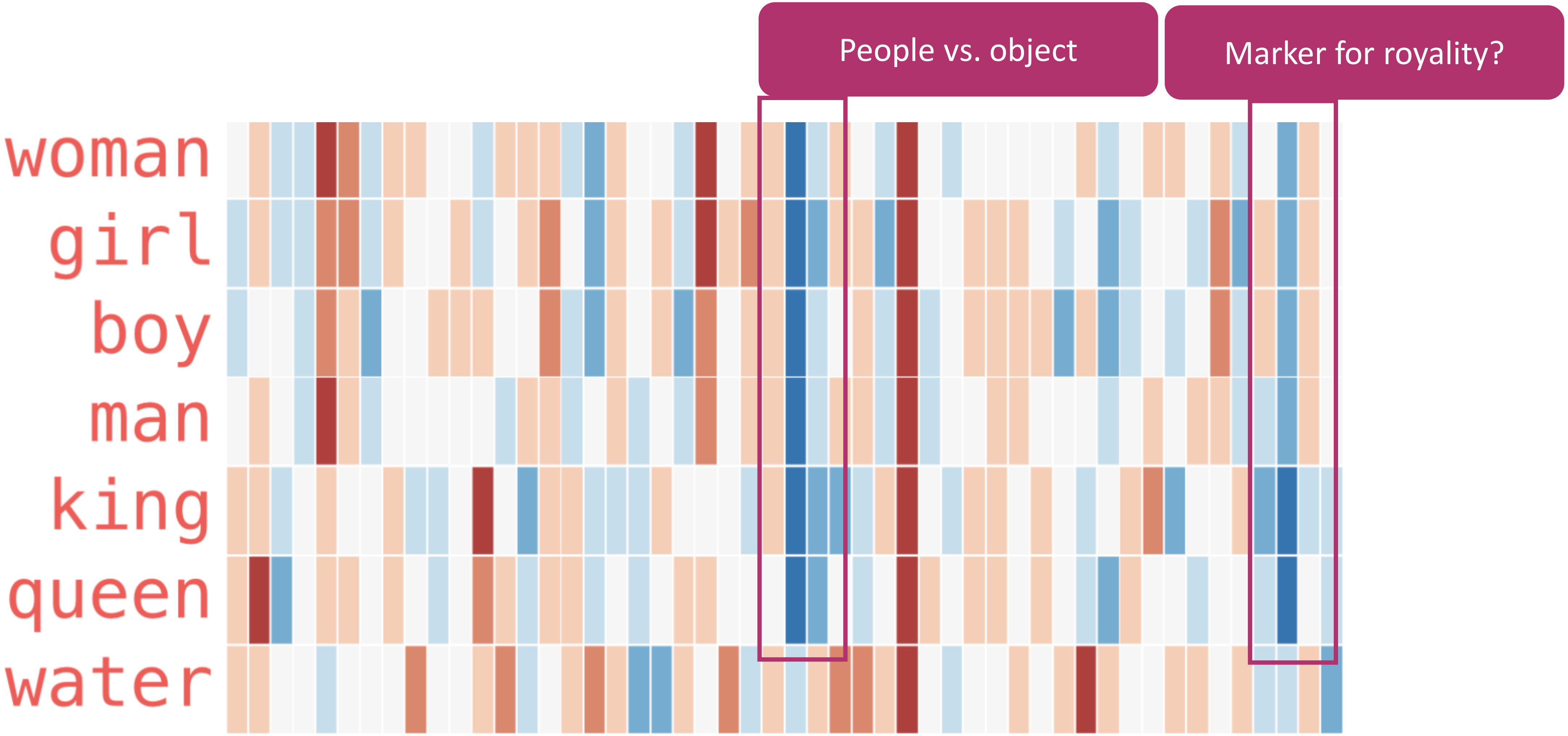
[3]



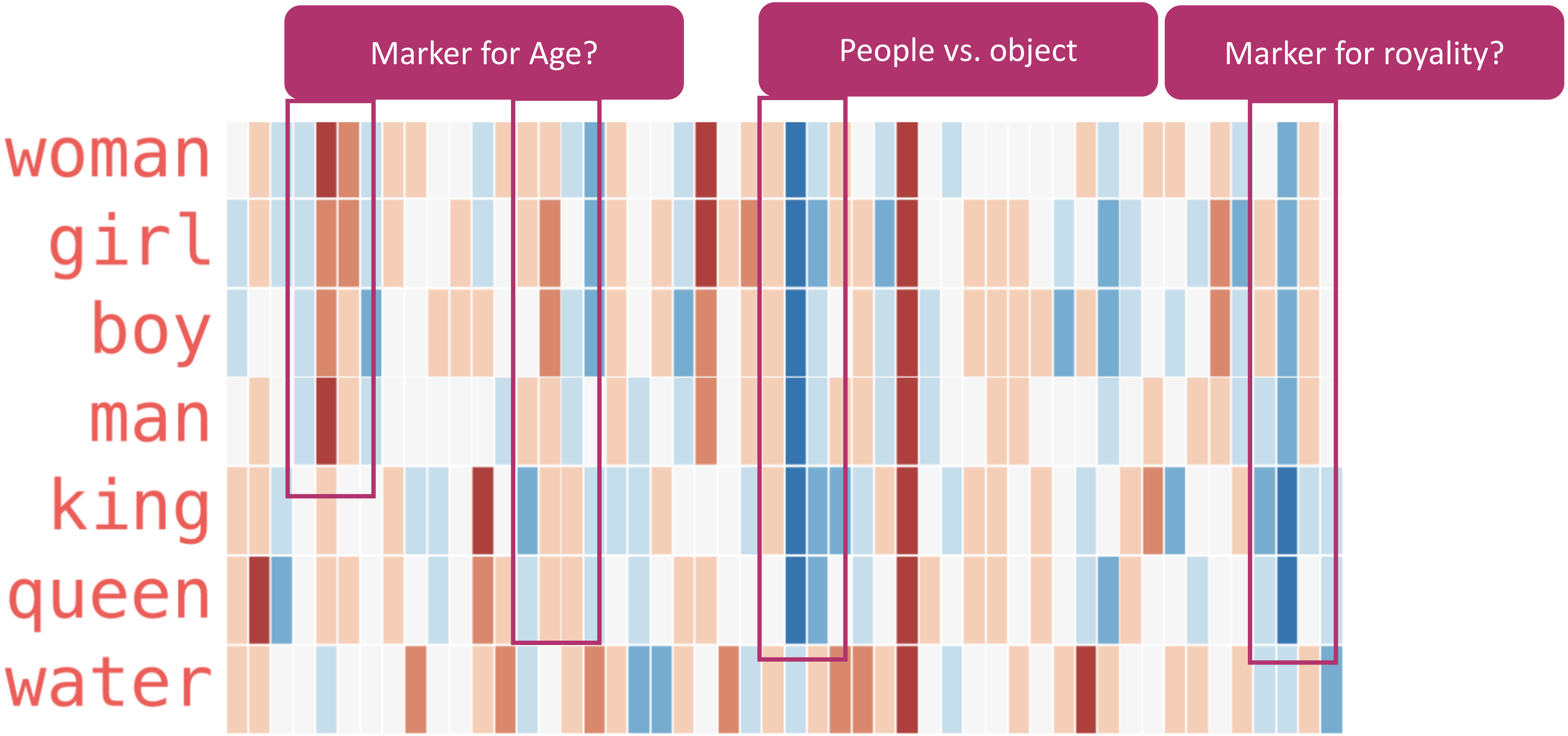
[3]



[3]

