



**VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ**

BRNO UNIVERSITY OF TECHNOLOGY

**FAKULTA INFORMAČNÍCH TECHNOLOGIÍ**

FACULTY OF INFORMATION TECHNOLOGY

**ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ**

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

**ANALÝZA DĚJOVÝCH LINIÍ NA ZÁKLADĚ SHRNU  
TÍ OBSAHU KNIH A UŽIVATELSKÝCH RECENZÍ**

PLOT ANALYSIS FROM BOOK SUMMARIES AND USER REVIEWS

**BAKALÁŘSKÁ PRÁCE**

BACHELOR'S THESIS

**AUTOR PRÁCE**

AUTHOR

**PETER RÚČEK**

**VEDOUcí PRÁCE**

SUPERVISOR

**doc. RNDr. PAVEL SMRŽ, Ph.D.**

BRNO 2022

## Zadání bakalářské práce



Student: **Růček Peter**  
Program: Informační technologie  
Název: **Analýza dějových linií na základě shrnutí obsahu knih a uživatelských recenzí**  
**Plot Analysis from Book Summaries and User Reviews**  
Kategorie: Web

### Zadání:

1. Prostudujte postupy a nástroje pro shromažďování obsahů knih a uživatelských recenzí z populárních služeb typu Goodreads
2. Seznamte se s pokročilými metodami strojového učení, používanými pro automatickou tvorbu shrnutí (sumarizaci) textů, reprezentaci významu a určování podobnosti
3. Navrhněte a implementujte systém, který dokáže pravidelně získávat, indexovat a analyzovat stahovaná data
4. Na základě získaných znalostí navrhněte a realizujte systém pro analýzu dějových linií a určování podobnosti a odlišnosti příběhů a reprezentaci výsledků na základě extrahovaných klíčových slov, týkajících se děje
5. Vytvořte testovací sadu pro vyhodnocení přesnosti a úplnosti hledání a vyhodnořte vytvořený systém pomocí standardních metrik
6. Vytvořte stručný plakát prezentující práci, její cíle a výsledky

### Literatura:

- dle dohody s vedoucím

Pro udělení zápočtu za první semestr je požadováno:

- funkční prototyp řešení

Podrobné závazné pokyny pro vypracování práce viz <https://www.fit.vut.cz/study/theses/>

Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**

Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.

Datum zadání: 1. listopadu 2021

Datum odevzdání: 11. května 2022

Datum schválení: 1. listopadu 2021

## Abstrakt

Cielom tejto práce je vytvoriť systém pre analýzu a klasifikáciu kľúčových dejových línií zo zhrnutých dejových zápletiiek a užívateľských recenzií v anglickom jazyku. Zvolený problém je riešený pomocou techniky strojového učenia založenej na transformeroch. Vo vytvorenom riešení je implementované aj sťahovanie dát a bol vytvorený dataset užívateľských recenzií a informácií o knihách prevyšujúci 23 miliónov recenzií a takmer 900 tisíc informácií o knihách. Systém dokáže predikovať aké typy dejových zápletiiek sa v dátach nachádzajú.

## Abstract

The aim of this work is to create a system for analysis and classification of plot keywords from summarized storylines and user reviews in English. The chosen problem is solved using a transformer-based machine learning technique. The created solution also implements data downloading and a dataset of user reviews and information about books was created, exceeding 23 million reviews and 900 thousand information about books. The system can predict what plot keywords the data contains.

## Kľúčové slová

spracovanie prirodzeného jazyka, strojové učenie, neurónové siete, bert, dejové línie, užívateľské recenzie, klasifikácia, multi-label, extrakcia dát z webu

## Keywords

natural language processing, machine learning, neural networks, bert, plots, user reviews, classification, multi-label, webscraping

## Citácia

RÚČEK, Peter. *Analýza dějových linií na základě shrnutí obsahu knih a užívateľských recenzí*. Brno, 2022. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

# Analýza dějových linií na základě shrnutí obsahu knih a uživatelských recenzí

## Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána docenta RNDr. Pavla Smrža Ph.D. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....  
Peter Rúček  
8. mája 2022

## Podakovanie

Ďakujem vedúcemu mojej bakalárskej práce docentovi RNDr. Pavlovi Smržovi Ph.D. za odbornú pomoc, rady a vedenie pri vypracovaní tejto bakalárskej práce.

# Obsah

<b>1</b>	<b>Úvod</b>	<b>3</b>
<b>2</b>	<b>Rozbor riešenej problematiky</b>	<b>5</b>
2.1	Príbeh . . . . .	5
2.2	Kľúčové dejové línie . . . . .	7
2.3	Extrakcia dát . . . . .	7
2.4	Spracovanie prirodzeného jazyka . . . . .	8
2.5	Klasifikácia . . . . .	9
2.6	Klasifikačné metódy . . . . .	10
2.7	Transformer . . . . .	12
2.8	BERT . . . . .	14
2.8.1	Fine-tuning . . . . .	15
2.8.2	Spracovanie dlhých refazcov . . . . .	15
2.9	Metriky vyhodnocovania . . . . .	15
2.10	Metriky podobnosti . . . . .	19
<b>3</b>	<b>Dáta</b>	<b>20</b>
3.1	Dátová sada z IMDb . . . . .	20
3.2	Dátová sada Booksummaries . . . . .	21
3.3	Dátová sada z Goodreads . . . . .	21
<b>4</b>	<b>Návrh systému</b>	<b>23</b>
4.1	Motivácia . . . . .	23
4.2	Analýza dát . . . . .	23
4.3	Schéma systému . . . . .	25
4.4	Zhrnutie návrhu . . . . .	26
<b>5</b>	<b>Implementácia systému</b>	<b>27</b>
5.1	Architektúra projektu . . . . .	27
5.2	Dáta . . . . .	28
5.3	Klasifikácia . . . . .	29
<b>6</b>	<b>Experimentálne vyhodnotenie a diskusia</b>	<b>32</b>
6.1	Dátová sada . . . . .	32
6.2	Najlepšie dosiahnuté štatistiky . . . . .	33
6.3	Príklady . . . . .	33
6.4	Analýza výsledkov a chyby . . . . .	40
6.5	Možné vylepšenia a využitie . . . . .	41

<b>7 Záver</b>	<b>42</b>
<b>Literatúra</b>	<b>43</b>
<b>A Obsah priloženého pamäťového média</b>	<b>46</b>
<b>B Zoznam kľúčových dejových línií</b>	<b>48</b>
<b>C Zhrnutia dejov</b>	<b>50</b>
<b>D Plagát</b>	<b>56</b>

# Kapitola 1

## Úvod

So stále narastajúcim množstvom digitalizovaných príbehov a sumarizovaných dejov narastá potreba analýzy a interpretácie nie len literárnych textov. Ich potenciál sa v prvom rade zameriava na širší pohľad pochopenia príbehu počítačom, pričom sľubuje analýzu viacerých literárnych textov súčasne.

Analýza príbehov má pomôcť jednému z pradávných cieľov umelej inteligencie: *pochopeniu prirodzeného jazyka*. Umeľá inteligencia, ktorá by dokázala pochopiť prirodzený jazyk so všetkými nuansami, a vedela by okrem iného aj prerozprávať príbeh, zhrnúť ho, pochopiť zmysel príbehu, atď., by bola obrovským prínosom pre spoločnosť.

Využitie analýzy príbehov má aj komerčné účely, napr. pri cielej reklame alebo pri určovaní úspešnosti diela. Detekcia kliše, spoilerov či určovanie podobnosti diel sú ideálnymi príkladmi pre systémy, ktoré sa zameriavajú na analýzu dejov. Ďalším veľmi praktickým a zaujímavým použitím analýzy príbehov je určenie konkrétneho diela len na základe strohého opisu udalostí alebo postáv v diele. Každému sa určite stalo, že si nevedel spomenúť na názov konkrétneho diela, ale pamätal si o ňom iba nepatrné množstvo údajov. Z tohoto podnetu vznikli rôzne fóra, ktoré sa snažia nájsť odpovede na tieto otázky. Aj na takéto originálne prípady je možné nasadiť analýzu príbehov.

V oblasti sociálnych sietí sa pri popise diela bežne používajú takzvané *tagy*, ktoré určujú kľúčové prvky deja. Pomocou týchto *tagov* a stručného deja príbehu sa dá získať lepšia predstava o danom diele. Správnou extrakciou dát je možné získať takzvané kľúčové dejové línie, ktoré popisujú hlavné témy, ktorými sa daný príbeh vyznačuje (milostný trojuholník, rodinné vzťahy či morálna dilema).

Táto práca sa zameriava na hľadanie týchto kľúčových dejových línií v zhrnutiach dejov. Systém je navrhnutý tak, že dokáže pracovať s akýmkoľvek sumarizovaným príbehom, a na základe strojového učenia systém predpovedá, o aký typ deja by sa podľa jeho názoru mohlo jednať. Takisto je systém schopný určovať podobnosť diel na základe týchto dejových línií. Popri implementácii systému je potrebné zozbierať dáta týkajúce sa kníh, zhrnutí kníh a užívateľských recenzií.

Poznatky z tejto práce by sa dali využiť v sociálnych sieťach orientovaných na knihy, filmy či časopisy, kde by diela mohli byť klasifikované podľa týchto kľúčových dejových línií. Systém by mohol hľadať podobnosti v jednotlivých dielach, a užívateľovi tak navrhovať ďalšie diela s obdobnou tematikou.

Práca je členená do siedmych kapitol. Kapitole 2 je zameraná na všeobecný teoretický pohľad, potrebný pre pochopenie problematiky. Zaoberá sa popisom deja, prehľadávaním webu, klasifikačnými metódami a následne uvedením doterajších a aj súčasných prístupov strojového učenia pri spracovaní prirodzeného jazyka. Dôležitou časťou je popis dát, čomu

je venovaná kapitola 3. Výsledný systém je navrhnutý v kapitole 4. Na základe návrhu sa v kapitole 5 detailne rozoberie implementácia funkčného systému. Kapitola 6 popisuje jednotlivé experimenty nad dátami. Záverečná 7. kapitola zhodnocuje ciele systému a dosiahnuté výsledky.



## Kapitola 2

# Rozbor riešenej problematiky

Táto kapitola obsahuje základné teoretické znalosti, z ktorých sa vychádza pri návrhu systému a následne pri jeho realizácii. Táto práca sa zaoberá vytvorením systému pre analýzu dejových línií zo zhrnutí dejov. Jednotlivé podkapitoly sa viažu k samotnému návrhu systému. Najskôr je popísané, čo je to dej, aké existujú kľúčové dejové línie a ako ich rôzni autori definujú. Následne je popísané, ako sa získavajú dáta z webu. Potom je bližšie priblížený problém klasifikácie. Ďalej sú uvedené základné klasifikačné metódy, hlbšie sú potom spomenuté neuronové siete a konkrétne Transformer a BERT. Záverečné podkapitoly sú venované metrikám vyhodnocovania a metrikám podobnosti.

### 2.1 Príbeh

Základom pre pochopenie kľúčových dejových línií a ich analýzu je pochopenie, čo je to príbeh, z čoho sa skladá a ako to súvisí s dejom. Nasledujúce podkapitoly vychádzajú z [17] a [16].

#### Dej

Dej je základ pre príbeh, založený na protichodných ľudských motiváciách, s činmi vyplývajúcimi z reálnej aj vymyslenej ľudskej reakcie. To znamená, že konflikt je základná časť, ktorú je potrebné vytvoriť, aby sa vytvoril súbor udalostí pri formovaní príbehu. Konflikt určí ďalšiu akciu alebo situáciu. Bude to určujúci faktor pre vytvorenie hlavnej štruktúry príbehu.

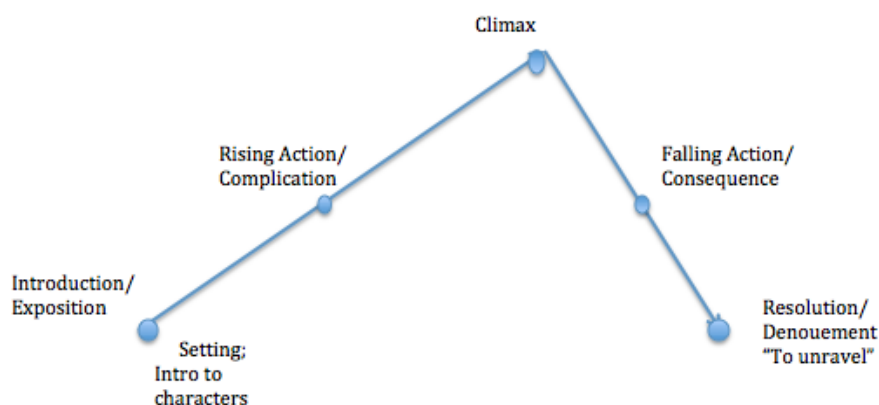
Dejom príbehu je teda nastolenie konfliktu a dôsledky, variácie a vývoj, ktoré z neho pramenia. Ďalej ním myslíme myšlienku, ktorá určuje, ako bude príbeh plynúť. Teda bude spájať jednu akciu s druhou, aby bol príbeh dobre organizovaný. V dobre vykreslenom príbehu nie je nič irelevantné; všetko spolu súvisí. V príbehu je čas dôležitý nielen preto, že jedna vec sa deje za druhou, ale preto, že jedna vec sa deje kvôli druhej.

Tak ako dej plynie, možno väčšinu dejov zaradiť do tejto tradičnej dejovej štruktúry, viď obrázok 2.1:

- **Expozícia (Úvod):** Expozícia je rozloženie a uvedenie základných prvkov v príbehu: hlavné postavy, ich pozadie, ich charakteristiky, ciele, obmedzenia a potenciál. Predstavuje všetko, čo bude v príbehu dôležité.
- **Kolízia (Zápletka):** Kolízia označuje začiatok veľkého konfliktu v príbehu. Účastníkmi sú protagonisti a antagonisti, spolu s akýmikoľvek myšlienkami alebo hodno-

tami, ktoré predstavujú, akými sú dobro a zlo, individualizmus a kolektivismus, láska a nenávisť, inteligencia a hlúposť, vedomosti a nevedomosť a podobne.

- **Kríza (Vyvrcholenie, Climax):** Kríza je bod zlomu, oddeľujúci medzi tým, čo bolo predtým, a tým, čo príde potom. Naplno sa v ňom prejaví konflikt a následné napätie. Ďalším spôsobom, ako myslieť na vyvrcholenie, je definovať ho ako bod v príbehu, v ktorom sa všetka ostatná akcia stáva nevyhnutnou.
- **Peripetia (Dejový obrat):** Peripetiou rozumieme nečakaný dejový zvrät, nepredpokladanú zmenu situácie. Dej sa zrýchľuje a smeruje ku koncu.
- **Katastrofa (Rozuzlenie):** Rozuzlenie je súbor akcií, ktorými sa príbeh končí. Hlavné akcie sú dokončené a posledná akcia podčiarkuje tón konečnosti.



Obr. 2.1: Štruktúra deja. Prebrané z [10].

## Postavy

Postavy sú osoby prezentované v diele, ktoré čitateľ interpretuje ako osoby obdarené morálnymi a dispozičnými vlastnosťami, ktoré sú vyjadrené tým, čo hovoria a čo robia. Na základe dôležitosti možno postavu rozdeliť do dvoch kategórií: hlavná postava a vedľajšia postava. V celom príbehu sa zvyčajne objavuje hlavná postava, stáva sa stredobodom príbehu. Udalosti, ktoré sa v príbehu dejú, sa ho vždy priamo či nepriamo týkajú. Na druhej strane, u vedľajších postáv sú úlohy menej dôležité ako u hlavnej postavy, pretože nie sú plne rozvinutými postavami a ich úlohy v príbehu sú len na podporu vývoja hlavnej postavy.

## Prostredie

Prostredie diela sa vzťahuje na prírodnú a umelú scenériu, v ktorom postavy žijú a pohybujú sa. Znamená to, že všetko, čo súvisí s prostredím, ako napr. denné svetlo, stromy, zvieratá, spoločnosť, popisované zvuky, pachy a počasie sú súčasťou prostredia. Dejiskom

diela je opis predmetov a fyzického vzhladu miesta, kde sa príbeh odohráva. Popis prostredia pomáha pri vytváraní dôveryhodnosti; môže pomôcť vysvetliť postavy aj situáciu; môže prispieť k atmosfére alebo prevládajúcej nálade; môže byť aktívny v predpovedaní; môže byť symbolickým. Okrem symboliky hlavných postáv sa prostredie používa aj ako prostriedok na posilnenie témy, znamená to, že prostredie sa považuje za dôležitú úlohu v príbehu a analýze.

