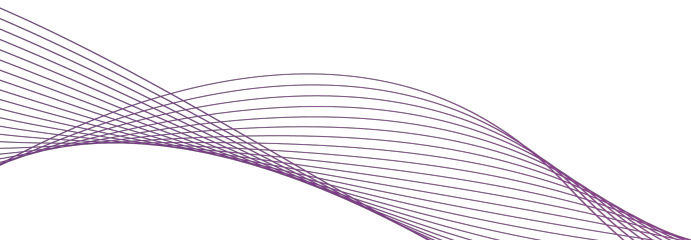


# Sodobna obdelava naravnega jezika: BERT prek praktičnih primerov



# Kdo sva?




Luka Vranješ



Andrej Miščič

## VALIRA AI

Connect via **LinkedIn**

Write us an 

# Akademija umetne inteligence za poslovne aplikacije

## **Danes:**

*Sodobna obdelava naravnega jezika: BERT prek praktičnih primerov*

*Komuniciranje aplikacij in jezikovnih modelov: GPTs*

## **Naslednjič:**

*Komuniciranje aplikacij in jezikovnih modelov: ChatGPT funkcije*

*Sodobna obdelava naravnega jezika: nadgradnja ChatGPT-ja s knjižnico LangChain*

# Agenda

## 00 Teoretični uvod:

- naloge obdelave naravnega jezika;
- predstavitve besedil;
- BERT (vs. GPT).

## 01 HuggingFace ekosistem

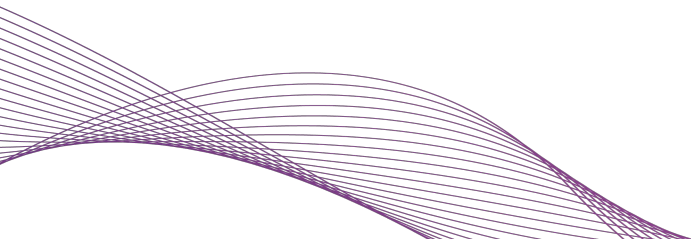
## 02 Praktični del:

- iskanje podobnih besedil;
- klasifikacija;
- iskanje imenskih entitet.

**Kasneje:** GPTs

# Obdelava naravnega jezika

*(natural language processing - NLP)*




# Obdelava naravnega jezika

Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.

# Obdelava naravnega jezika

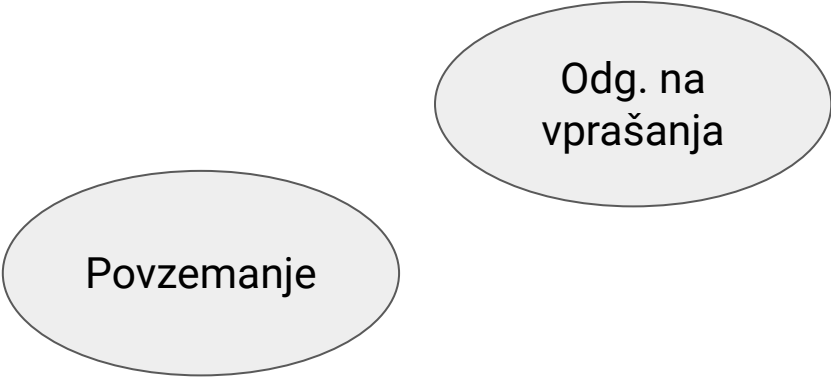
Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.



Povzemanje

# Obdelava naravnega jezika

Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.



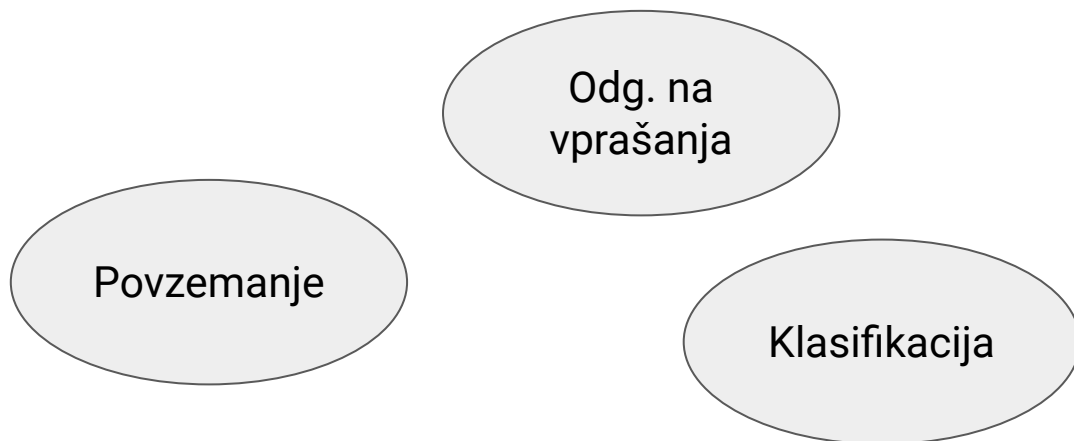
Povzemanje

Odg. na  
vprašanja



# Obdelava naravnega jezika

Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.



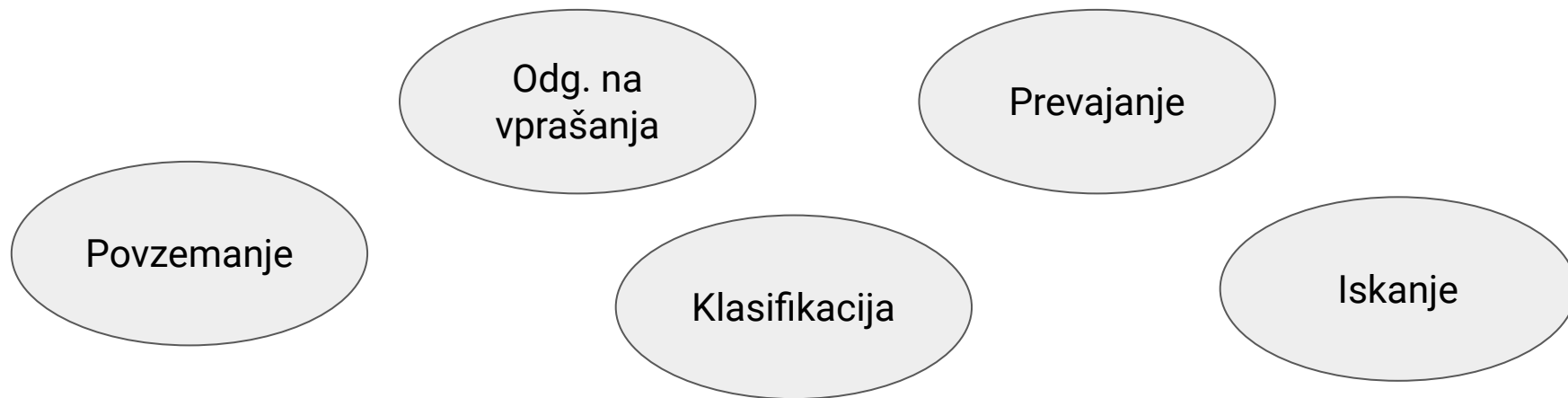
# Obdelava naravnega jezika

Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.



