VŠB TECHNICKÁ UNIVERZITA OSTRAVA

www.vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Zpracování textu pomocí hlubokých neuronových sítí

Text Processing using Neural Networks

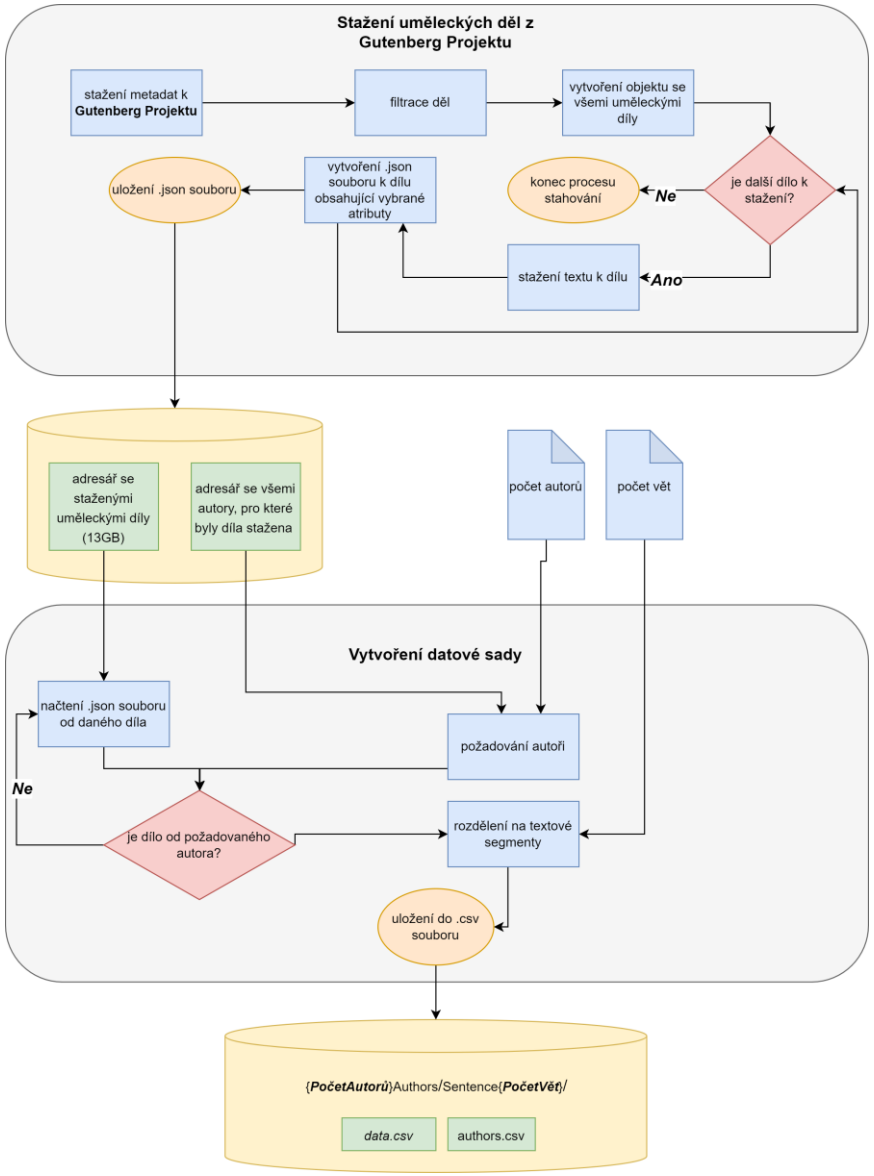
Bc. Vojtěch Prokop
prof. Ing. Jan Platoš, Ph.D.

Cíl práce

- zpracování přirozeného jazyka pomocí dostupných metod
- rozpoznání autora (*autorství*)
- **klasifikační problém**
- prozkoumat možnosti **Transformer** modelu

Vytvoření datové sady

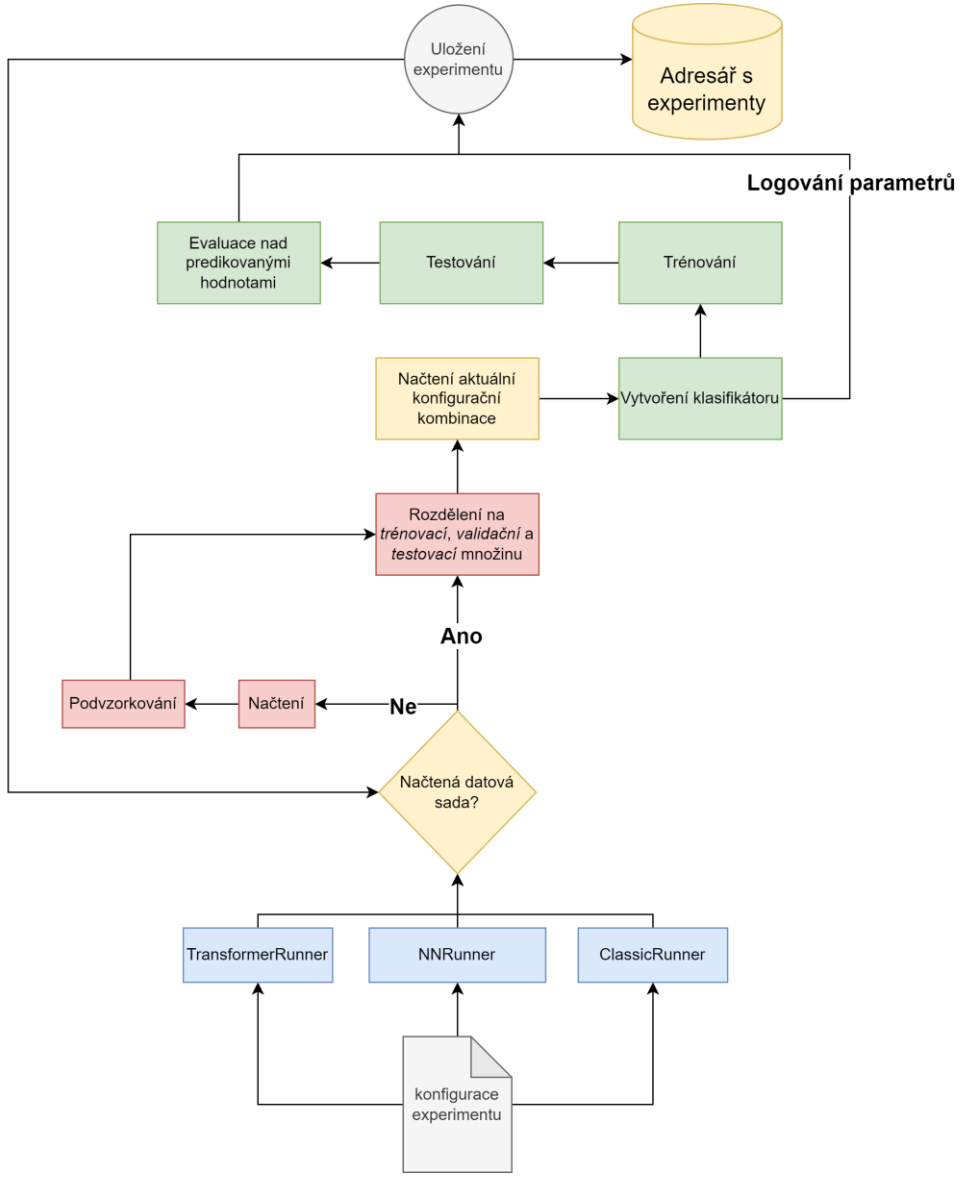
- zdroj dat *(Projekt Gutenberg)*
- stažení uměleckých děl
  - *.json pro každé umělecké dílo*
  - *13GB*
- tvorba datových sad
  - *počet autorů*
  - *počet vět*
  - *rozličná velikost souborů*