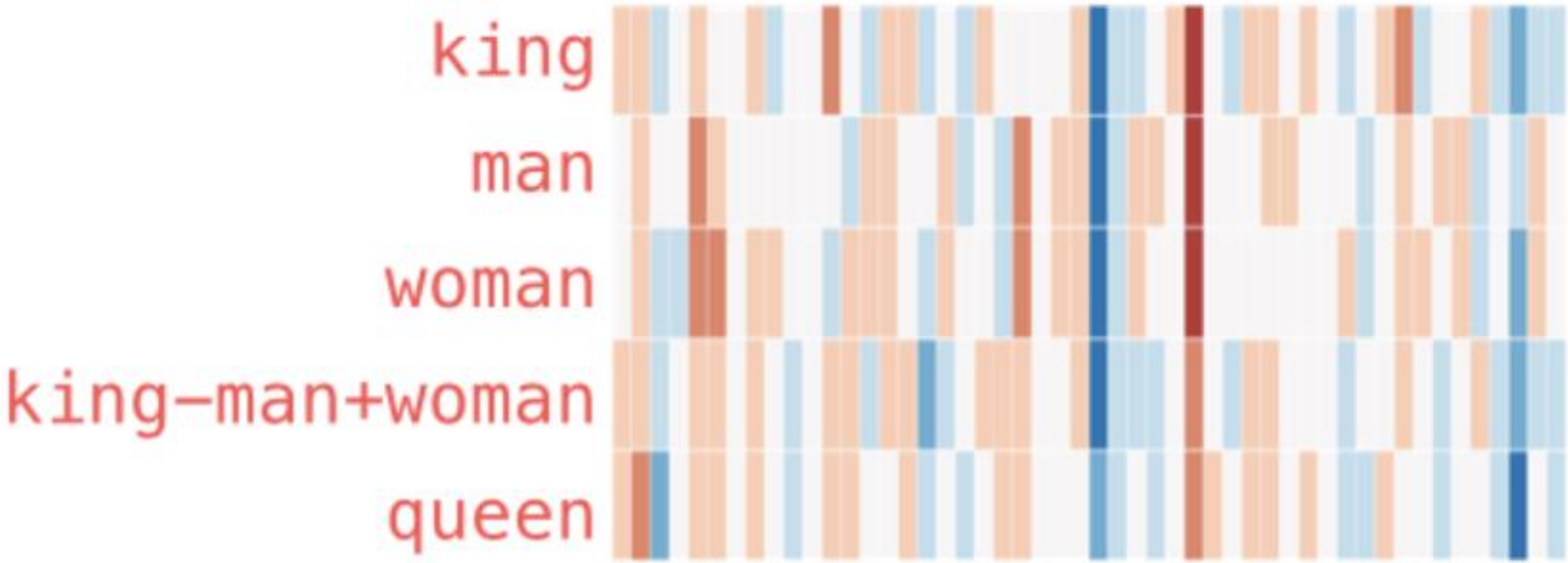

People vs. object
Marker for royality?
woman
girl
boy
man
king
queen
water

[3]


Marker for Age?
People vs. object
Marker for royality?
woman
girl
boy
man
king
queen
water

[3]

king – man + woman ~= queen



[3]

```python
def solve_analogies(A, B, C):
    fasttext = WordEmbeddings('crawl')
    result = compute_embedding_D(A, B, C, fasttext)
    vocab = get_embedding_english_vocab(fasttext)
    D = find_closest_matching_word(result, vocab, {A, B, C})

    return f'{A} is to {B} as {C} is to {D}'

#anal_solv = pn.Row(solve_analogies)
solve_analogies('king', 'man', 'queen')
```

```
'king is to man as queen is to woman'
```

Word A is to Word B
As
Word C is to Word D

```python
solve_analogies('Amsterdam', 'Netherlands', 'Paris')
```

```
'Amsterdam is to Netherlands as Paris is to France'
```

Different Embedding methods

Word2Vec (Mikolov et al.)

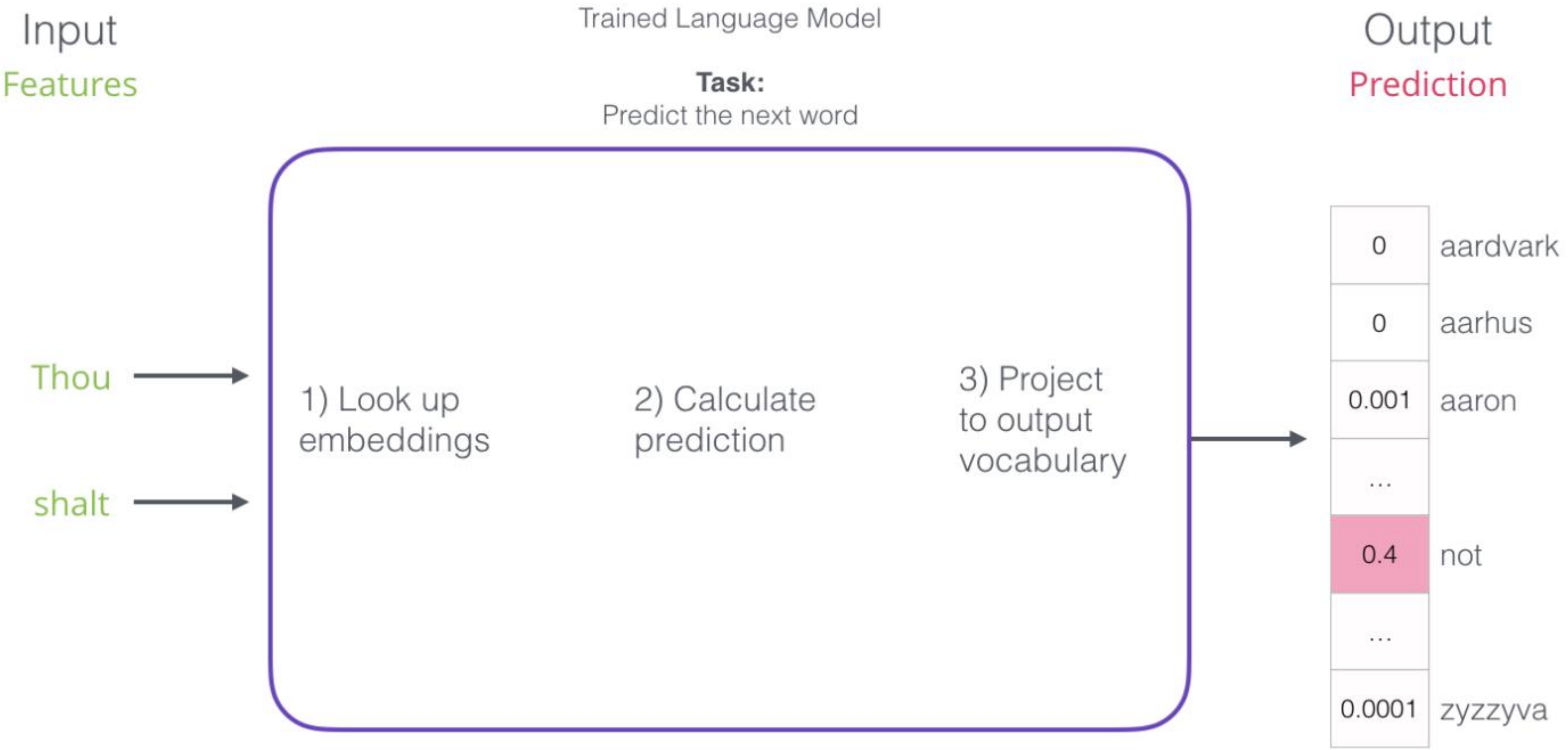
- Embedding for every word in corpus
- Sematics: consider direct neighbors
- Out of vocabulary words

FastText (Bojanowski et al.)

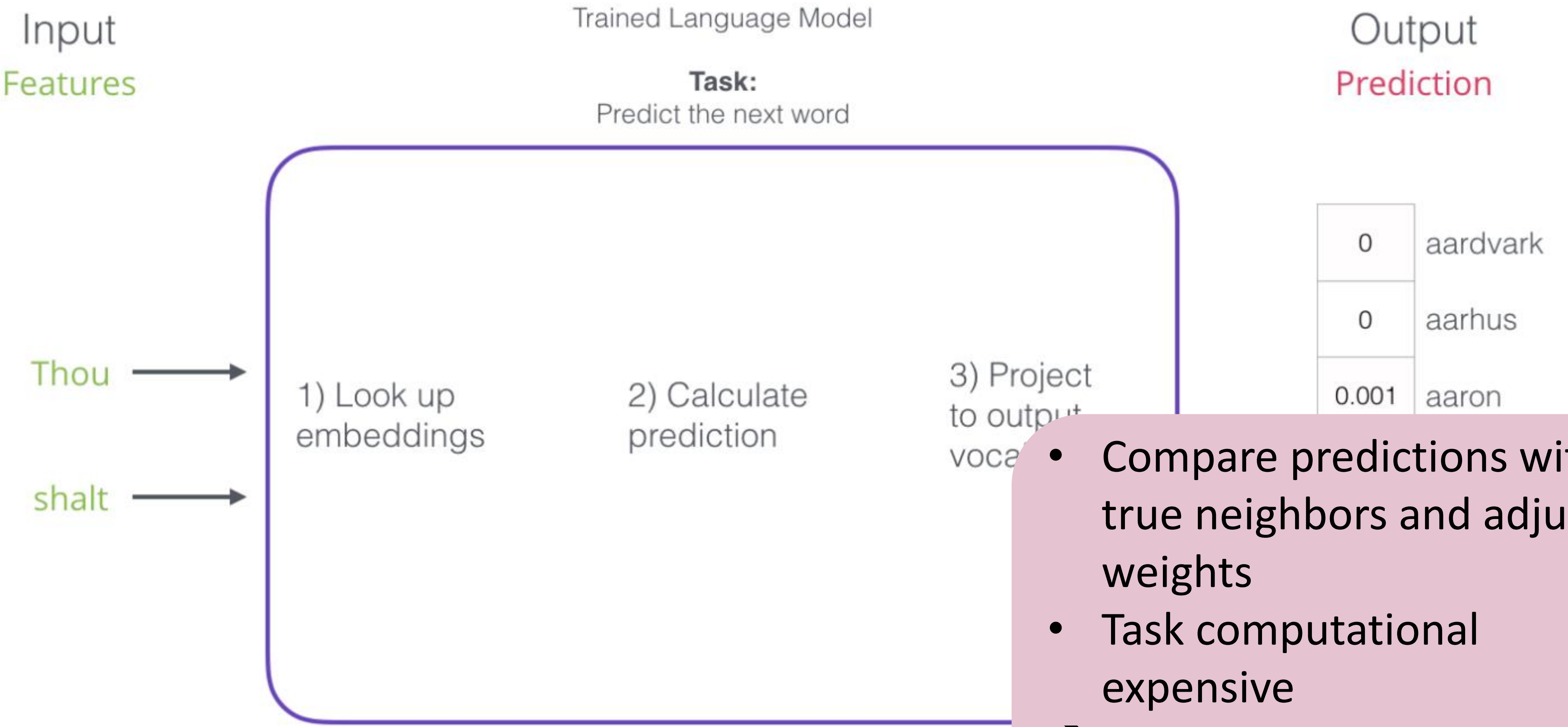
- Embedding for every word in corpus extended by subwords
- Sematics: consider direct neighbors

