Metódy transformérov a ich aplikácia v oblasti textovej analýzy

Študent: Martin Bača

Vedúci: RNDr. Šimon Horvát, PhD.

Konzultant: doc. RNDr. Ľubomír Antoni, PhD.

Výsledky

- Vytvorenie datasetu
- Tokenizácia a predspracovanie
- Trénovanie a evaluácia modelov

Vytvorenie datasetu

- Model Llama2 generoval otázky a odpovede
- Dôležitá úloha: Zadať modelu správne dopyty
- Kategórie odpovedí:
  - Text
  - Číslo
  - Pravda/Nepravda
- Odpovede typu text - ak odpovede nemali presnú zhodu v texte tak sme ich vymazali

Ukážka datasetu

Text


```
{
    "id": "1",
    "question": "Where are the goods being picked up from?",
    "answers": [
        {
            "text": "Wuhan Iron & Steel at Apple 1969 in Durham, 48348, Austria",
            "answer_type": "string",
            "answer_start": 178
        }
    ]
},
```


Číslo


```
{
    "id": "7",
    "question": "What is the total weight of the goods being transported?",
    "answers": [
        {
            "text": "92",
            "answer_type": "number"
        }
    ]
},
```


Pravda/Nepravda


```
{
    "id": "12",
    "question": "Are there four different items to be transported?",
    "answers": [
        {
            "text": "True",
            "answer_type": "boolean"
        }
    ]
},
```

Tokenizácia a predspracovanie

- Prevod textov na tokeny – rozdelenie na menšie časti (tokeny) a priradenie číselnej hodnoty každému tokenu
- Dlhé texty je potrebné rozdeliť na chunky
- Formát vstupných chunkov: [CLS]Question[SEP]Context[SEP]

Tokenizácia a predspracovanie

- Potrebné nájsť začiatky a konce odpovedí v textoch
- Transformovať ich na začiatočné a koncové tokeny

```
Answer text: Small lorry
Answer context: Small lorry
Answer start, end: 307, 318
Context start, end: 14, 310
Window start, end: 1, 1139
Start and end token (beginning): 14, 310
Start and end token (end): 97, 99
(307, 312)
(314, 318)
Window 1
```

...indexy odpovede v texte (string)