Obr. 1 Automatizovaná tvorba datové sady

Návrh experimentů v závěrečné práci

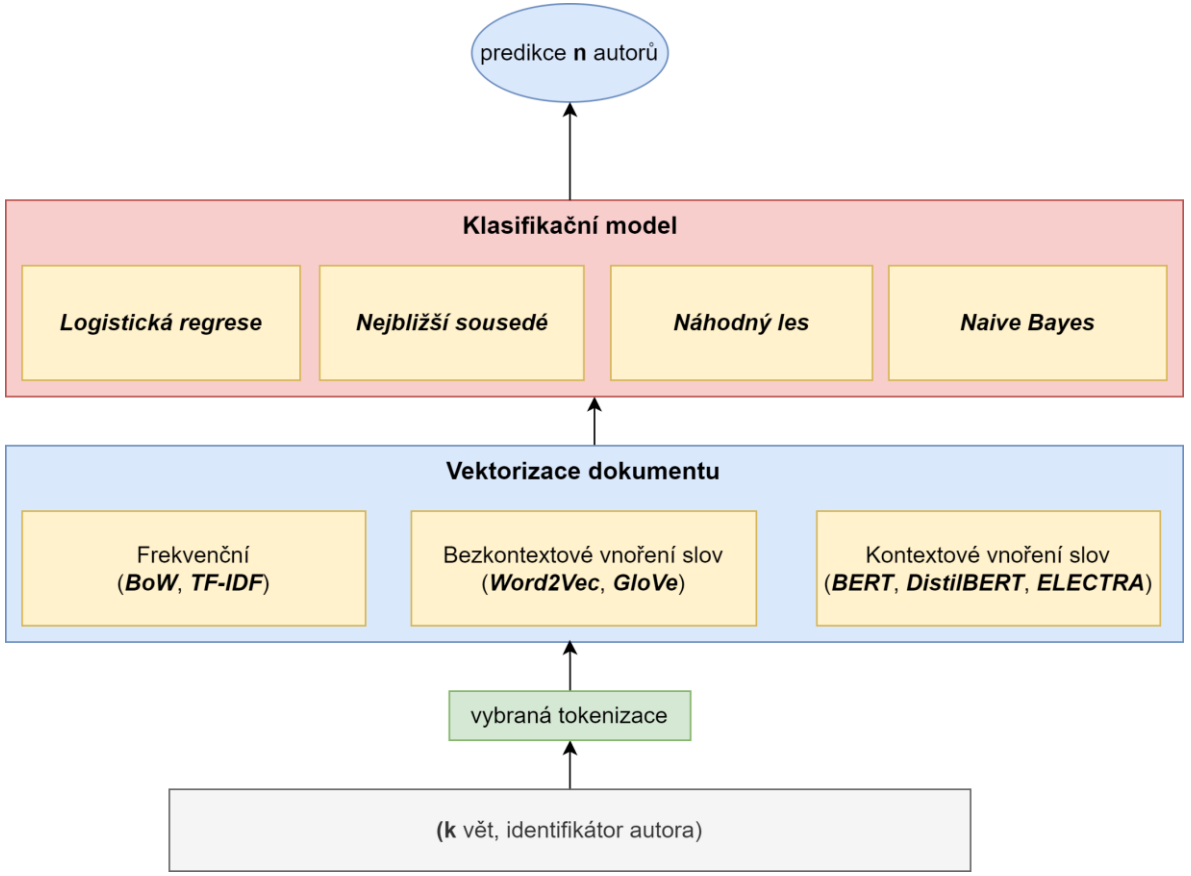
- explorační analýza
  - *nalezení optimálních hyperparemetrů pro navržené modely*
  - *návrh předzpracování textových dat*
  - *nalezení hodnoty k podvzorkování*
- typ experimentu
  - *vektorizace textu ve spojení s klasickými modely*
  - *neuronové sítě*
  - *BERT (Transformer)*
- automatizace a logování



Obr. 2 Automatizované spouštění experimentů

Vektorizace textu a klasické modely

- tokenizace
- vektorizace
  - *frekvenční*
  - *bezkontextové a kontextové vnoření slov*
- klasifikátor
  - *log. regrese*
  - *nej. sousedé*
  - *náh. les*
  - *naive bayes*