# Obdelava naravnega jezika

Področje, ki se ukvarja s problemi, kjer so  
**vhodi** in/ali **izhodi** v obliki besedil.



# Naloge obdelave naravnega jezika

## Razumevanje jezika

- klasifikacija besedil;
- zaznava imenskih entitet;
- bralno razumevanje;
- ...

## Generiranje jezika

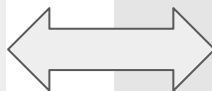
- strojno prevajanje;
- povzemanje;
- odgovarjanje na vprašanja;
- ...

# Naloge obdelave naravnega jezika

## Razumevanje jezika

- klasifikacija besedil;
- zaznava imenskih entitet;
- bralno razumevanje;
- ...

Lahko rešujemo tudi z generativnimi pristopi.



## Generiranje jezika

- strojno prevajanje;
- povzemanje;
- odgovarjanje na vprašanja;
- ...

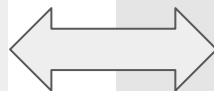
Zahteva razumevanje jezika.

# Naloge obdelave naravnega jezika

## Razumevanje jezika

- klasifikacija besedil;
- zaznava imenskih entitet;
- bralno razumevanje;
- ...

Lahko rešujemo tudi z generativnimi pristopi.



## Generiranje jezika

- strojno prevajanje;
- povzemanje;
- odgovarjanje na vprašanja;
- ...

Zahteva razumevanje jezika.

LLM-ji so odlični pri obeh tipih nalog.

# Praktični primeri

1. Iskanje (podobnih besedil)
2. Klasifikacija besedil
3. Zaznava imenskih entitet

# Praktični primeri

1. **Iskanje (podobnih besedil)**
2. Klasifikacija besedil
3. Zaznava imenskih entitet

## ISKANJE:

Pridobivanje dokumentov, ki so relevantni za uporabnikovo poizvedbo.

- **vhod:**  
poizvedba

- **izhod:**  
relevantni dokumenti (ali kaj drugega)



# Praktični primeri

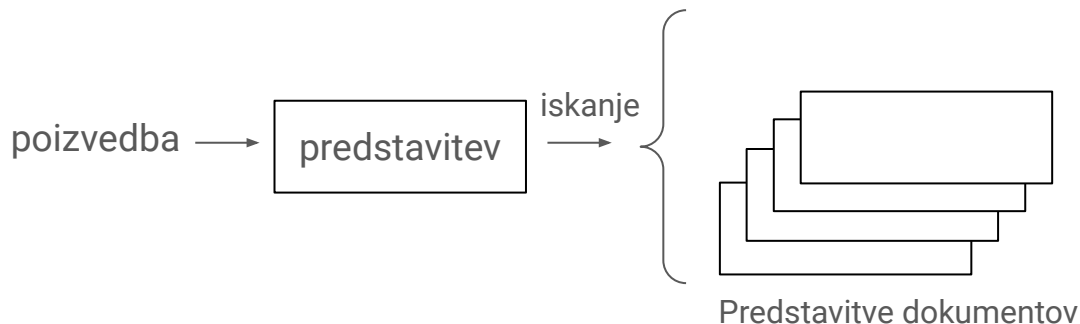
1. **Iskanje (podobnih besedil)**
2. Klasifikacija besedil
3. Zaznava imenskih entitet

## ISKANJE:

Pridobivanje dokumentov, ki so relevantni za uporabnikovo poizvedbo.

- **vhod:**  
poizvedba

- **izhod:**  
relevantni dokumenti (ali kaj drugega)



# Praktični primeri

1. **Iskanje (podobnih besedil)**
2. Klasifikacija besedil
3. Zaznava imenskih entitet

## ISKANJE:

### 1. primer:

- Spletna trgovina ponuja nekaj tisoč izdelkov.
- Uporabnik vpiše poizvedbo, iskanje vrne relevantne izdelke - tiste, katerih opis je najbolj podoben uporabnikovi poizvedbi.

### 2. primer:

- Uporabni si ogleduje enega od izdelkov.
- Okence *podobni izdelki* vsebuje izdelke, ki imajo podoben opis kot dotičen izdelek.

# Praktični primeri

1. Iskanje (podobnih besedil)
2. **Klasifikacija besedil**
3. Zaznava imenskih entitet

## KLASIFIKACIJA:

Določanje kategorije nekemu besedilu.

- **vhod:**

besedilo

- **izhod:**

kategorija / razred

# Praktični primeri

1. Iskanje (podobnih besedil)
2. **Klasifikacija besedil**
3. Zaznava imenskih entitet

## KLASIFIKACIJA:

Določanje kategorije nekemu besedilu.

- **vhod:**

besedilo

- **izhod:**

kategorija / razred



# Praktični primeri

1. Iskanje (podobnih besedil)
2. **Klasifikacija besedil**
3. Zaznava imenskih entitet

## KLASIFIKACIJA:

Določanje kategorije nekemu besedilu.

- **vhod:**

besedilo

- **izhod:**

kategorija / razred



# Praktični primeri

1. Iskanje (podobnih besedil)
- 2. Klasifikacija besedil**
3. Zaznava imenskih entitet

## KLASIFIKACIJA:

### 1. primer:

- Podjetje prejema večje število e-mailov svojih strank.
- Za vsako prejeto sporočilo klasificirajo odgovoren oddelek v podjetju (e.g. tehnična pomoč, splošna pomoč, računovodstvo, ...).

### 2. primer:

- Podjetje želi oceniti priljubljenost svoje znamke.
- Komentarje na družbenih omrežjih klasificirajo v pozitivne in negativne.

# Praktični primeri

1. Iskanje (podobnih besedil)
2. Klasifikacija besedil
3. Zaznava imenskih entitet

## ZAZNAVA IMENSKIH ENTITET:

Določanje kategorije vsaki besedi v besedilu.

- **vhod:**

besedilo

- **izhod:**

kategorija / razred vsake besede

Sva **Luka** in **Andrej**. Delava za **Valiro AI**.

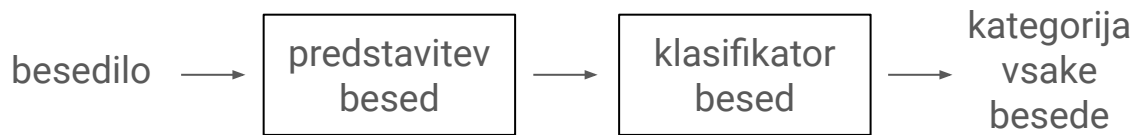
## Praktični primeri

1. Iskanje (podobnih besedil)
2. Klasifikacija besedil
- 3. Zaznava imenskih entitet**

### ZAZNAVA IMENSKIH ENTITET:

Določanje kategorije vsaki besedi v besedilu.