BERT

- Contextual embeddings
- Sematics: consider pairs of sentences


Input
Features
Trained Language Model
Task:
Predict the next word
Output
Prediction
Thou
shalt
1) Look up embeddings
2) Calculate prediction
3) Project to output vocabulary
0 aardvark
0 aarhus
0.001 aaron
...
0.4 not
...
0.0001 zyzzyva

[2]

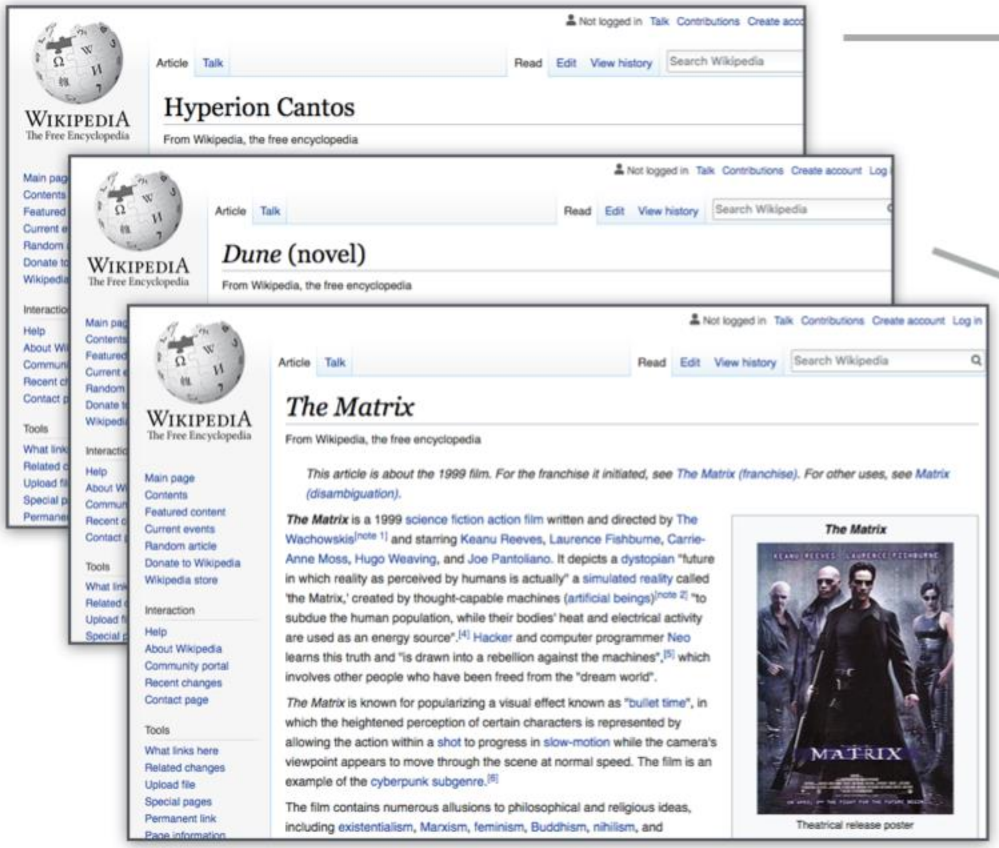


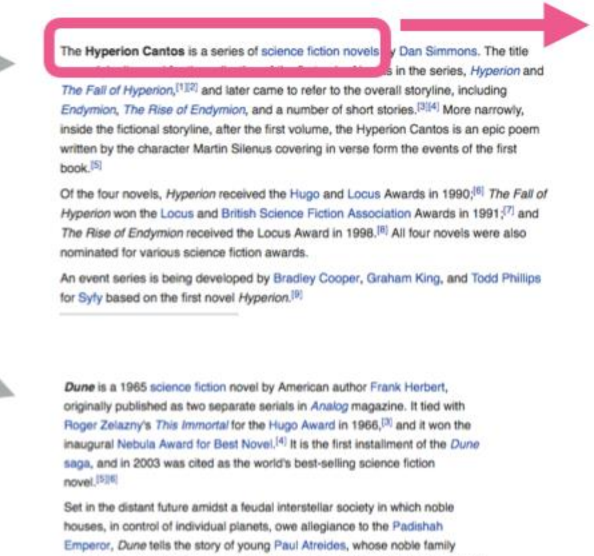
- Compare predictions with true neighbors and adjust weights
- Task computational expensive

➔ We want this to be a binary classification task


not
thou
Untrained Model
Task:
Are the two words neighbours?
0.90

[2]



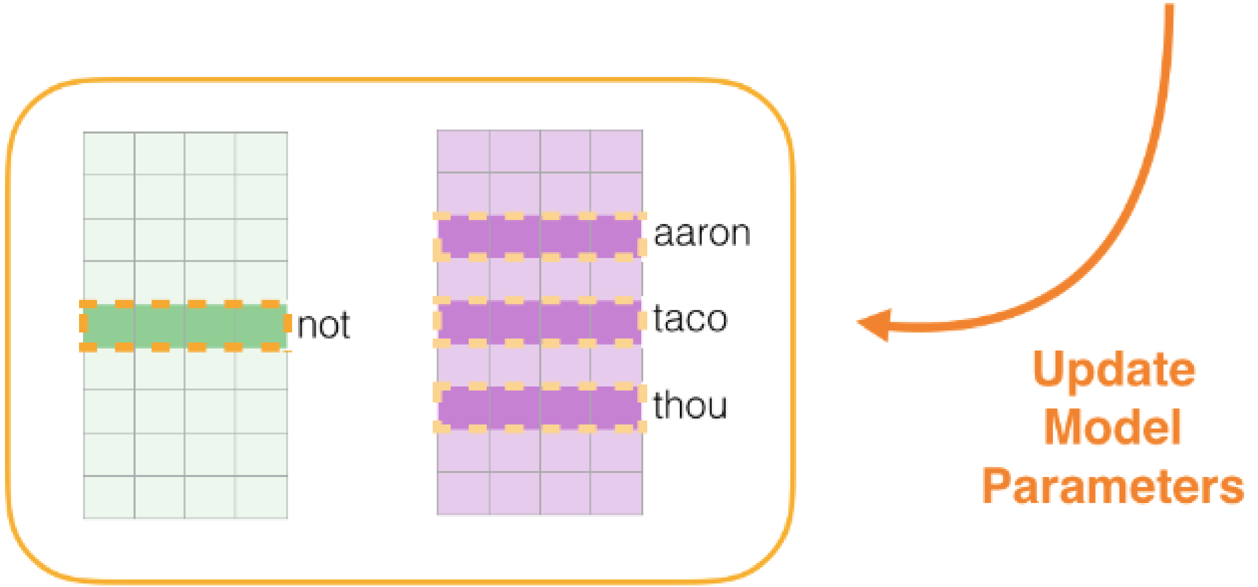


[2]

- Get a big corpus
- Slide window over corpus = positive samples
- Random negative samples

➔ Training examples

input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68



[2]

Continuous Bag of words (CBOW)

A quick brown fox jumps over the lazy dog

[3]

Predict central word based on neighbors

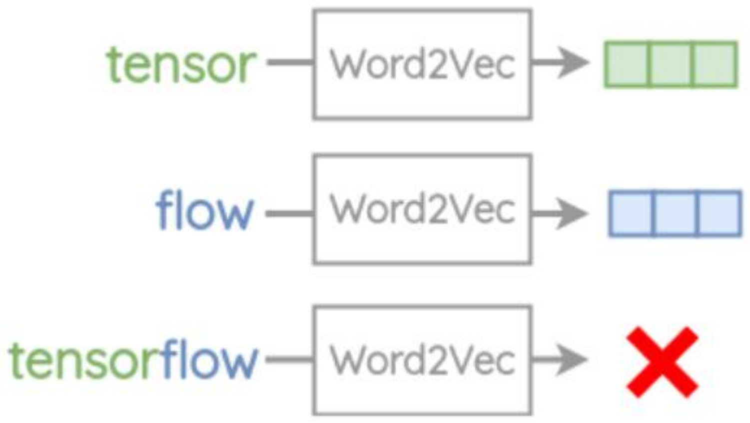
Skip-gram

A quick brown fox jumps over the lazy dog

[3]

Predict neighbor words from central word

Limitation: Out of Vocabulary Words cause problems



[2]

Limitation: Morphology, no Parameter sharing

[2]