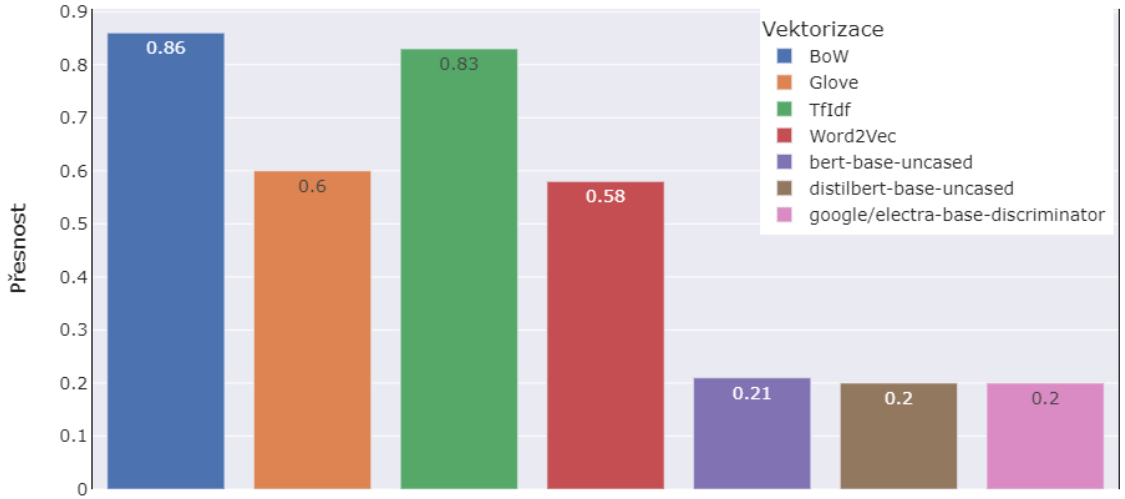


Obr. 3 Návrh experimentů klasických modelů

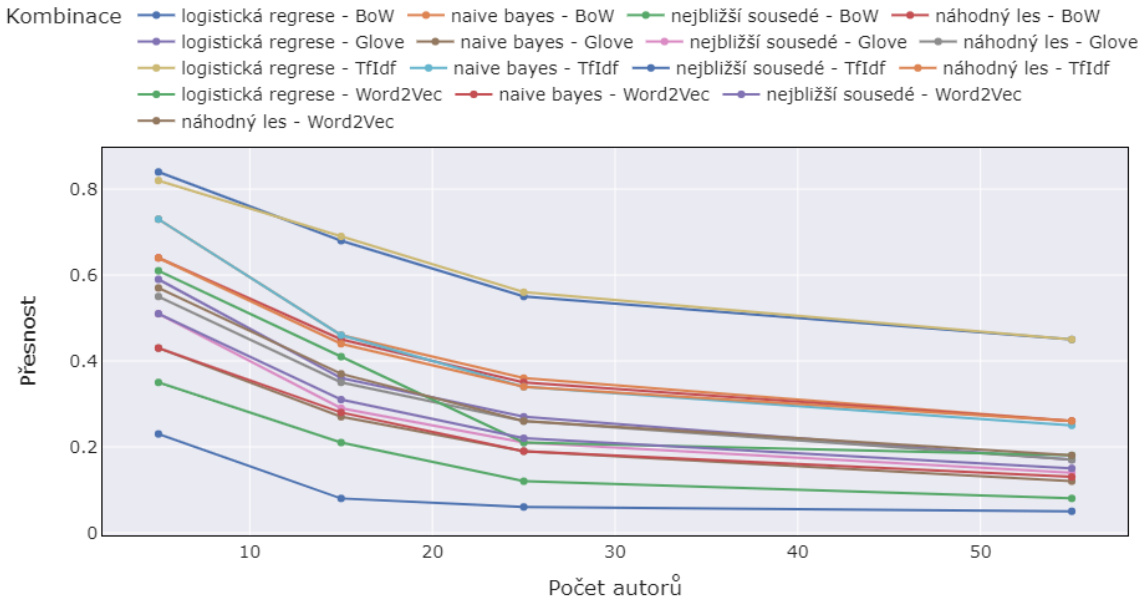
Experiment - Vektorizace textu a klasické modely

- různé kombinace mezi vektorizací a typem klasifikátoru
- **nejlepší varianta** - log. regrese ve spojení s *BoW* a *TF-IDF* vektorizací (***až 86% přesnost***)
- testování většího počtu autorů nad nejlepšími modely
- rozdíl mezi 5 a 55 autory průměrně ***30 %***