Sva **Luka** in **Andrej**. Delava za **Valiro AI**.





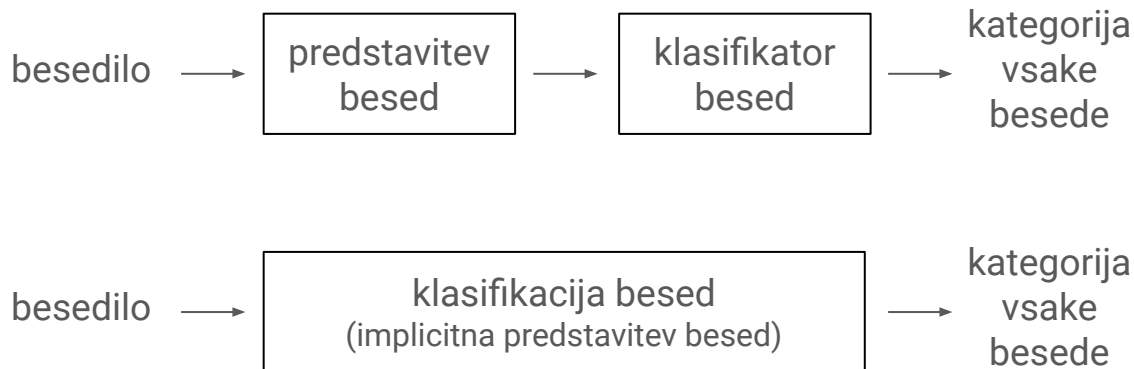
## Praktični primeri

1. Iskanje (podobnih besedil)
2. Klasifikacija besedil
- 3. Zaznava imenskih entitet**

### ZAZNAVA IMENSKIH ENTITET:

Določanje kategorije vsaki besedi v besedilu.

Sva **Luka** in **Andrej**. Delava za **Valiro AI**.



# Praktični primeri

1. Iskanje (podobnih besedil)
2. Klasifikacija besedil
- 3. Zaznava imenskih entitet**

## ZAZNAVA IMENSKIH ENTITET:

### 1. primer:

- Podjetje obdeluje besedila, ki jih želi javno objaviti - javna objava zahteva skritje osebnih podatkov.
- V besedil zaznajo imena oseb ter ostale osebne podatke ter jih anonimizirajo.

### 2. primer:

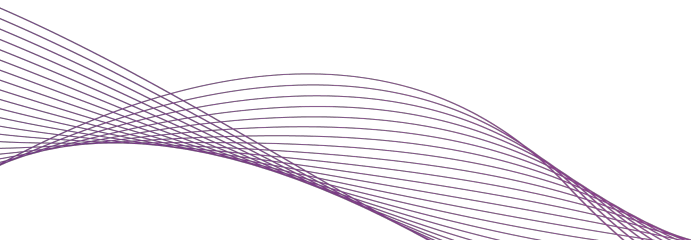
- Podjetje prejema večje število računov v PDF-jih, podatke želi v strukturirani obliki pripeljati v svoje ostale sisteme.
- Z zaznavo entitet v besedil ekstrahira strukturane informacije iz besedil računov.

# Povzetek

1. Iskanje (podobnih besedil) - na ravni **besedil**
2. Klasifikacija besedil - na ravni **besedil**
3. Zaznava imenskih entitet - na ravni **besed**

# Predstavitve

(besed, besedil, ...)



# Predstavitev besedil

= metode, ki pretvorijo besedilo v obliko, ki jo razume (in lahko obdeluje) računalnik

## **CILJ:**

Predstavitev besedila ohrani njegov (semantični) pomen.

*= dve besedili, ki imata podoben pomen, imata tudi podobno predstavitev.*

# Ilustrativen primer (iskanje podobnih besedil)

*Uporabnikovo vprašanje:*

- Kje leži mačka?

*Dokumenti:*

- Mačka leži na preprogi.
- V gozdu se nahaja medved.

# Predstavitev besedil (podobnost)

- Kje leži mačka?
- Mačka leži na preprogi.
- V gozdu se nahaja medved.

# Predstavitev besedil (podobnost)

Preštujemo, kolikokrat se posamezna beseda pojavi v besedilu.

	kje	leži	mačka	na	preprogi	v	gozdu	se	nahaja	medved
Kje leži mačka?	1	1	1	0	0	0	0	0	0	0
Mačka leži na preprogi.	0	1	1	1	1	0	0	0	0	0
V gozdu se nahaja medved.	0	0	0	0	0	1	1	1	1	1



## Predstavitev besedil (podobnost)

Kje leži mačka?      **in**      Mačka leži na preprogi.

*Ujemanje: 2 / 10*

Kje leži mačka?      **in**      V gozdu se nahaja medved.

*Ujemanje: 0 / 10*

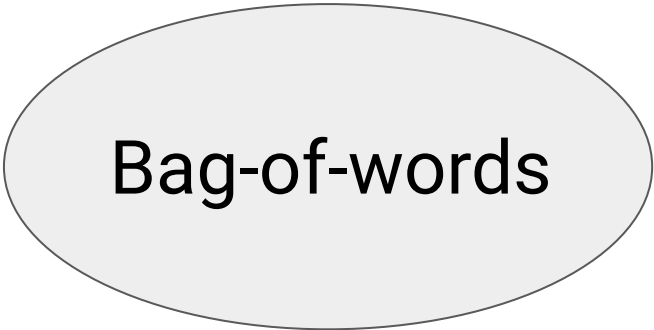
# Predstavitev besedil (podobnost)

Kje leži mačka?      **in**      Mačka leži na preprogi.

*Ujemanje: 2 / 10*

Kje leži mačka?      **in**      V gozdu se nahaja medved.

*Ujemanje: 0 / 10*



Bag-of-words

# Bag-of-words (*vreča besed*) v praksi

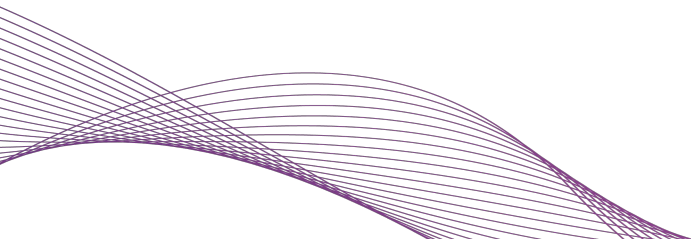
Besede v besedilu imajo različno pomembnost:

- odstrani nepomembne besede (t.i. stop words);
- uteži besede glede na njihovo "*pomembnost*", e.g. tf-idf ali BM25.

tf-idf = tf \* idf



# Praktiční primer (Jupyter notebook)



# Slabosti vreč *besed*

- sopomenke;
- nerazumevanje pomena;
- neupoštevanje vrstega reda;
- ...