...tokeny zodpovedajuce odpovedi

Trénovanie a evaluácia modelov

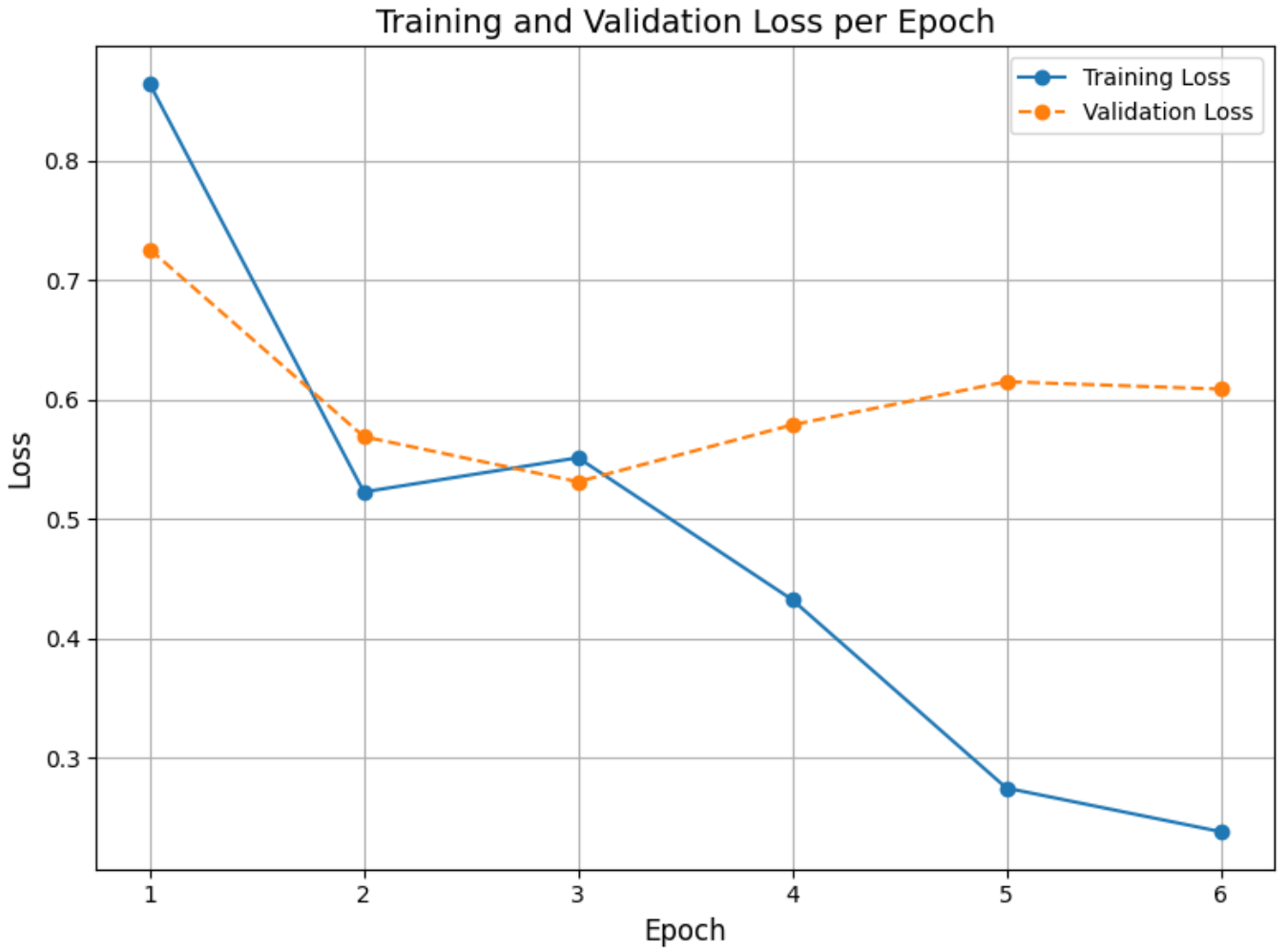
- Vstupy modelu:
  - **Input_ids** – tokenizované otázky a kontexty
  - **attention_mask** – indikuje ktoré tokeny sú vstup a ktoré padding
  - **offset_mapping** – umožňuje prevod tokenov naspäť na text
  - **overflow_to_sample_mapping** – indikuje, ku ktorému vstupnému textu patrí chunk
  - **start_positions** - začiatok odpovede (zač. token) v texte
  - **end_positions** - koniec odpovede (kon. token) v texte

Trénovanie a evaluácia modelov

- Natrénované modely:
  - **deepset/tinyroberta-squad2** (6 mil. parametrov)
  - **distilbert-base-uncased-distilled-squad** (66 mil. parametrov)
  - **deepset/roberta-base-squad2** (125 mil. parametrov)
- Každý z modelov poskytuje vlastný tokenizér
- EarlyStopping – zastaví tréning keď sa model začne preúčať
- Evaluácia po chunkoch a agregovanie pravdepodobností do jednej finálnej odpovede