## 2.2 Kľúčové dejové línie

Pojem *klúčové dejové línie* nemožno exaktne definovať. V starovekom Grécku rozlišovali dva typy deja: komédiu a tragédiu, no odvtedy sa delenie dejových línií posunulo ďalej. V minulosti sa o nájdenie rozdelenia deja pokúšalo veľa významných dramatikov, či spisovateľov, medzi nimi aj Shiller<sup>1</sup> či Gozzi<sup>2</sup>. Na posledného menovaného nadväzuje kniha „The Thirty Six Dramatic Situations“ [15], ktorá je jednou z mála kníh, ktorá sa zaoberá klasifikovaním deja. Ako z názvu vyplýva, predkladá 36 dramatických situácií, ktoré sú aj na dnešnú dobu aktuálne, no kniha udáva aj konkrétne deje pod každou dramatickou situáciou, tie sú však z väčšiny zastaralé a v dnešnej dobe nepoužiteľné.

Ludia odpradáva hľadajú idey pre vymýšľanie nových kníh, či filmov, preto možno hľadať kľúčové dejové línie najmä v knihách, ktoré sa zaoberajú inšpirovaním o témach a dejových líniách, o ktorých by mohli ich čitatelia písať.

Jednou z takýchto kníh je „Plotto“ [4], ktorá mala autorom pomôcť *poskladať si* príbeh. Toto dielo prezentuje tisíce veľmi podrobných dejových línií, tzv. *Masterplot*. Každá táto dejová línia pozostáva z 3 *klauzúl*. Kde klauzula A reprezentuje druh protagonistu, klauzula B uvádza základnú myšlienku deja a klauzula C ukončuje dej. Keďže B klauzuly sú veľmi všeobecné, každá B klauzula pozostáva z desiatok podrobnejších dejov. Tieto podrobné deje, môžu na seba rôzne navzájom nadväzovať a vytvoriť tak celkom konkrétny podrobný dej. Celkovo počet dejov, ktorý sa týmto mechanizmom dokáže vytvoriť možno rátať v státisícoch.

Ďalšou knihou z tejto kategórie je „20 Master Plots: And How to Build Them“ [22] v ktorej autor prednáša, ako opäť z názvu vyplýva, 20 oblastí ktorých sa môže týkať dej. Sú nimi: Hľadanie, Dobrodružstvo, Prenasledovanie, Záchrana, Únik, Pomsta, Hádanka, Rivalita, Smoliar, Pokušenie, Metamorfóza, Transformácia, Dozrievanie, Láska, Zakázaná láska, Obetovanie, Objavenie, Úbohý prebytok, Vzostup a Zostup.

Je nutné podotknúť, že kľúčové dejové línie sa týkajú iba deja, netýkajú sa ani prostredia ani hlavných postáv a ani žánrov. Preto napr. *Sherlock Holmes type* alebo *Happy ending* nemožno považovať za kľúčové dejové línie, lebo sa skôr týkajú niečoho iného ako deja. Aj na toto je treba dbať pri výbere kľúčových dejových línií.

## 2.3 Extrakcia dát

Dáta sú nevyhnutnou súčasťou každého výskumu, či už to je akademický, marketingový alebo vedecký. Ľudia chcú zbierať a analyzovať údaje z viacerých webových stránok. Rôzne webové stránky, ktoré patria do konkrétnej kategórie zobrazuje informácie v rôznych formátoch. Údaje môžu byť rozložené na viacerých stránkach a v rôznych sekciách. Väčšina webových stránok neumožňuje uložiť kópiu údajov, zobrazených na ich webových strán-

<sup>1</sup>Friedrich Schiller – nemecký dramatik (1759–1805)

<sup>2</sup>Carlo Gozzi – taliansky dramatik (1720–1806)

kach do lokálneho úložiska. *Web scraping* je technikou extrakcie neštruktúrovaných údajov z webových stránok a taktiež transformáciou týchto údajov do štruktúrovaných údajov, ktoré možno uložiť do zrozumiteľnej štruktúry ako napr. tabuľka, databáza alebo súbor vo formáte CSV.

Proces získavania údajov z internetu možno rozdeliť do dvoch po sebe nasledujúcich krokov: získavanie webových zdrojov a následné extrahovanie požadovaných informácií zo získaných údajov. Webové údaje sa bežne extrahujú pomocou protokolu HTTP (Hypertext Transfer Protocol) alebo cez web prehliadač. To sa buď vykonáva manuálne pomocou užívateľa alebo automaticky robotom alebo webovým prehliadačom. Konkrétne, program začína vytvorením HTTP požiadaviek na získanie zdrojov z cieľovej webovej stránky. Po úspešnom prijatí a spracovaní žiadosti cieľovou webovou stránkou sa požadovaný zdroj získa z webovej lokality a potom sa odošle späť do programu. Zdroj môže byť vo viacerých formátoch, akými sú webové stránky vytvorené z HTML, dátové kanály vo formáte XML alebo JSON alebo multimediálne dáta, akými sú obrázky, audio alebo video súbory. Po stiahnutí webových údajov proces extrakcie pokračuje v analýze, preformátovaní a usporiadaní údajov štruktúrovaným spôsobom.

Vzhľadom na to, že sa na internete neustále generuje obrovské množstvo heterogénnych údajov, je web scraping široko uznávaný ako efektívna a výkonná technika na zber veľkých dát.

Aj keď je web scraping účinnou technikou pri zhromažďovaní veľkých súborov údajov, je taktiež veľmi kontroverznou a môže vyvolať právne otázky súvisiace s autorskými právami a zmluvnými podmienkami. V rámci osobného použitia sa jedná o legálne využitie, problém však nastáva v prípade komerčného využitia. Takisto, ak takýto program posiela žiadosti o získavanie údajov príliš často, je to funkčne ekvivalentné útoku odmietnutia služby (DoS – Denial-of-Service), pri ktorom môže byť vlastníkovi odmietnutý vstup a môže byť zodpovedný za vzniknuté škody, pretože vlastník webovej aplikácie má majetkovú účasť na fyzickom webovom serveri, ktorý je hosťiteľom aplikácie.

Táto podkapitola vychádza z publikácií [26] a [18].

## 2.4 Spracovanie prirodzeného jazyka

Spracovanie prirodzeného jazyka (NLP – angl. Natural Language Processing ) je počítačový prístup k analýze textu a je založený na súbore teórií aj na súbore technológií. Je to veľmi aktívna oblasť výskumu a vývoja v dnešnej dobe. NLP možno definovať asi nasledovne: Spracovanie prirodzeného jazyka je teoreticky motivované rozpätie výpočtových techník na analýzu a reprezentáciu prirodzene sa vyskytujúcich textov na jednej alebo viacerých úrovniach lingvistickej analýzy s cieľom dosiahnuť podobnosť s človekom pri spracovaní jazyka pre celý rad úloh alebo aplikácií.

Viacere prvky tejto definície možno podrobnejšie rozviesť. Existuje viacero metód resp. techník, z ktorých si vybrať na vykonanie konkrétneho typu jazykovej analýzy. „Prirodzene sa vyskytujúce texty“ môžu byť akéhokoľvek jazyka, žánru a pod., jedinou požiadavkou je, aby boli v jazyku, ktorý používajú ľudia na komunikáciu medzi sebou. Analyzovaný text by tiež nemal byť špecifický vytvorený na účely analýzy, ale získaný zo skutočných použití. Pojem „úrovne lingvistickej analýzy“ odkazuje na skutočnosť, že existuje viacero typov jazykového spracovania, o ktorých je známe, že ľudia používajú pre pochopenie jazyka. NLP sa považuje za vnútornú disciplínu Umelej inteligencie (AI – angl. Artificial Intelligence ), keďže NLP sa snaží o výsledky podobné človeku, je vhodné považovať NLP za disciplínu AI.

Cieľom NLP, ako je uvedené vyššie, je dosiahnuť spracovanie ľudského jazyka. Výber slova „spracovanie“ je veľmi zámerný a nemal by sa nahradiť výrazom „pochopeie“. Oblasť NLP bola pôvodne označovaná ako Porozumenie prirodzeného jazyka (NLU – angl. Natural Language Understanding ) v začiatkoch AI, a keďže dodnes nebol jazyk skutočne „pochopeie“, tak sa NLU považuje za cieľ NLP.

Úplný NLU systém by bol schopný:

1. **Parafrázovať vstupný text**
2. **Preložiť text do iného jazyka**
3. **Odpovedať na otázky z textu**
4. **Vyvodzovať závery z textu**

Zatiaľ čo NLP urobilo vážne zásahy do dosiahnutia cieľov 1 až 3, skutočnosť, že NLP systémy nedokážu samy o sebe vyvodzovať závery z textu spôsobuje, že cieľom NLP stále zostáva NLU [12].

## 2.5 Klasifikácia

Námety pre túto kapitolu vychádzali z [8]. Klasifikácia (classification) je jednou z najpopulárnejších tém Spracovania prirodzeného jazyka a Hĺbkovej analýzy dát (Data Mining). Je to zvyčajne prediktívna úloha vedená pomocou techník učenia pod dohľadom (supervised learning). Klasifikácia má za cieľ naučiť sa z *označovaných* vzorov model schopný predpovedať značky (labels) pre budúce, nikdy predtým nevidené, ukážky údajov.

Sada atribútov v súbore údajov klasifikácie (classification dataset) je rozdelená do dvoch podmnožín. Tá prvá obsahuje vstupné vlastnosti (features), premenné, ktoré budú fungovať ako prediktory. Druhá podmnožina obsahuje výstupné atribúty, takzvané *značky* (labels), ktoré sú priradené pre každú inštanciu. Klasifikačné algoritmy indukujú model analyzujúci koreláciu medzi vstupnými vlastnosťami a výstupnými značkami. Keď sa získa natrénovaný model, môže byť použitý na spracovanie množiny vstupných vlastností nových vzoriek údajov a tým získať predikciu značiek. V závislosti od povahy druhej podmnožiny atribútov, ktorá obsahuje značky, možno identifikovať niekoľko druhov klasifikačných problémov, v závislosti od počtu výstupov a ich typov.

Medzi základné patrí:

- **Binárna klasifikácia** (Binary Classification) Toto je najjednoduchší klasifikačný problém, ktorému možno čeliť. Inštanície v binárnej dátovej sade majú iba jeden výstupný atribút a môže mať iba dve rôzne hodnoty. Tieto sú zvyčajne známe ako pozitívne a negatívne, ale možno ich interpretovať aj ako pravda a nepravda, 1 a 0 alebo akákoľvek iná kombinácia dvoch hodnôt. Klasický príklad tejto úlohy je filtrovanie spamu, pri ktorom sa klasifikátor učí zo správ obsah, ktorý možno považovať za spam.
- **Viac-triedna klasifikácia** (Multi-class Classification) Viac-triedna dátová sada má tiež iba jeden výstupný atribút, ako binárna klasifikácia, ale môže obsahovať ktorúkoľvek z určitého súboru preddefinovaných hodnôt. Vo viac-triednej klasifikácii sa *značky* nazývajú *triedy* (class). Význam každej z týchto tried a samotná hodnota sú

špecifické pre každú aplikáciu, súbor tried je však konečný a diskrétny. Jedným z najznámejších príkladov klasifikácie viacerých tried je identifikácia druhov dúhovky, kde sa klasifikátor učí, ako klasifikovať nové inštancie do zodpovedajúcej rodiny (triedy). Mnoho viac-triednych klasifikačných algoritmov závisí na binarizácii, metóda, ktorá iteratívne trénuje binárny klasifikátor pre každú triedu zvlášť. Viac-triednu klasifikáciu možno považovať za zovšeobecnenie binárnej klasifikácie. Výstup je len jeden, ale môže nadobudnúť akúkoľvek hodnotu, zatiaľ čo v binárnom prípade je obmedzený na podmnožinu dvoch hodnôt.

- **Viac-značková klasifikácia** (Multi-label Classification) Na rozdiel od dvoch predchádzajúcich klasifikačných modelov, vo viac-značkovej klasifikácii má každá z inšancií údajov priradený vektor výstupov, nie iba jednu hodnotu. Dĺžka tohto vektora je pevná a má dĺžku podľa počtu značiek v dátovej sade. Každý prvok vektora je binárna hodnota označujúca, či príslušná značka je alebo nie je relevantná pre vzorku. Aktívnych môže byť niekoľko značiek naraz. Viac-značková klasifikácia sa v súčasnosti používa v mnohých oblastiach, z ktorých väčšina súvisí na automatické označovanie zdrojov zo sociálnych médií, akými sú obrázky, hudba, video, správy, či blogové príspevky. Algoritmy použité na túto úlohu musia byť schopné vytvoriť niekoľko predpovedí naraz, či už ide o transformáciu pôvodných dátových sád alebo ich prispôbenie na existujúce binárne/viac-triedne klasifikačné algoritmy.

## 2.6 Klasifikačné metódy

Kapitoly týkajúce sa klasifikačných metód sú prevzaté z publikácie [11]. V ére informačnej explózie môže byť zdĺhavé a náročné spracovávať a klasifikovať veľké množstvo textových údajov manuálne. Okrem toho môže byť kvalita manuálnej klasifikácie textu ovplyvnená ľudskými faktormi, akými sú únava a neodbornosť. Je preto žiaduce použiť metódy strojového učenia na automatizáciu procesu klasifikácie textu, aby boli výsledky spoľahlivejšie a menej subjektívne. Okrem toho to môže tiež pomôcť zvýšiť efektívnosť vyhľadávania informácií a zmierniť problém informačného preťaženia.

Textové údaje sa líšia od číselných, obrazových alebo signálových dát, vyžadujú NLP techniky pre ich spracovanie. Prvým dôležitým krokom je predspracovanie textových údajov pre model. Tradičné modely zvyčajne potrebujú získať dobré vlastnosti (features) umelými metódami a potom ich klasifikovať pomocou klasických algoritmov strojového učenia. Preto účinnosť týchto metód je do značnej miery obmedzená extrakciou vlastností (feature extraction). Preto väčšina výskumných prác zameraných na klasifikáciu textu sú zamerané na hlboké neuronové siete, čo sú prístupy založené na údajoch (data-driven approaches) s vysokou výpočtovou náročnosťou.

Klasifikáciou textu sa myslí extrahovanie prvkov z nespracovaných textových údajov a predpovedanie kategórie textových údajov na základe takýchto vlastností (features). Množstvo modelov bolo navrhnutých za posledných niekoľko desaťročí na klasifikáciu textu. Z tradičných modelov, Naïve Bayes (NB) bol prvým použitým modelom na túto úlohu. Potom sa používajú generické klasifikačné modely, ako napríklad K-Nearest Neighbor (KNN), Support Vector Machine (SVM) a Random Forest (RF). V poslednej dobe eXtreme Gradient Boosting (XGBoost) a Light Gradient Boosting Machine (LightGBM) majú potenciál poskytnúť vynikajúci výkon. Modely hlbokého učenia takisto išli veľmi do popredia, odkedy bola použitá konvulčná neuronová sieť po prvýkrát na klasifikáciu textu. Aj keď nie je špeciálne navrhnutý na prácu s textom, je Bidirectional Encoder Representation from

Transformers (BERT) široko používaný pri návrhu modelov pre klasifikáciu textu, vzhľadom na jeho efektívnosť vo viacerých dátových sadách zameraných na textovú klasifikáciu.

## Naïve Bayes

Naïve Bayes (NB) je najjednoduchší a najrozšírenejší model založený na použití Bayesovho teorému. Algoritmus NB primárne využíva podmienenú pravdepodobnosť. Výhodou NB je, že na odhad parametrov potrebných na klasifikáciu vyžaduje len malý počet tréningových dát. Parametre NB sú menej citlivé na chýbajúce údaje a algoritmus je jednoduchý. Predpokladá však, že vlastnosti sú na sebe nezávislé. Keď počet vlastností je veľký, alebo korelácia medzi vlastnosťami je významná, výkon NB klesá. Napriek „naïvnému“ dizajnu a zjednodušeným predpokladom je NB široko používaný aj v zložitých problémoch.

