Metódy Transformerov a ich aplikácia v oblasti textovej analýzy

Obhajoba diplomovej práce

Martin Bača

Vedúci práce: RNDr. Šimon Horvát, PhD.
Konzultant práce: doc. RNDr. Ľubomír Antoni, PhD.

Prírodovedecká fakulta
Univerzita Pavla Jozefa Šafárika v Košiciach

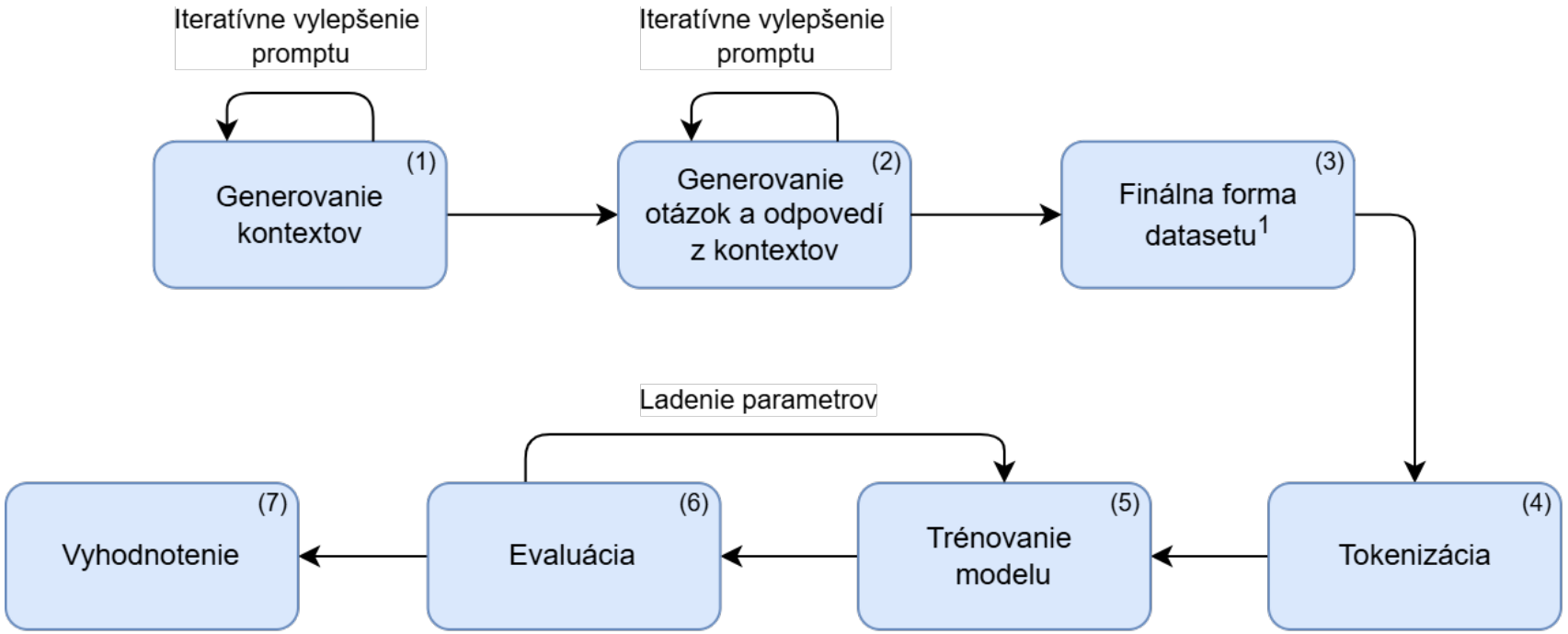
Košice, 29. máj 2025

Motivácia

- **Odpovedanie otázok** (Question Answering - QA) sa stáva dôležitou súčasťou moderných AI systémov
- Prudký vývoj v oblasti výskumu jazykových modelov
- Väčšina výskumu v oblasti QA je realizovaná na datasetoch, ktoré sú často len monolingválne alebo obsahujú len jeden typ otázok
- Skúmanie schopností QA systémov v odpovedaní na binárne otázky

Postup riešenia

- Generovanie datasetu (kontext – otázky – odpovede)
- Predspracovanie a tokenizácia
- Trénovanie QA modelov (BERTa based a XLM-R based)
- Vyhodnotenie úspešnosti modelov pre rôzne typy otázok



Štruktúra počiatočných dát

- Základné dáta predstavujú 500 objednávok na prepravu
- Dáta sú náhodne generované a uložené v JSON formáte


```
"pickup": {
    "company_name": "COMPANY_PICKUP",
    "postal_code": "ZIP_PICKUP",
    "street_name": "STREET_PICKUP",
    "datetime": "2044-04-18 18:30:00"
},
"delivery": {
    "company_name": "COMPANY_DELIVERY",
    "postal_code": "ZIP_DELIVERY",
    "street_name": "STREET_DELIVERY",
    "datetime": "2044-04-20 23:30:00"
},
"goods": {
    "weight": "94 kg",
    "dimensions": [
        {
            "length": "32m",
            "width": "86m",
            "height": "43m",
            "weight": "94 kg"
        },
    ]
},
"required vehicle": "Van",
"special request": "Double driver required"
```