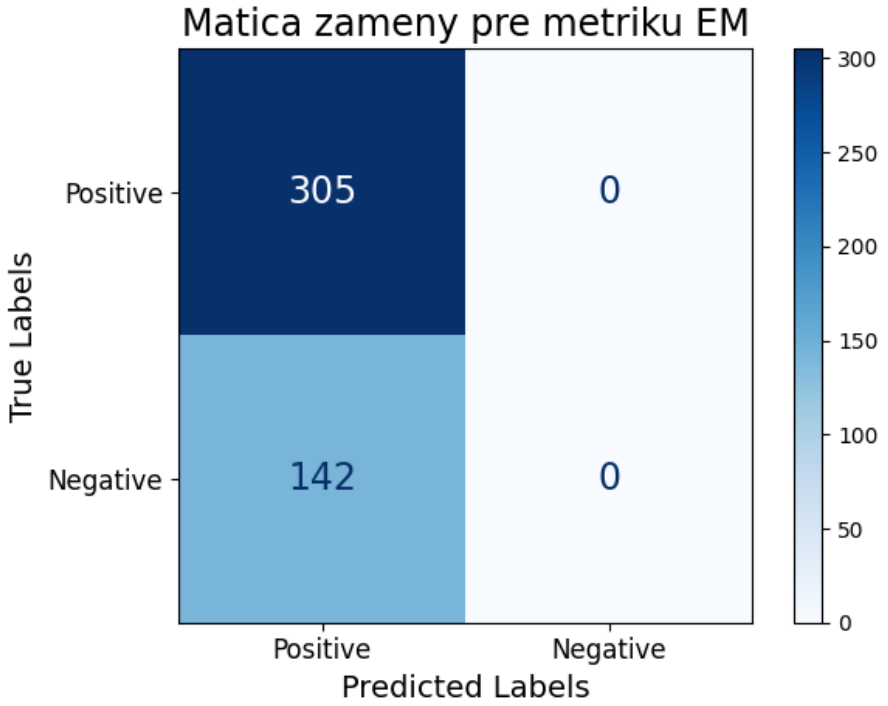
deepset/tinyroberta-squad2

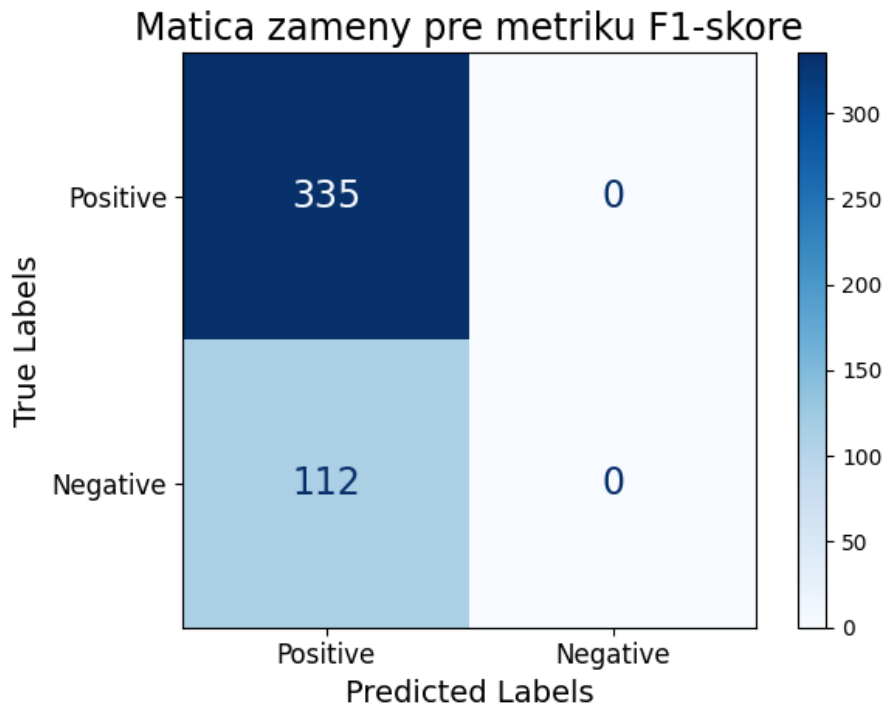
Epoch	Training Loss	Validation Loss
1	0.864400	0.725888
2	0.522800	0.569030
3	0.551400	0.531322
4	0.432600	0.578897
5	0.275000	0.615059
6	0.238500	0.608854