## K-najbližších susedov

Jadrom algoritmu k-najbližších susedov (angl. KNN – K-Nearest Neighbors) je klasifikácia neoznačenej vzorky nájdením kategórie s najväčším počtom vzoriek na  $k$  najbližších vzorkách. Je to jednoduchý klasifikátor bez nutnosti vytvárania modelu a môže znížiť zložitosť pomocou jednoduchého procesu získavania KNN. Avšak v dôsledku pozitívnej korelácie medzi časovou/priestorovou zložitou modelu a množstvom údajov, je algoritmus KNN na rozsiahlych dátových sadách nezvyčajne pomalý. KNN závisí hlavne od okolitých susedných vzoriek a pre dátové sady s väčším prekrytím tried je vhodnejší ako iné metódy.

## Support Vector Machine

Prístupy založené na SVM menia klasifikáciu textu na viaceré úlohy binárnej klasifikácie. SVM sa snaží vytvárať optimálnu nadrovinu vo vektorovom priestore pre maximalizáciu vzdialenosti medzi triedami a určiť vzdialenosť hranice kategórie v smere kolmom na nadrovinu čo najväčšiu, čo bude mať za následok nižšiu chybovosť klasifikácie. Využíva predchádzajúce znalosti na vytvorenie vhodnejšej štruktúry a rýchlejšie štúdium. SVM dokáže vyriešiť vysokorozmerné a nelineárne problémy. Má vysokú generalizačnú schopnosť, ale je citlivý na chýbajúce údaje.

## Rozhodovacie stromy

Rozhodovacie stromy (DT – Decision Trees) sú metódy učenia s učiteľom stromovej štruktúry a sú konštruované rekurzívne, odrážajú myšlienku *rozdeľuj a panuj*. Rozhodovacie stromy môžu byť všeobecne rozdelené do dvoch odlišných etáp: stavba stromov a prerezávanie stromov. Začína sa na koreňovom uzly kde testuje vzorky údajov, a rozdeľuje dátovú sadu do rozdielnych podmnožín podľa rôznych výsledkov. Podmnožiny dátovej sady tvoria synovské uzly a každý listový uzol v rozhodovacom strome predstavuje kategóriu. Konštrukcia rozhodovacieho stromu má určiť koreláciu medzi triedami a atribútmi, ktoré sa ďalej využívajú na predpovedanie kategórie neznámych budúcich typov. DT je ľahké pochopiť a interpretovať. Z pozorovaného modelu je ľahké odvodiť zodpovedajúci logický výraz z vygenerovaného rozhodovacieho stromu.

## Neurónové siete

Hlboké neurónové siete pozostávajú z umelých neurónových sietí, ktoré simulujú ľudský mozog, aby sa automaticky učil vysokoúrovňové vlastnosti z údajov, ktoré dosahujú lepšie

výsledky ako tradičné modely spomenuté vyššie. Konvolučné neurónové siete (CNN – Convolutional Neural Networks ) môže súčasne aplikovať konvolúcie definované rôznymi jadrami na viaceré časti sekvencie. Preto sa CNN používajú na mnohé úlohy NLP vrátane klasifikácie textu. Pre klasifikáciu textu sa vyžaduje, aby bol text reprezentovaný ako vektor. Najskôr sú všetky vektory slov vstupného textu spojené do matice. Matica sa potom privádza do konvolučnej vrstvy, ktorá obsahuje niekoľko filtrov s rôznymi rozmermi. Nakoniec výsledok konvolučnej vrstvy prechádza cez rozhodovaciu vrstvu a zrežazí všetky výsledky, aby sa získala konečná vektorová reprezentácia, ktorá určuje výslednú triedu.

Hlboké učenie pozostáva z viacerých skrytých vrstiev v neurónovej sieti s vyššou úrovňou zložitosti a môže byť trénovaný na neštruktúrovaných dátach. Hlboké učenie sa dokáže naučiť jazykové vlastnosti a dokáže zvládať aj abstraktnejšie jazykové prvky založené na slovách a vektoroch na vysokej úrovni. Architektúra hlbokého učenia sa vie naučiť reprezentovať vlastnosti priamo zo vstupu bez príliš veľkého manuálneho zásahu a predchádzajúcich znalostí. Technológia hlbokého učenia je však metódou založenou na dátach, ktorá si vyžaduje obrovské množstvo údajov na dosiahnutie vysokého výkonu.

## 2.7 Transformer

Transformer [23] je prominentný model hlbokého učenia, ktorý je široko používaný v rôznych oblastiach, ako napr. NLP, počítačové videnie alebo spracovanie reči. Transformer bol pôvodne navrhnutý ako *Sequence-to-Sequence* model pre strojový preklad. Sequence-to-Sequence (alebo Seq2Seq) je neurónová sieť, ktorá transformuje danú sekvenciu prvkov, ako napríklad sekvenciu slov vo vete, na inú sekvenciu. Modely Seq2Seq pozostávajú z kodéra a dekodéra. Kóder vezme vstupnú sekvenciu a namapuje ju do priestoru vyššej dimenzie ( $n$ -rozmerný vektor). Tento abstraktný vektor sa privedie do dekodéra, ktorý ho zmení na výstupnú sekvenciu. Výstupná sekvencia môže byť v inom jazyku, symboloch, kópii vstupu atď.

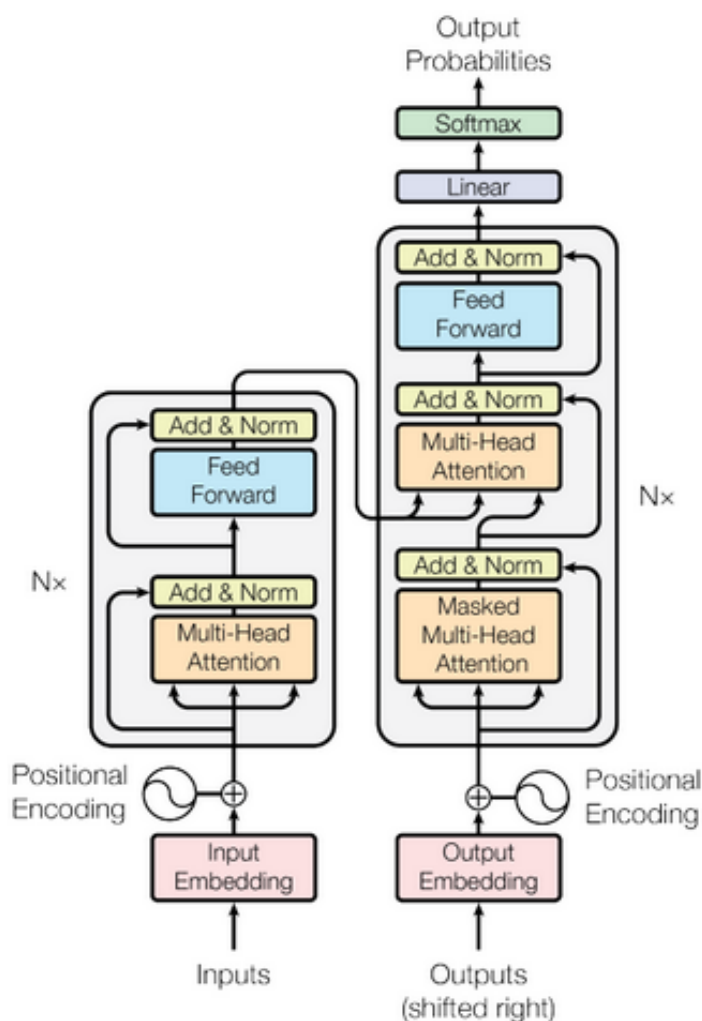
Čistý Transformer takisto teda pozostáva z kodéra a dekodéra, z ktorých každý z nich je zložený z  $n$  rovnakých blokov, ktoré je možné na seba viackrát naskladať, čo je na obrázku 2.2 popísané pomocou  $Nx$ . Kóder je vľavo a dekodér vpravo. Vidíme, že moduly pozostávajú hlavne z vrstiev *Multi-Head Attention* a *Feed Forward*.

Pri Transformeroch sa stretávame s pojmom: *Pozornosť*. Model si musí pamätať, ako sa sekvencie vkladajú do modelu, je nutné teda nejako dať každému slovu v sekvencii relatívnu polohu, pretože sekvencia závisí od poradia jej prvkov. Tieto pozície sa pridávajú do  $n$ -rozmerného vektora každého slova. V transformeroch sa toto nazýva *mechanizmus pozornosti* (angl. Attention-mechanism). Mechanizmus pozornosti sa pozerá na vstupnú sekvenciu a v každom kroku rozhoduje, ktoré ďalšie časti sekvencie sú dôležité.

Inými slovami, pre každý vstup, ktorý kóder načíta, mechanizmus pozornosti berie do úvahy niekoľko ďalších vstupov súčasne a rozhoduje, ktoré z nich sú dôležité, priradením rôznych váh týmto vstupom. Dekodér potom vezme ako vstup zakódovanú vetu a váhy poskytnuté mechanizmom pozornosti.

Funkciu pozornosti možno opísať ako mapovanie dotazu (Q) a dvojice kľúč (K) – hodnota (V) na výstup, kde dotazy, kľúče, hodnoty a výstup sú všetko vektory. Výstup sa vypočíta ako vážený súčet hodnôt, kde váha priradená každej hodnote je vypočítaná funkciou kompatibility dotazu s príslušným kľúčom. Vstup pozostáva z dotazov a kľúčov dimenzie  $d_k$ . Mechanizmus pozornosti možno popísať rovnicou (2.1):





Obr. 2.2: Transformer – architektúra modelu [23].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

$Q$  je matica, ktorá obsahuje dotaz (vektorová reprezentácia jedného slova v sekvencii),  $K$  sú všetky klúče (vektorové reprezentácie všetkých slov v sekvencii) a  $V$  sú hodnoty, ktoré sú opäť vektorové reprezentácie všetkých slov v sekvencii.

Váhy sú definované tým, ako je každé slovo v sekvencii ( $Q$ ) ovplyvnené všetkými ostatnými slovami v sekvencii ( $K$ ). Okrem toho je na váhy aplikovaná funkcia *SoftMax*, aby mali rozdelenie medzi 0 a 1. Tieto váhy sa potom aplikujú na všetky slová v sekvencii ( $V$ ).

Architektúra transformerov používa komplexnejšiu *Multi-Head attention* vrstvu, ktorá paralelizuje tento mechanizmus a umožňuje to systému učiť sa z rôznych reprezentácií  $Q$ ,  $K$  a  $V$ .

Neskoršie práce ukazujú, že pred-trénované modely založené na transformeroch môžu dosiahnuť *state-of-the-art performance* pri rôznych úlohách. V dôsledku toho sa Transformer stal *go-to* architektúrou v NLP. [21] [1]

## 2.8 BERT

BERT [6] (Bidirectional Encoder Representations from Transformers) je zástupca z rodiny transformerov. Klúčovou technickou inováciou BERT je aplikácia obojsmerného učenia Transformera. To je v kontraste s predchádzajúcimi snahami, ktoré sa zaoberali textovou sekvenciou buď zľava doprava, alebo kombinovaným učením zľava doprava a sprava doľava. Výsledky ukazujú, že jazykový model, ktorý je trénovaný obojsmerne, môže mať hlbší zmysel pre jazykový kontext a tok ako jednosmerné jazykové modely. V článku vedci podrobne opisujú nové techniky, ktoré umožňujú obojsmerný tréning v modeloch, v ktorých to bolo predtým nemožné.

BERT využíva mechanizmus pozornosti, ktorý sa učí kontextové vzťahy medzi slovami v texte. Vo svojej základnej podobe obsahuje Transformer dva samostatné mechanizmy: kóder, ktorý číta textový vstup, a dekodér, ktorý vytvára predpoveď pre danú úlohu. Keďže cieľom BERT je vygenerovať jazykový model, je potrebný iba kóder. Na rozdiel od smerových modelov (directional models), ktoré čítajú textový vstup postupne (zľava doprava alebo sprava doľava), kóder Transformera načíta celú sekvenciu slov naraz. Preto sa považuje za obojsmerný, aj keď presnejšie by bolo povedať, že je nesmerný (non-directional). Táto charakteristika umožňuje modelu naučiť sa kontext slova na základe celého jeho okolia (vľavo a vpravo od slova).

Pri trénovaní jazykových modelov existuje výzva, ako definovať cieľ predikcie. Mnoho modelov predpovedá ďalšie slovo v sekvencii, čo je smerový prístup (directional approach), ktorý vo svojej podstate obmedzuje kontextové učenie. Na prekonanie tejto výzvy používa BERT dve tréningové stratégie:

- **Maskovaný jazykový model** (MLM – Masked Language Model ) Nanešťastie, štandardné jazykové modely môžu byť len trénované zľava doprava alebo sprava doľava, pretože obojsmerné by každému slovu umožnili nepriamo „vidieť samého seba“ a model by mohol triviálne predpovedať cieľové slovo vo viacvrstvovom kontexte. Preto sa jednoducho maskuje určité percento vstupných tokenov a potom sa tieto maskované tokeny predpovedajú. Tento postup sa označuje ako MLM, a BERT je trénovaný tak, že je náhodne maskovaných asi 15 % všetkých tokenov. Hoci táto stratégia umožňuje získať obojsmerný pred-trénovaný model, nevýhodou je nesúlad medzi pred-tréningom (pre-training) a doladovaním (fine-tuning), pretože maskované tokeny sa počas doladovania nevyskytujú.
- **Predpoveď ďalšej vety** (NSP – Next Sentence Prediction ) Mnohé dôležité nadväzujúce úlohy (downstream tasks) sú založené na pochopení vzťahu medzi dvoma vetami, ktorý nie je priamo zachytený jazykovým modelovaním. Aby bolo možné trénovanie modelu, ktorý rozumie väzbám viet, používa BERT binarizovanú NSP úlohu, ktorú možno triviálne vygenerovať z akéhokoľvek jednojazyčného korpusu. V tréningovom procese model BERT dostáva páry viet ako vstup a učí sa predpovedať, či druhá veta v páre je vetou nasledujúcou v pôvodnom dokumente. Počas tréningu tvorí 50 % vstupov dvojica, kde je druhá veta vetou nasledujúcou v pôvodnom dokumente, zatiaľ čo v ostatných 50 % je ako druhá veta zvolená náhodná veta z korpusu. Predpokladom je, že náhodná veta bude odpojená od prvej vety.

Pri trénovaní modelu BERT sa MLM a NSP trénujú spoločne s cieľom minimalizovať kombinovanú stratovú funkciu (loss function) týchto dvoch stratégií. [9] [19]

### 2.8.1 Fine-tuning

Použitie BERT na konkrétnu úlohu je pomerne jednoduché. BERT možno použiť na širokú škálu jazykových úloh, pričom do základného modelu stačí pridať iba malú vrstvu. Doladovanie (fine-tuning) je proces adaptovania BERT na konkrétnu nadväzujúcu úlohu (downstream task), či už je vstup jedno-textový alebo sa jedná o textové páry. Pre každú úlohu stačí zapojiť vstupy a výstupy špecifické pre danú úlohu do BERT a doladiť všetky parametre end-to-end. Na vstupe namiesto vety A a vety B z pred-tréningu je pri klasifikácii pár text-značky. V porovnaní s pred-tréningom je doladovanie relatívne lacné, čo sa týka času. Pre vhodné doladenie BERT na konkrétnu úlohu je nutné okrem iného aj správne nastavenie hyper-parametrov alebo pridanie extra vrstvy. Pre viac-značkovú klasifikáciu (multi-label) je nutné pridať vrstvu s aktivačnou funkciou sigmoid. Často sa používa softmax aktivačná funkcia vo viac-triednych úlohách (multi-class), čo je predvolená aktivačná funkcia tradičného BERT. Funkcia softmax však nie je vhodná pre úlohy s viacerými značkami, pretože súčet pravdepodobností výstupu funkcie softmax je 1. To sa nehodí v situáciách, keď môže byť prítomných viacero značiek súčasne. Výstupy sigmoidnej funkcie sa pohybujú od 0 do 1, ktoré sú na sebe nezávislé a predstavujú pravdepodobnosť pre každú značku zvlášť. Súčet výstupov sigmoidnej funkcie nie je 1, takže viac ako jedna značka má možnosť získať skóre nad určitú hranicu prahu (threshold). Taktoto možno každú značku klasifikovať na 0 a 1 podľa hodnoty prahu. [6] [19] [20]

### 2.8.2 Spracovanie dlhých reťazcov

Ďalším faktorom pri doladovaní BERT na klasifikáciu textu je dĺžka vstupnej sekvencie a spracovanie dlhých reťazcov, keďže maximálna dĺžka vstupu BERT je 512 tokenov. Prípadajú do úvahy nasledujúce spôsoby zaoberajúce sa dlhými reťazcami.