Tvorba datasetu pomocou jazykových modelov

- Prevod štruktúrovaných JSON dát do prirodzeného jazyka
- Využitie veľkých jazykových modelov (LLMs) pri generovaní kontextov

```
We need to transport a shipment of goods
from COMPANY_PICKUP in ZIP_PICKUP,
STREET_PICKUP on May 11th at 19:00
to COMPANY_DELIVERY in ZIP_DELIVERY,
STREET_DELIVERY between May 13th 23:15
and May 17th 00:15. The shipment consists
of two items with the following dimensions:

* Item 1: Length 21cm, Width 21cm, Height 58
cm, Weight 88kg
* Item 2: Length 49cm, Width 14cm, Height 41
cm, Weight 55kg

The total weight of the shipment is 143kg.
We require a Sprinter vehicle with ADR
(dangerous goods) capabilities to transport
these goods. Please note that the pickup
and delivery times are specific, so please
ensure the driver is available during those
hours.
```

Tvorba datasetu pomocou jazykových modelov

LLMs generovali aj otázky a k nim prislúchajúce odpovede. Kladený dôraz, aby sa odpovede nachádzali priamo v texte.

Príklady otázok a odpovedí generovaných modelom LLaMa:

- **Otázka 1:** Where do we need to pick up the goods?
- **Otázka 2:** What type of vehicle is needed for the transport?
- **Otázka 3:** How many packages are being transported?

Text: We need to pick up some goods from **J.C. Penney at Maple 8904 in Chicago, 38250, Bahrain** between 2021-03-04 13:30:00 and 2021-03-13 05:15:00. We'll be using a Small lorry to transport these goods, which include three packages with the following dimensions...

Tvorba datasetu pomocou jazykových modelov

LLMs generovali aj otázky a k nim prislúchajúce odpovede.
Kladený dôraz, aby sa odpovede nachádzali priamo v texte.

Príklady otázok a odpovedí generovaných modelom LLaMa:

- **Otázka 1:** Where do we need to pick up the goods?
- **Otázka 2:** What type of vehicle is needed for the transport?
- **Otázka 3:** How many packages are being transported?

Text: We need to pick up some goods from J.C. Penney at Maple 8904 in Chicago, 38250, Bahrain between 2021-03-04 13:30:00 and 2021-03-13 05:15:00. We'll be using a **Small lorry** to transport these goods, which include three packages with the following dimensions...

Tvorba datasetu pomocou jazykových modelov

LLMs generovali aj otázky a k nim prislúchajúce odpovede.
Kladený dôraz, aby sa odpovede nachádzali priamo v texte.

Príklady otázok a odpovedí generovaných modelom LLaMa:

- **Otázka 1:** Where do we need to pick up the goods?
- **Otázka 2:** What type of vehicle is needed for the transport?
- **Otázka 3:** How many packages are being transported?

Text: We need to pick up some goods from J.C. Penney at Maple 8904 in Chicago, 38250, Bahrain between 2021-03-04 13:30:00 and 2021-03-13 05:15:00. We'll be using a Small lorry to transport these goods, which include **three packages** with the following dimensions...

Tvorba datasetu pomocou jazykových modelov

Generovanie binárnych otázok predstavuje väčšiu výzvu.
Model musí porozumieť kontextu.

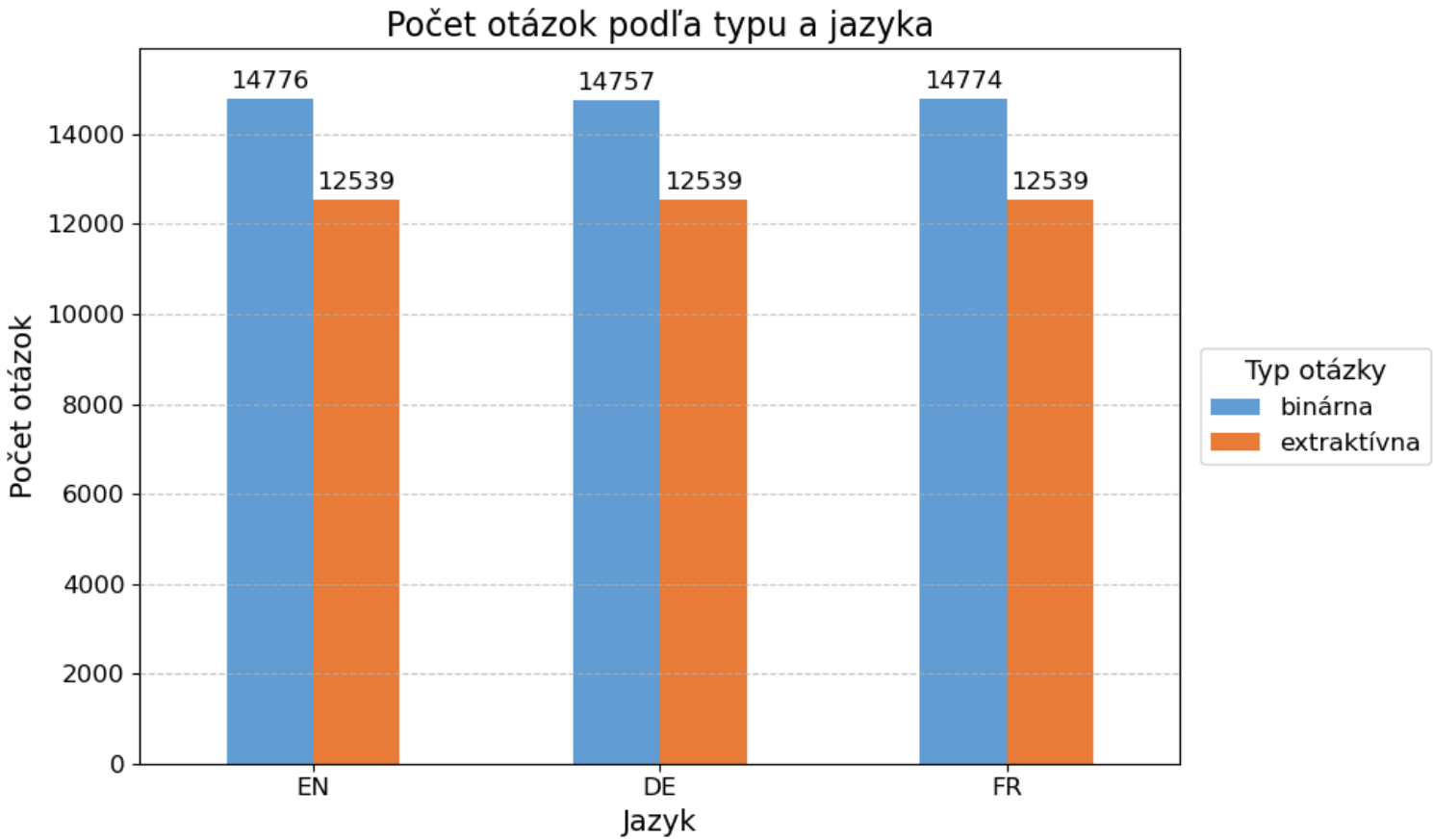
Príklad nesprávneho generovania:

Text: We need your assistance with shipping some goods from Tyson Foods ... We have a total of four pallets of goods that need to be transported, each with the following dimensions: ... Pallet 4: Length - 192 m, Width - 48 m, Height - 12 m, Weight - 18 kg. **All pallets are stackable** and require a Sprinter vehicle with a tilt function (curtain sided van)...

Otázka: Are the pallets stackable?
Generovaná Odpoveď: Nie
Správna odpoveď: Áno

Charakteristiky datasetu

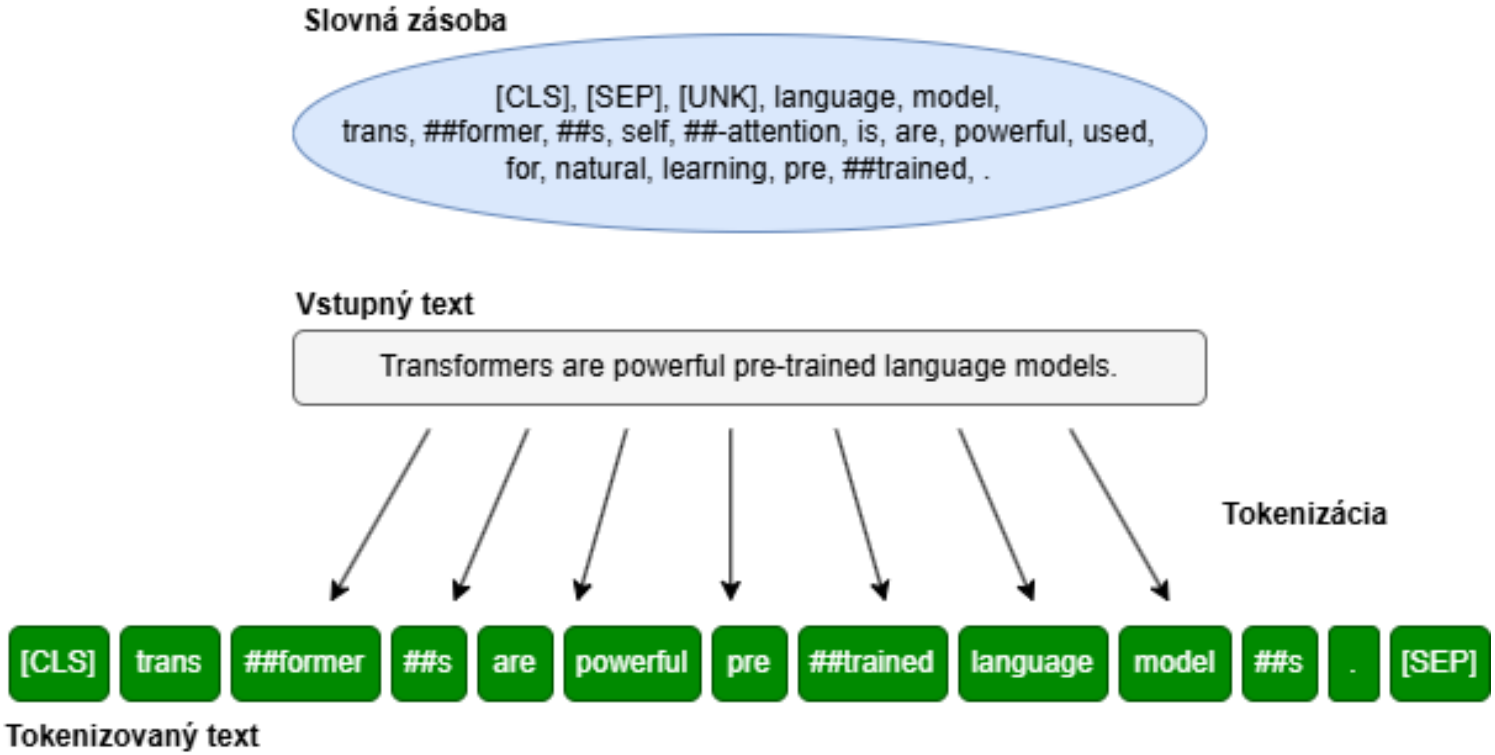
Výsledný dataset obsahuje spolu 81 924 otázok (37 617 extraktívny typ + 44 307 binárny typ).