# Slabosti *bag-of-words*

- **sopomenke;**
- nerazumevanje pomena;
- neupoštevanje vrstega reda;
- ...

- *advokat* in *odvetnik*
- *aerodrom* in *letališče*
- *hiter* in *uren*
- *ideja* in *zamisel*

Enak pomen, vendar povsem drugačna predstavitev z *bag-of-words* pristopi.

## Slabosti *bag-of-words*

- sopomenke;
- **nerazumevanje pomena;**
- neupoštevanje vrstega reda;
- ...

Predstavitev je zgrajena na podlagi besed v besedilu, kar ne nujno ohrani pomena.

Prejšnji primer:

- ~~Kje leži mačka?~~ Kje se nahaja mačka?
- Mačka leži na preprogi.
- V gozdu se nahaja medved.

# Kako “dodati” pomen v predstavitev?

*“You shall know a word by the company it keeps.”*

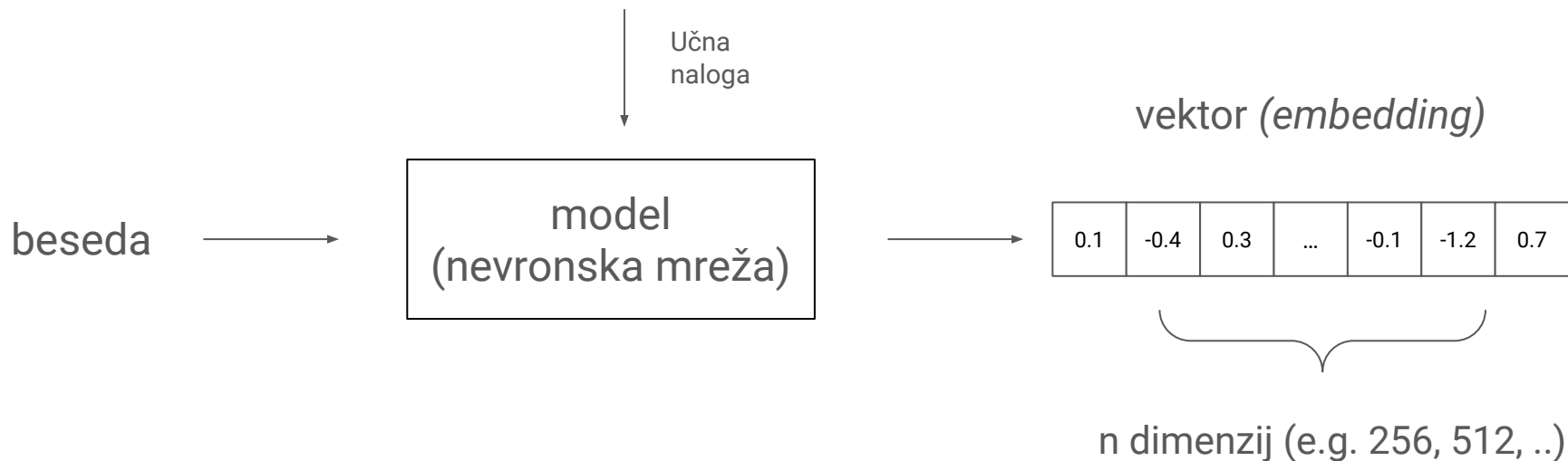
*- John Rupert Firth (linguist)*



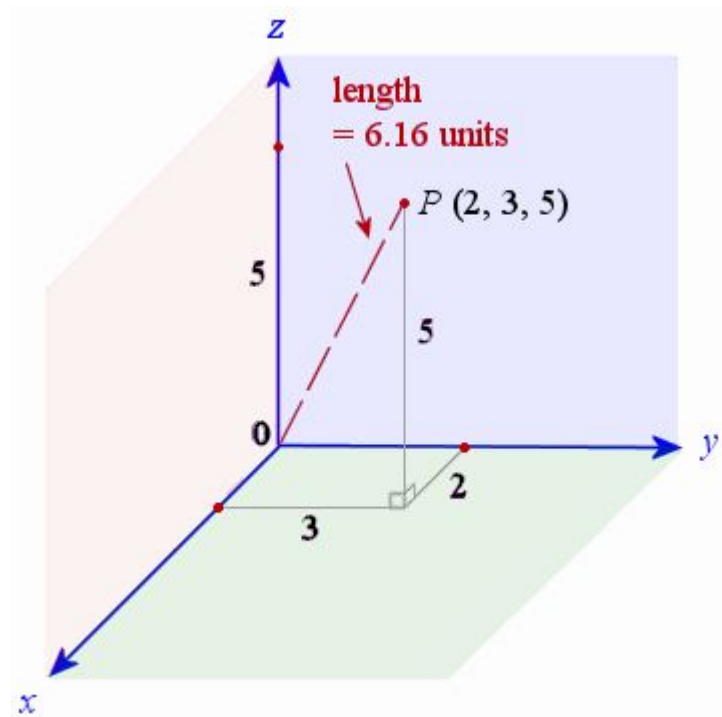
# Kako “dodati” pomen v predstavitev?

*“You shall know a word by the company it keeps.”*

*- John Rupert Firth (linguist)*



# Intermezzo: vektorji



# word2vec

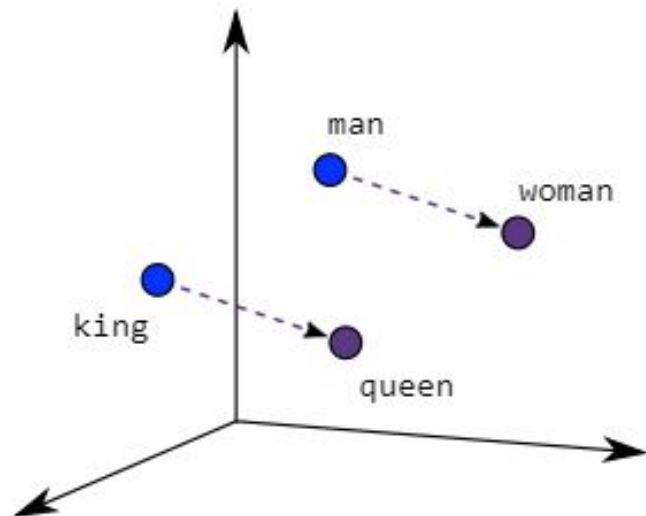
- Model je naučen na veliki zbirki besedil.
- Za vsako besedo imamo vektor (t.i. *word embedding*)

0.1	-0.4	0.3	...	-0.1	-1.2	0.7
-----	------	-----	-----	------	------	-----

- Če delamo z besedili, moramo nekako združiti vektorje besed.

# word2vec

- *embedding* nosi semantični pomen besed.



king - man + woman = queen

# Slabost(i) *word2vec*-a

Ne upošteva konteksta:

- ista beseda je lahko različno *pomembna* v različnih kontekstih.
- enaka beseda ima lahko različen *pomen* v različnih kontekstih.

