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).
- Porovnanie z metodou RESNET-50.

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 :**

  “Veta1.”

  “Veta2.”

Vytvaranie anomalnych textov

ŠTATISTICKÁ ANALÝZA

- Počet dokumentov: 200 (100 +100)
- Počet viet: 42056
- Priemerná dĺžka viet: 17.54 slov na vetu
- Minimálna dĺžka vety: 7.98
- Maximálna dĺžka vety: 24.45

GENEROVANIE EM-VEKTOROV

Cieľ:

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

Cieľ:

Naučiť model rekonštruovať normálne embeddingy textov na detekciu odchýlok.

Architektúra modelu:

- Vstup: Vektor embeddingu (768 dimenzií).
- Encoder: Znižovanie dimenzií cez vrstvy 768 → 512 → 256 → 128 → 32.
- Decoder: Rekonštrukcia embeddingu späť na 768 dimenzií cez vrstvy 32 → 128 → 256 → 512 → 768.

TRÉNING AUTOENCODERU

Priebeh tréningu:

- **Dáta: Em-vektory 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 embeddingu textu a jeho rekonštruovanej verzie.

Postup:

1. **Výpočet kosínovej podobnosti**

Kozinová vzdialenosť vyjadruje, ako dobre sieť dokáže rekonštruovať vetu na základe podobnosti medzi pôvodným a rekonštruovaným vektorom vety.

STANOVENIE HRANÍC PRE ANOMÁLIE

2. Agregácia metrík pre každý dokument
   - Priemerná podobnosť (mean): Poskytujúci celkový obraz o kvalite rekonštrukcie.
   - Mediánová podobnosť (median): Stredná hodnota podobností, odolná voči extrémnym hodnotám.
   - Minimálna a maximálna podobnosť (min/max): Ukazujú rozsah podobností, identifikujú najväčšie rozdiely a zhody v dokumente.

STANOVENIE HRANÍC PRE ANOMÁLIE

3. Stanovenie prahu anomality

- **<u>Výpočet prahu:</u>**

Priemerná kosínová podobnosť normálnych textov = 0.892, slúžila ako referenčná hodnota pre vyhodnocovanie anomálií.

- **<u>Prahové hodnoty :</u>**
  - **min_threshold = 0.755: Dokument je anomálny, ak minimálna podobnosť klesne pod túto hodnotu.**

DVOJÚROVŇOVÁ KLASIFIKÁCIA DOKUMENTOV

1. Anomaly Status:
   - Anomália: Dokument je označený ako anomálny, ak minimálna podobnosť v dokumente klesne pod hodnotu 0.755.
2. Suspicion Level:
   - Tento krok sa zameriava na hodnotenie celkovej kvality rekonštrukcie dokumentu pomocou priemernej a mediánovej kosínovej podobnosti.
   - Suspicion Level: Dokumenty sú hodnotené podľa týchto metrík, ktoré odrážajú celkovú konzistenciu rekonštrukcie.

KLASIFIKÁCIA NA ZÁKLADE METRIKY

Logika AND:

Priemerná podobnosť < 0.89 AND Mediánová podobnosť < 0.90

Tento prístup zachytáva dokumenty s konzistentným a výrazným narušením kvality rekonštrukcie.

Logika OR:

Priemerná podobnosť < 0.89 OR mediánová podobnosť < 0.90

Tento prístup je citlivejší a zachytáva aj mierne odchýlky v rekonštrukcii.

ÚSPEŠNOSŤ MODELU

Logika AND:

- **Pri použití logiky AND sme dosiahli 78% úspešnosť.**
- **Podozrivé dokumenty: 35**
- **OK dokumenty: 22**
- **Anomálne dokumenty: 43**

Logika OR:

- **Pri použití logiky OR sa úspešnosť modelu zvýšila na 98%.**
- **Podozrivé dokumenty: 55**
- **OK dokumenty: 2**
- **Anomálne dokumenty: 43**

KLASIFIKÁCIA TEXTOV POMOCOU RESNET-50

Cieľ: Detekcia anomálií v slovenských akademických textoch

Postup:

1. SlovakBERT → em-vektory
2. Redukcia dimenzie → 256D
3. Prevod em-vektorov na obraz (256×256)
4. Klasifikácia cez ResNet-50 (CNN)

Typy výstupu: normálny / anomálny text

TECHNICKÝ POPIS

Em-vektory:

- **SlovakBERT → 768D**
- **Redukcia: Dense(256, activation='relu')**

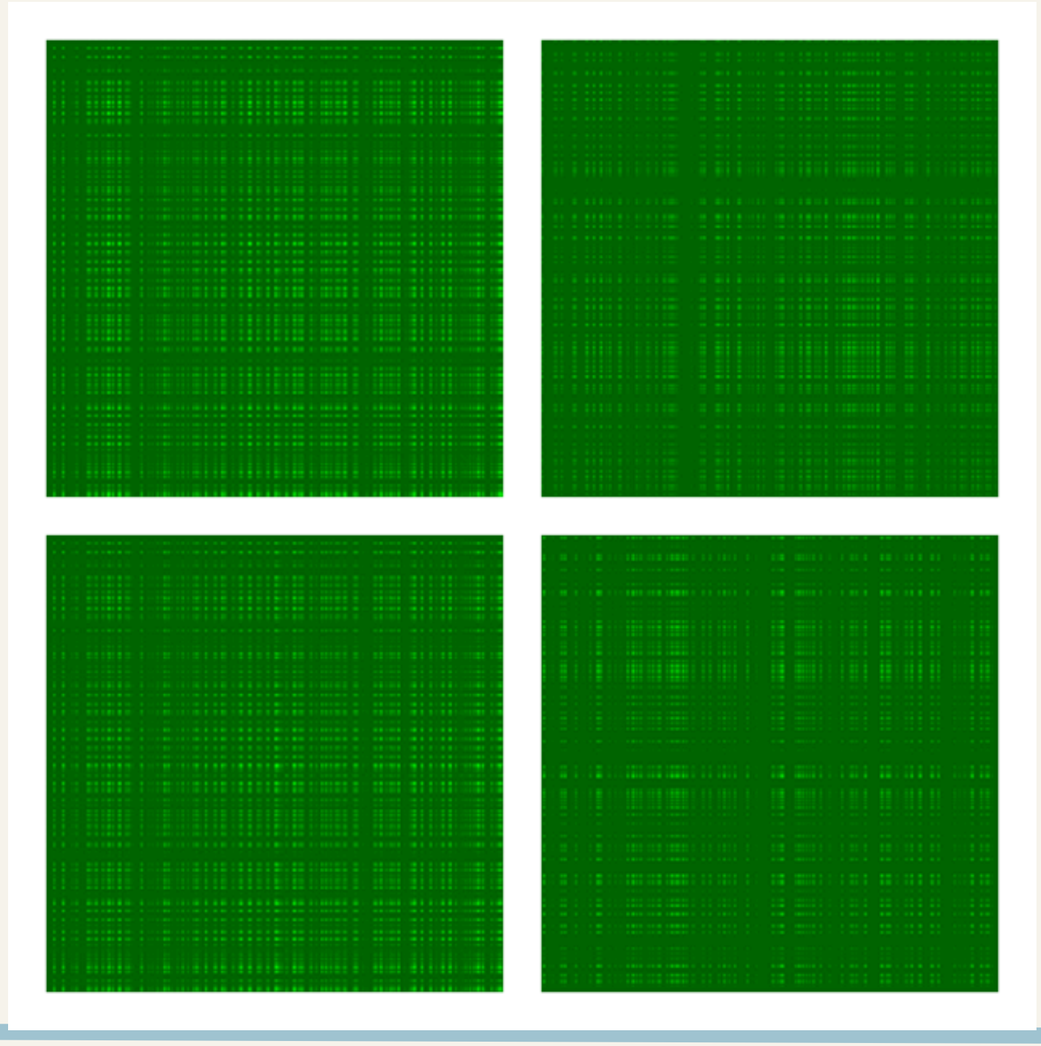
Obraz z textu:

- **Každý text = [počet viet] x 256**
- **Výpočet A^TA → 256x256 matica**
- **Normalizácia na [0,1], prevod na RGB**

Model:

- **ResNet-50**
- **Upravený výstup (softmax, 2 triedy)**
- **Vstupné obrázky: 256x256x3**

VIZUALIZÁCIA TRANSFORMOVANÝCH EMBEDDINGOV



VÝSLEDKY A ZHODNOTENIE

Výsledky a zhodnotenie

- Dataset: 200 textov → 100 normálnych, 100 anomálnych
- Tréning: 80/20 split
- Presnosť: 100 % na testovacej množine
- Výhody:
  - Transformácia textu na obraz umožňuje využitie CNN
  - Model efektívne detekuje podozrivé práce
- Nevýhody:
  - Malý dataset, treba testovať na reálnych dátach

ĎAKUJEM ZA POZORNOSŤ