deepset/tinyroberta-squad2

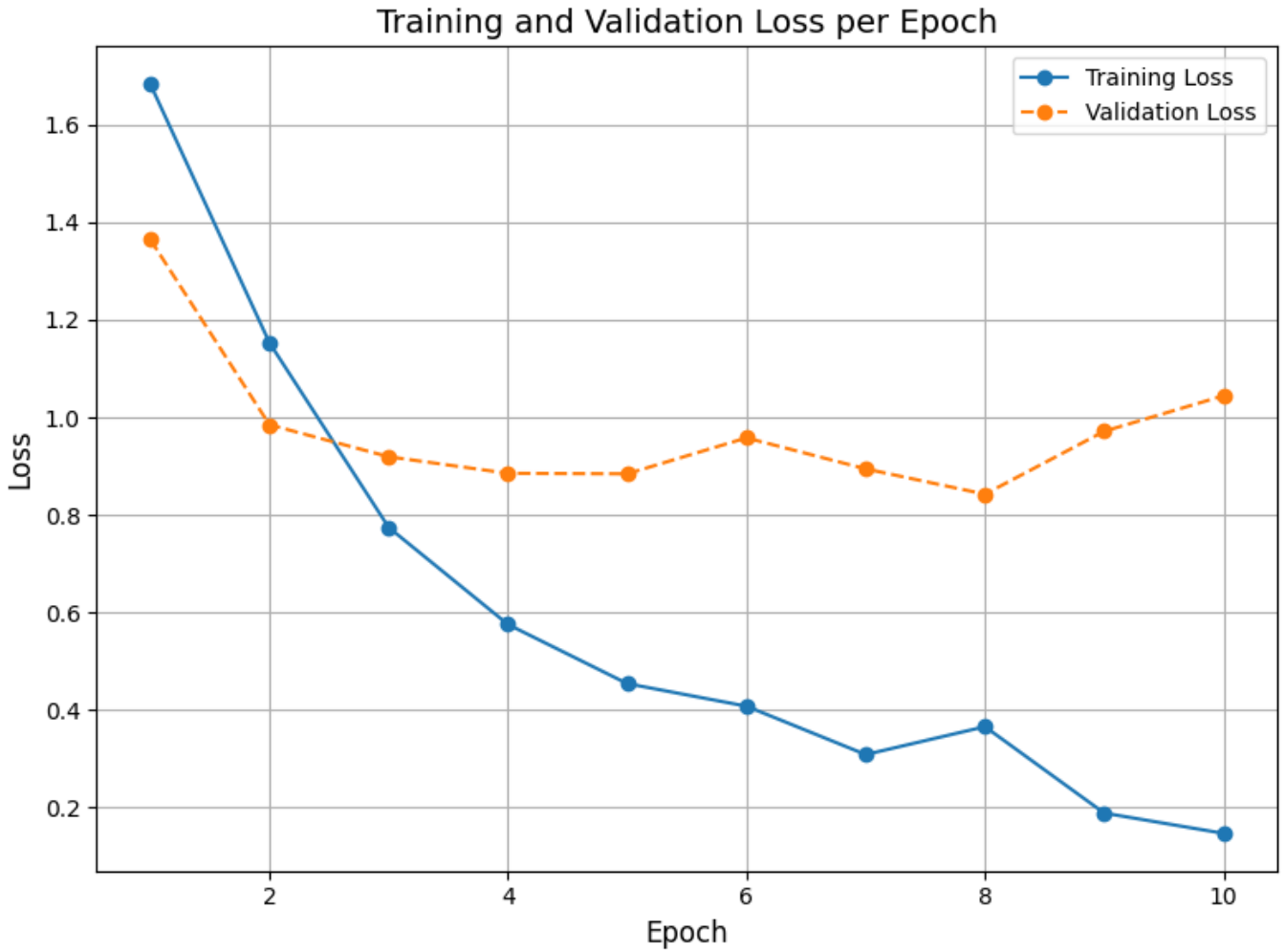
- Average Exact Match Score: 0.682326621923937 3
- Average F1 Score: 0.807460126755413 8
- Average Precision: 0.867638917135561 4
- Average Recall: 0.822211947229608 8