## Primer enakopisnic:

Fižol je posadil v ***prst***.

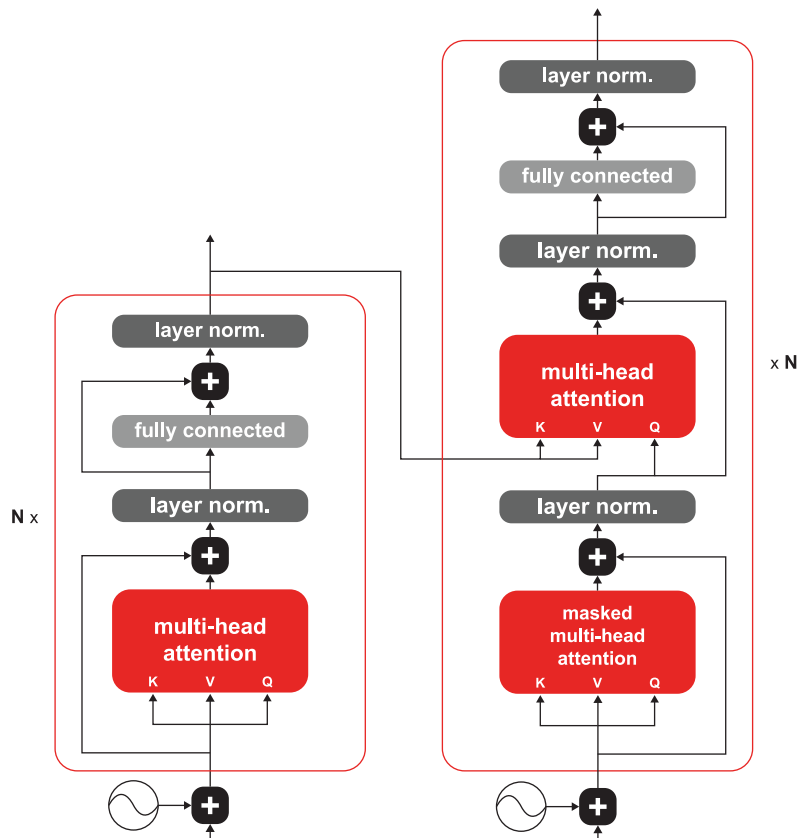
Med kuhanjem se je vrezal v ***prst***.

# Rešitev

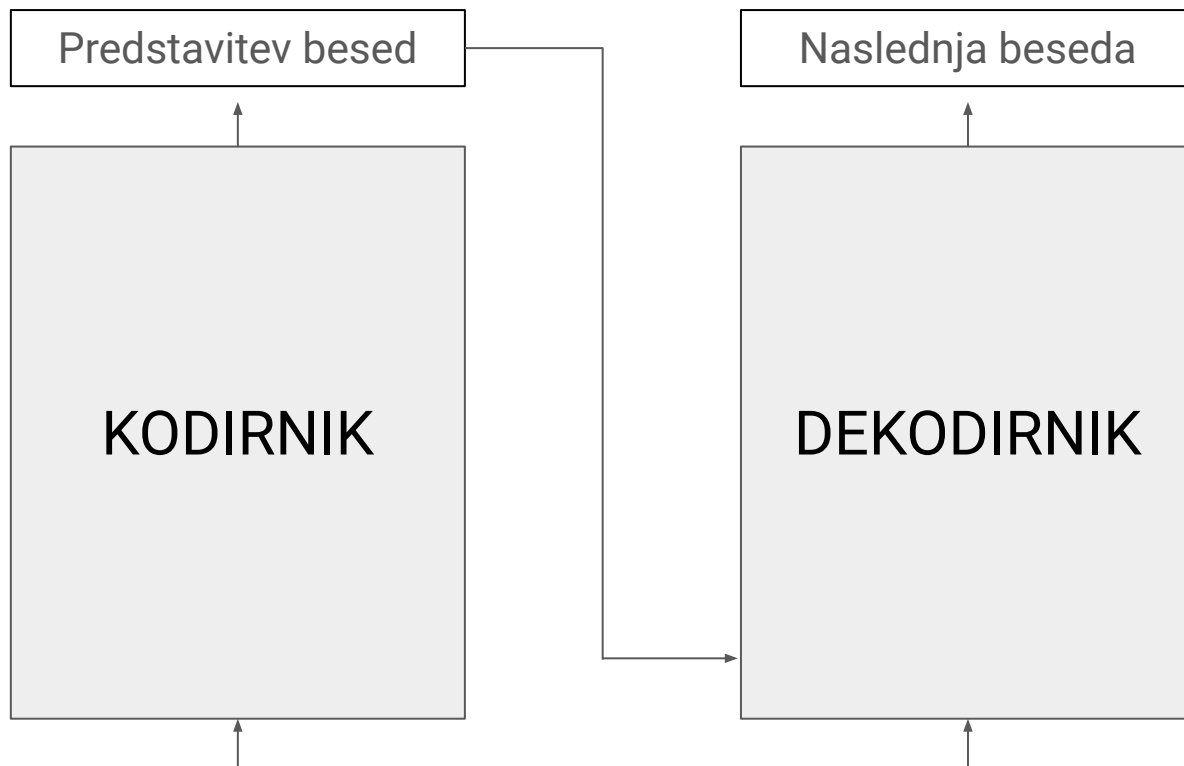
## Kontekstualne predstavitve besedil:

- Predstavitev besed/besedila je odvisna od konteksta, v katerem se pojavi.
- Danes bazirajo na arhitekturi transformer.