**Metódy skrátenia** Zvyčajne sú kľúčové informácie článku na začiatku a na konci. Preto možno použiť tri rôzne metódy skrátenia textu.

1. Iba úvodných 510 tokenov<sup>3</sup>
2. Iba konečných 510 tokenov
3. 128 tokenov z úvodu a 382 tokenov z konca textu

**Hierarchické metódy** Vstupný text je najskôr rozdelený na  $k = L/510$  frakcií, ktoré sa privádzajú do BERT s cieľom získať zastúpenie  $k$  textových zlomkov. Potom sú použité metódy *mean pooling*, *max pooling* a seba-pozornosť na kombináciu reprezentácie všetkých zlomkov. [19]

**Longformer** [3] Ďalšou možnosťou ako spracovávať dlhšie reťazce je použiť upravenú verziu BERT, s názvom Longformer, ktorý dokáže spracovávať vstup o veľkosti 4096 tokenov v relatívne krátkom čase.

## 2.9 Metriky vyhodnocovania

Táto podkapitola je prebraná z [8]. Výstup akéhokoľvek klasifikátora s viacerými značkami pozostáva z predpovedaného vektora značiek pre každú testovaciu inštanciu. Pri práci v tradičnom scenári s iba jednou triedou ako výstupom, predpoveď môže byť správna alebo

<sup>3</sup>Treba odpočítat tokeny [CLS] a [SEP].

nesprávna, naproti tomu predpovede viacerých značiek môžu byť úplne správne, čiastočne správne/nesprávne (v rôznych stupňoch) alebo úplne nesprávne. Použitie rovnakých metrík používaných v tradičnej klasifikácii je možné, ale zvyčajne prehnane prísne. To je dôvod na použitie špecifických hodnotiacich metrík, kde sa berú do úvahy aj prípady medzi týmito dvoma extrémami. V súčasnosti je v literatúre definovaných viac ako dvadsať rôznych výkonnostných metrík pre viac-značkovú klasifikáciu. Všetky metriky vyhodnocovania viacerých značiek možno zoskupiť podľa dvoch kritérií:

- **Podľa toho ako sa predpoveď vypočítava:** Meranie môže byť vykonané inštanciou alebo pomocou značiek, čím sa získajú dve rôzne skupiny metrík:
  - **Metriky založené na príkladoch:** Tieto metriky sa počítajú samostatne pre každú inštanciu a potom sú spriemerované počtom vzoriek.
  - **Metriky založené na značkách:** Na rozdiel od predchádzajúcej skupiny metrík založené na značkách sa vypočítajú nezávisle pre každú značku pred ich spriemerovaním. Pritom možno použiť dve rôzne stratégie:
    - \* **Makro-priemerovanie:** Metrika sa vypočítava individuálne pre každú značku a výsledok je spriemerovaný počtom značiek.
    - \* **Mikro-priemerovanie:** Počítadlá trafených a netrafených predpovedí pre každú značku sú najskôr agregované a potom sa metrika vypočíta iba raz.
- **Podľa toho ako sa výsledok sprostredkuje:** Výstup produkovaný viac-značkovým klasifikátorom môže byť binárna bipartícia značiek alebo poradie značiek. Niektoré z nich poskytujú oboje výsledky.
  - **Binárna bipartícia:** Binárne bipartícia je vektor označujúci 0 a 1, ktoré indikujú, že značka je relevantná pre spracovanú vzorku. Existujú metriky, ktoré fungujú nad týmito bipartíciami a počítajú počet skutočne pozitívnych (angl. True Positive – TP), skutočne negatívnych (angl. True Negative – TN), falošne pozitívnych (angl. False Positive – FP) a falošne negatívnych (angl. False Negative – FN) vzoriek.
  - **Poradie značiek:** Výstupom je zoznam značiek zoradených podľa nejakej relevantnosti. Binárnu bipartíciu možno z tohoto poradia získať použitím prahu, zvyčajne daný samotným klasifikátorom. Existujú však metriky, ktoré namiesto toho pracujú s nespracovaným poradím na výpočet vyhodnotenia.

## Hammingova strata

Hammingova strata (angl. Hamming loss) je jednou z metrík založených na príkladoch, je pravdepodobne najbežnejšie používanou metrikou viac-značkovej klasifikácie, asi aj pretože je ľahké ju vypočítať ako možno vidieť v rovnici (2.2). Operátor  $\Delta$  vracia symetrický rozdiel medzi  $Y_i$ , reálnym vektorom značiek  $i$ -tej inštancie a tým predpovedaným  $Z_i$ . Operátor  $|r|$  počíta počet 1 v tomto rozdiel, inými slovami počet nesprávnych predpovedí. Celkový počet chýb v  $n$  inšanciách sa agreguje a potom normalizuje počtom značiek  $k$  a počtom inšancií.

## Správnosť

Vo viac-značkovej oblasti je správnosť (angl. Accuracy) definovaná ako (2.3) pomer medzi hodnotami počtu správne predpovedaných značiek a celkového počtu značiek v oboch

vektoroch. Metrika je počítaná pre každú inštanciu a potom je spriemerovaná ako všetky metriky založené na príkladoch.

## Presnosť

Presnosť (angl. Precision)(2.4), naproti tomu, sa považuje za jednu z intuitívnejších metrík na hodnotenie viac-značkových klasifikátorov. Vypočíta sa ako podiel medzi počtom skutočne pozitívnych značiek a celkovým počtom pozitívnych značiek. Teda môže byť interpretovaná ako percento predpovedaných značiek, ktoré sú skutočne relevantné pre danú inštanciu. Táto metrika sa zvyčajne používa v spojení s funkciou úplnosť alebo senzitivita (angl. Recall) (2.5), ktorá vracia percento správne predpovedaných značiek spomedzi všetkých skutočne relevantných značiek, teda pomer správnych značiek je výstupom klasifikátora.

## F-skóre

Spoločné používanie presnosti a senzitivity je tak bežné, že je definovaná metrika, ktorá ich kombinuje. Je známa ako F-skóre (angl. F-measure/ F-score) (2.6) a vypočíta sa ako harmonický priemer týchto dvoch. Týmto spôsobom je vyvážená miera toho, koľko relevantných značiek je predpovedaných a koľko predpovedaných značiek je relevantných.

## Priemerná presnosť

Všetky metriky založené na príkladoch popísané vyššie fungujú cez binárnu bipartíciu značiek, takže potrebujú sadu značiek ako výstup z klasifikátora. Naproti tomu, nasledujúca metrika potrebujú poradie značiek, takže buď stupeň spoľahlivosti alebo pravdepodobnosť príslušnosti každej zo značiek je potrebná na jej výpočet.

Metrika priemerná presnosť (angl. Average precision) (2.7) určuje pre každú značku v inštanciách, podiel relevantných značiek, ktoré sú v predpokladanom poradí nad ňou. Cieľom tejto metriky je zistiť, koľko pozícií treba v priemere skontrolovať, predtým, než sa nájde nerelevantné označenie. Čím väčšia je hodnota tejto metriky, tým lepší je výkon klasifikátora. V rovnici (2.7) je  $rank(x_i, l)$  definovaná ako funkcia ktorá pre inštanciu  $x_i$  a príslušnú značku  $l$ , ktorej poloha je známa, vráti stupeň spoľahlivosti.

## Metriky založené na značkách

Všetky výkonnostné metriky vymenované v predchádzajúcich častiach sa vyhodnocujú jednotlivo pre každú inštanciu a potom sú spriemerované počtom inštancií. Preto má každá vzorka údajov v konečnom výsledku rovnakú váhu. Na druhej strane, metriky založené na značkách možno vypočítať pomocou dvoch rôznych stratégií spriemerovania. Tieto sú zvyčajne známe ako makro-priemerovanie a mikro-priemerovanie. Ktorúkoľvek z metrík popísaných v tejto kapitole je možné spriemerovať pomocou týchto stratégií.

V prístupe makro-priemerovania (2.8) sa metrika vyhodnocuje raz za značku pomocou akumulovaného počítadla a priemer sa potom získa vydelením počtom značiek. Takto je každej značke priradená rovnaká váha, či už je častá alebo zriedkavá. Naopak, stratégia mikro-priemerovania (2.9) najskôr pripočítava počítadla pre všetky značky a potom vypočíta metriku iba raz. Preto prínos každej značky vo finále nie je rovnaký.

$$HammingLoss = \frac{1}{n} \frac{1}{k} \sum_{i=1}^n |Y_i \triangle Z_i| \quad (2.2)$$

$$Accuracy = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

$$Precision = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i \cap Z_i|}{|Z_i|} = \frac{TP}{TP + FP} \quad (2.4)$$

$$Recall = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i \cap Z_i|}{|Y_i|} = \frac{TP}{TP + FN} \quad (2.5)$$

$$F-score = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (2.6)$$

$$AveragePrecision = \frac{1}{n} \sum_{i=1}^n \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' | rank(x_i, y') \leq rank(x_i, y), y' \in Y_i\}|}{rank(x_i, y)} \quad (2.7)$$

$$MacroMetric = \frac{1}{k} \sum_{l \in L} EvalMetric(TP_l, FP_l, TN_l, FN_l) \quad (2.8)$$

$$MicroMetric = EvalMetric\left(\sum_{l \in L} TP_l, \sum_{l \in L} FP_l, \sum_{l \in L} TN_l, \sum_{l \in L} FN_l\right) \quad (2.9)$$

$n$  = počet inštancií

$k$  = počet značiek

$Y_i$  = reálny vektor značiek i-tej inštancie

$Z_i$  = predpovedaný vektor značiek i-tej inštancie

$\triangle$  = symetrický rozdiel

$|r|$  = počet nesprávnych predpovedí

$TP$  = True Positive

$TN$  = True Negative

$FP$  = False Positive

$FN$  = False Negative

$L$  = všetky značky v dátovej sade

## 2.10 Metriky podobnosti

Pre porovnanie dvoch vektorov a určenie ich podobnosti, existujú viaceré metriky. Tieto vektory môžu reprezentovať text, obraz a iné dáta a je možné teda takto určiť podobnosť medzi týmito dvoma vektormi. Metriky podobnosti fungujú na základe merania uhlu medzi týmito dvoma vektormi, aj keď sa jedná o viac-dimenzionálne vektory.

Jedny zo základných metrick sú Pearsonov korelačný koeficient, Spearmanov korelačný koeficient a kosínusová podobnosť, ktoré sú základom pre analýzu údajov.

- **Kosínusová podobnosť** Kosínusová podobnosť je štandardná metrika používaná pri získavaní informácií. Je to kosínus uhla medzi dvoma euklidovskými vektormi, a preto nie je ovplyvnený skalárnymi transformáciami v údajoch. Je definovaný nižšie v rovnici (2.10) pre vektory  $x$  a  $y$ .
- **Pearsonov korelačný koeficient** Pearsonove a Spearmanove koeficienty merajú silu asociácie medzi dvoma premennými  $X$  a  $Y$ . Pearsonov koeficient, bežne označovaný ako  $\rho$ , je definovaný ako kovariancia dvoch premenných delená súčinom ich príslušných štandardných odchýlok. (2.11)
- **Spearmanov korelačný koeficient** Spearmanov koeficient sa získa aplikáciou Pearsonovho koeficientu na dáta transformované podľa poradia. Obe nie sú ovplyvnené lineárnymi transformáciami údajov. Z vektorov  $x$  a  $y$ , respektíve premenných  $X$  a  $Y$ , každý s dĺžkou  $n$ , sa Spearmanov koeficient  $r_{x,y}$  získa odhadom kovariancie populácie a štandardných odchýlok od vzoriek, ako je definované v rovnici (2.12).  $\bar{x}$  a  $\bar{y}$  tu označujú priemer vzoriek. [7]

$$\cos(\theta) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} \quad (2.10)$$

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (2.11)$$

$$r_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (2.12)$$

## Kapitola 3

# Dáta

Dáta sú nevyhnutnou súčasťou každej dátovej analýzy a je rovnako dôležitá ich kvantita ako aj ich kvalita. V tejto kapitole sú popísané všetky dátové sady použité v tejto práci. V nasledujúcej sekcii je popísané získavanie a výber kľúčových dejových línií, tento dataset slúži ako hlavný tréningový dataset. Ďalej je popísaná populárna dátová sada z internetu obsahujúca zhrnutia kníh. Nakoniec sú predstavené dátové sady zo sociálnej siete Goodreads, ktoré boli vytvorené v rámci tejto práce.

### 3.1 Dátová sada z IMDb

Veľmi dôležitou súčasťou klasifikácie dejov je zdefinovanie si, s akými kľúčovými dejovými líniami sa pracuje a vybrať teda konkrétne, ktoré sú to, určiť či ich je zopár alebo stovky. Ako je napísané v kapitole 2.2, pohľady na toto sú rôzne. Rozhodol som sa rozšíriť prístup, ktorý bol zvolený v [22], kde je spomenutých 20 kľúčových dejových línií. Chcel som, aby konkrétne kľúčové dejové línie vypovedali niekedy aj trochu viac o príbehu ako jednoduché jednoslovné pomenovania v spomínanej knihe.

Možnosť označovať si dáta ručne nepripadá do úvahy, lebo na učenie je potrebné kvantum dát, preto je nutné hľadať zdroje, ktoré ponúkajú takto označované dáta. Možnosti nie je až tak veľa, no jednou vhodnou sú kľúčové slová z webu IMDb<sup>1</sup>, kde ku najmä populárnym filmom a seriálom možno nájsť zhrnutia dejov a kľúčové slová. Jedinou nevýhodou je, že sa jedná o filmy a nie o knihy, no dej sa rovnako nachádza aj vo filme aj v knihe. Primárne sa práca zameriava na knihy no môže sa jednať o akékoľvek zhrnutie deja. Tieto kľúčové slová hovoria o všetkých možných aspektoch, ktoré sa v danom diele nachádzajú. Pre konkrétne dielo môžu byť aj stovky kľúčových slov. Ak vo filme prší je kľúčovým slovom dážď, ak je vo filme dôležitý nejaký predmet alebo udalosť, tak je kľúčovým slovom práve to. Týkajú sa takisto aj témy, prostredia, žánru ale aj príbehu.

Tak som si teda vybral viac ako 100 kľúčových slov, týkajúcich sa deja spomedzi všetkých týchto kľúčových slov, a tie budú reprezentovať kľúčové dejové línie. Nachádzajú sa medzi nimi aj jednoduché jednoslovné ako napr. *love*, *revenge*, *escape*, *racism* a pod., ale aj špecifickejšie ako *time travel*, *abuse of power*, *one against many* alebo *love triangle*. Takto spojené aj jednoduché aj zložitejšie môžu už dať dobrý obraz o tom, o čom príbeh je. Celý po experimentoch upravený zoznam je možné si prezrieť v prílohe B.

Tieto dáta boli získané zo serverov výskumnej skupiny KNOT, ktorá IMDb stránky už mala stiahnuté v rámci iných projektov. V rámci tejto práce bola z týchto HTML súborov

---

<sup>1</sup><https://www.imdb.com/>



extrahovaná dátová sada obsahujúca názvy filmov/seriálov, zhrnutia deja a kľúčové dejové línie. Táto dátová sada obsahuje 172 789 záznamov. Vzorka týchto dát sa nachádza v tabuľke 3.1.

### Vzorka dát

Tabuľka 3.1: Vzorka dát z IMDb dátovej sady.

IMDb ID	Movie/Series title	Plot keywords	Plot summary
2140373	Saving Mr. Banks	loss of father, family relationships, death, flashback	A doting father, Walter Elias Disney (Tom Hanks)...

## 3.2 Dátová sada Booksummaries

Okrem kľúčových dejových línií je nutné zozbierať aj knižné zhrnutia dejov. Jednou z voľne dostupných dátových sád je *CMU Book Summary Dataset* [2], ktorá obsahuje zhrnutia dejov pre 16 559 kníh extrahovaných z Wikipédie spolu s metadátami z Freebase, vrátane autora knihy, názvu a žánrov. Ukážku týchto dát je možné si pozrieť v tabuľke 3.2.

### Vzorka dát

Tabuľka 3.2: Vzorka dát z Booksummaries.

Wikipedia ID	Book title	Author	Book genres	Plot summary
620	Animal Farm	George Orwell	Roman à clef, Satire, Children's literature, Speculative fiction, Fiction	Old Major, the old boar on the Manor Farm, calls the animals on the farm for a meeting, where...