Use internal structure to improve embeddings

Do skip-gram embeddings and obtain subwords for central word

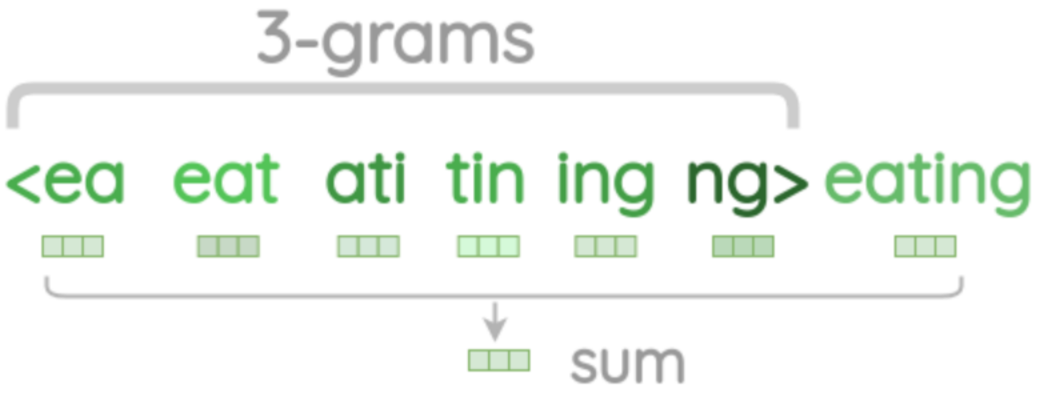
[2]

1.) Embed central word

Example Sentence

[2]

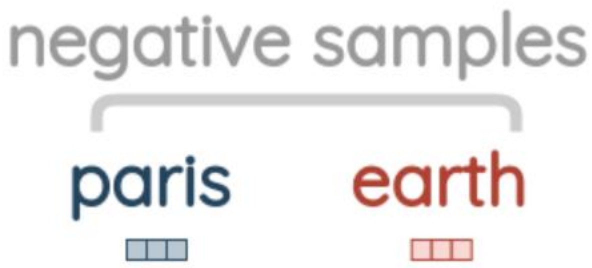
Central word embedding



[2]

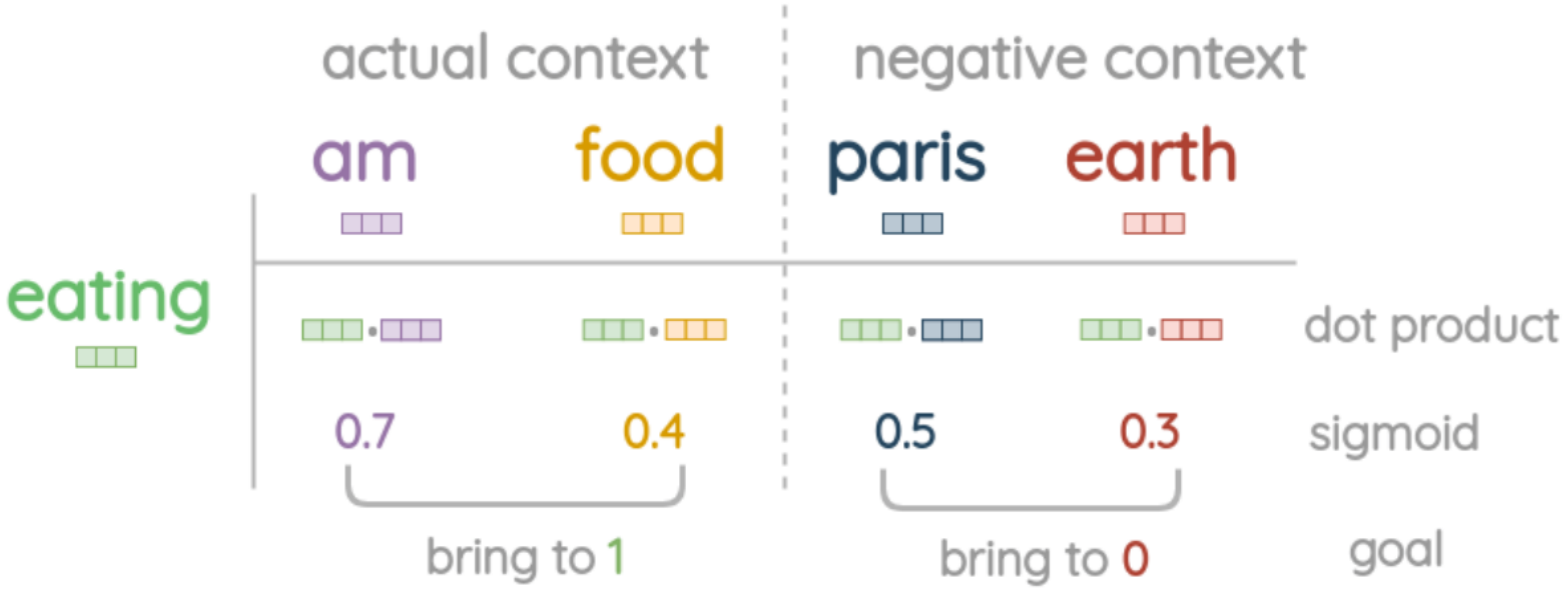
2.) Sampling

[2]



[2]

2.) Train the model



[2]

BERT embeddings: Bidirectional Encoder Representations from Transformer

BERT embeddings: Bidirectional Encoder Representations from Transformer

Contextual embeddings: the same word can have different embeddings based on context

BERT embeddings: Bidirectional Encoder Representations from Transformer

Contextual embeddings: the same word can have different embeddings based on context

Embeddings based on whole sentences

BERT embeddings: Bidirectional Encoder Representations from Transformer

Contextual embeddings: the same word can have different embeddings based on context

Embeddings based on whole sentences

Whole word embedding, subword embedding, character embedding

Multi-layer model → how to define the final embedding?

BERT embeddings: Bidirectional Encoder Representations from Transformer

- Word embedding:
  - Concatenate last four layers (3.072 dimensions)
  - Sum last for layers (768 dimensions)

BERT embeddings: Bidirectional Encoder Representations from Transformer

- Word embedding:
  - Concatenate last four layers (3.072 dimensions)
  - Sum last for layers (768 dimensions)

- Sentence embedding:
  - Average second – last hidden layer (768 dimensions)

BERT embeddings: Bidirectional Encoder Representations from Transformer

- Word embedding:
  - Concatenate last four layers (3.072 dimensions)
  - Sum last for layers (768 dimensions)

- Sentence embedding:
  - Average second – last hidden layer (768 dimensions)

Purpose:

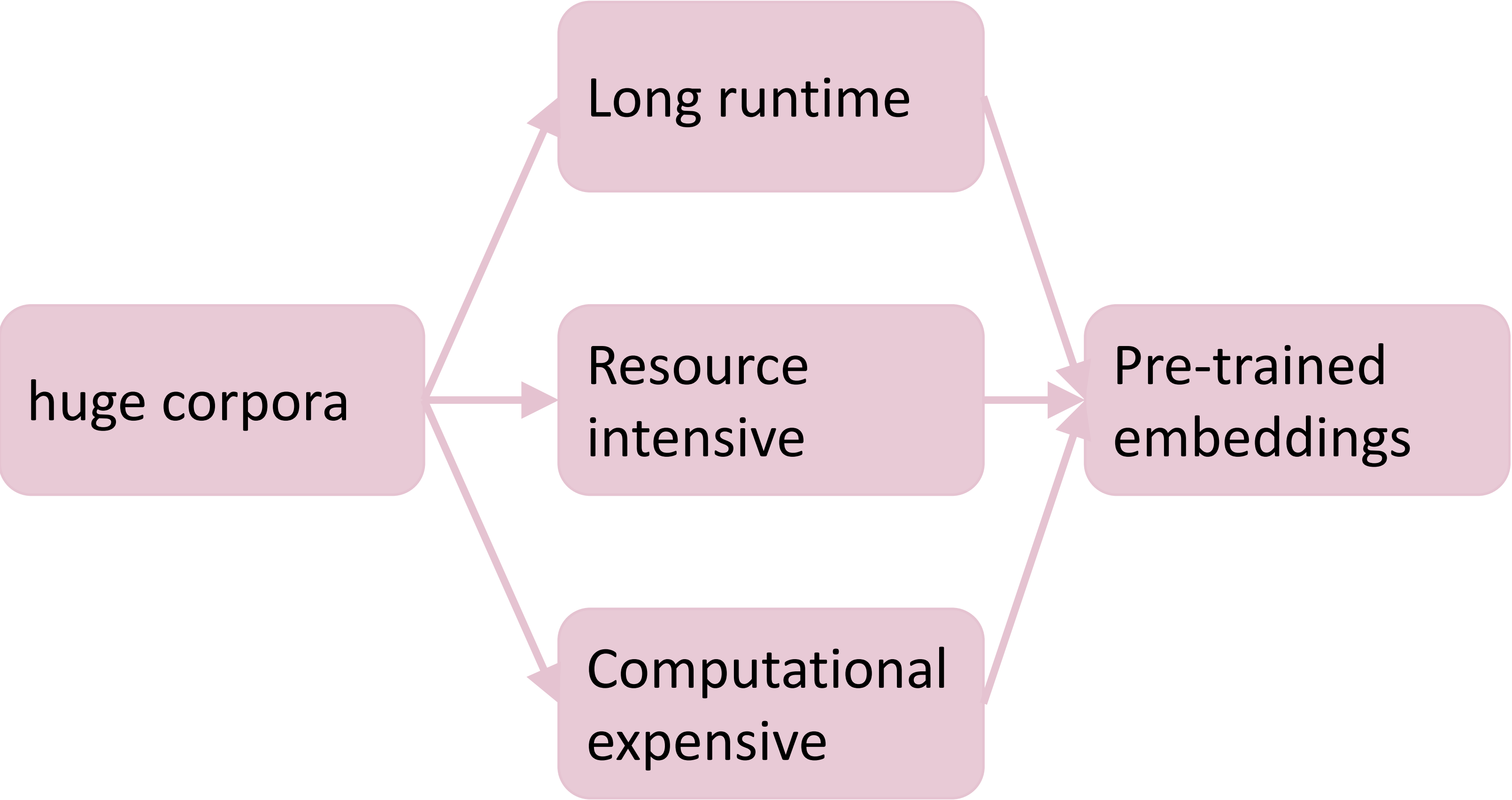
- Information retrieval without keyword or phrase overlap
- High-quality input features for downstream NLP tasks

