

HĽADANIE ANOMÁLIÍ V TEXTOCH POMOCOU NEURÓNOVÝCH SIETÍ

Autor: Olha Koliesnikova

Školiteľ: doc. RNDr. Gabriela Andrejková, CSc

ABSTRACT

Detekcia anomálií v textoch predstavuje kľúčový nástroj s rozsiahlym uplatnením v rôznych oblastiach, ako sú odhaľovanie podvodov, zlepšovanie kvality dát, monitorovanie komunikácie či identifikácia nekonzistentností v dokumentoch.

PREHĽAD ZNÁMYCH METÓD

1. Štatistické metódy

- TF-IDF, n-gramy: odchýlky na základe frekvencie slov a pravdepodobnosti.
- Rýchle, ale málo presné pre komplexné texty.

2. Klasifikácia textov

- SVM, Naive Bayes: klasifikácia na základe datasetu.
- Obmedzené malým množstvom dát.

PREHĽAD ZNÁMYCH METÓD

3. Neurónové siete

- Autoencodery: detekcia anomálií pomocou rekonštrukčných chýb.

4. Klasterizácia

- K-means: skupiny na základe podobnosti, odľahlé texty = anomálie.

5. Kombinované prístupy

- Transformátory (SlovakBERT) + klasifikátory (CNN, LSTM) = vysoká presnosť.

VLASTNÝ PRÍSTUP A RIEŠENIE

Kľúčové kroky:

- Preprocessing dát a štatistická analýza.
- Generovanie embeddingov pomocou SlovakBERT.
- Tréning autoencoderu na normálnych dátach.
- Stanovenie hraníc pre anomálie na základe rekonštrukčných chýb (L1, L2, MSE).
- Klasifikácia textov (anomalita/neanomalita).

PREPROCESSING DÁT

Čistenie textov:

- Odstránenie špeciálnych znakov, čísel a nadbytočných medzier.

Filtrácia textov:

- Odstránenie irelevantných textov alebo viet (podľa kľúčových slov).

Formátovanie dát:

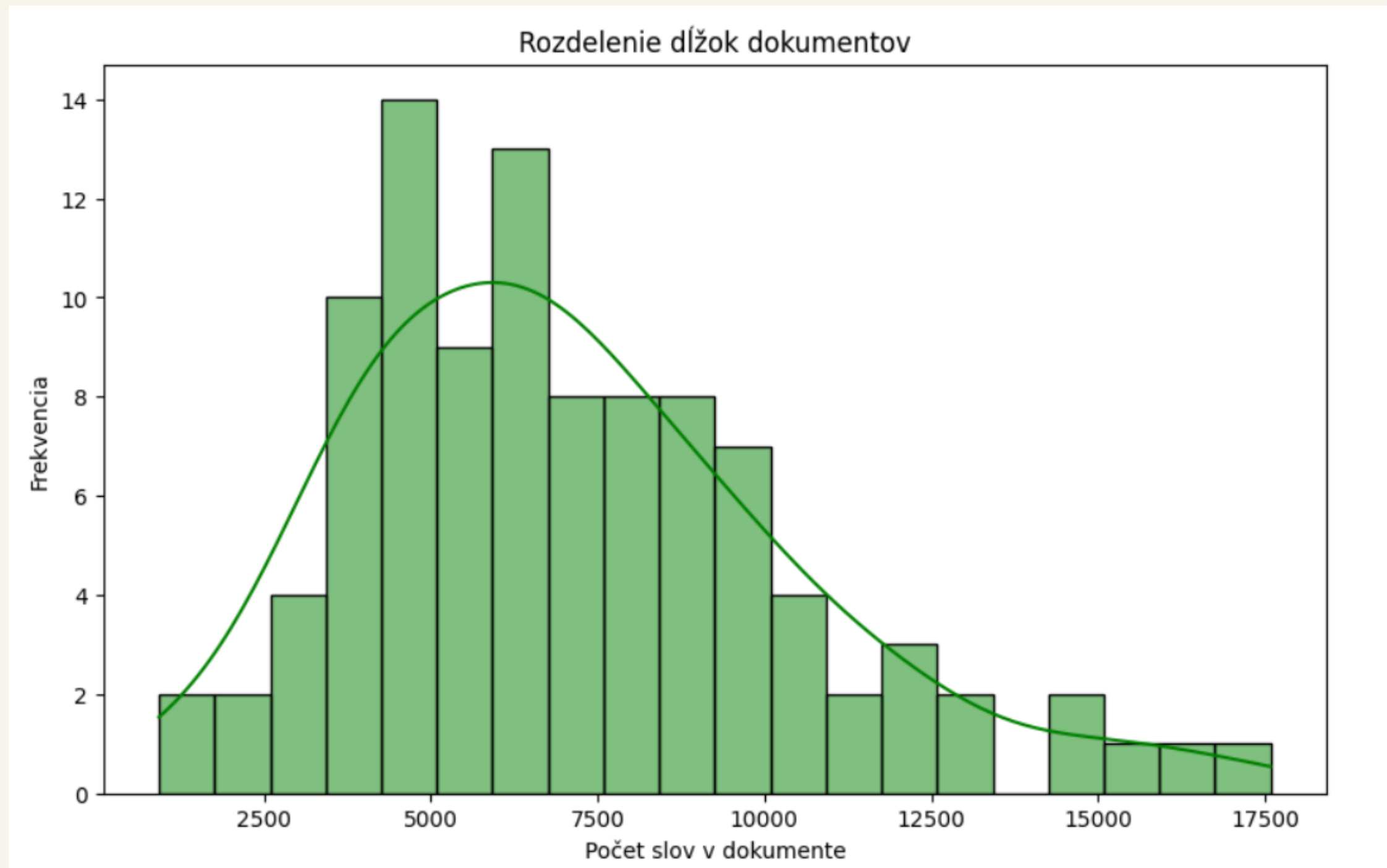
- Texty vo formáte :
"Vetal."
"Veta2."