Dataset sme rozdelili stratifikovane v pomere 80/20.

Tokenizácia

- Proces rozdeľovania textu na menšie podjednotky nazývané tokeny
- Tokeny môžu byť znaky, podslová, slová alebo celé vety.
- Väčšina modelov používa tokenizáciu na báze slov alebo podslov



Reprezentácia vstupu

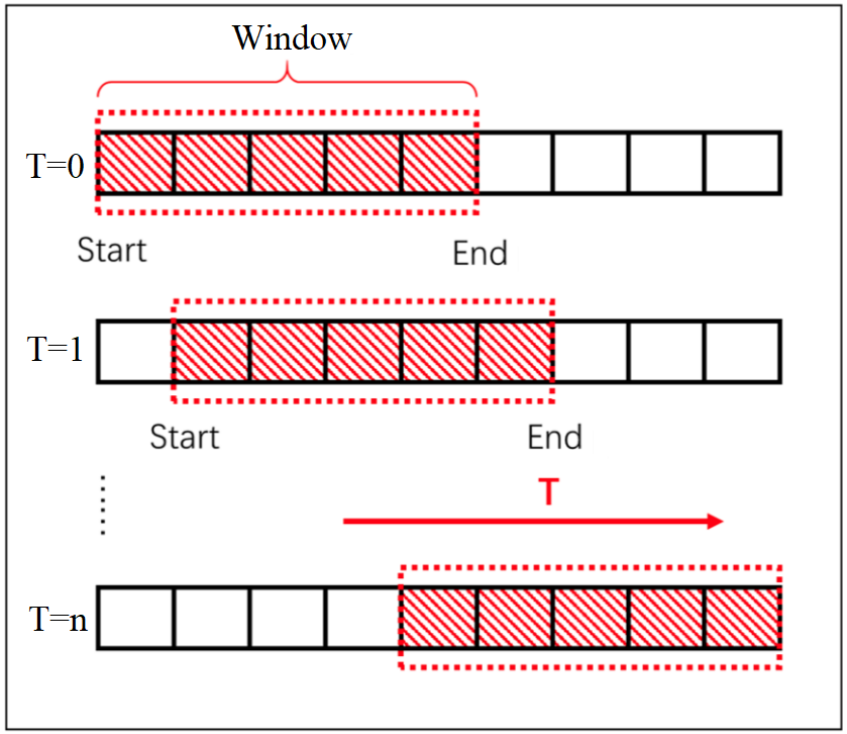
- Model na vstup dostáva tokenizovanú otázku a kontext oddelené špeciálnymi tokenmi

[CLS] Otázka [SEP] Kontext [SEP]

- [CLS] token sa pridáva na začiatok vstupu, jeho výstupná reprezentácia sa používa na klasifikačné úlohy
- [SEP] token slúži na oddelenie rôznych častí vstupu

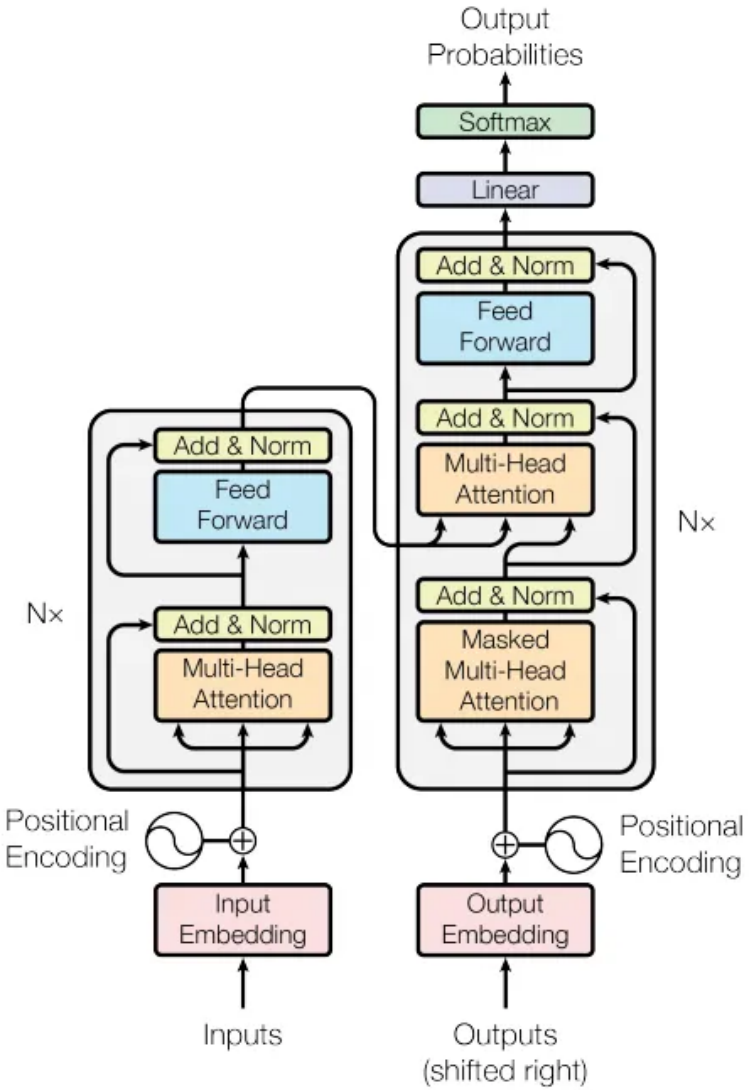
Sliding window

- Pri dlhých textoch sme narazili na limit vstupnej dĺžky
- Použili sme techniku sliding window, ktorá vytvára prekrývajúce sa segmenty
- Tento prístup maximalizuje pravdepodobnosť, že správna odpoveď sa bude nachádzať aspoň v jednom segmente celá