Different Embedding methods, different performance

Word2Vec (Mikolov et al.)	FastText (Bojanowski et al.)	BERT
• Good performance on semantic analogy	• Improved performance on syntactic analogy • Worse performance on semantic analogy	• similarity comparison for words invalid • Similarity comparison for sentences valid

state-of-the-art performance:

- No out-of-vocabulary words
- Capture syntax and semantics



How does the magic of NLP comes to life?

- Text vectorization
- **Powerful NLP models**
- Application for my project

ELMO	BERT	Open-GPT
Allen AI	Google	OpenAI

[4]

[5]

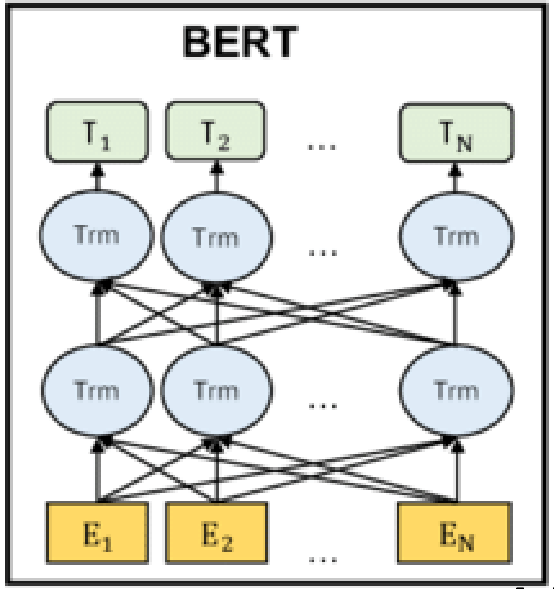
[6]

Powerful NLP models

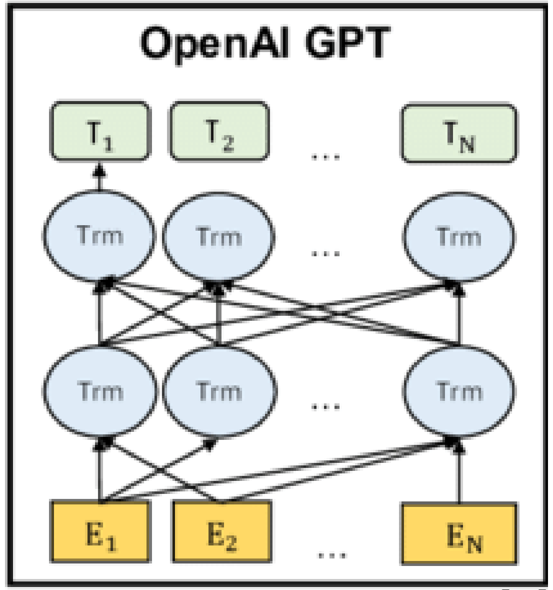

[5]

	BERT	Open-GPT
Model Architecture	• Transformers • bidirectional	• Transformers • Unidirectional (left)

[6]



[7]



[7]

 [5]

	BERT	Open-GPT [6]
Model Architecture	• Transformers • bidirectional	• Transformers • Unidirectional (left)
Task Type	Supervised e.g. text classification	Unsupervised e.g. text generation


[5]

[6]

	BERT	Open-GPT
Model Architecture	• Transformers • bidirectional	• Transformers • Unidirectional (left)
Task Type	Supervised e.g. text classification	Unsupervised e.g. text generation
Training data	masked language modelling, next sentence prediction	Language modelling


[5]

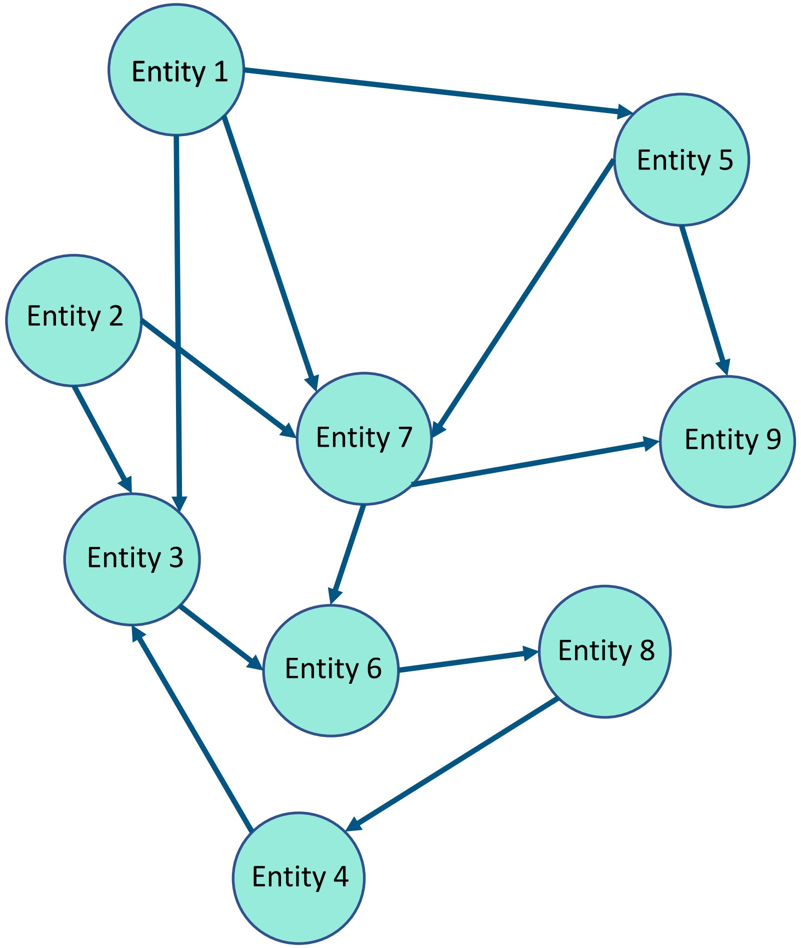
[6]

	BERT	Open-GPT
Model Architecture	• Transformers • bidirectional	• Transformers • Unidirectional (left)
Task Type	Supervised e.g. text classification	Unsupervised e.g. text generation
Training data	masked language modelling, next sentence prediction	Language modelling
Output	Fixed length embeddings for downstream NLP tasks	Sequence of tokens (variable length)

How does the magic of NLP comes to life?

- Text vectorization
- Powerful NLP models
- **Application for my project**

[8]



Biomedical Knowledge Graph

Entity 1 (Subject) – Verb → Entity 2 (Object)

Knowledge Filtering and Priorization tools:

- Certainty an Author expresses
- Polarity of the verb

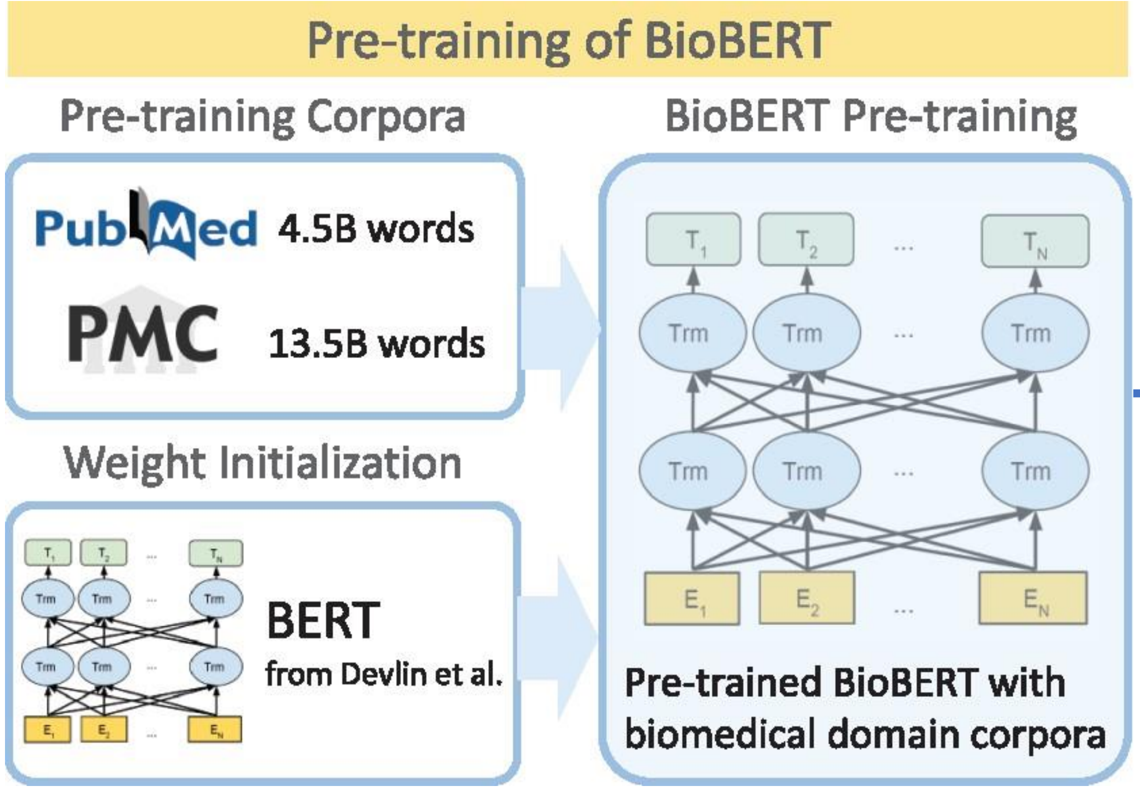
Previous studies have **shown** that entity 1 inhibits entity 2

Entity 1 **may** inhibit entity 2.

Certainty:
- Experimentally proven
- Proven by previous studies
- Always observed in setting

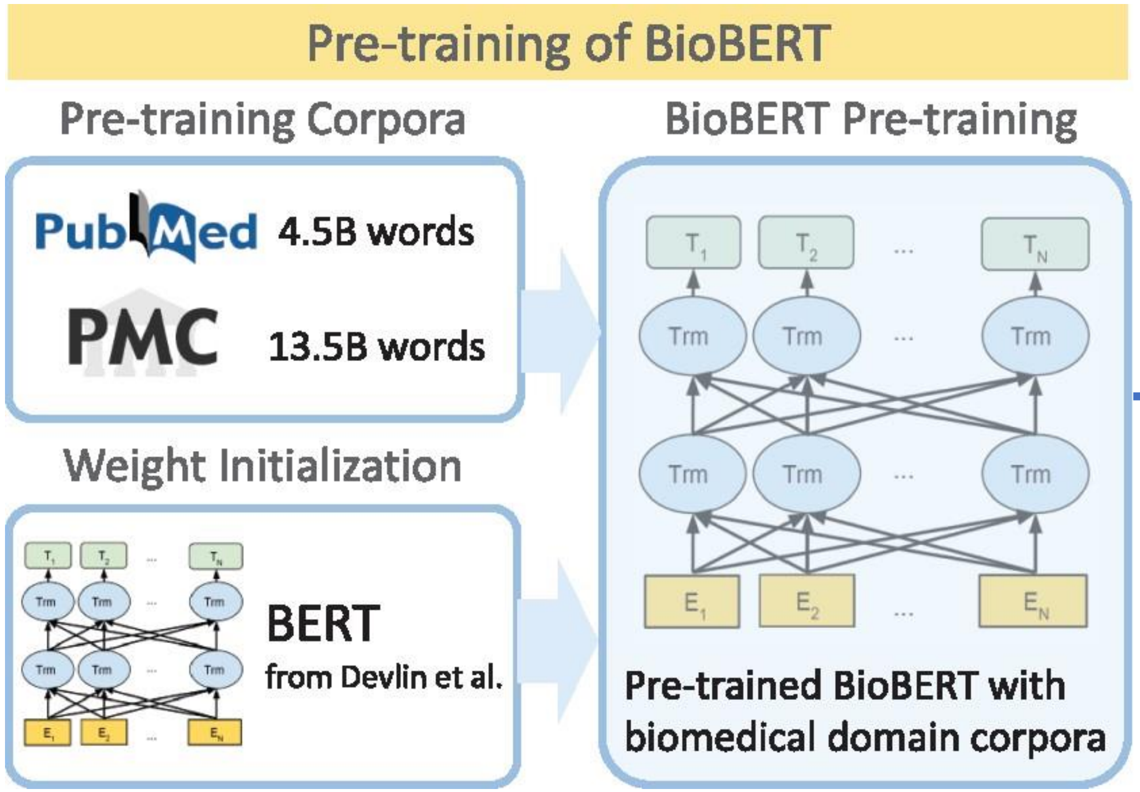
Uncertainty:
- Speculative terms
- Suggestive terms
- Usually observed but not always

TEXT CLASSIFICATION TASK:
- High-quality Sentence embeddings for biomedical text
- Downstream task: Classification
- Tagged dataset for classifier training



[9]

Document Embeddings



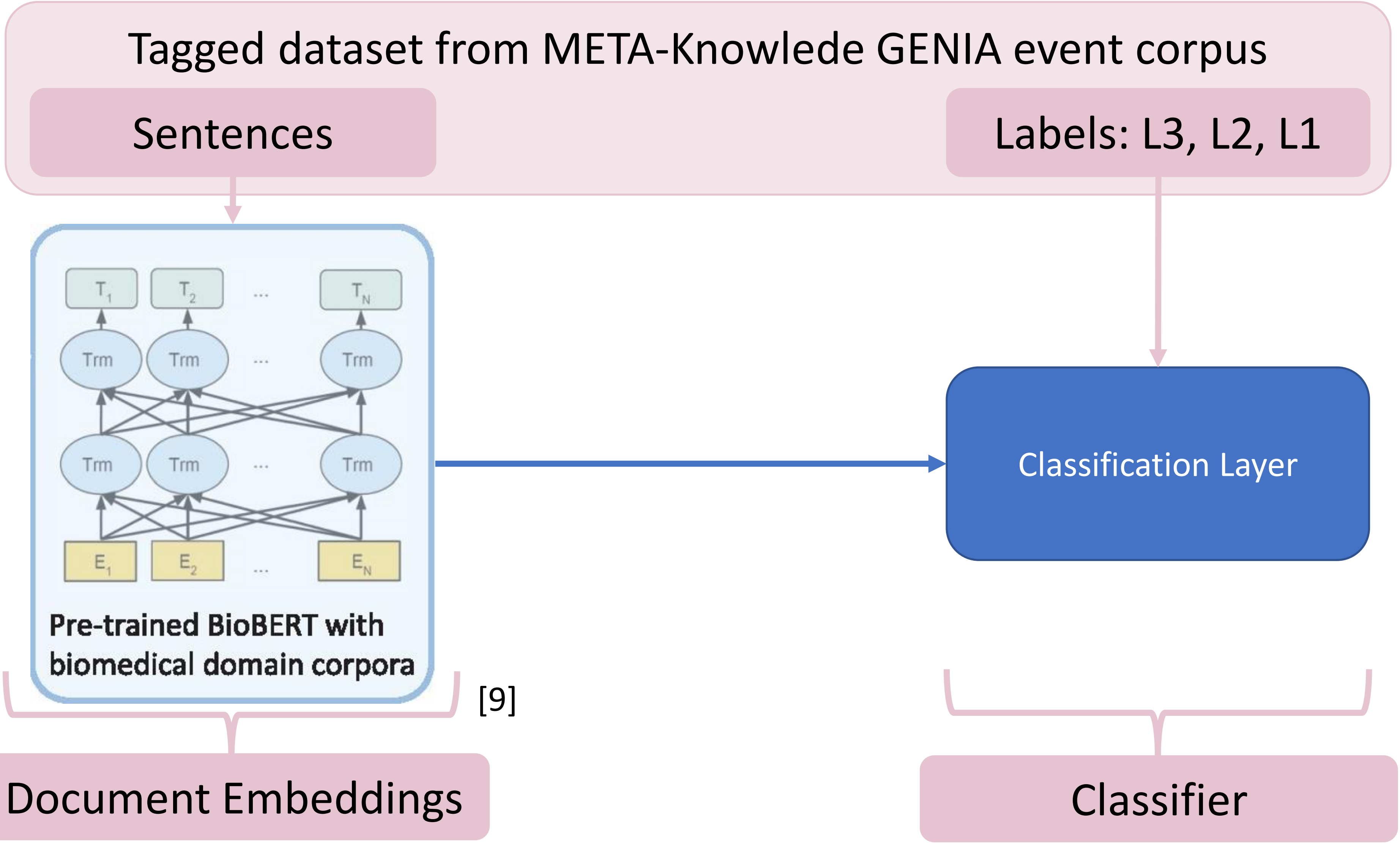
[9]