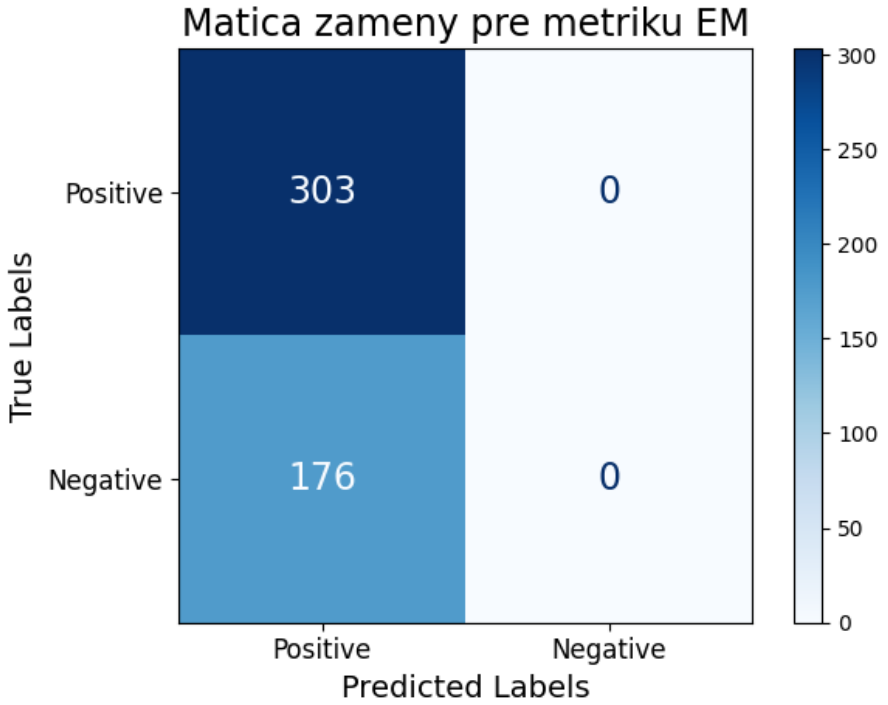
distilbert-base-uncased-distilled-squad

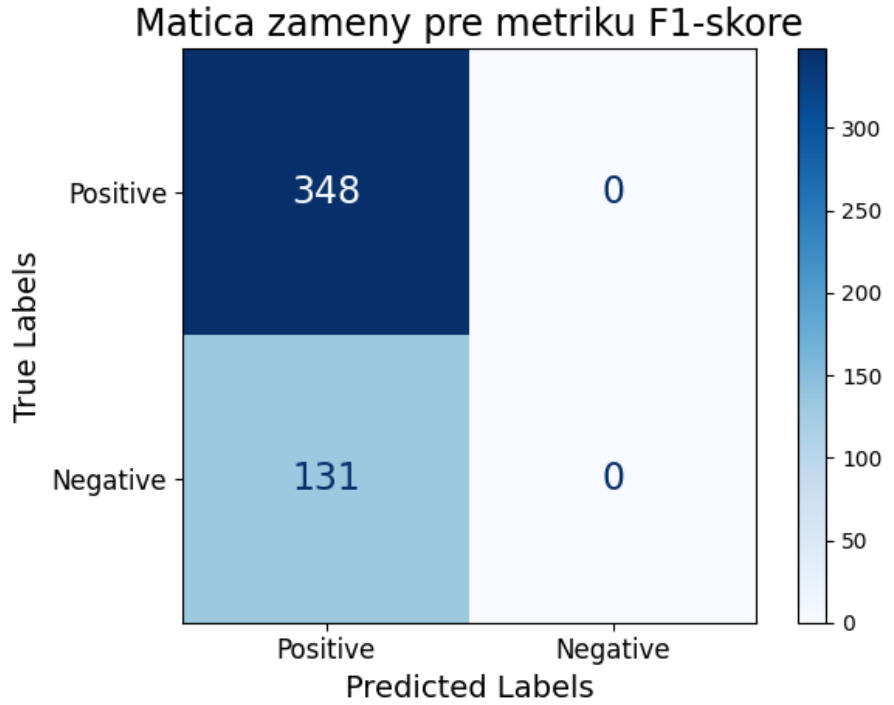
Epoch	Training Loss	Validation Loss
1	1.683800	1.365056
2	1.153100	0.985049
3	0.774500	0.919472
4	0.575700	0.885489
5	0.454400	0.884436
6	0.408100	0.958166
7	0.308900	0.894583
8	0.366100	0.842664
9	0.189000	0.971633
10	0.147400	1.044379