Zdroj: Gu et al., 2021, The Sliding Window and SHAP Theory

Architektúra Transformer



Zdroj: Vaswani et al., 2017, Attention is All You Need

Výpočet attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Hlavy pozornosti:

$$MultiHead(Q, K, V) = Concat(h_1, h_2, \ldots, h_n)W^O,$$

$$h_i = Attention(QW_i^Q, KW_i^K, VW_i^V).$$

Pozičné kódovanie:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right),$$

$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right),$$

Monolingválna úloha

- Extrahujeme anglický subset vygenerovaného datasetu
- Len otázky extraktívneho charakteru
- cca 4200 otázok distribuovaných medzi 500 kontextov

Použité modely založené na architektúre BERT:

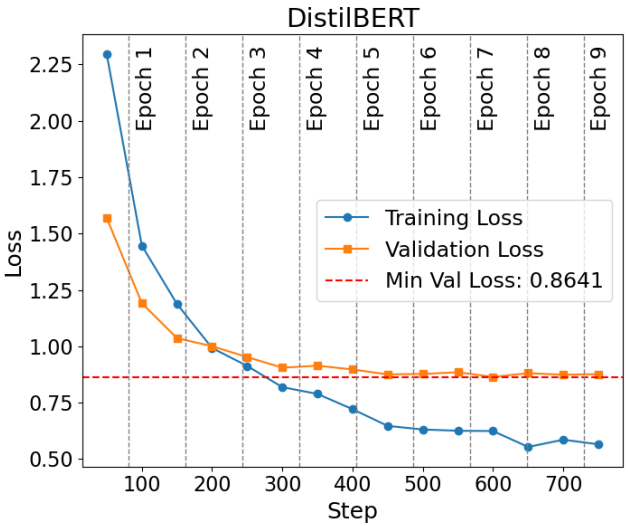
- DistilBERT - 66,4 mil. parametrov
- TinyRoBERTa - 81,5 mil. parametrov
- RoBERTa base - 125 mil. parametrov

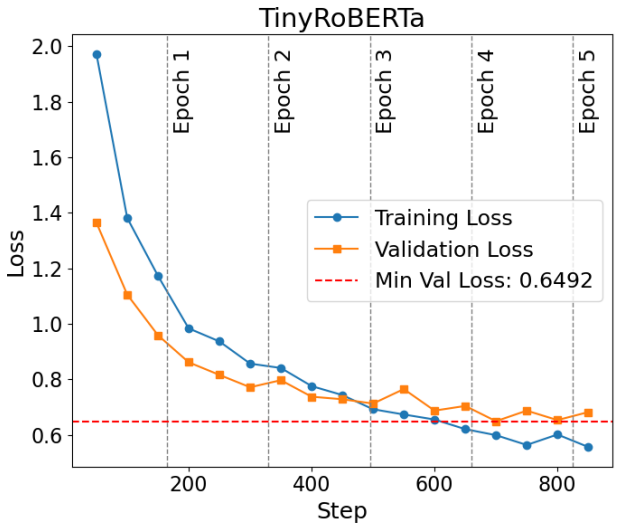
Trénovanie modelov

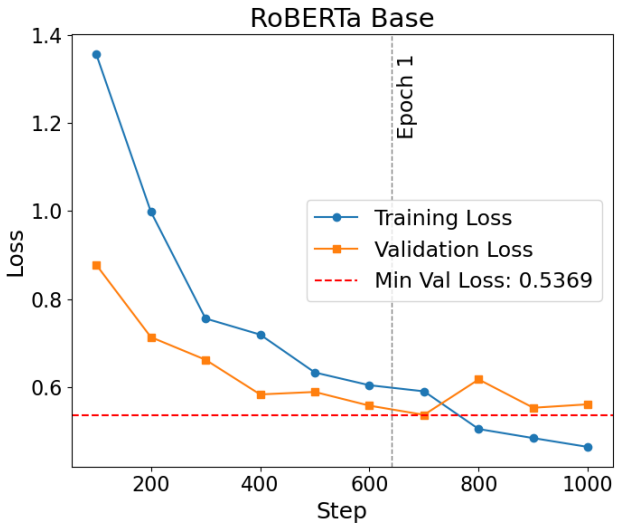
Tréningové parametre zvolené pre malý dataset:

- Malý batch-size: 8-16
- Menšie hodnoty učiaceho pomeru: 1e-5 – 2e-5
- Regularizácia váh: 0.01 - 0.1
- Warmup ratio: 10 % - stabilizuje učenie na začiatku
- Vyšší dropout: 20 % (default 10 %)
- Zmrazené vrstvy: 1 - 3 (v závislosti od veľkosti modelu)

Trénovanie modelov a výsledky







Model	EM	F1-skóre	Precision	Recall
DistilBERT	62,82 %	75,82 %	82,96 %	79,85 %
TinyRoBERTa	66,28 %	79,03 %	84,35 %	80,86 %
RoBERTa Base	68,42 %	80,21 %	85,25 %	82,00 %