Graf. 1 Výsledky log. regrese ve spojení s různými typy vektorizace

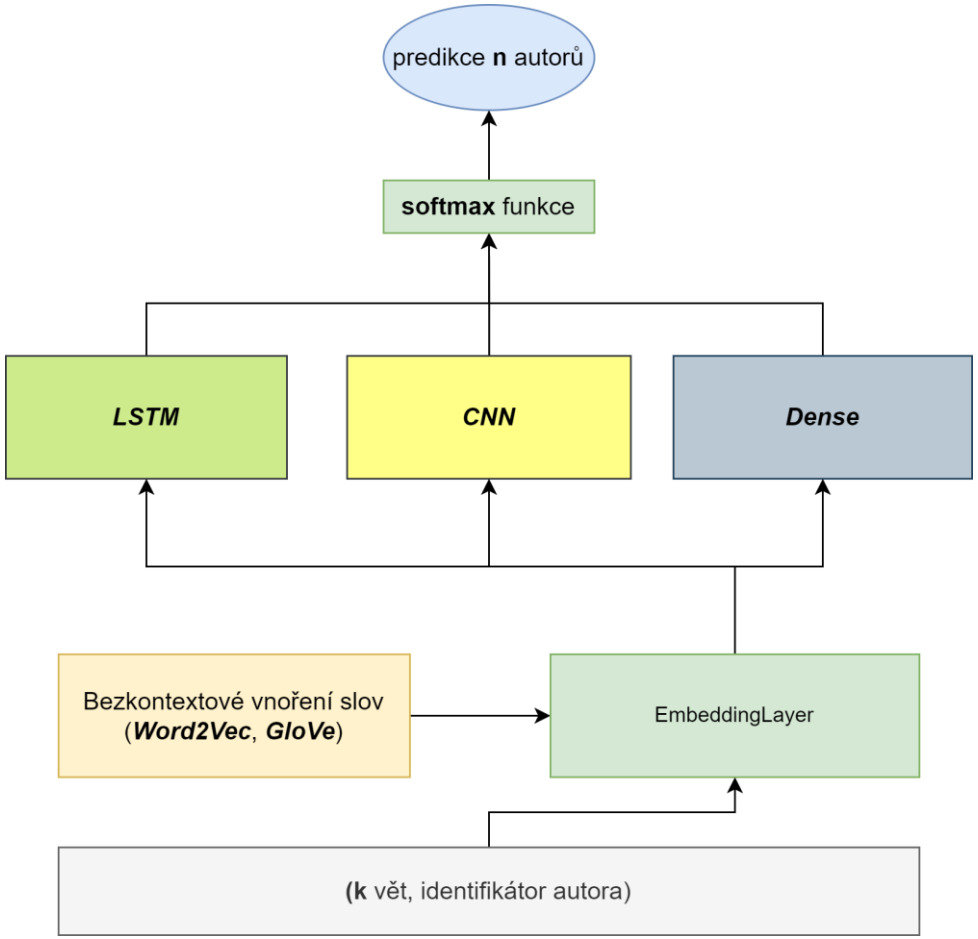


Graf. 2 Výsledky log. regrese ve spojení s různými typy vektorizace



Neuronové sítě

- vnoření slov
  - *prázdná inicializace*
  - *Word2Vec*
  - *GloVe*
- architektura
  - *LSTM*
  - *CNN*
  - *Dense*

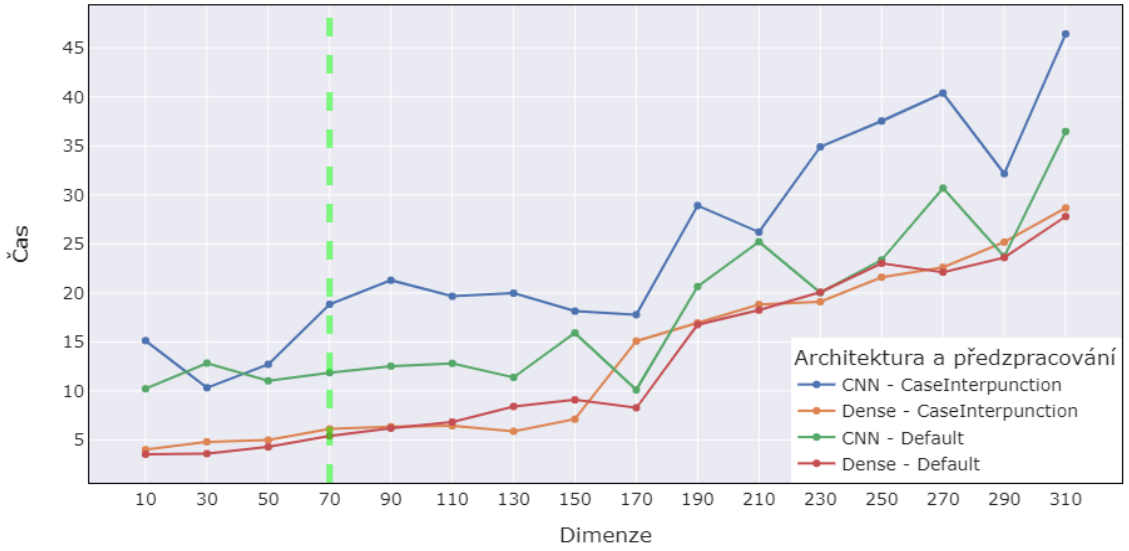


Obr. 4 Návrh experimentů neuronových sítí

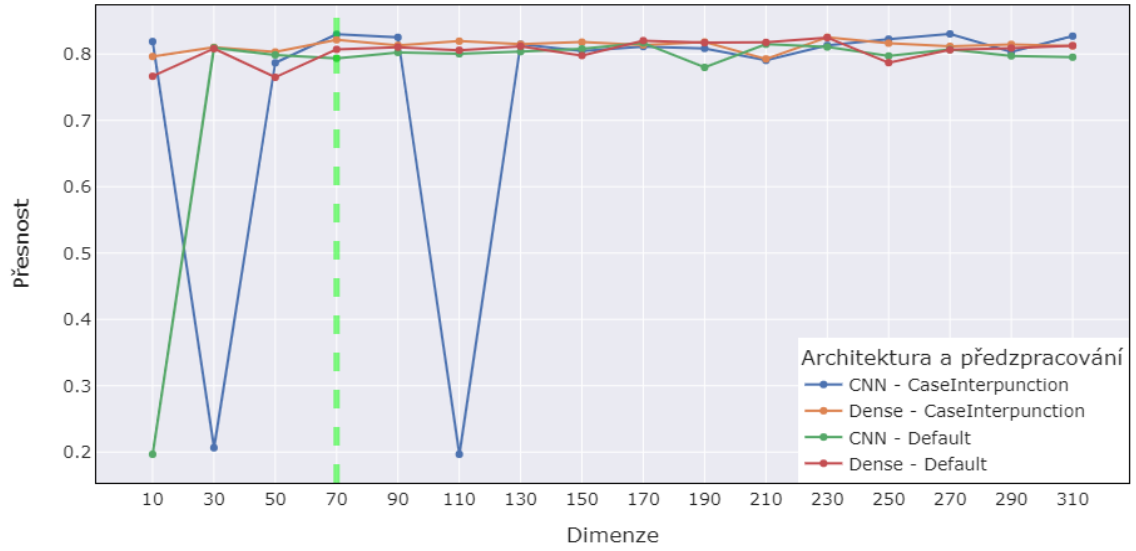
Experiment – neuronové sítě

- rozličné kombinace typu inicializace a architektury
- *nejlepší výsledek experimentálního bloku – **CNN + GloVe (82,65 %)***
- modelování potřebné velikosti vektoru (***hodnota 70***)
- přesnost kolem ***82 %***