## 3.3 Dátová sada z Goodreads

Jednou z najväčších sociálnych sietí o knihách je určite Goodreads<sup>2</sup>, kde si užívatelia môžu prezerat viac než 60 miliónov kníh, zistiť si o nich základné údaje, písať k nim recenzie a veľa ďalšieho. Okrem iného obsahuje každá knižná stránka zopár užívateľských recenzií k danej knihe a krátky text opisujúci určitým spôsobom danú knihu. Niekedy sa jedná o citácie z knihy, niekedy o to čo iný povedali o tejto knihe, niekedy je to abstrakt a niekedy dokonca aj zhrnutie deja. Všetky tieto dáta boli pomocou extrakcie dát z webu 2.3 stiahnuté na servery výskumnej skupiny KNOT. Bol to proces veľmi zdĺhavý, keďže ako je

<sup>2</sup><https://www.goodreads.com/>

spomenuté, knižných stránok je veľmi veľa. Sťahovanie trvalo vyše mesiaca. Cieľom bolo vytvoriť pravdepodobne najväčšiu dátovú sadu informácií o knihách a užívateľských recenziách o knihách vôbec. Všetky tieto dáta teda bolo nutné ďalej spracovať a odfiltrovať. Boli vybrané iba knihy, ktoré obsahujú viac ako 10 užívateľských recenzií, teda knihy aspoň trochu čítané, aby výsledná dátová sada bola prehľadnou a prínosnou, bez prebytočných dát. A z týchto knižných stránok boli extrahované iba dôležité informácie a boli tak vytvorené 2 dátové sady: o informáciách o knihe a o recenziách. Výsledné datasety obsahujú 23 748 238 recenzií a 891 300 informácií o knihe. V tabuľkách 3.3 a 3.4 sa nachádzajú príklady z týchto dátových sád.

Nepýtal som si žiadne povolenie od stránky Goodreads na takéto sťahovanie a keďže oblasť extrakcie dát z webu je veľmi háklivou témou, tak tieto dátové sady samozrejme nebudú nikde zverejnené a sú určené len na študijné účely a výskumná skupina KNOT ich samozrejme môže ďalej využívať v ďalších projektoch.

## Vzorka dát

Tabuľka 3.3: Vzorka dát z Goodreads datasetu informácií o knihe.

Goodreads ID	Book title	Author	No. of Ratings	No. of Reviews
18007564	The Martian	Andy Weir	931 685	77 600
Avg. Rating	Year	Description		
4.4	2011	Six days ago, astronaut Mark Watney became...		

Tabuľka 3.4: Vzorka dát z Goodreads datasetu recenzií.

Book ID	Book title	User ID	Username	Rating (/5)
18007564	The Martian	25375513	Rick Riordan	5
Date	Has spoiler	Review		
July 24, 2015	False	Adult science thriller. Love it, love it!...		

## Kapitola 4

# Návrh systému

Táto kapitola sa venuje návrhu riešenia systému pre analýzu dejových línií. Pre návrh podobného systému sú alfou a omegou dáta. Je ním zvlášť venovaná kapitola 3, kde sú popísané všetky dátové sady, s ktorými sa v práci pracuje. V tejto kapitole je popísaný proces analýzy dát, kde je priblížené ako sa dá klasifikátor naučiť predikovať kľúčové dejové línie na základe získaných dát a ako tieto výsledky vyhodnocovať a interpretovať. Nakoniec sú zhrnuté ciele tejto práce a bližšie popísané čo sa chce touto prácou dokázať.

### 4.1 Motivácia

Analýza dejových línií nie je oblasťou veľmi prebádanou, no existuje veľa prác a článkov, dokonca aj tutoriálov, o analýze žánrov. Ako je naznačené aj v úvode, v kľúčových dejových líniách 2.2 je množstvo skrytého potenciálu, ktorý sa táto práca snaží priblížiť. Sú zaujímavou alternatívou k žánrom a vedia niekedy vypovedať viac o deji ako žánre. Pomocou nich by sa v sociálnych sieťach mohol definovať nový spôsob delenia obsahu.

Oproti ostatným prácam je táto inou aj tým, že miesto častejšie používanej viac-triednej klasifikácie sa používa klasifikácia viac-značková (viď. 2.5). Je to kvôli tomu, že kľúčových dejov sa môže v zhrnutí deja nachádzať viac než len jeden, čo opäť pridáva väčšiu hodnotu pri ich klasifikácií.

Práca sa zameriava na menej bežné, no nemenej dôležité a zaujímavé dejové línie a snaží sa analyzovať zhrnutia dejov a užívateľských recenzií v anglickom jazyku a následne predpovedať k nim kľúčové dejové línie. Ďalej je systém schopný porovnať dva deje na základe týchto kľúčových dejových línií a určiť tak či sú si diela, čo sa deja týka, podobné.

Bola aj snaha zozbierať čo najväčšie množstvo dát o knihách a k nim užívateľských recenzií, keďže takéto dátové sady sú väčšinou malé, alebo neposkytujú dostatok informácií. Preto bola vytvorená pravdepodobne najväčšia dátová sada informácií o knihách a k nim užívateľských recenzií v anglickom jazyku.

### 4.2 Analýza dát

Zo spracovanými dátami a vytvorenými dátovými sadami, je už možné tieto dáta ďalej analyzovať a prejsť k úlohe klasifikácie. Najlepšie výsledky dosahujú metódy založené na strojovom učení, a najmä tie založené na neurónových sieťach, preto je asi najvhodnejšie vybrať BERT [6], prípadne nejakú jeho variantu, keďže dominuje v rebríčkoch<sup>1</sup> viacerých úloh

---

<sup>1</sup><http://nlpprogress.com/>

spracovania prirodzeného jazyka. Je nutné stiahnuť nejakú variantu BERT a adaptovať ho na dáta, s ktorými bude pracovať. Bude sa teda jednať o *fine-tuning* 2.8.1, ktorému sa v tomto kontexte bude hovoriť tréning, učenie či ladenie. Pred samotným tréningom je nutné sa uistiť aby mal BERT v poslednej vrstve aktivačnú funkciu sigmoid a nie softmax, pretože sa bude jednať o viac-značkovú klasifikáciu, bez tohoto by na výstupe boli úplne iné hodnoty. Predtým, než sa model začne učiť, je treba správne nastaviť parametre učenia, napr. rýchlosť učenia, hodnotiacia funkcia, po koľkých krokoch sa má model znova vyhodnotiť a pod. Budú sa takisto musieť využiť nejaké spôsoby pre spracovanie dlhých reťazcov 2.8.2, lebo veľa zhrnutí dejov je dlhších ako 512 tokenov. Je vhodné vyskúšať viaceré metódy a zistiť, ktorá bude fungovať najlepšie. Takisto je potrebné rozdeliť dataset do 3 častí: tréningovej, testovacej a validačnej. Každá z týchto 3 podmnožín je určitou percentuálnou časťou pôvodného datasetu a je nutné určiť aby toto rozdelenie bolo čo najlepšie. Ďalej, by mali všetky tieto podmnožiny mať zhruba rovnaké zastúpenie všetkých prípadov. V rámci tejto práce to znamená, aby čo možno všetky značky boli zastúpené v každej podmnožine, preto je vhodné aby dáta boli do týchto podmnožín rozdelené náhodne. Tréningová sada býva zo všetkých najväčšia. Z týchto dát sa bude model učiť a snažiť sa nájsť správne nastavenia vnútorných stavov tak, aby sa čo najviac priblížil pôvodným dátam. Testovacia sada slúži na výpočet vyhodnocovacej metriky. Po každej fáze učenia, po určitom počte krokov, nasleduje fáza vyhodnotenia, teda kde model predpovedá ako by určil kľúčové dejové línie a porovnávajú sa so skutočnými údajmi. Spôsob porovnania závisí od vyhodnocovacích metrick 2.9, ktoré je nutné vybrať správne, aby metriky neboli príliš prísne, ale zároveň prínosné. Po tejto fáze model dostane spätnú väzbu a môže pokračovať v učení lepším spôsobom. Validačná sada slúži na to isté čo testovacia, až na to, že sa táto sada nepoužíva počas tréningu, ale až na finálnu validáciu modelu. Jedná sa teda o dáta, ktoré model „nikdy nevidel“ a vyhodnocovacie metriky nie sú nimi žiadno ovplyvnené.

Po tomto môže byť započatý proces učenia. Je zvykom nechať modelom prejsť tréningovú sadu viackrát, zvlášť keď je počet dát veľký. Jedno takéto „prejdenie“ sa nazýva epocha. IMDb dátová sada patrí medzi tie väčšie, preto bude nutné pustiť učenie na desiatky možno až stovky epoch, aby boli výsledky kvalitné. Treba podotknúť, že modely typu BERT sú pamäťovo veľmi náročné a spolu z veľkou dátovou sadou bude výpočet veľmi náročný nie len na pamäť grafickej karty ale aj čo sa týka času. Preto je nutné tréningovať takúto úlohu na výkonných grafických kartách, no napriek tomu môže výpočet trvať aj niekoľko dní. V praxi sa miesto klasických GPU používajú tzv. TPU<sup>2</sup>, ktoré sú určené presne na tréningovanie neurónových sietí a úlohy podobného typu.

Pre naučenie klasifikátora odhadovať kľúčové dejové línie sa použije dátová sada z IMDb, kde vstupom budú zhrnutia dejov a výstupom budú pravdepodobnosti každej značky, ktoré budú znamenať na koľko si model myslí, že sa v zhrnutí nachádza daná kľúčová dejová línia. Zvolí sa prah, od ktorého sa budú kľúčové dejové línie klasifikovať na tie, ktoré sa v diele nachádzajú a na tie ktoré nie. Predpokladané výsledky sa porovnajú s reálnymi a BERT sa bude snažiť nachádzať súvislosti medzi kľúčovými dejovými líniami a slovami v zhrnutí. Nepredpokladá sa, že výsledky vyhodnocovacích metrick budú vysoké, lebo všeobecne viac-značková klasifikácia z veľa značkami nemá vynikajúce výsledky, takisto nemusia byť dáta dobre označované a rovnako aj skracovanie textov bude znižovať hodnoty metrick. Je ale dôležitejšie ako bude model fungovať nad dátami, ktoré nie sú označované a či dokáže aj v takých dátach nájsť kľúčové dejové línie, ktoré budú odrážať skutočnosť. Pre toto je vhodné vybrať si knihy (zhrnutia), ktoré poznáme, alebo sú nám známe a nechať model

---

<sup>2</sup>TPU – tensor processing unit

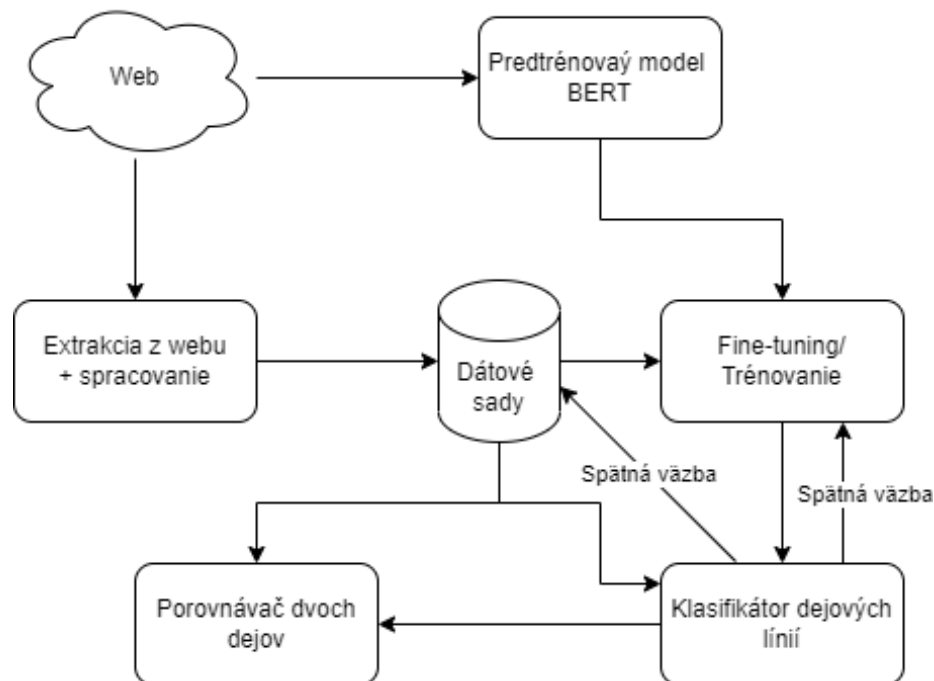
nech nad nimi predpovedá kľúčové dejové línie a výsledky ohodnotí podľa vlastných predpokladov a očakávaní. Toto je vlastne cieľom tejto práce, či dokáže klasifikátor odhadnúť aké deje sa v príbehu nachádzajú a na koľko sú relevantné. Na toto budú využité zvyšné dátové sady.

Takýto klasifikátor možno takisto použiť na porovnávanie dvoch dejov a pomocou metrík podobnosti 2.10 porovnať získané predpovede dvoch dejov a na základe tohoto porovnania sa zistí, či dané pravdepodobnostné vektory sú si podobné a teda či sú podobné aj deje. Úspešnosť a presnosť tohoto porovnania však do veľkej miery závisí od toho ako dobre sa budú predpovedať kľúčové dejové línie.

### 4.3 Schéma systému

Na obrázku 4.1 je zobrazená navrhovaná schéma výsledného systému. Šípky v tomto diagrame reprezentujú tok dát a obdĺžniky predstavujú konkrétne moduly.

Z webu je potrebné stiahnuť potrebné dátové sady, či už pomocou web scrapingu, alebo už priamo pripravené sady. Tieto dáta treba spracovať (HTML parsing) a odfiltrovať a vytvoriť z nich dátové sady, ktoré môžu byť použité na tréningovanie neurónovej siete. Z internetu bude prevzatý aj pretrénovaný model BERT, ktorý sa doladí z vytvorených dátových sád. Výsledkom bude funkčný klasifikátor dejových línií. Bude však mať svoje chyby a proces tréningovania sa bude opakovať viac-krát s tým, že sa budú upravovať dátové sady a parametre učenia, aby sa docielilo čo najlepších výsledkov. Z klasifikátora bude vychádzať aj porovnávač dejových línií.



Obr. 4.1: Navrhovaná schéma výsledného systému.

## 4.4 Zhrnutie návrhu

Hlavným cieľom práce je vytvoriť systém schopný zo zhrnutí dejov určiť o aké kľúčové deje [2.2](#) sa v ňom jedná. Bude sa jednať o viac-značkovú klasifikáciu [2.5](#), teda viac ako jeden dej sa môže vyskytnúť v jednom zhrnutí. Klasifikátorom kľúčových dejových línií bude predtrénovaný model typu BERT [2.8](#), ktorý bude adaptovaný na túto úlohu. Kvalitu modelu budú určovať metriky vyhodnotenia [2.9](#), no takisto vlastný názor.

Jedným z vedľajších cieľov je vytvoriť porovnávač dvoch dejových línií, ktorý bude z klasifikátora vychádzať. Tento porovnávač najskôr zistí pravdepodobnosti každej z kľúčových dejových línií v zhrnutí v oboch dielach a potom tieto vektory porovná pomocou podobnostných metrik [2.10](#) a určí podobnosť týchto diel.

V neposlednom rade je cieľom zozbierať viacero zdrojov knižných dát a vytvoriť systém schopný extrahovať informácie o knihách a užívateľských recenziách z obľúbenej sociálnej siete zameranej na knihy – Goodreads.

## Kapitola 5

# Implementácia systému

V tejto kapitole bude priblížená implementácia systému popísaného v kapitole 4 založená na dátach z kapitoly 3. Najskôr bude popísané, z akých modulov sa riešenie skladá a z akých knižníc je riešenie vytvorené. Následne sú popísané jednotlivé moduly projektu. Je v nich objasnené čo všetko muselo byť naprogramované, konkrétne využité parametre a metriky vybrané po viacerých experimentoch. Možné námety na vylepšenia sú popísané v záverečnej podkapitole. V tejto kapitole je len popísané, ako systém funguje a ako bol implementovaný, no konkrétne príklady a experimenty budú popísané v kapitole 6.

### 5.1 Architektúra projektu

Implementačná fáza projektu pozostáva z 3 častí:

- Sťahovanie a spracovanie dát
- Trénovanie a adaptácia BERT
- Načítanie BERT pre vyhodnotenie

Systém je implementovaný v jazyku *Python 3*, ktorý je veľmi obľúbený a často používaný v úlohách týkajúcich sa analýzy dát a aplikáciach strojového učenia.