Multilingválna úloha

- Používame celý dataset (všetky jazyky a typy otázok)
- Jazyky kontextov a otázok sú rôzne kombinované (napr. ang. kontext - nem. otázka)

Použité modely založené na architektúre X-LMR:

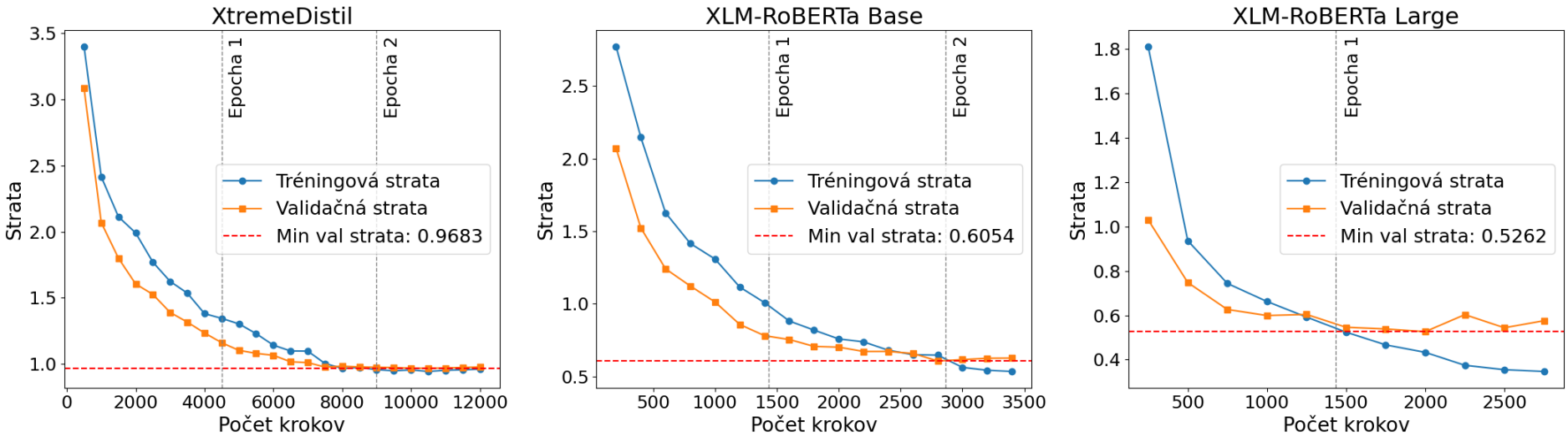
- XtremeDistil - 33 mil. parametrov
- XLM-RoBERTa Base - 279 mil. parametrov
- XLM-RoBERTa Large - 560 mil. parametrov

Trénovanie modelov

Zvolili sme vhodné tréningové parametre pre veľký dataset:

- Väčší batch-size: 16-32
- Menšie hodnoty učiaceho pomeru: 1e-5 – 2e-5
- Regularizácia váh: 0.01
- Warmup ratio: 10% - stabilizuje učenie na začiatku
- Nižší dropout: 10 % - 15 % (default 10 %)
- Zmrazené vrstvy: žiadne

Trénovanie modelov



Obr. 1: Grafy loss funkcie

Výsledky

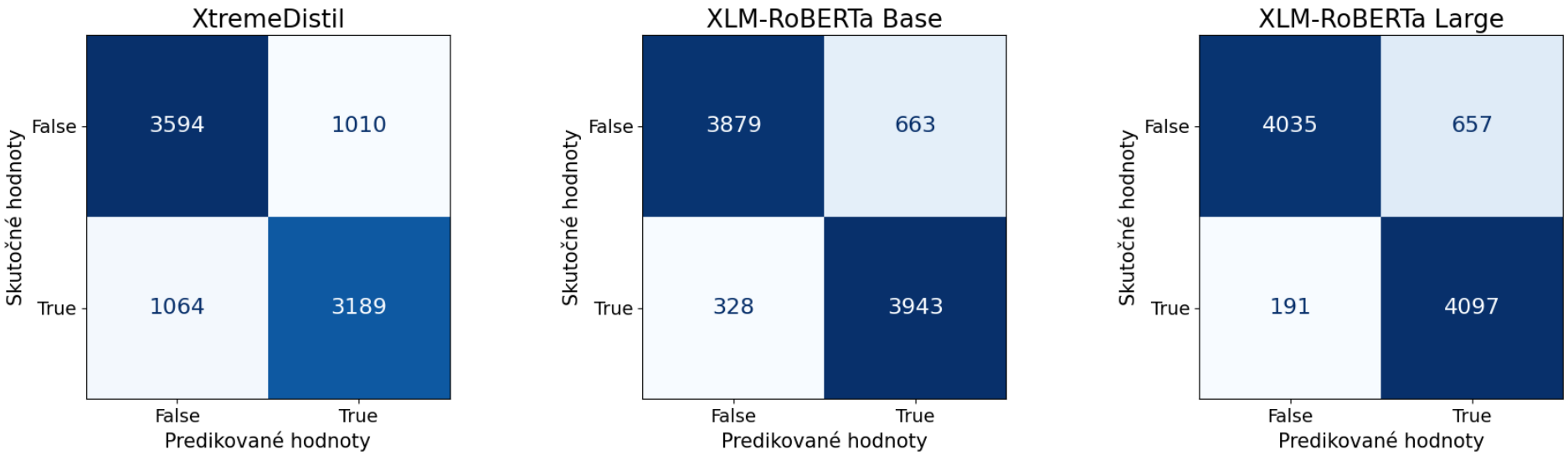
Model	EM	F1-skóre	Precision	Recall
XtremeDistil	59,71 %	77,20 %	74,77 %	79,80 %
XLM-RoBERTa Base	76,78 %	89,37 %	89,11 %	89,65 %
XLM-RoBERTa Large	81,40 %	91,21 %	91,44 %	90,99 %