Graf. 3 Modelování mnohodimenzionálního vektoru – **čas** *(minuty)*

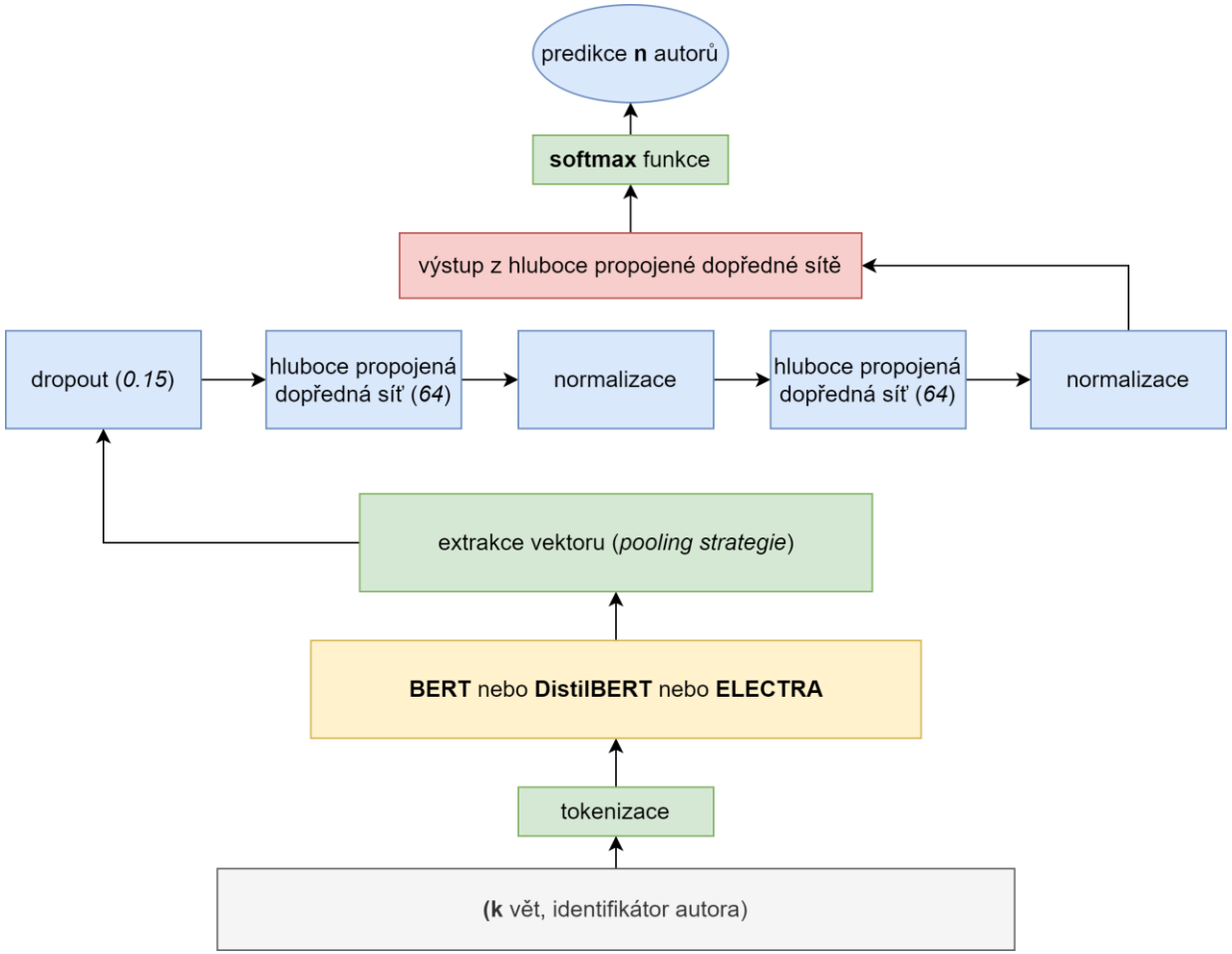


Graf. 4 Modelování mnohodimenzionálního vektoru - **přesnost**



BERT

- typ **Transformeru**
- extrakce dokumentové reprezentace
- hluboká dopředná síť