Pre sťahovanie dát bola využitá knižnica *Selenium*<sup>1</sup>, pre následné spracovanie stiahnutých súborov v HTML bola použitá knižnica *BeautifulSoup*<sup>2</sup>. Pre vytvorenie dátových sád sa používali *Pandas*<sup>3</sup>, *Numpy*<sup>4</sup> a knižnica *Datasets*<sup>5</sup>, vytvorená priamo spoločnosťou HuggingFace<sup>6</sup> na vytváranie datasetov pre pred-trénované modely. Skripty na trénovanie neurónovej siete boli takisto používajú knižnicu od HuggingFace – *Transformers*<sup>7</sup>, ktorá ponúka API pre jednoduché sťahovanie a trénovanie najmodernejších pred-trénovaných modelov pre viaceré frameworky určené pre strojové učenie. V tejto práci bol *Transformers* použitý spolu s *PyTorch*<sup>8</sup>. Už viac-krát spomenutý HuggingFace je web, poskytujúci veľké množstvo voľne dostupných pred-trénovaných modelov pre rôzne typy úloh, z ktorého bol

---

<sup>1</sup><https://www.selenium.dev/>

<sup>2</sup><https://beautiful-soup-4.readthedocs.io/>

<sup>3</sup><https://pandas.pydata.org/>

<sup>4</sup><https://numpy.org/>

<sup>5</sup><https://huggingface.co/docs/datasets>

<sup>6</sup><https://huggingface.co/>

<sup>7</sup><https://huggingface.co/docs/transformers>

<sup>8</sup><https://pytorch.org/>

prevzatý aj model BERT pre účely tejto práce. Pre vyhodnocovanie úspešnosti klasifikácie a pre vyhodnocovanie podobnosti boli použité knižnice *Sklearn*<sup>9</sup> a *Scipy*<sup>10</sup>.

## 5.2 Dáta

Ako bolo spomenuté v kapitole 3, v práci sa používajú 3 dátové sady, ale iba dátové sady z Goodreads je nutné sťahovať, a spolu s dátovou sadou z IMDb aj určitým spôsobom spracovať. V tejto podkapitole je bližšie popísané ako sa dáta sťahovali a spracovávali.

### Sťahovanie

Sťahovaná bola iba dátová sada z Goodreads, lebo dátová sada z IMDb sa už nachádzala na serveroch KNOT, preto nebolo nutné tieto dáta znova sťahovať. Pre sťahovanie dát z Goodreads bolo nutné pripraviť zoznam URL, ktoré je treba stiahnuť. Tento zoznam následne sťahoval modul `download`. Jedná sa o skript sťahujúci údaje z internetu a je v ňom využitá knižnica *Selenium*, ktorá simuluje webový prehliadač, pomocou tzv. *Chromedriver*. Proces sťahovania bol paralelizovaný medzi všetkých 77 serverov skupiny KNOT pre urýchlenie procesu sťahovania, no aj tak tento proces trval viac ako mesiac. Výstupom boli súbory typu HTML všetkých knižných stránok, ktoré je treba ďalej spracovať. Z neznámych dôvodov však servery Goodreads vracali dva druhy stránok typu HTML.

### Spracovanie súborov typu HTML

Takto stiahnuté dáta treba spracovať do formy, z ktorou sa ľahko manipuluje, preto všetky dátové sady sú súbory typu TSV<sup>11</sup>.

Keďže dáta z Goodreads sú dvojakeho druhu, boli vytvorené aj dva skripty pre ich spracovanie. Jedná sa o moduly `extract_book_info` a `extract_fuzzy_book_info`, ktoré pomocou *BeautifulSoup4* extrahujú najdôležitejšie informácie najskôr o knihe samotnej a potom aj o recenziách danej knihy. Výstupom sú potom dva súbory typu TSV, jeden o informáciách o knihe, druhý obsahujúci užívateľské recenzie. V tabuľkách 3.3 a 3.4 je popísané ako vyzerá konkrétny príklad jedného riadku oboch takto zostrojených dátových sád.

Čo sa týka spracovania súborov typu HTML z IMDb, bol použitý rovnaký mechanizmus ako pri Goodreads dátach, pričom `extract_keywords_imdb` je skript, ktorý túto extrakciu zabezpečuje. Ako je možné si všimnúť v tabuľke 3.1, ktorá opäť ukazuje názornú ukážku z vytvorenej dátovej sady, nachádza sa tu len málo položiek dát. Je to kvôli tomu, že táto dátová sada slúži len na tréningovanie a netreba k nej dodatočné informácie. Tieto dáta bolo nutné ešte odfiltrovať, pretože najskôr bol takto spracovaný dataset so všetkými kľúčovými slovami a so všetkými filmami/seriálmi. Filtrácia spočívala v tom, že najskôr boli vybraté kľúčové dejové línie (viď. príloha B) a pokiaľ sa medzi kľúčovými slovami nachádzala aspoň jedna kľúčová dejová línia, tak bol záznam ponechaný a všetky kľúčové slová, mimo kľúčových dejových línií boli vymazané, inak bol záznam vymazaný celý.

### Spracovanie dátovej sady

Predtým ako sú dáta predstavené neurónovej sieti je nutné ich pripraviť na formát, ktorým sa dokáže sieť učiť. Takisto je to aj v tomto prípade. Keďže je využitá knižnica *Transformers*,

---

<sup>9</sup><https://scikit-learn.org/>

<sup>10</sup><https://scipy.org/>

<sup>11</sup>TSV – Tab-separated values (hodnoty oddelené tabulátorom)



tak bola využitá aj knižnica *Datasets*, ktoré spolu dobre komunikujú, keďže sú vytvorené jednou spoločnosťou. Modul zabezpečujúci túto prípravu je `prepare_imdb_data`.

Jedná sa hlavne o prípravu kľúčových dejových línií do správneho formátu. Keďže neuronová sieť sa text na díva ako na čísla, je treba ich aj takto reprezentovať. Pre viac-značkovú klasifikáciu v tejto práci to znamená, že namiesto položky kľúčové dejové línie, ktorá obsahuje slovný zoznam kľúčových dejových línií platných pre dané zhrnutie, budú všetky kľúčové dejové línie (nie len tie, ktoré sú platné pre dané zhrnutie) mať vlastný záznam obsahujúci informáciu o tom, či je v danom zhrnutí platná (0,1).

Okrem tohoto sú upravené aj zhrnutia. Existujú totiž siete, ktoré rozlišujú malé a veľké písmena, no ich veľkosť a výpočetná náročnosť je samozrejme väčšia. U zhrnutí deja nehrá veľkú úlohu veľkosť počiatočných písmen, preto bol použitý model, ktorý nerozlišuje veľké a malé písmená, preto boli všetky zhrnutia prevedené do malých písmen. Okrem toho, je nutné vyriešiť problém toho, že väčšina zhrnutí je dlhších ako 512 tokenov, teda je nutné istým spôsobom skrátiť tieto zhrnutia. Najskôr bolo použité odstránenie tzv. *stopwords*, ktoré reprezentujú slová, ktoré nemajú emocionálny význam a nemajú teda čo dočinenia s dejom a sú viac-menej zbytočné. Jedná sa o slová ako napr. *of, it, the, a*. Nemusí to byť najlepšou voľbou, ale aj takéto nepatrné zmeny môžu pomôcť skrátiť zhrnutie. Nepomôže to však všetkým zhrnutiam, preto je nutné ešte skrátiť tieto zhrnutia. Boli vyskúšané viaceré metódy, no najlepšou a zároveň jednou z najjednoduchších sa javí odseknutie iba posledných 510 tokenov 2.8.2. Je to kvôli tomu, že rozuzlenie zápletky a teda určenie typu deja sa deje na konci zhrnutia. Rovnako sú zo zhrnutí odstránené interpunkcie, aby model pracoval iba s čistým textom. Tieto zhrnutia sú neskôr takisto prevedené na čísla, na identifikátory tokenov, pomocou tzv. *tokenizer*. Ako spomenuté aj v návrhu, je nutné túto dátovú sadu rozdeliť na tréningovú, testovaciu a validačnú. Konkrétne sa použilo rozloženie 90:5:5, ktoré je často používané vo viacerých úlohách strojového učenia.

Výsledné dáta ktorými je model „kŕmený“ možno vidieť v tabuľke 5.1.

Tabuľka 5.1: Vzorka dát z IMDb dátovej sady spracovaná na tréningovanie.

Movie/Series title	loss of father	moral dilemma	...	death	Plot summary
Saving Mr. Banks	1	0	...	1	doting father walter elias disney tom hanks...

### 5.3 Klasifikácia

V tejto podkapitole je popísaný konkrétny pred-trénovaný model použitý na tréningovanie, ako prebiehal proces tréningu, parametre tréningu, konkrétne použité metriky a ako sa pracovalo s výsledným modelom.

#### Stiahnutý model

Z webu HuggingFace je možné si vybrať z veľkého množstva voľne dostupných pred-trénovaných modelov využiteľných na rôzne úlohy. Pre jednoduchosť bol zvolený `roberta-base`. Jedná sa o model založený na princípoch BERT, ako z názvu vyplýva – RoBERTa [13], značí *A Robustly Optimized BERT Pretraining Approach*, teda robustne optimalizovaný prístup pred-tréningovania BERT. Dosahuje takisto trochu lepšie výsledky ako klasický BERT. Bola

použitá „menšia“ varianta, lebo už aj táto varianta je dostatočne pamäťovo náročná a prínos väčšieho modelu by bol zanedbateľný.

## Trénovanie/adaptácia

Na trénovanie slúži modul `trainer_imdb`, ktorý používa *Trainer API* z knižnice *Transformers*. Pre proces trénovania je nutné nastaviť parametre trénovania – spôsob ako bude tréning prebiehať a ako sa trénovanie bude vyhodnocovať. Toto všetko je v *Trainer API* zabezpečené veľmi priamočiaro, aj vďaka spolupráci s *Datasets*. Všetky parametre tréningu je možné si pozrieť v kóde, no tie dôležitejšie sú uvedené v tomto odstavci. Po 2000 krokoch vždy nasleduje vyhodnocovanie, miera učenia (angl. learning rate) bola po viacerých experimentoch zvolená na  $3e-5$ , veľkosť dávok (batch size) bola zvolená 4, aj preto, že väčšia už nemohla byť kvôli pamäti na grafickej karte. Táto nízka veľkosť dávky bola ale vyvážená gradientovými akumuláčnými krokmi (angl. gradient accumulation steps). Počet epoch vo finálnej verzii je 30, čo aj tak reprezentuje tréning, ktorý trvá niekoľko dní. Vyhodnocovacie metrikou, ktoré určovali úspešnosť modelu boli viaceré. Konkrétne klasické F1-skóre, jeho makro a mikro varianta, správnosť a varianty priemernej presnosti. Hlavnou, ktorá určovala zlepšenie bolo klasické F1-skóre. Dátovou sadou pre trénovanie je spomínaná trénovacia dátová sada a pre priebežne vyhodnotenie to je testovacia dátová sada. Validačná dátová sada je použitá po tréningu na celkový výpočet úspešnosti.

Trénovanie bežalo na grafickej karte NVidia RTX3090 TURBO na jednom zo strojov skupiny KNOT. Karta má 24 GB veľkosť operačnej pamäte, no aj tak to v niektorých prípadoch experimentov bolo málo.

Počas tréningu sa modely priebežne ukladajú, aby mohli byť ľahko načítané a použité.

## Načítanie modelu

Najlepší model je načítaný z pamäte a s ním sú potom vykonávané experimenty a vyhodnotenia. Model sa načítava pre jednu z troch vecí: výpočet vyhodnocovacích metrík, určenie kľúčových dejových línií, porovnanie dvoch dejov. Modul `load_model_imdb` je zodpovedný za výpočet vyhodnocovacích metrík. Modelom sa nechá prejsť validačná dátová sada a všetky predpovede modelu sa ukladajú a nakoniec sa porovnajú s očakávanými výsledkami, vypočítajú sa a vypíšu sa výsledky. Metriky, ktoré sa používajú sú: F1-skóre, mikro F1-skóre, makro F1-skóre, správnosť, priemerná presnosť, mikro priemerná presnosť a makro priemerná presnosť. Iné metriky sú dosť prísne, preto boli použité práve tieto. Výsledky modelu sa dajú priebežne získavať počas trénovania, teda určitá predstava o výsledkoch je už počas adaptácie. Avšak prácou s dátami, ktoré model nikdy nevidel je získané objektívnejšie zhodnotenie modelu. Toto načítanie trvá aj niekoľko minút, keďže model musí všetky položky z validačnej dátovej sady vyhodnotiť, to zaberá najviac času, potom už výpočet metrík je časovo nenáročný.

Ďalej sa model načítava pre určenie kľúčových dejových línií ľubovoľného zhrnutia. Vstupom je ľubovoľné zhrnutie, ktoré je opäť pomocou *tokenizer* prevedené do jazyka, ktorému model rozumie a rovnako musí byť zhrnutie upravené a skrátené tak ako pri spracovaní dátovej sady. Výstupom nie je nič iné ako predpovedané kľúčové dejové línie. Takisto sú k výstupom vypísané jednotlivé pravdepodobnosti kľúčových dejových línií, teda na koľko si bol model istý, že sa jedná o danú kľúčovú dejovú líniu. Toto zabezpečuje skript `predict_plot`.

Pre porovnanie dvoch dejov slúži modul `compare_plots`, ktorý robí to isté ako modul popísaný v predošlom odstavci, akurát to robí pre dve zhrnutia. Vypočítané pravdepodobnosti oboch zhrnutí pre všetky kľúčové dejové línie sú porovnané pomocou metrík podob-

nosti 2.10. Všetky tri metriky, teda Spearmanov korelačný koeficient, Pearsonov korelačný koeficient a Kosínusová podobnosť sú spolu spriemerované a výsledok určuje percento podobnosti týchto dvoch dejov.

## Kapitola 6

# Experimentálne vyhodnotenie a diskusia

V tejto kapitole je najskôr zhodnotený stiahnutý dataset a porovnaný s inými voľne dostupnými dátovými sadami podobného typu. Ďalej sa popisuje úspešnosť výsledného klasifikátora pomocou viacerých príkladov a vyhodnocovacích metrík. Následne je popísaná využiteľnosť takéhoto klasifikátora a nakoniec sú navrhnuté námety na zlepšenie tejto práce.

### 6.1 Dátová sada

Stiahnuté dátové sady z Goodreads patria určite medzi tie najväčšie svojho druhu. Dátová sada informácií o knihe obsahuje 891 300 záznamov, v porovnaní z voľne dostupnou dátovou sadou *Best Books Ever Dataset* [14], ktorá obsahuje len niečo cez 50 000 záznamov, je omnoho väčšou, no neobsahuje toľko informácií. *Best Books Ever Dataset* obsahuje 25 položiek v každom zázname, oproti 8 v stiahnutej Goodreads dátovej sade. Obsahuje navyše napr. cenu, žáner, ISBN, ocenenia a iné. Dátová sada *Goodreads-books* [5] obsahuje takmer všetko čo stiahnutá sada, ale bez popisu knihy. Takisto obsahuje táto sada len viac ako 11 000 informácií o knihe. Tieto dve dátové sady výrazne zaostávajú čo sa kvantily týka, ale prvá spomenutá obsahovala viac informácií. Ďalšie dátové sady *Goodreads Datasets* [24] [25] sú veľmi podobné tým stiahnutým. Rovnako majú zvlášť dátovú sadu pre informácie o knihe a zvlášť pre užívateľské recenzie. *Goodreads Datasets* s informáciami o knihe obsahuje viac než 2.3 milióna záznamov, obsahuje aj viac informácií, ako napr. podobné knihy, jazyk knihy, vydavateľa, počet strán a pod. *Goodreads Datasets* s užívateľskými recenziami obsahuje viac ako 15 miliónov viac-jazyčných recenzií. Obsahuje tie isté dáta ako stiahnutá dátová sada recenzií, ale navyše informácie o počte komentárov, o tom kedy bola kniha recenzentom prečítaná a kedy bola recenzia upravená. Stiahnutá dátová sada z Goodreads obsahuje 23 748 238 záznamov, recenzie v nej sú prevažne v anglickom jazyku, no knihy, ktoré nemajú anglické recenzie, obsahujú tieto ne-anglické recenzie. *Goodreads Datasets* sú dátové sady veľmi podobné tým stiahnutým dátovým sadám z Goodreads, v niečom obsahujú aj viac záznamov a v oboch sadách obsahujú viac informácií. Informácie navyše však nemusia byť vždy prínosom. *Goodreads Datasets* boli pravdepodobne získané cez Goodreads API, ktoré bolo v roku 2019 zrušené, teda tieto sady nemajú tendenciu sa zväčšovať, či upravovať. Stiahnuté dátové sady z Goodreads boli získané extrakciou z webu, preto je možné ľahko pridať dodatočné informácie a sťahovania po čase opakovať.