# Transformer

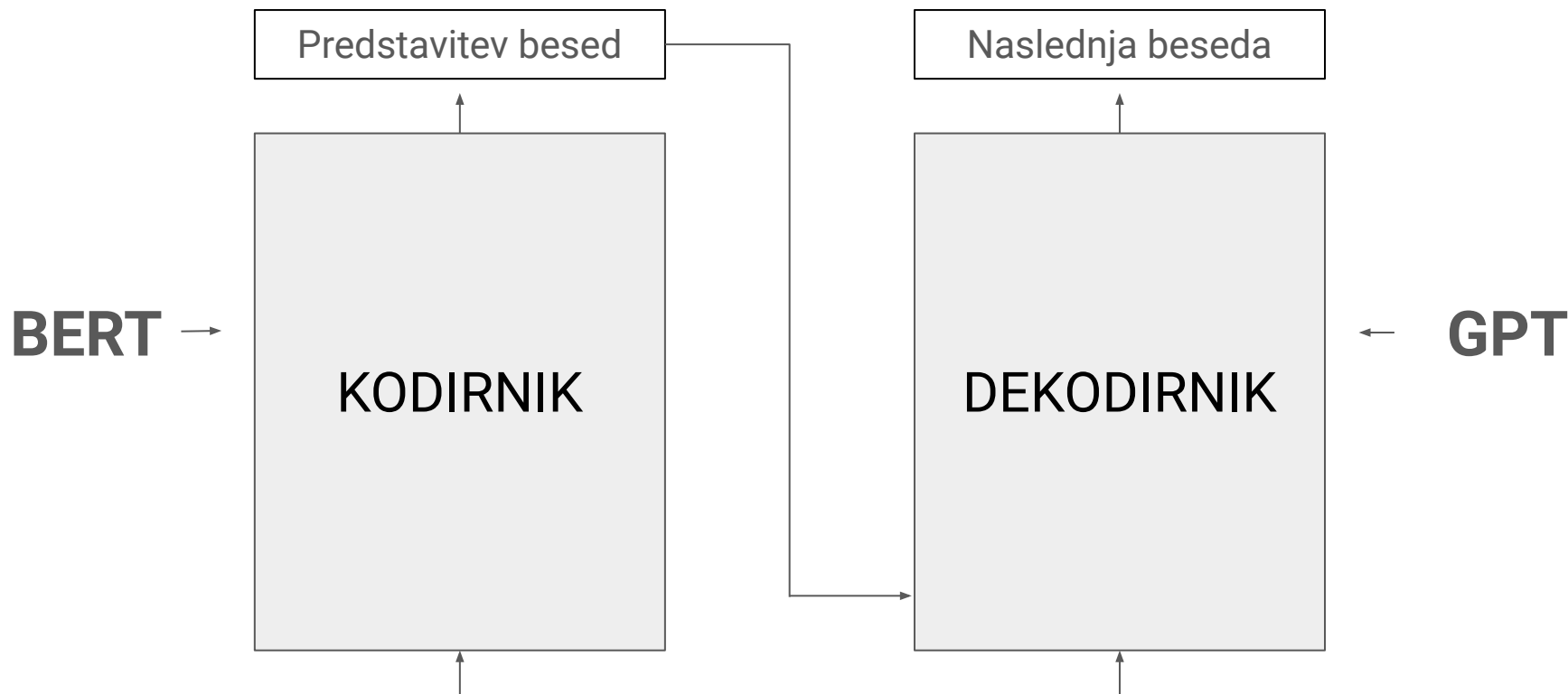


# Transformer



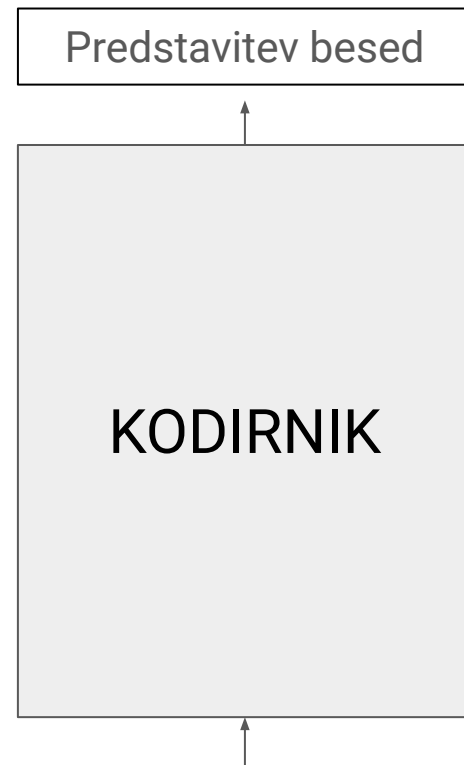


# Transformer

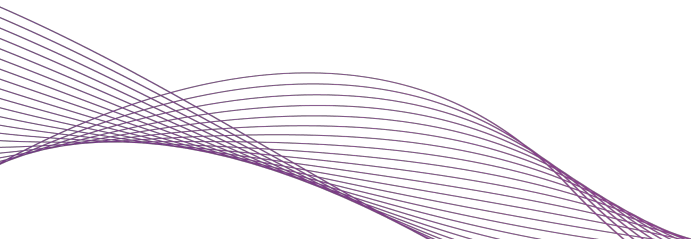


# BERT

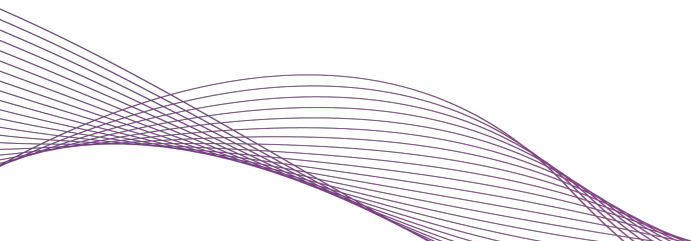
- Naučen za predstavitve besed in besedil.
- Kot tak se lahko douči za naloge tako nad besedami:
  - zaznava imenskih entitet;
  - ...;
- kot besedili:
  - klasifikacija;
  - iskanje podobnih besed;
  - ...



# Praktični del (Jupyter notebook-i)



# Dodatek



# Fine-tuning approaches

- can we develop models that adapt to many NLP tasks with little modification?
- training has two phases:
  - **Pre-training:** using a large unlabeled corpus and an auxiliary task, pre-train a model for general language representation;
  - **Fine-tuning:** using a (possibly smaller) labeled dataset, further train the pre-trained model for a specific downstream task.
- GPT [1], BERT [2]

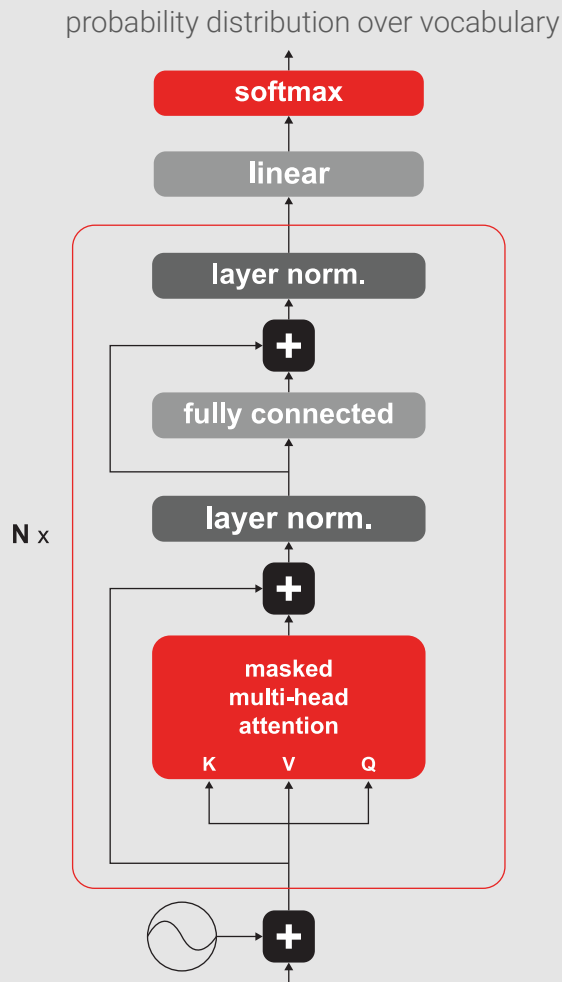
[1] [Radford et al.: Improving Language Understanding by Generative Pre-Training, 2018.](#)

[2] [Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.](#)

# GPT [1]

- Generative Pre-trained Transformer;
- using only the decoder part of Transformer;
- pre-trained for language modelling, i.e. predicting next word given the context.