distilbert-base-uncased-distilled-squad

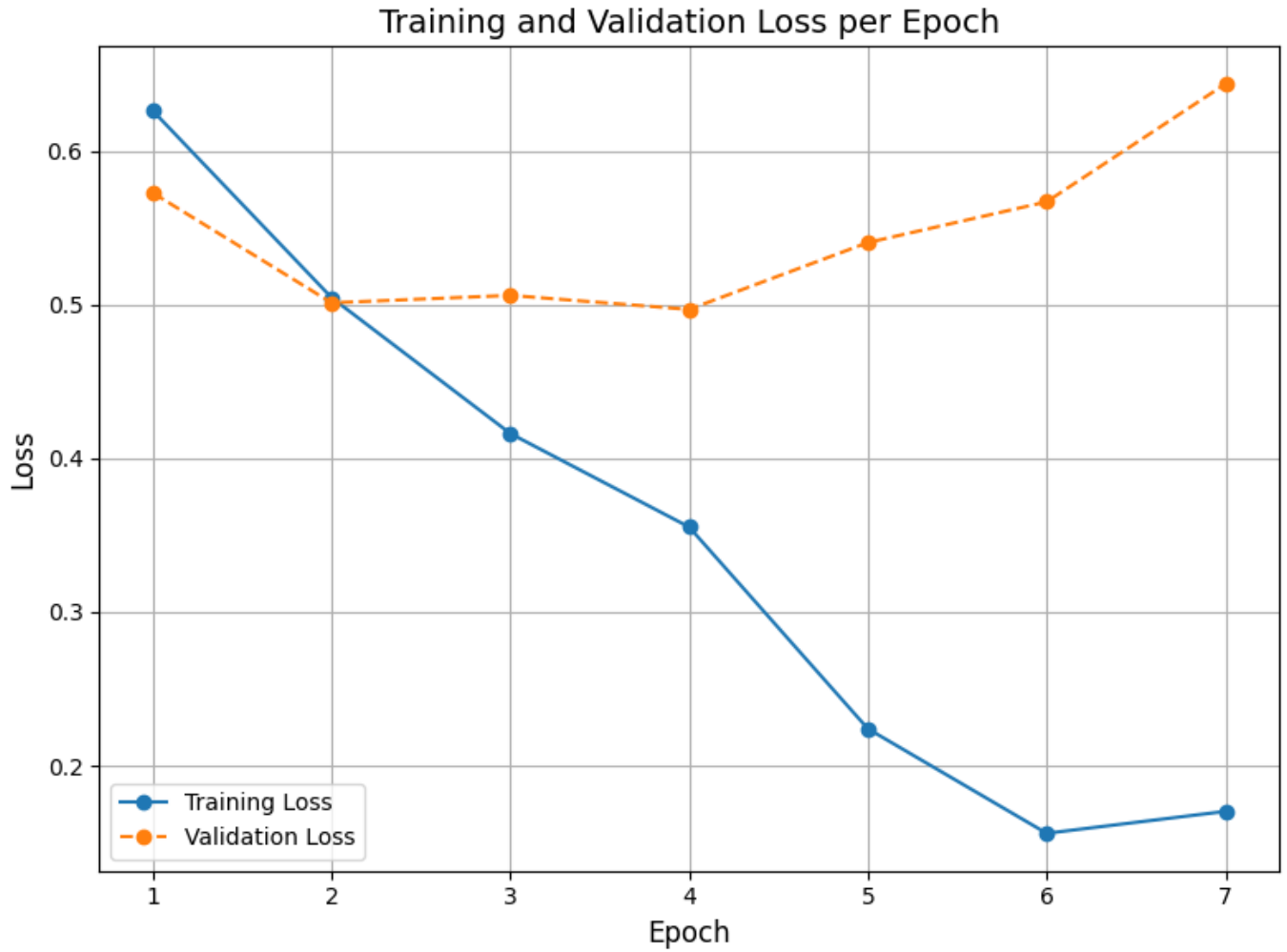
- Average Exact Match Score: 0.6325678496868476
- Average F1 Score: 0.7851342626741173
- Average Precision: 0.8570227325446532
- Average Recall: 0.8118079147240547