## 6.2 Najlepšie dosiahnuté štatistiky

V tabuľke 6.1 sú zobrazené vybrané metriky najlepšieho natrénovaného modelu na validačnej dátovej sade z IMDb. Model bol adaptovaný na dáta tridsiatimi epochami. Bolo by možné model trénovať aj dlhšie, no to by neprineslo markantne lepšie výsledky klasifikácie, možno len nepatrne lepšie metriky vyhodnocovania.

Z dát v tabuľke si možno všimnúť, že výsledky nevyzerajú veľmi dobré, v príkladoch však bude ukázané, že výsledky vôbec nie sú také zlé, ako sa na prvý pohľad zdá. Je pochopiteľné prečo je tomu tak. Pre konkrétne zhrnutie môže byť správnych aj viac než 10 kľúčových dejových línií a niekedy aj iba 1. Preto správne určenie všetkých značiek je úlohou veľmi obtiažnou aj pre BERT, z toho vyplýva aj nízka hodnota správnosti. F1-skóre už trochu lepšie reprezentuje výsledky a je vidieť na hodnote mikro metriky, že keď sú hodnoty spriemerované globálne pre všetky značky, sú výsledky ešte o niečo lepšie. To značí, že niektoré kľúčové dejové línie klasifikuje lepšie ako iné. Najlepšie si vedie v priemernej presnosti, čo značí, že pravdepodobnosti skutočných značiek sú v zozname všetkých pravdepodobností vyššie ako tie nerelevantné. To hovorí o systéme, že dokáže aspoň čiastočne určovať kontext a zaradiť relevantné kľúčové dejové línie vyššie.

Tabuľka 6.1: Výsledná úspešnosť klasifikátora určujúci kľúčové dejové línie.

Acc.	F1	$F1_{Micro}$	$F1_{Macro}$	Avg.Pre.	Avg.Pre.Micro	Avg.Pre.Macro
23.3	30.7	36.9	27.1	45.4	33.8	24.5

## 6.3 Príklady

V tejto časti budú podrobne popísané výstupy modelu na viacerých príkladoch, či už na tých kde si model vedie dobre, ale aj na tých, ktoré vyhodnotí zle. Budú použité dáta zo všetkých dátových sád spomenuté v kapitole 3. Príklady sú vykonávané na najlepšie adaptovanom modeli.

### Príklady z dátovej sady z IMDb

Prvým príkladom je zhrnutie z IMDb z filmu *Taken*, po slovensky známe pod názvom *96 hodín*. Zhrnutie je nasledovného znenia:

*"A retired CIA agent travels across Europe and relies on his old skills to save his estranged daughter, who has been kidnapped while on a trip to Paris. Seventeen year-old Kim is the pride and joy of her father Bryan Mills. Bryan is a retired agent who left the Central Intelligence Agency to be near Kim in California. Kim lives with her mother Lenore and her wealthy stepfather Stuart. Kim manages to convince her reluctant father to allow her to travel to Paris with her friend Amanda. When the girls arrive in Paris they share a cab with a stranger named Peter, and Amanda lets it slip that they are alone in Paris. Using this information an Albanian gang of human traffickers kidnaps the girls. Kim barely has time to call her father and give him information. Her father gets to speak briefly to one of the kidnappers and he promises to kill the kidnappers if they do not let his daughter go free. The kidnapper wishes him 'good luck', so Bryan Mills travels to Paris to search for his daughter and her friend."*

Kto videl film vie, že hlavný hrdina má na záchranu, ako zo slovenského názvu vyplýva, len 96 hodín a počas celého filmu sa všade strieľa zo zbraní. To však v zhrnutí povedané

nie je, a predsa na obrázku 6.1 je vidieť, že model z kontextu zistil obe tieto veci (race against time, gunfight, shot to death). Na obrázku sa nachádza 15 kľúčových dejových línií s najväčšou pravdepodobnosťou, ktoré sa podľa klasifikátora v zhrnutí nachádzajú. Dovolím si tvrdiť, že všetky z nich sú správne, aj s určitou znalosťou o filme. Klasifikátor má nastavený prah na hodnotu 0.5, preto finálne určené kľúčové dejové línie sú: ['death', 'murder', 'violence', 'kidnapping', 'chase', 'shot to death', 'gunfight', 'flashback', 'corruption', 'one man army', 'race against time', 'rescue', 'revenge']

```

Taken
Predicted:
('kidnapping', 0.9963363409042358)
('murder', 0.9520508646965027)
('rescue', 0.9314407110214233)
('chase', 0.9166193604469299)
('one man army', 0.9074912071228027)
('violence', 0.8842470049858093)
('death', 0.7916989922523499)
('corruption', 0.6554111838340759)
('shot to death', 0.6456857323646545)
('revenge', 0.580049455165863)
('gunfight', 0.550561249256134)
('race against time', 0.5262531638145447)
('flashback', 0.5108363628387451)
('one against many', 0.45202288031578064)
('beating', 0.4232092499732971)

```

Obr. 6.1: Pravdepodobnosti 15 kľúčových dejových línií s najväčšou pravdepodobnosťou pre zhrnutie deja filmu *96 hodín*.

Na ďalšom príklade bude ukázané porovnanie 2 dejov. Bude sa porovnávať rozprávka *Nemo* a vyššie spomínaný film *96 hodín*. Na prvý pohľad sa to môže zdať ako vtip, no v oboch filmoch ide o to, že otec hľadá svojho uneseného potomka, pričom musí prejsť cez rozličné prekážky. Z filmu *96 hodín* bolo zobrazené rovnaké zhrnutie ako v príklade vyššie a z filmu *Nemo* išlo o následovné zhrnutie:

*" After his son is captured in the Great Barrier Reef and taken to Sydney, a timid clownfish sets out on a journey to bring him home. A clown fish named Marlin lives in the Great Barrier Reef and loses his son, Nemo, after he ventures into the open sea, despite his father's constant warnings about many of the ocean's dangers. Nemo is abducted by a boat and netted up and sent to a dentist's office in Sydney. While Marlin ventures off to try to retrieve Nemo, Marlin meets a fish named Dory, a blue tang suffering from short-term memory loss. The companions travel a great distance, encountering various dangerous sea creatures such as sharks, anglerfish and jellyfish, in order to rescue Nemo from the dentist's office, which is situated by Sydney Harbour. While the two are searching the ocean far and wide, Nemo and the other sea animals in the dentist's fish tank plot a way to return to the sea to live their lives free again. "*

Pre film *nemo* určil klasifikátor nasledujúce dejové línie: ['friendship', 'kidnapping', 'flashback', 'rescue']. Na obrázku 6.2 je vidieť zvyšné pravdepodobnosti. Podľa mňa si v tomto prípade viedol model dobre, naozaj ide o priateľstvo, únos a záchranu.

```

Nemo
Predicted:
('flashback', 0.9710222482681274)
('friendship', 0.8978711366653442)
('kidnapping', 0.8814532160758972)
('rescue', 0.6166682243347168)
('escape', 0.2998323142528534)
('chase', 0.22296494245529175)
('courage', 0.1757747232913971)
('violence', 0.14620810747146606)
('death', 0.13336676359176636)
('trauma', 0.11240385472774506)
('lie', 0.08635343611240387)
('memory loss', 0.08482448011636734)
('fight', 0.08413172513246536)
('fear', 0.07524951547384262)
('drunkenness', 0.06318911164999008)

```

Obr. 6.2: Pravdepodobnosti 15 klúčových dejových línií s najväčšou pravdepodobnosťou pre zhrnutie deja filmu *Nemo*.

Na obrázku 6.3 je možné vidieť vypočítané podobnostné metriky pre tieto dve diela. Boli vypočítané zo všetkých 88 pravdepodobností klúčových dejových línií a porovnané medzi sebou metrikami na obrázku. Model je na hranici váhania, či sú filmy podobné, ale v priemere určil, že nie sú. Možno sa môže zdať, že by filmy mohli byť podobné, ale model počíta koreláciu všetkých dejov, no a vo filme *96 hodín* sú relevantné deje ako napr. *one man army*, *gunfight*, *beating* atď., čo vo filme *Nemo* samozrejme nevyskytujú tak vysoko.

```

Correlation:
Cosine similarity: 0.5022087652711409
Spearman correlation: 0.4926561234193935
Pearson correlation: 0.40692395125457187
=====
Average : 0.4672629466483688

```

Obr. 6.3: Vypočítané metriky podobnosti pre pravdepodobnosti filmov *Nemo* a *96 hodín*.

## Príklady z dátovej sady Booksummaries

Pre kvalitu klasifikátora je potrebné vyskúšať aj knižné zhrnutia. Jednou z klasických kníh je kniha *Na západe nič nové*. Zhrnutie deja tejto knihy je možné si prečítať v prílohe C. Predpovede modelu je možné vidieť na obrázku 6.4. Model určil tieto kľúčové dejové línie: ['friendship', 'death', 'murder', 'violence']. Každý, kto čítal knihu vie, že dej sa odohráva vo vojne, teda zabíjanie, násilie a smrť možno označiť za témy v knihe. Priateľstvo je veľmi kľúčovou dejovou líniou, ale nevystihuje to dostatočne pointu deja. Možno však povedať, že skoro všetky kľúčové dejové línie, ktoré je vidieť na obrázku však istým spôsobom vystihujú dej, ale nemajú dostatočnú pravdepodobnosť. Chýba stále asi hlavná pointa, možno by mohla byť zaradená kľúčová dejová línia *hľadanie zmyslu života* alebo *nezmyselnosť*.

```
Erich Maria Remarque - All Quiet on the Western Front
Predicted:
('death', 0.987436056137085)
('violence', 0.6199653744697571)
('friendship', 0.6198490262031555)
('murder', 0.5922976136207581)
('battle', 0.25238141417503357)
('suicide', 0.2504725456237793)
('shot to death', 0.15569864213466644)
('courage', 0.10911519080400467)
('ambush', 0.07036996632814407)
('drinking', 0.059596892446279526)
('fight', 0.05748329311609268)
('family relationships', 0.048228126019239426)
('pregnancy', 0.04457623511552811)
('love', 0.04427250847220421)
('marriage', 0.04324689880013466)
```

Obr. 6.4: Pravdepodobnosti 15 kľúčových dejových línií s najväčšou pravdepodobnosťou pre zhrnutie deja knihy *Na západe nič nové*.



Nasledujúci príklad je skôr hrou ako knihou, ale je na tomto príklade od Williama Shakespeara vidieť ďalšie fenomény modelu. Zhrnutie hry *Rómeo a Júlia* je možné is prezrieť v prílohe C. Model určil kľúčové dejové línie následovne : ['murder', 'suicide', 'revenge']. Zvyšné sa nachádzajú na obrázku 6.5. Nedá sa povedať, že zabitie, samovražda a pomsta popisujú tento dej. Avšak veľmi vysoko v zozname je kľúčová dejová línia *forbidden love*, ktorá akoby vychádzala z tejto klasickej hry, avšak má príliš nízku pravdepodobnosť. Rovnako by *family relationships*, *rivalry* či *marriage* mohli byť zaradené medzi správne.

```

William Shakespeare - Romeo and Juliet
Predicted:
('murder', 0.8249778151512146)
('revenge', 0.7284821271896362)
('suicide', 0.5880095958709717)
('death', 0.464555561542511)
('forbidden love', 0.28789207339286804)
('battle', 0.2830275297164917)
('love', 0.27093255519866943)
('family relationships', 0.24258042871952057)
('jealousy', 0.18859781324863434)
('rivalry', 0.14176155626773834)
('violence', 0.10342272371053696)
('marriage', 0.0927327424287796)
('betrayal', 0.08640135824680328)
('romantic rivalry', 0.07478997111320496)
('unrequited love', 0.07169553637504578)

```

Obr. 6.5: Pravdepodobnosti 15 kľúčových dejových línií s najväčšou pravdepodobnosťou pre zhrnutie deja hry *Rómeo a Júlia*.

Ďalší príklad ukazuje podobnosť dvoch známych kníh : *Zločin a trest* (angl. *Crime and Punishment*) a *Mechanický pomaranč* (angl. *A Clockwork Orange*). Zhrnutia týchto dejov je možné si prečítať v prílohe C. Na obrázku 6.6 je vidieť predpovede dejov oboch kníh, spolu s vypočítanými metrikami podobnosti. Knihu *Mechanický pomaranč* určil model dobre, nedá sa povedať, že skvele, lebo *marriage* by nemalo mať takú vysokú pravdepodobnosť a zase iné ako napr. *fight* a *vandalism* by ju mohli mať vyššiu. *Zločin a trest* určil asi rovnako dobre/zle. *Justice* a *lie* by však mali byť vyššie v zozname a *marriage* a *family relationship* zase nižšie.

Anthony Burgess - A Clockwork Orange

Predicted:

('violence', 0.7444061040878296)  
('paranoia', 0.5723904967308044)  
('marriage', 0.5109282732009888)  
('suicide', 0.49975669384002686)  
('friendship', 0.3422143757343292)  
('beating', 0.29962360858917236)  
('murder', 0.2814376652240753)  
('family relationships', 0.23992189764976501)  
('love', 0.1898125559091568)  
('flashback', 0.16344062983989716)  
('jealousy', 0.14349576830863953)  
('fight', 0.14211896061897278)  
('vandalism', 0.11525020748376846)  
('infidelity', 0.1048455759882927)  
('trauma', 0.10236535221338272)

With threshold 0.5:

['marriage', 'violence', 'paranoia']

Fyodor Dostoyevsky - Crime and Punishment

Predicted:

('murder', 0.9475010633468628)  
('suicide', 0.9089058041572571)  
('family relationships', 0.8105157017707825)  
('guilt', 0.5827731490135193)  
('marriage', 0.5345286726951599)  
('death', 0.5033612251281738)  
('revenge', 0.3263147175312042)  
('flashback', 0.154094398021698)  
('lie', 0.13657838106155396)  
('love', 0.12998023629188538)  
('jealousy', 0.11375859379768372)  
('investigation', 0.11245767027139664)  
('divorce', 0.10266073048114777)  
('justice', 0.10146995633840561)  
('violence', 0.0808754712343216)

With threshold 0.5:

['family relationships', 'marriage', 'death', 'murder', 'suicide', 'guilt']

Correlation:

Cosine similarity: 0.5557419667886223  
Spearman correlation: 0.7842450072205982  
Pearson correlation: 0.47330912495780425  
=====  
Average : 0.6044320329890083

Obr. 6.6: Vypočítané metriky podobnosti pre pravdepodobnosti knihy *Zločin a trest* a *Mechanický pomaranč*.

Napriek trocha nepresným predpokladom však model určil, podľa mňa správne, že knihy si sú trocha podobné. Ústrednými témami oboch diel je vražda a spravodlivosť, preto si myslím, že sú diela čiastočne podobné a rovnako veľmi podobné kľúčové dejové línie je aj vidieť medzi oboma dielami.

## Príklady z dátovej sady z Goodreads

V stiahnutej dátovej sade informácií o knihe z Goodreads sa nachádza popis knihy. Aj tento popis môže byť použitý pre určenie deja. V tomto prípade sa jedná o knihu *Martan*, s nasledujúcim popisom:

*” Six days ago, astronaut Mark Watney became one of the first people to walk on Mars. Now, he’s sure he’ll be the first person to die there. After a dust storm nearly kills him and forces his crew to evacuate while thinking him dead, Mark finds himself stranded and completely alone with no way to even signal Earth that he’s alive—and even if he could get word out, his supplies would be gone long before a rescue could arrive. Chances are, though, he won’t have time to starve to death. The damaged machinery, unforgiving environment, or plain-old “human error” are much more likely to kill him first. But Mark isn’t ready to give up yet. Drawing on his ingenuity, his engineering skills — and a relentless, dogged refusal to quit — he steadfastly confronts one seemingly insurmountable obstacle after the next. Will his resourcefulness be enough to overcome the impossible odds against him? ”*

Na prvý pohľad sa nezdá, že je v takomto texte veľa informácií, no klasifikátor určil že sa v ňom nachádzajú nasledovné deje: ['death', 'fear', 'courage', 'race against time', 'near death experience', 'rescue']. Na obrázku 6.7 sa nachádzajú ďalšie predpovede modelu. Všetci čo čítali knihu, alebo videli film potvrdia, že určené dejové línie sú všetky správne, asi by ale ešte mohlo byť zaradené *escape* medzi relevantné deje. Smrť je zaradená medzi správne, no ťažko určiť či je správnou, keďže v diele nikto nezomrel, ale na hlavného hrdinu číha na každom kroku.

```
Andy Weir - The Martian
Predicted:
('rescue', 0.9275297522544861)
('race against time', 0.8681074380874634)
('near death experience', 0.822539746761322)
('fear', 0.7161270976066589)
('courage', 0.5933665633201599)
('death', 0.5842492580413818)
('self sacrifice', 0.48343586921691895)
('flashback', 0.3594614565372467)
('escape', 0.3569062650203705)
('moral dilemma', 0.3356294333934784)
('sacrifice', 0.1237850934267044)
('love', 0.06318236887454987)
('pregnancy', 0.05537949874997139)
('paranoia', 0.05240940675139427)
('friendship', 0.041137661784887314)
```

Obr. 6.7: Pravdepodobnosti 15 kľúčových dejových línií s najväčšou pravdepodobnosťou pre popis knihy *Martan*.