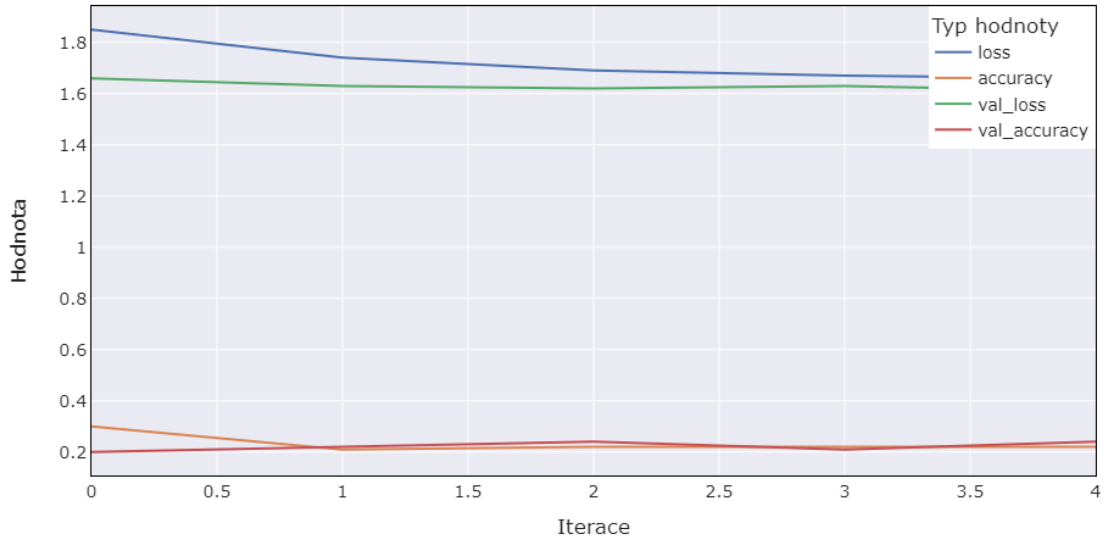


Obr. 5 Návrh experimentů BERT modelu

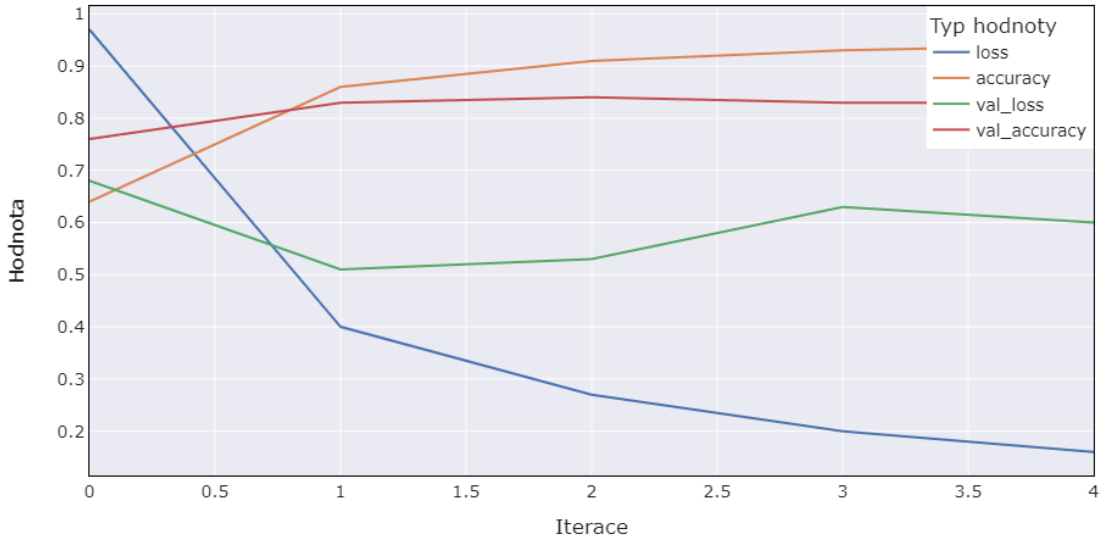
Experiment kvality obecného jazykového modelu

- nutnost doučení – stagnace na **20 %**
- nízký počet iterací k schopnosti porozumět problému ***autorství***