[1] [Radford et al.: Improving Language Understanding by Generative Pre-Training, 2018.](#)

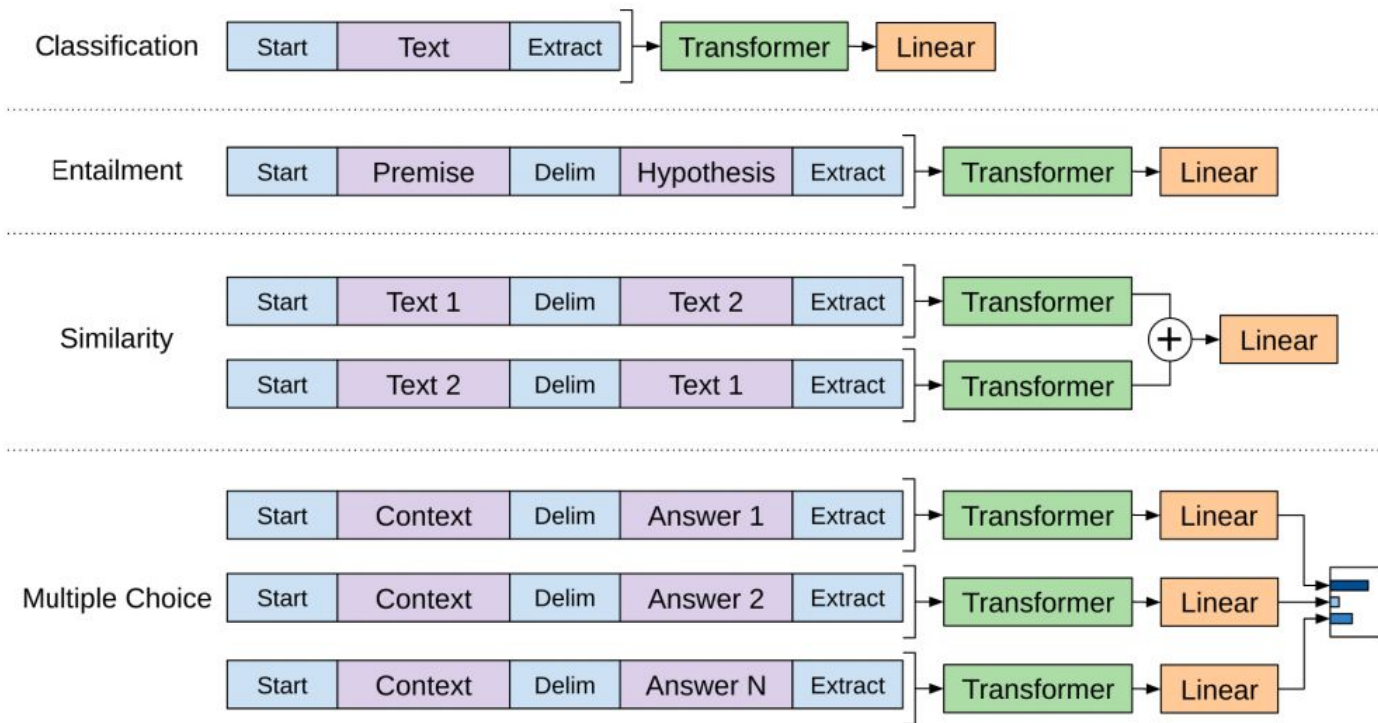


# Fine-tuning

$$L(X) = L_{LM}(X) + \lambda L_D(X)$$

language  
modelling  
loss

downstream  
task  
loss



# GPT shortcomings

- language modelling is an unidirectional task, models predict the next word given the left context:

*'What are those?' he said while looking at my **[?]***

- better language understanding requires incorporating bidirectionality:

*'What are **those**?' he said while looking at my crocs.*

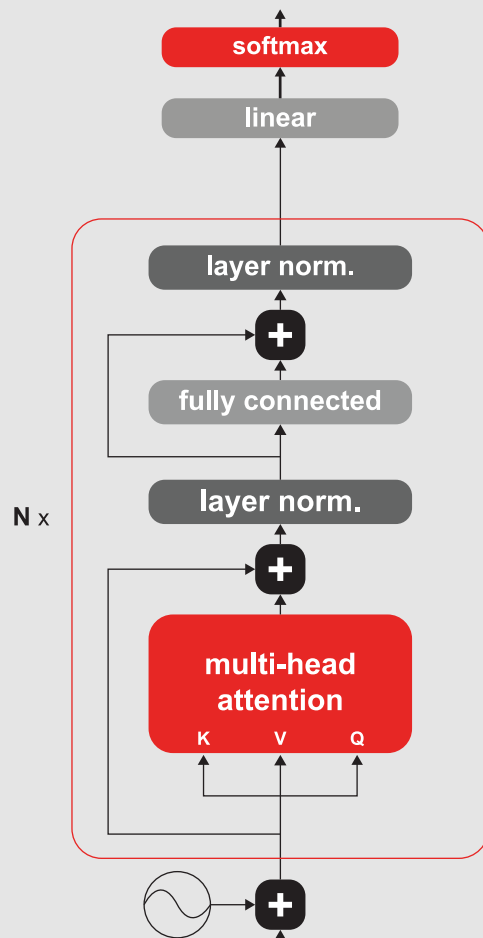


# BERT <sup>[1]</sup>

- Bidirectional Encoder Representations from Transformers;
- using only the encoder part of Transformer.

[1] [Devlin et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.](#)

probability distribution over vocabulary



# Masked language modelling

- 15% of input words are masked, the model learns to predict the missing words

What looking  
↑ ↑  
'[MASK] are those?' he said while [MASK] at my crocs.

## Too much masking:

Model is not provided with enough context.

## Too little masking:

Learning becomes very slow.

# Next sentence prediction

- given a pair of sentences predict if they follow one another;
- aims to learn sentence relationships that are important in certain downstream tasks (e.g. question answering).