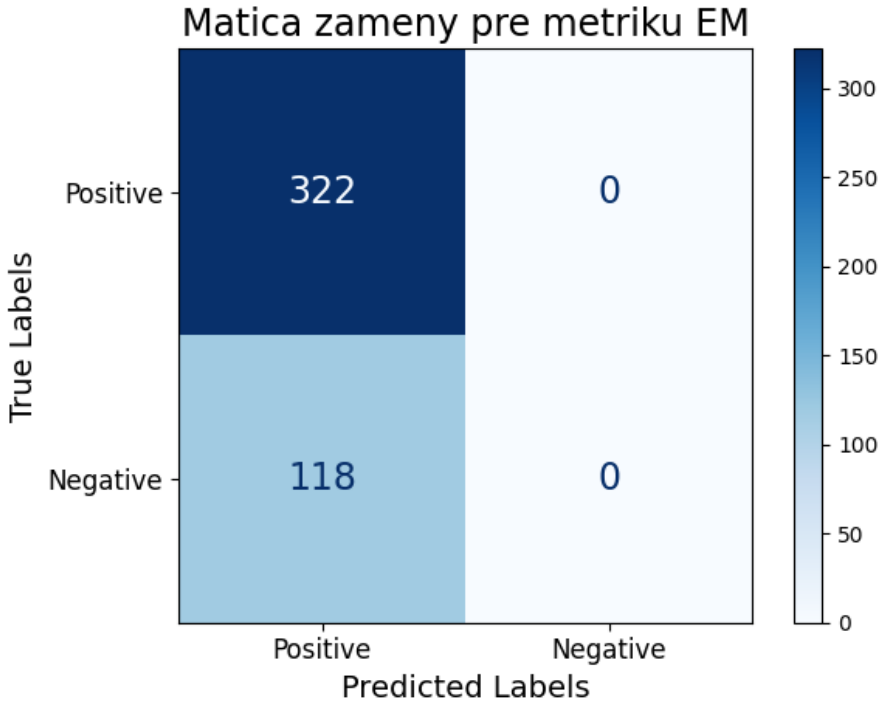
deepset/roberta-base-squad2

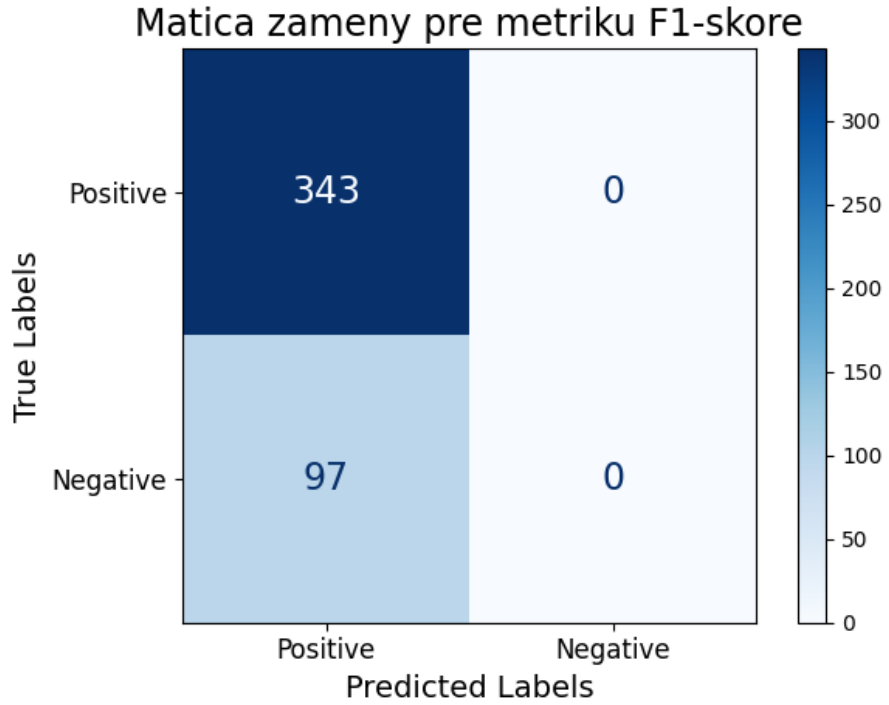
Epoch	Training Loss	Validation Loss
1	0.626500	0.573012
2	0.504800	0.501424
3	0.416300	0.506264
4	0.355400	0.497132
5	0.224000	0.540664
6	0.155900	0.567217
7	0.170200	0.644001



deepset/roberta-base-squad2

- Average Exact Match Score: 0.7318181818181818
- Average F1 Score: 0.8317352221923987
- Average Precision: 0.8958360389610388
- Average Recall: 0.8471047727193062





Záver

- Roberta base – Najlepšie výsledky podľa očakávaní
- Distilbert – prekvapivo horšie výsledky oproti Tiny Roberte aj napriek väčšiemu počtu parametrov
- Tiny Roberta – schopná efektívne vykonať Q&A úlohu
  - Rýchlejší beh dopytov ako Roberta base – pomer efektivity a časovej náročnosti je lepší pri Tiny Roberte

Ďalšie kroky

- Rozšírenie datasetu o ďalšie otázky a odpovede
- Pridanie ďalších jazykov (GER, FRA, ITA, ESP,...)
- Natrénovanie modelu na odpovede typu číslo a boolean

Ďakujem za pozornosť