Tabuľka 1: Výsledky modelov na extraktívnych otázkach

Model	Accuracy	F1-skóre	Precision	Recall
XtremeDistil	76,58 %	75,45 %	75,95 %	74,98 %
XLM-RoBERTa Base	88,76 %	88,84 %	85,61 %	92,32 %
XLM-RoBERTa Large	90,56 %	90,62 %	86,18 %	95,55 %

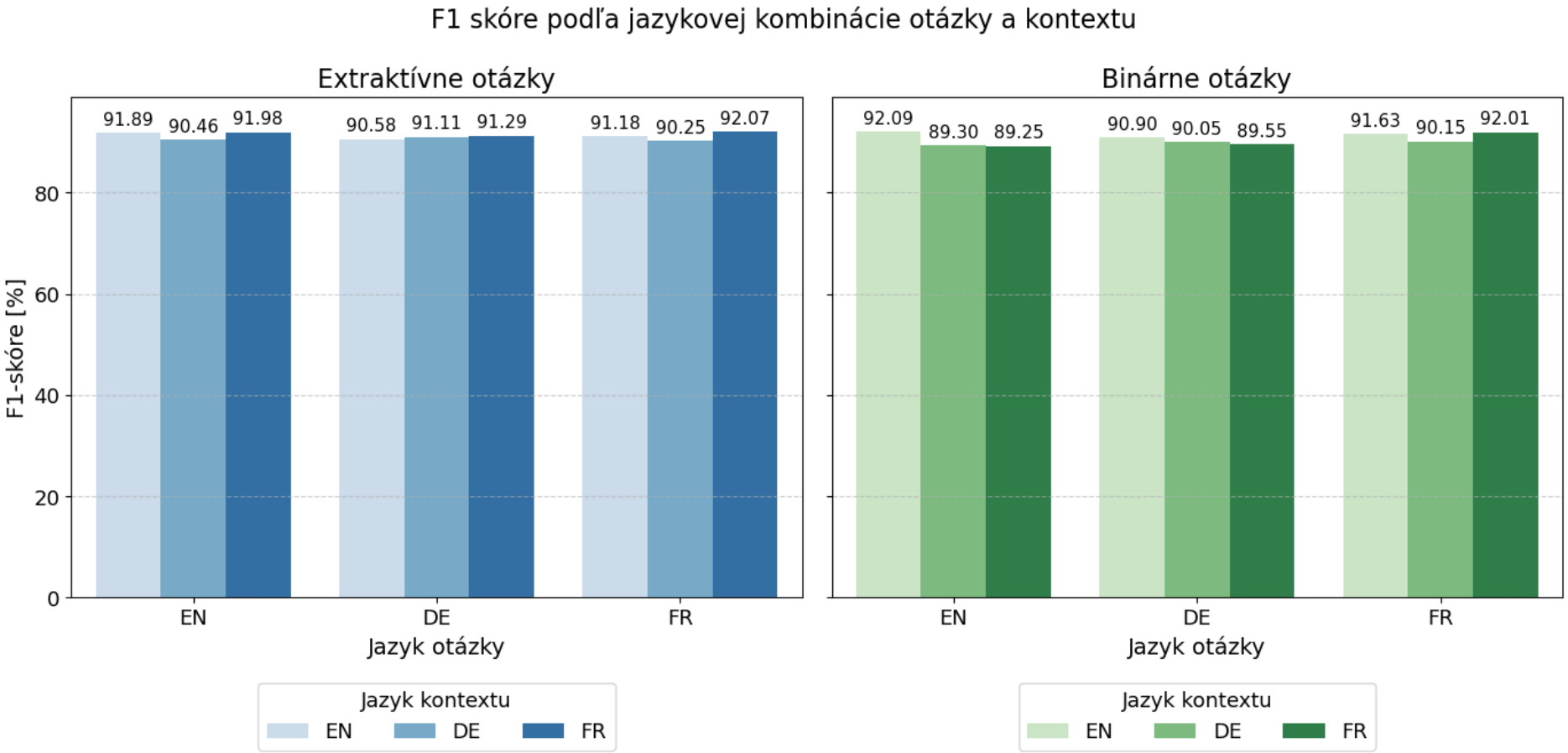
Tabuľka 2: Výsledky modelov na binárnych otázkach