ŠTATISTICKÁ ANALÝZA

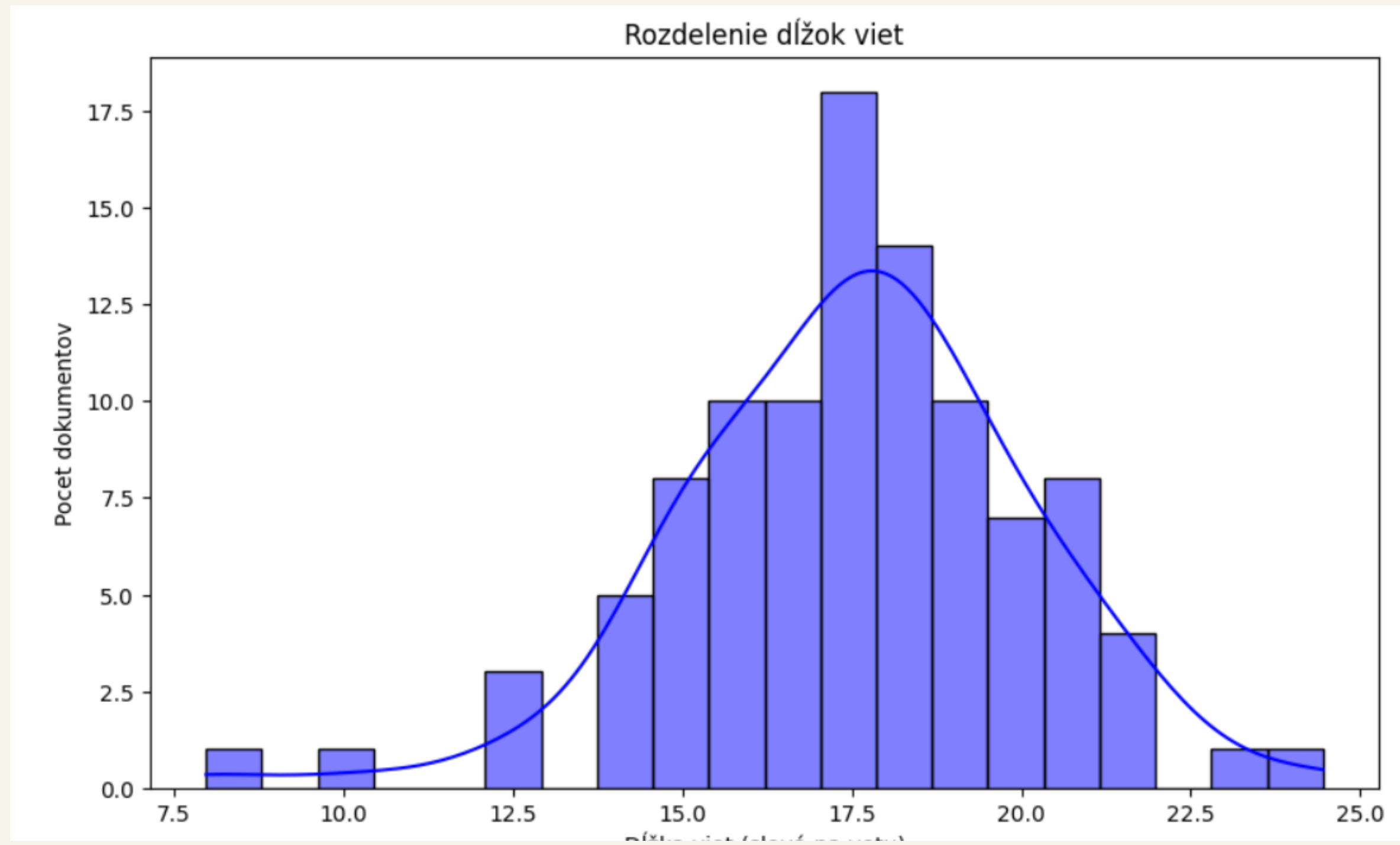
- Počet dokumentov: 101
- Počet viet: 42056
- Priemerná dĺžka viet: 17.54 slov na vetu
- Minimálna dĺžka vety: 7.98
- Maximálna dĺžka vety: 24.45