Graf. 5 Trénování dopředné vrstvy



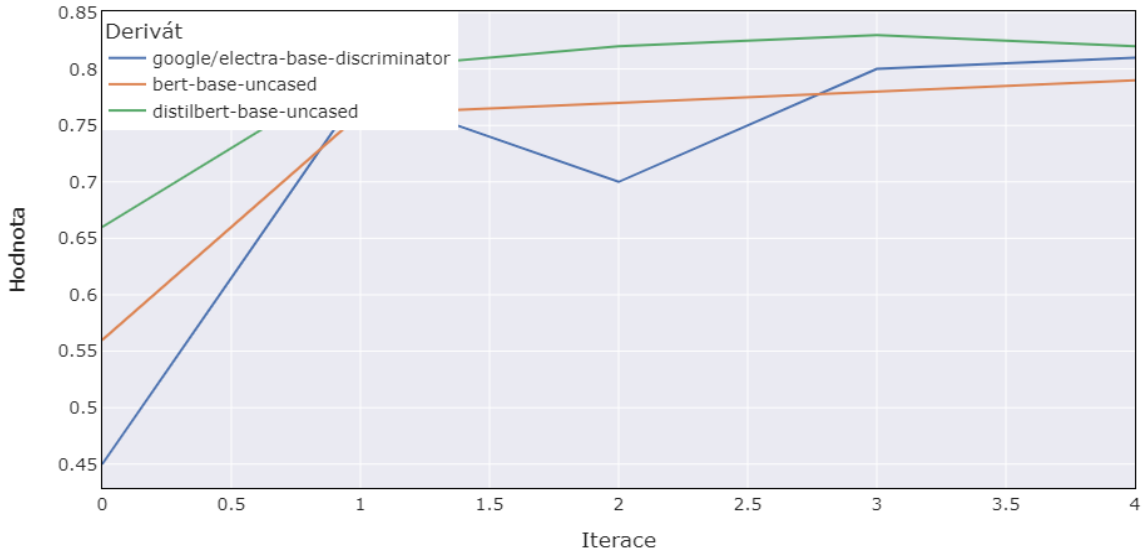
Graf. 6 Trénování celého modelu



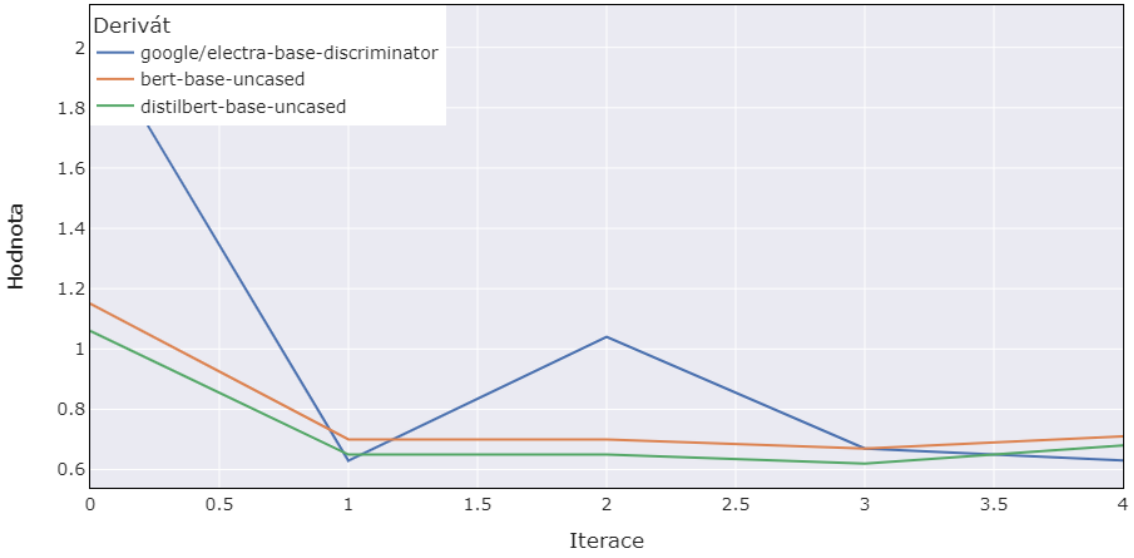
Experiment typu derivátu BERT modelu

- ELECTRA, BERT, DistilBERT
- nejlepší výsledek DistilBERT derivátu - **kolem 84 %**
- možnost zlepšení výsledku u ELECTRA modelu vzhledem k snižující se chybě

Graf. 7 Průběh derivátu - **přesnost**



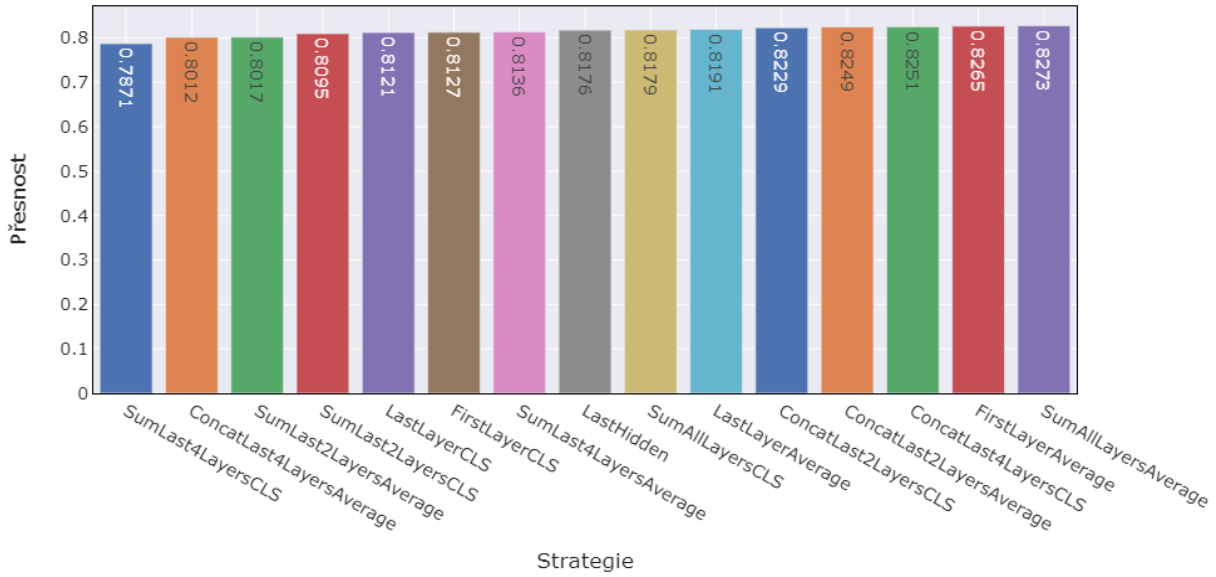
Graf. 8 Průběh derivátu - **chyba**



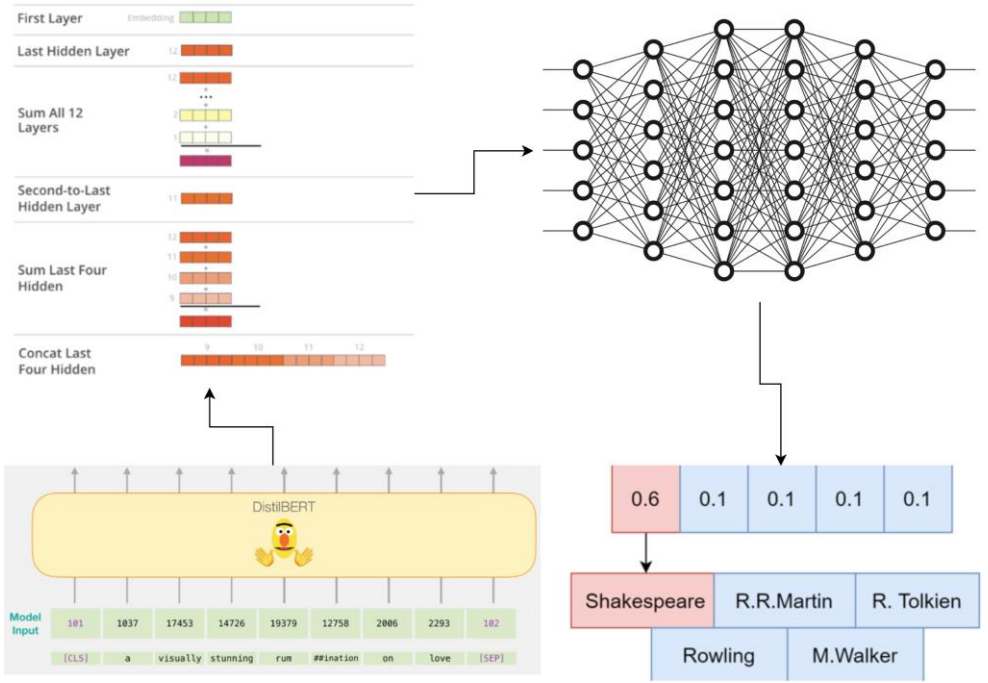
Experiment strategie využití vektorů z jazykového modelu DistilBERT

- 4% zlepšení přesnosti vzhledem k nejlepší a nejhorší strategii
- zlepšení oproti „doporučené“ strategii o 1 %
- sečtení všech vrstev – **nejlepší strategie**
- neschopnost výběru nejlepší obecné strategie

Graf. 9 Výsledky přesnosti nad 5 autory, 3 větami a různými typy strategií



Obr. 6 Vizualizace extrakce číselných vektorů (https://jalammar.github.io)