Výsledky

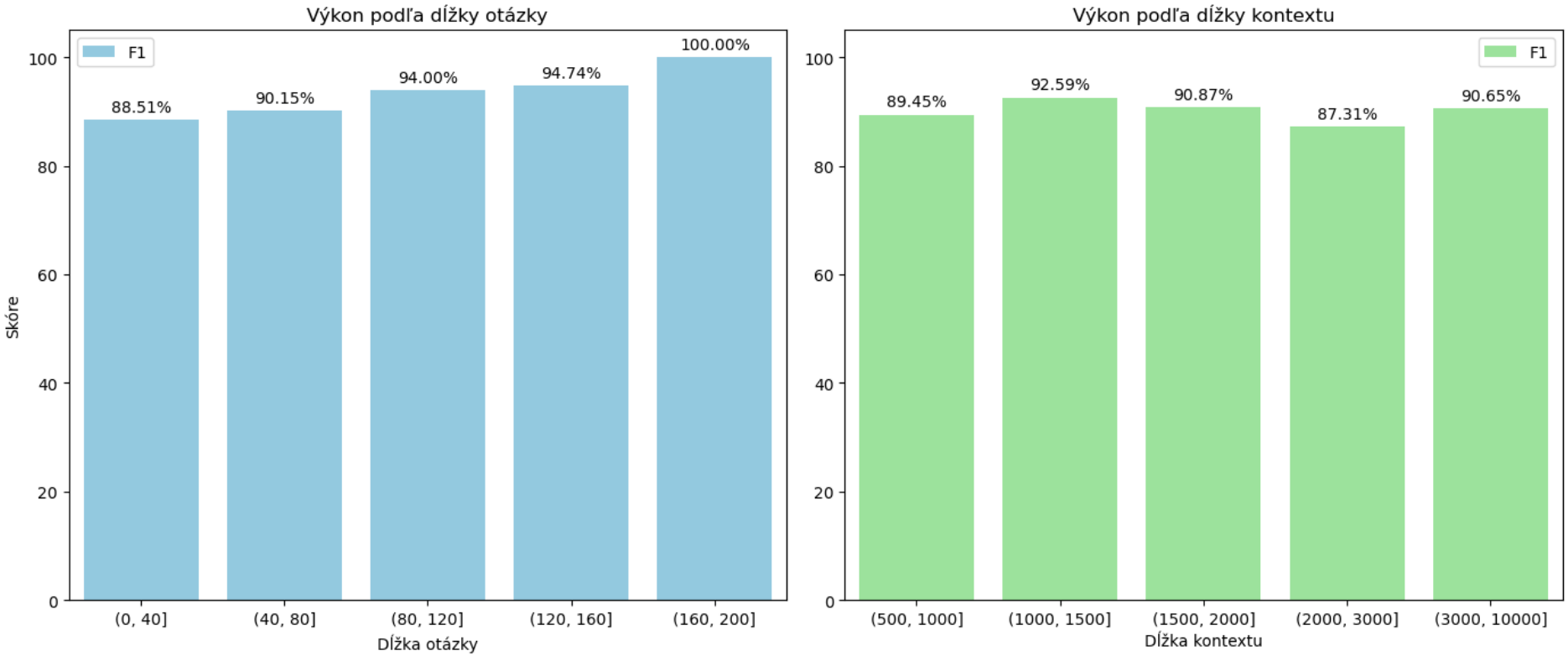


Obr. 2: Matice zámeny pre binárne otázky

Výkon modelu XLM-RoBERTa Large



Výkon modelu XLM-RoBERTa Large (binárne otázky)



Wordcloud chybne klasifikovaných binárnych otázok

Obr. 3: Wordcloud chybne klasifikovaných binárnych otázok

Model má problém klasifikovať otázky ohľadom rozmerov a hmotnosti balíčku.
Taktiež má problém správne identifikovať čas a miesto vyzdvihnutia/doručenia

Príklad chybne klasifikovanej binárnej otázky

Príklad falošne negatívne klasifikovanej otázky:

Text: We have a delivery for you! Our team will be picking up a shipment of goods from Deutsche Bank ... We'll be using a LKW vehicle with a tail lift, as requested. The goods include four packages with the following dimensions: * **Package 1**: 25cm x 46cm x 52cm, **weighing 52 kg** * Package 2: 61cm x 66cm x 22cm, weighing 52 kg ... We'll make sure to handle your shipment with care and get it to its destination safely.

Otázka: Does Package 1 weigh more than 50 kg?
Predikovaná odpoveď: Nie
Správna odpoveď: Áno

Interpretácia výsledkov a záver

- Tréning všetkých modelov prebiehal stabilne a skonvergoval,
- Pri binárnych otázkach presnosť klasifikácie rástla s dĺžkou otázky.

- TinyRoBERTa mal porovnateľné výsledky s RoBERTa Base,
- XLM-RoBERTa Base a Large mali veľmi podobné výsledky,
- Extrémne distilovaná verzia X-LMR si poradila slušne.

Ďalšia práca:

- Experimentovať s predspracovaním dát
- Rozšíriť dataset o ďalšie jazyky a porovnať jazykové rodiny (románske, germánske, . . .)

Zdroje

[1] Gu, X. et al., 2021. The Sliding Window and SHAP Theory—An Improved System with a Long Short-Term Memory Network Model for State of Charge Prediction in *Electric Vehicle Application*. Energies, Vol. 14, No. 12, Article No. 3692. ISSN: 1996-1073.

[2] Vaswani, A. et al., 2017. Attention Is All You Need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Vol. 30. ISSN: 1049-5258.

Ďakujem za pozornosť

Otázky

1. Dali by sa použiť LLM/Foundation modely aj na samotnú úlohu odpovedania otázok namiesto bežných transformerových modelov ako sú vairanty BERT-u? Ak áno, aké sú výhody a nevýhody takéhoto prístupu?

2. Ako ste riešili kontrolu kvality automaticky generovaného datasetu? Nepredstavuje riziko nekonzistentnosť generovaných otázok a odpovedí?

3. Aký vplyv má veľkosť modelu na výpočtové nároky v porovnaní s prínosom vo výkone? Je podľa vás vždy výhodné používať veľké modely?