ŠTATISTICKÁ ANALÝZA



ŠTATISTICKÁ ANALÝZA



GENEROVANIE EMBEDDINGOV

Ciel':

Prevod textu do numerických vektorov zachytávajúcich význam a kontext.

Postup:

- Vstup: Textové súbory → vetná štruktúra.
- Model: SlovakBERT (gerulata/slovakbert) + pooling vrstva.

Výstup:

- Každý text reprezentovaný ako [počet viet] x 768 vektorov.

TRÉNING AUTOENCODERU

Ciel':

- Naučiť model rekonštruovať normálne embeddingy textov na detekciu odchýlok.

Architektúra modelu:

- Vstup: Vektor embeddingu (768 dimenzií).
- Skryta vrstva: 20 dimenzií (zachytáva podstatu textu).
- Výstup: Rekonštrukcia vstupu (768 dimenzií).

TRÉNING AUTOENCODERU

Priebeh tréningu:

- **Dáta:** Embeddingy normálnych textov ([počet viet] x 768).
- **Strata:** Mean Squared Error (MSE).
- **Parametre:** 100 epoch, batch size 10, optimalizátor Adam.
- **Výsledok:** Model úspešne rekonštruuje dáta.

STANOVENIE HRANÍC PRE ANOMÁLIE

Cieľ:

- Rozpoznať anomálie porovnaním originálneho a rekonštruovaného embeddingu.

Postup:

- Normálne dáta: Vypočítať chyby (L1, L2, MSE).
- Stanoviť prah na základe štatistickej analýzy.
- Anomálne dáta: Overiť, či chyby presahujú stanovené prahy.

Výsledok:

- Text je anomálny, ak chyba (L1/L2/MSE) prekročí stanovený prah.

PROBLÉMY A PERSPEKTÍVY

Výzvy počas práce:

- Texty bolo potrebné manuálne prechádzať a upravovať -> časovo náročné.
- Hľadanie optimálnej architektúry autoencoderu (počet vrstiev, veľkosti skrytých vrstiev) + nastavenie hyperparametrov.
- Tréning modelov (SlovakBERT, autoencoder) vyžadoval značný čas a výpočtový výkon.

Plán na dokončenie:

- Testovanie na väčšom množstve dát.
- Porovnávanie určenia anomálií s natrénovanou sieťou ResNet50
- Finalizácia písania práce.

The background features three vertical stripes on the left: a wide pink stripe, a narrower blue stripe, and a narrow beige stripe. The rest of the background is a light beige color with two rectangular areas of a pink dot pattern, one in the top right and one in the bottom right.

**DAKUJEM ZA
POZORNOST**