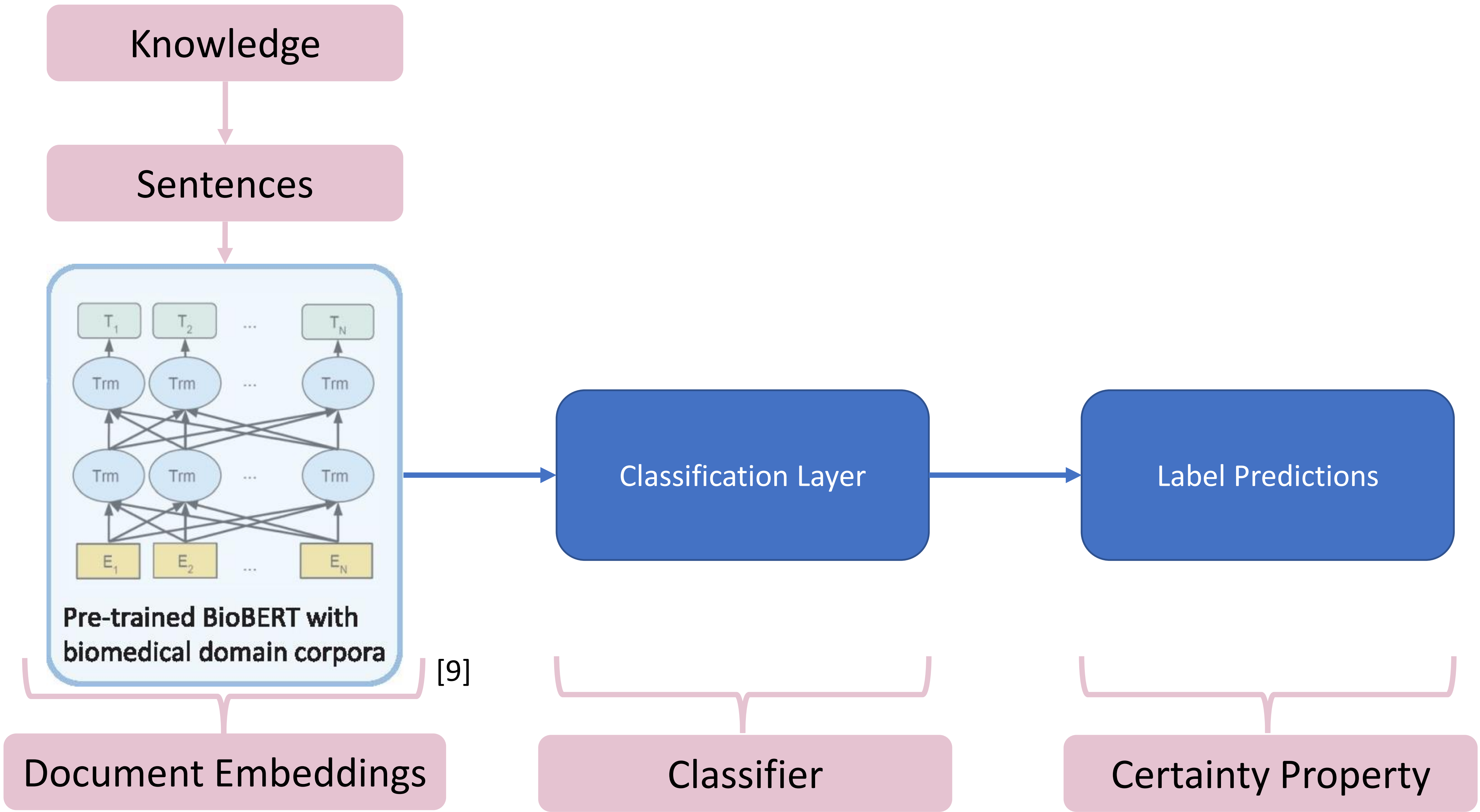
Linear Neural Network Layer

Class Label

Document Embeddings

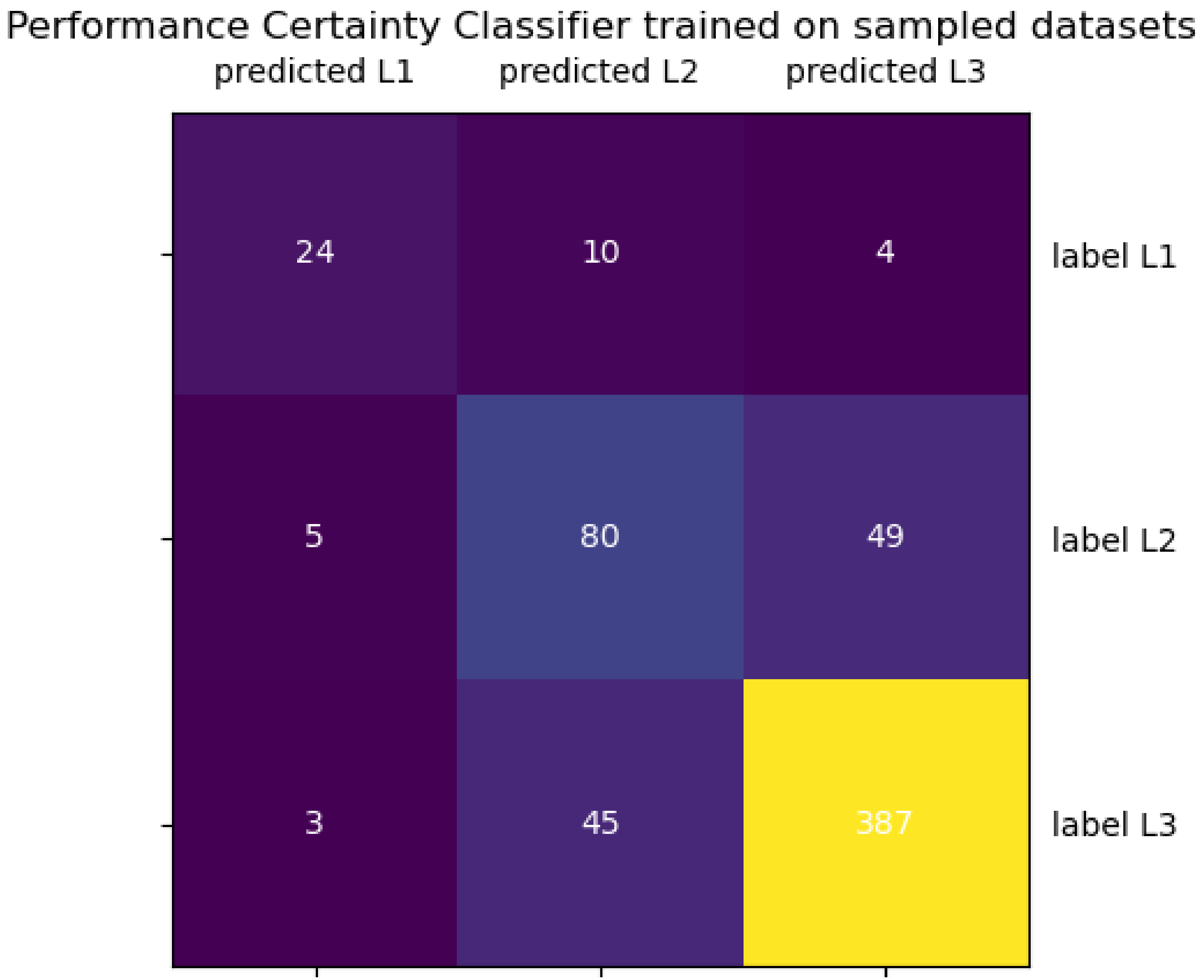
Classifier


Tagged dataset from META-Knowlede GENIA event corpus
Sentences
Labels: L3, L2, L1
T_1
T_2
...
T_N
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E_1
E_2
...
E_N
Pre-trained BioBERT with biomedical domain corpora
[9]
Classification Layer
Document Embeddings
Classifier



Knowledge
Sentences
T1 T2 ... TN
Trm Trm ... Trm
Trm Trm ... Trm
E1 E2 ... EN
Pre-trained BioBERT with biomedical domain corpora
[9]
Classification Layer
Label Predictions
Document Embeddings
Classifier
Certainty Property

- Total number of Sentences: 3.588
- Label Distribution:
    - L3: 37,21 %
    - L2: 52,15 %
    - L1: 10,65 %
- Overall Accuracy: 80,89 %



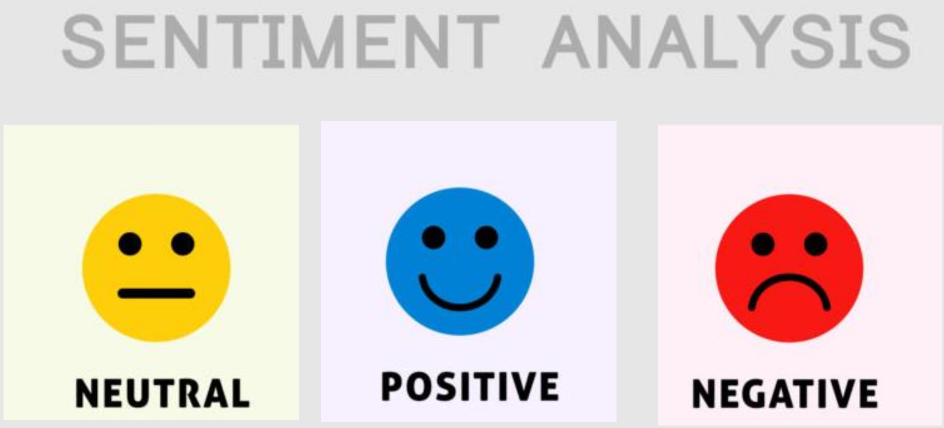
Bad Performance on L2

Good performance comes from many examples

Improve Dataset:
➔ Merge L1 and L2
➔ More restrictive with L3 sampling
➔ Different Corpus