**A:** 'What are those?' he said while looking at my crocs.

**B:** My new shoes.

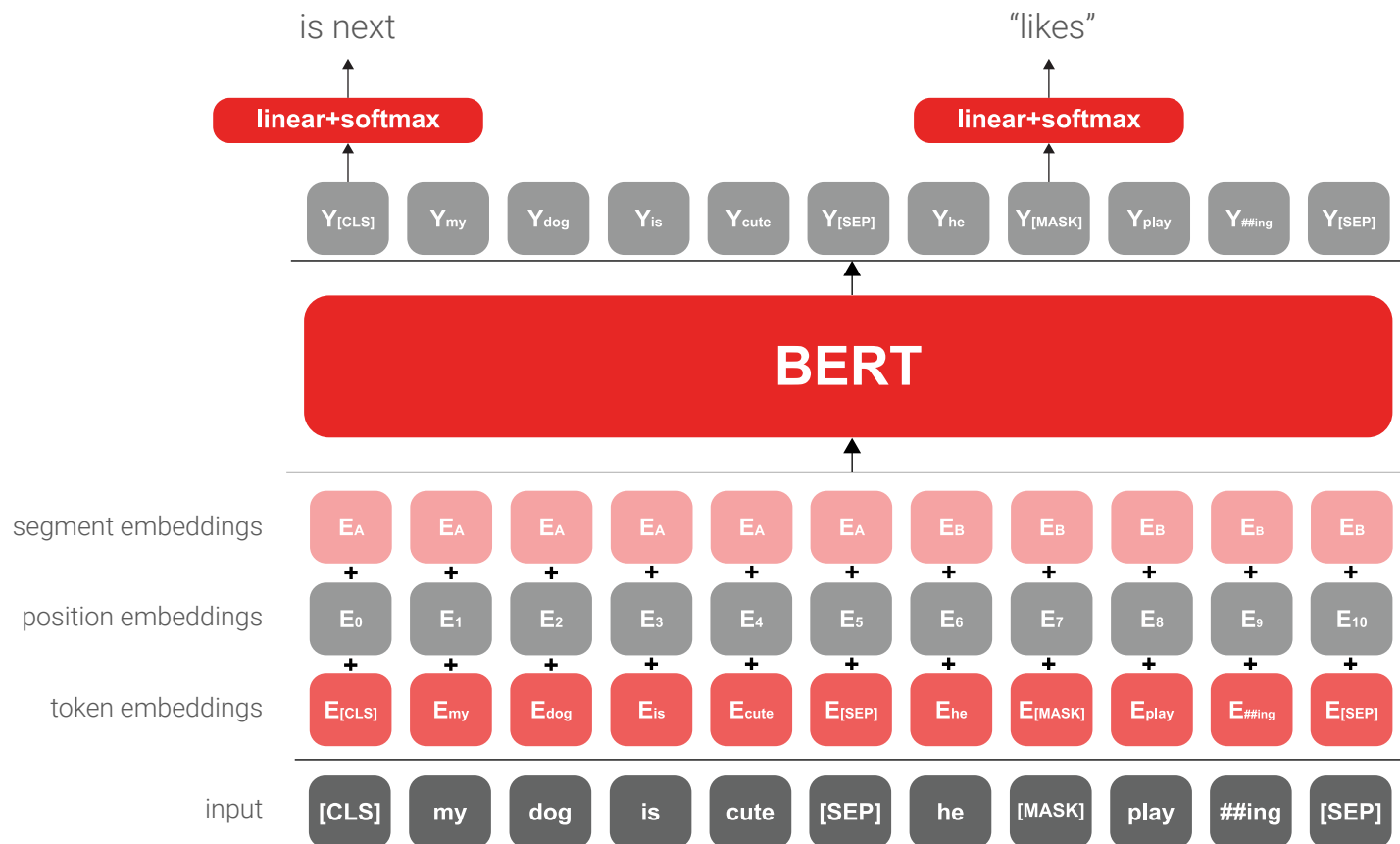
**Ground truth:** next

**A:** 'What are those?' he said while looking at my crocs.

**B:** The sky is blue.

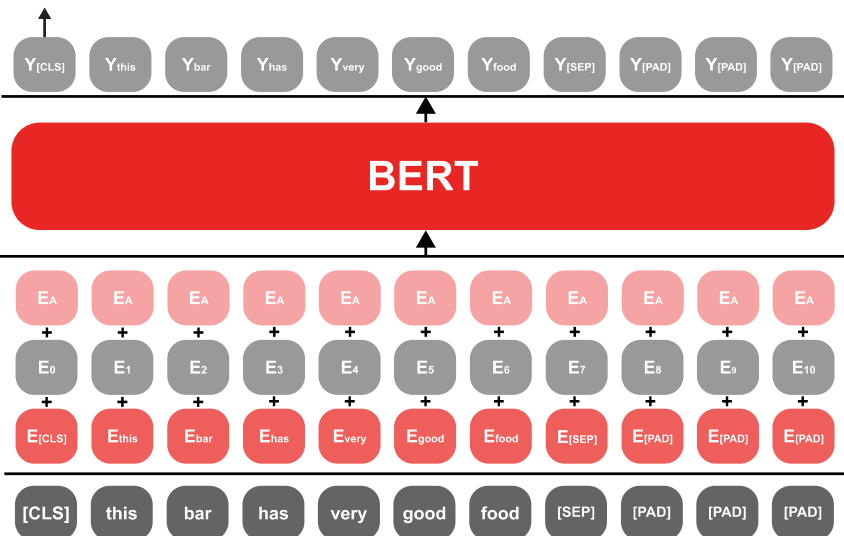
**Ground truth:** not next

# Pre-training

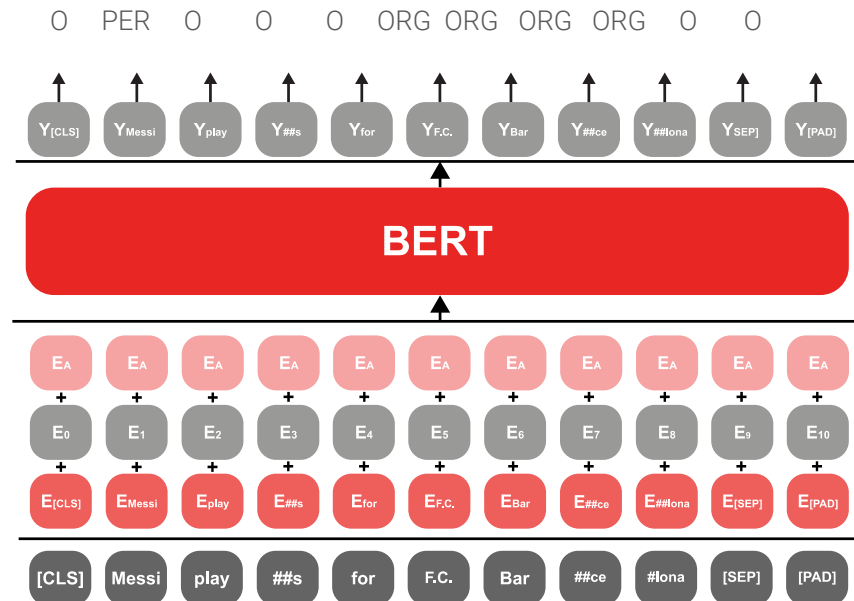


# Fine-tuning

positive



classification



named entity recognition