Experiment variabilního větného okna u DistilBERT modelu

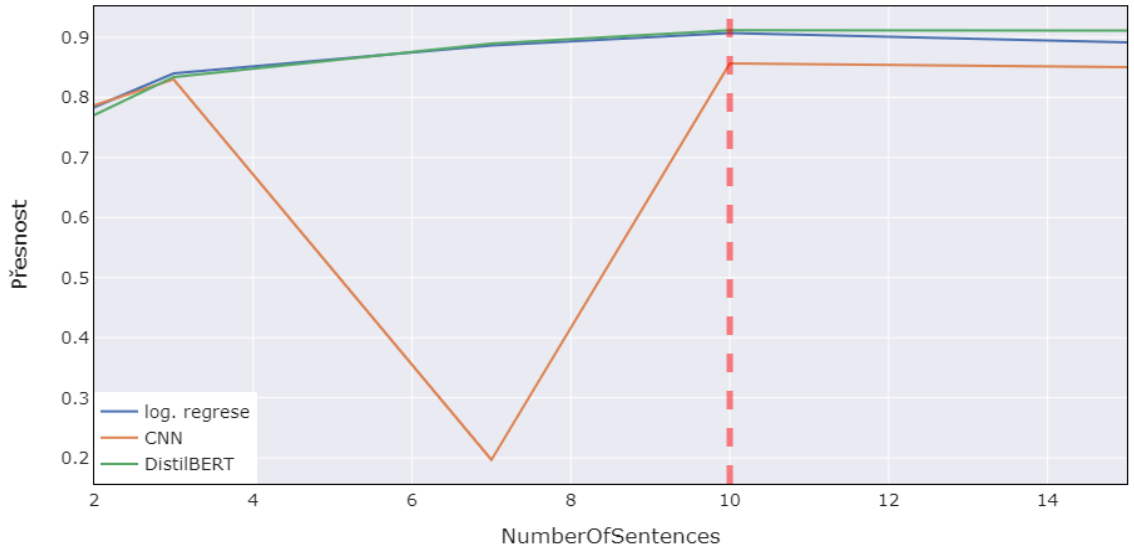
	počet záznamů v množině					
počet vět	*trénovací*	*validační*	*testovací*	*hod. podvzorkování*	***přesnost (%)***	*velikost vstupu*
1	162 562	28 688	33 750	45 000	65,80	30
2	81 281	14 344	16 875	22 500	77,04	60
3	54 187	9 563	11 250	15 000	83,40	100
7	23 481	4 144	4 875	6 500	88,94	190
10	16 256	2 869	3 375	4 500	**91,17**	260
15	10 837	1 913	2 250	3 000	91,11	380

Tab. 1 Výsledky experimentů s větným oknem

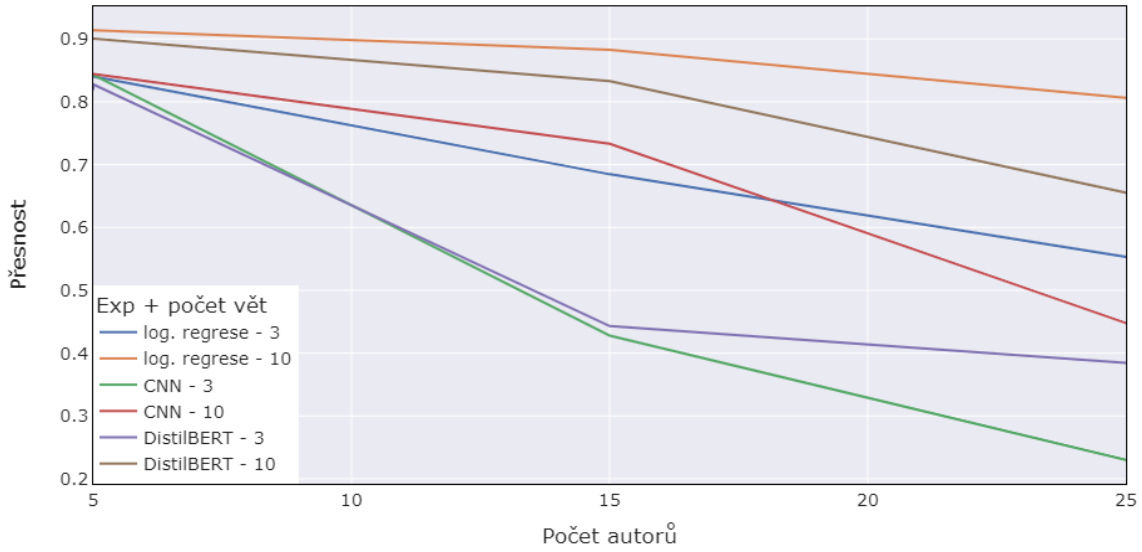
Srovnání nejlepších modelů

- srovnání nejlepších modelů z experimentálních bloků – **log. reg., CNN, DistilBERT**
- testování variabilního větného okna – **zvýšení** přesnosti modelů, schopnost konkurence **CNN** u nízké hodnoty
- testování většího počtu autorů – podobné výsledky u **BERT derivátu** a **log. reg**.

Graf. 10 Výsledky modelů u variabilního větného okna (2, 3, 7, 10, 15)



Graf. 11 Výsledky vybraných modelů u více autorů (5, 15, 25)



Závěr

- metody a přístupy u zpracování textu – vektorizace, neuronové sítě, klasické modely, Transformer
- navržení systému schopného rozpoznání autora
- srovnání přesnosti a výpočetní náročnosti zkoušených modelů
- nekompatibilita klasických klasifikátorů s *BERT* vektorizací
- prospěch variabilního větného okna
- smysl výběru strategie pro extrakci vektorové reprezentace dokumentu z *BERT* modelu
- vysoká úspěšnost *BERT* modelu

Otázky z hodnocení

- Jak je to se zpracováním češtiny? - *doc. Mgr. Jiří Dvorský, Ph.D.*

Děkuji za pozornost

Otázky ?