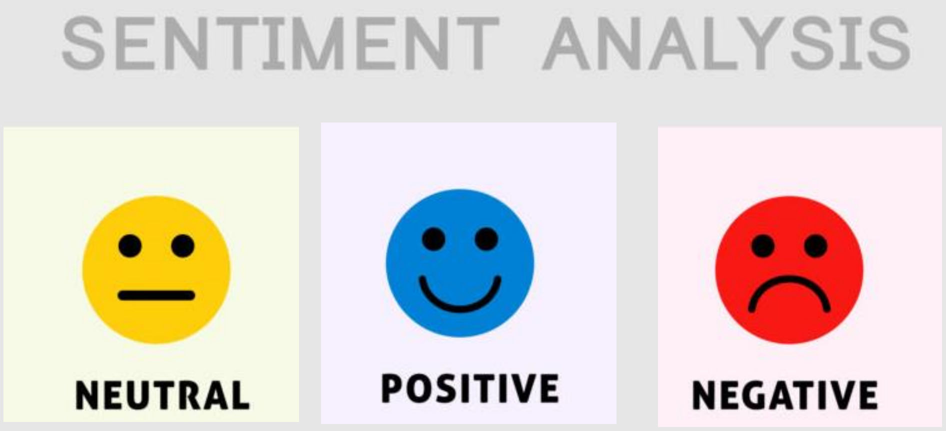
Improve Model:
➔ fine-tuning method instead of training
➔ Hyperparametertuning

Main limitations: domain and language specific, state-of-the-art results difficult to achieve



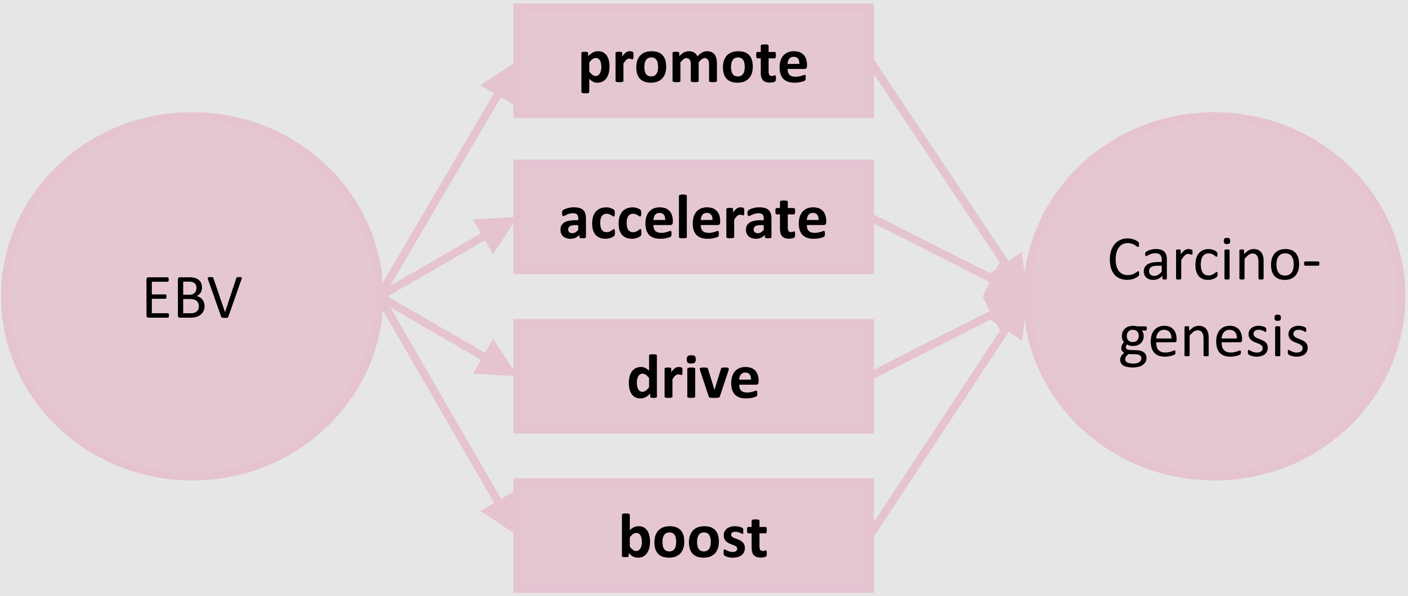
[10]

- Sentiment carrier:
  - Adjectives
  - Adjverbs
- Granularity:
  - Document
  - Sentence
  - Aspects
- Domains of application
  - Opinion mining from social media posts
  - Feedback from customer reviews
- Algorithmic approach:
  - Rule-based
  - ML
  - hybrid



- Adjective driven
- Tweets, Customer Reviews

[10]



- Verb based
- Scientific Vocabulary

Application for my project

Custom Polarity tagging Algorithm

Verb Categorisation

Verb tagging Rate: 99%

Misclassifications

KNOWLEDGE

PubMedID	source	verb	target	polarity
31316604	foxp3 silencing	inhibit	cell growth	-

PubMedID	source	verb	target	polarity
35532158	actin assembly sites	promotes	er autophagy	+

PubMedID	source	verb	target	polarity
35498032	coenzyme q	decreased	myocardial infarct size	-

PubMedID	source	verb	target	polarity
35534864	glycolysis	accelerated	apoptosis cells	-

Improve Polarity Resource:

→ Ruleset for manual Annotations
→ Restrict synonym antonym Resource

Quality of Resource is key for correct tagging

Domain Transfer is challenging

Concluding Remarks

- **Limitations**
- My favourite NLP tools

Custom solutions too expensive and
no state-of-the-art performance

Work with the Resources given:
Domain transfer challenging, suboptimal results

Language is not precise: exceptions to be handled

Concluding Remarks

- Limitations
- **My favourite NLP tools**

spaCy [11]

Sci spaCy [12]

Outstanding performance for Lemmatization
Biomedical sequence tagging with scientific models

flair [13]

Own embedding method
Interface for third-party model use especiall for text classification

TextBlob [14]

NLTK

GENSIM
topic modelling for humans

[15]

Do you see me now?

References:
[1] spam filter image: https://i0.wp.com/www.metronetworksllc.com/wp-content/uploads/2018/08/iStock-538057636.jpg?fit=2510%2C1194&ssl=1
[2] A visual Guide to FastText Embeddings: https://amitness.com/2020/06/fasttext-embeddings/
[3] The illustrated Word2vec: https://jalammar.github.io/illustrated-word2vec/
[4] ELMO image: https://static.smalljoys.me/2020/04/img_5e8f13ed41e91.png
[5] BERT image: https://i1.wp.com/jacobiem.org/wp-content/uploads/2020/10/Bert.jpg
[6] GPT image: https://mixed.de/wp-content/uploads/2019/03/open_ai_lp_logo.jpg
[7] BERT vs GPT image: https://www.researchgate.net/publication/340797092_Recent_Trends_in_Deep_Learning_Based_Open-Domain_Textual_Question_Answering_Systems/figures?lo=1
[8] PubMed: http://gomerpedia.org/images/thumb/1/10/PubMed_Logo.jpg/600px-PubMed_Logo.jpg
[9] BioBERT: https://academic.oup.com/view-large/figure/394146824/BIOINFORMATICS_36_4_1234_f1.png
[10] Sentiment analysis: https://www.expressanalytics.com/wp-content/uploads/2021/06/sentimentanalysishotelgeneric-2048x803-1.jpg
[11] spaCy: https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Fcdn.analyticsvidhya.com%2Fwp-content%2Fuploads%2F2020%2F03%2Flogo.jpg&f=1&nofb=1&ipt=e302d95cf8cf666fd7b986920eb64ad92db863a3c7e15207b864f217a3ceeca4&ipo=images
[12] ScispaCy: https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Fraw.githubusercontent.com%2Fallenai%2Fscispacy%2Fmaster%2Fdocs%2Fscispacy-logo.png&f=1&nofb=1&ipt=74c2da01bc4c1d01842d993131ef45b3309c5bb97269d4067336c536d1738a37&ipo=images
[13] flair: https://i.pinimg.com/originals/b3/76/fa/b376fa02b4699f22b4f9ec2d314a4f13.png
[14] textblob: https://unipython.com/wp-content/uploads/2018/03/An%C3%A1lisis-de-sentimientos-con-Python-min-1316x547.png
[15] GENSIM: https://tech.clickdo.co.uk/wp-content/uploads/2021/07/Gensim.jpg