## 6.4 Analýza výsledkov a chyby

Ako celok model funguje relatívne dobre. Dokáže odhadnúť väčšinu kľúčových dejových línií správne, a keď nie, sú radené medzi vysoké priečky zoznamu pravdepodobností. Občas zaradí medzi správne dejovú líniu, ktorá tam „nemá čo robiť“ a niekedy naopak, dejová línia, ktorá sa pre zhrnutie „hodí“, má príliš nízku pravdepodobnosť. Metriky vyhodnocovania udávajú veľmi malé úspešnosti, ale nedá sa povedať, že by každá druhá dejová línia bola zle zaradená. Dovolím si tvrdiť, že model pracuje lepšie, ako ho metriky hodnotia.

Model robí najčastejšie chyby pravdepodobne v dejových líniách, ktoré v tréningu nevidel tak často. Ako vidieť v príklade, *Rómeo a Júlia* je typická dejová línia *forbidden love*, no model jej neudelil dostatočnú pravdepodobnosť. Väčšina ostatných chýb podľa mňa súvisí s dátami. Veľmi záleží aké zhrnutie je modelu dané. Pokiaľ je príliš krátke, alebo neobsahuje dostatok informácií, tak logicky nedokáže model nič relevantné určiť. Rovnako dôležité sú aj označené dejové línie. Dané diela totiž značkujú ľudia. Ak sa zrovna dejová línia nehodí pre daný film, a predsa ju tam užívatelia zaradia, tak sa s tým nedá nič robiť. Aj spôsob akým sú kľúčové dejové línie odfiltrované z kľúčových slov môže mať svoje chyby. Viac krát sa vyskytujú kľúčové dejové línie *marriage*, *family relationship* alebo, *death*. Stačí aby v diele bola spomenutá manželka alebo svadba, pričom to vôbec nemusí byť kľúčové pre dej, ale užívatelia to tam správne zaradia, lebo to istú spojitosť má. Lenže keď sa extrahujú kľúčové dejové línie, je týmto kľúčovým slovám zaradená väčšia váha a to spôsobuje, že vo viacerých zhrnutiach je vidieť takéto anomálie.

Niekedy akoby modelu chýbala nejaká kľúčová dejová línia, alebo by ich mohlo byť viac, aby pokryli čo najväčší záber všetkých dejov. Z pôvodného návrhu bola urobená iba jedna revízia, ktorá skresala línie, ktoré neboli dobre určené, alebo ich bolo málo. Určite by sa dalo pridať a zlepšiť výber kľúčových dejových línií, to by výsledky modelu určite zdvihlo. Možno by však muselo byť nutné ručne značkovať zhrnutia, čo je práca veľmi zdĺhavá. Ďalej by model určite vylepšilo aj pretriedenie zhrnutí, aby v dátovej sade boli iba také, ktoré poskytujú dostatok informácií, ale na druhej strane nie sú príliš dlhé aby nevznikalo veľa chýb skracovaním týchto zhrnutí. Dáta sú rozhodujúce pri spracovaní prirodzeného jazyk, takže oba tieto námety na skvalitnenie dát by výrazne pomohli zlepšiť kvalitu modelu.

Isté obohatenie modelu by prinieslo aj dlhšie tréňovanie, no toto zlepšenie by bolo minimálne. Lepšie výsledky by mohli priniesť kvalitnejšie nastavené hyper-parametre učenia. Toto by však bol proces veľmi zdĺhavý, lebo by sa museli vykonať desiatky tréňovaní, pričom jedna adaptácia môže trvať aj niekoľko dní. Po každom by sa muselo skúšať lepšie nastaviť parametre pre danú dátovú sadu, a zároveň ich porovnávať s predošlými behmi. Oba tieto námety by mohli, ale len veľmi minimálne, zvýšiť kvalitu klasifikátora.

## 6.5 Možné vylepšenia a využitie

Čo sa týka zlepšenie systému ako celku, okrem vyššie spomínaného, by sa dalo tomuto všetkému prirobiť grafické užívateľské rozhranie, kde by sa dalo vložiť zhrnutie a systém by preňho vyhodnotil kľúčové dejové línie. Interakcia s užívateľom by sa dala skvalitniť aj prirobením porovnania s viacerými dejmi. Teda najskôr by celá dátová sada prešla modelom a boli by pre všetky deje vyhodnotené kľúčové dejové línie a boli by uložené. Potom by systém vedel pri zadaní zhrnutia predpovedať na aké dielo sa zhrnutie podobá.

Klasifikácia kľúčových dejových línií by sa dala využiť v sociálnych sieťach o knihách, filmoch a pod., kde by slúžila ako doplnok ku klasickým žánrom. Takisto by sa tento systém dal použiť pre navrhovanie podobných diel. Mohol by sa tým vytvoriť systém, ktorý dokáže filtrovať *spoilery* z užívateľských recenzií. Rozšírenie tejto práce by sa dali použiť na hľadanie prvotných príbehov alebo určenie konkrétneho príbehu len na základe malého množstva informácií o ňom.

# Kapitola 7

## Záver

Cieľom tejto práce bolo vytvoriť systém schopný určovať kľúčové dejové línie na základe opisu diela vo forme zhrnutia deja a určovať podobnosť diel na základe kľúčových dejových línií. Pre potreby analýzy a riešenia problematiky bolo nutné zozbierať dáta týkajúce sa informácií o knihe, zhrnutia deja a užívateľských recenzií. Všetky tieto ciele boli splnené.

V práci je načrtnutý teoretický základ potrebný pre zostavenie takéhoto systému, kde je vysvetlené, čo sú to kľúčové dejové línie, aj základná teória týkajúca sa klasifikácie a spracovania prirodzeného jazyka. Následne bolo objasnené s akými dátami sa v práci manipuluje, ďalej bol tento systém navrhnutý a implementovaný. Fungovanie systému je vysvetlené a vyhodnotené na ukážkach konkrétnych príkladov.

Pomocou extrahovania dát z internetu boli zozbierané 2 dátové sady zo sociálnej siete Goodreads. Jedna s informáciami o knihách obsahujúca takmer 900 tisíc záznamov a druhá pozostávajúca z viac než 23 miliónov užívateľských recenzií o knihách na tejto sociálnej sieti. Rovnako bola použitá aj predom stiahnutá dátová sada zo sociálnej siete IMDb, ktorá slúžila na tréningovanie klasifikátora určujúci kľúčové dejové línie a z ktorej boli tieto kľúčové dejové línie aj prebrané. Sumarizované obsahy kníh predstavuje dátová sada Booksummaries, ktorá bola využitá na testovanie klasifikátora.

Systém dokáže na základe stručnej voľnej formulácie deja určiť aké dejové línie sa v ňom nachádzajú. Celkovo model rozlišuje 88 kľúčových dejových línií, pričom k danému opisu môže prislúchať viac než len jedna dejová línia. Systém dosahuje priemernú presnosť na úrovni 45 % a F1-skóre na úrovni 31 %. Výsledný model však dokáže relatívne dobre určovať relevantné kľúčové dejové línie. Taktiež je možné určovať podobnosť 2 sumarizovaných dejov na základe pravdepodobností príslušnosti ku všetkým kľúčovým dejovým líniám.

Počas práce som sa oboznámil s extrakciou dát z webu, ozrejnil som si princípy analýzy a klasifikácie textu, ako aj praktickú skúsenosť s tréningovaním neurónovej siete a vyskúšal som si pracovať s novinkami zo sveta spracovania prirodzeného jazyka.

Úspešnosť klasifikátora dejových línií je možné naďalej zlepšovať pomocou skvalitnenia dátovej sady, na ktorej bol tréningovaný. Po doplnení porovnávaného deja ku všetkým dejom dátovej sady, by bolo možné určovať podobnosti k jednotlivým dejovým opisom. Rozšírenia tejto práce by mohli pomôcť. Pre zvýšenie komfortu práce so systémom a interakcie s užívateľom je možné doplniť grafické užívateľské rozhranie.

# Literatúra

- [1] ALLARD, M. What is a Transformer? *Medium* [online]. Január 2019 [cit. 2022-04-13]. Dostupné z: <https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04>.
- [2] BAMMAN, D. *CMU Book Summary Dataset* [online]. 2013. Dostupné z: <https://www.cs.cmu.edu/~dbamman/booksummaries.html>.
- [3] BELTAGY, I., PETERS, M. E. a COHAN, A. Longformer: The Long-Document Transformer. *ArXiv:2004.05150*. 1. vyd. 2020, č. 1. DOI: 10.48550/ARXIV.2004.05150. Dostupné z: <https://arxiv.org/abs/2004.05150>.
- [4] COOK, W. W. *Plotto: A New Method of Plot Suggestion for Writers of Creative Fiction*. 1. vyd. Brooklyn, New York: Tin House Books, 1928. ISBN 978-1935639183.
- [5] DASGUPTA, S. R. *Goodreads-books* [online]. 2019. Dostupné z: <https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks?resource=download>.
- [6] DEVLIN, J., CHANG, M.-W., LEE, K. a TOUTANOVA, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv, 2018. DOI: 10.48550/ARXIV.1810.04805. Dostupné z: <http://arxiv.org/abs/1810.04805>.
- [7] DONGEN, S. a ENRIGHT, A. Metric distances derived from cosine similarity and Pearson and Spearman correlations. *ArXiv:1208.3145*. 1. vyd. August 2012, č. 1. DOI: 10.48550/ARXIV.1208.3145. Dostupné z: <https://arxiv.org/abs/1208.3145>.
- [8] HERRERA, F. et al. *Multilabel Classification*. 1. vyd. Švajčairsko: Springer International Publishing, 2016. ISBN 978-3319411101.
- [9] HOREV, R. BERT Explained: State of the art language model for NLP. *Medium* [online]. Október 2018 [cit. 2022-04-13]. Dostupné z: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>.
- [10] KAYLEEB00821. *English: Terms used in Plot* [online]. December 2015 [cit. 2022-30-03]. Dostupné z: [https://commons.wikimedia.org/wiki/File:Intro-Rising\\_Action-Climax-Falling\\_Action-Resolution.png](https://commons.wikimedia.org/wiki/File:Intro-Rising_Action-Climax-Falling_Action-Resolution.png).
- [11] LI, Q. et al. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.* 1. vyd. USA: Association for Computing Machinery. Apríl 2021, zv. 37, č. 4. ISSN 21576912.
- [12] LIDDY, E. D. Natural Language Processing. In: MCDONALD, J. D. a LEVINE CLARK., M., ed. *Encyclopedia of Library and Information Sciences*. 2. vyd. NY. Marcel Decker, Inc., 2001. ISBN 978-0824720797.

- [13] LIU, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*. 1. vyd. 2019, abs/1907.11692, č. 1. Dostupné z: <http://arxiv.org/abs/1907.11692>.
- [14] LOZANO, L. C. a PLANELLAS, S. C. *Best Books Ever Dataset* [online]. 1.0.0. Zenodo, november 2020. DOI: 10.5281/zenodo.4265096. Dostupné z: <https://doi.org/10.5281/zenodo.4265096>.
- [15] POLTI, G. *The Thirty Six Dramatic Situations*. 1. vyd. 1921. ISBN 978-0871161093.
- [16] ROBERTS, E. V. a JACOBS, H. E. *Fiction: An Introduction to Reading and Writing*. 1. vyd. Englewood Cliffs: Prentice-Hall, Inc., 1987. ISBN 978-0133192605.
- [17] ROSARIA, L. I. *A study of plot, character, and setting to convey the theme as seen in hemingway's the garden of eden* [online]. Yogyakarta, Indonézia, 2004. [cit. 2022-30-03]. Bakalárska práca. Sanata Dharma University. Dostupné z: [https://repository.usd.ac.id/28275/2/994214147\\_Full1%5b1%5d.pdf](https://repository.usd.ac.id/28275/2/994214147_Full1%5b1%5d.pdf).
- [18] SIRISURIYA, D. S. A comparative study on web scraping. *8th International Research Conference, KDU*. 1. vyd. 2015, č. 1, s. 135–140.
- [19] SUN, C., QIU, X., XU, Y. a HUANG, X. *How to Fine-Tune BERT for Text Classification?* arXiv, 2019. DOI: 10.48550/ARXIV.1905.05583. Dostupné z: <https://arxiv.org/abs/1905.05583>.
- [20] TANG, T., TANG, X. a YUAN, T. Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text. *IEEE Access*. 1. vyd. 2020, zv. 8, č. 1, s. 193248–193256. DOI: 10.1109/ACCESS.2020.3030468.
- [21] TIANYANG, L. a OTHERS. A Survey of Transformers. *CoRR*. 1. vyd. 2021, abs/2106.04554, č. 1. Dostupné z: <https://arxiv.org/abs/2106.04554>.
- [22] TOBIAS, R. B. *20 Master Plots: And How to Build Them*. 1. vyd. Writer's Digest Books, september 1993. ISBN 978-1582972398.
- [23] VASWANI, A. a OTHERS. Attention is All you Need. In: GUYON, I. a OTHERS, ed. *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2017, sv. 30, s. 6000–6010. ISBN 978-1510860964. Dostupné z: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [24] WAN, M. et al. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In: KORHONEN, A. et al., ed. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 2019, s. 2605–2610. DOI: 10.18653/v1/p19-1248. ISBN 978-1950737482. Dostupné z: <https://doi.org/10.18653/v1/p19-1248>.
- [25] WAN, M. a MCAULEY, J. J. Item recommendation on monotonic behavior chains. In: PERA, S. et al., ed. *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. ACM, 2018, s. 86–94. DOI: 10.1145/3240323.3240369. ISBN 978-1450359016. Dostupné z: <https://doi.org/10.1145/3240323.3240369>.



- [26] ZHAO, B. Web Scraping. In: SCHINTLER, L. A. a MCNEELY, C. L., ed. *Encyclopedia of Big Data*. Švajčairsko: Springer International Publishing, 2017, s. 1–3. DOI: 10.1007/978-3-319-32001-4\_483-1. ISBN 978-3319320014.

## Príloha A

# Obsah priloženého pamäťového média

Na pamäťovom médiu sa nachádza:

- Zložku `thesis` so zdrojovými súbormi a ostatnými dátami pre vytvorenie technickej správy.
- Zložku `data`, ktorá obsahuje potrebné skripty pre prípravu dátových súbôrov:
  - Zložku `goodreads` obsahujúca skripty pre extrakciu dát z Goodreads: `download.py` pre sťahovanie HTML súborov, `extract_book_info.py` a `extract_fuzzy_book_info.py` pre extrahovanie kľúčových údajov a vytvorenie dátovej sady.
  - Zložku `processed_dataset` – spracovaná dátová sada z IMDb, aby sa ňou dal model trénovať.
  - `extract_keywords_imdb.py` – súbor extrahujúci kľúčové dáta z HTML súborov potrebné pre vytvorenie dátovej sady.
  - `prepare_imdb_data.py` – súbor, ktorý pripravujú dátovú sadu na tréning.
  - `plot_keywords.txt` – súbor, obsahujúci zoznam všetkých kľúčových dejových línií.
- Zložku `classifier`, ktorá obsahuje potrebné skripty pre tréning a prácu s klasifikátorom:
  - Zložku `best_trained_roberta`, ktorá obsahuje najlepšie natrénovaný model klasifikátora.
  - `trainer_imdb.py` – súbor, ktorý zabezpečuje tréning na dátovej sade z IMDb.
  - `load_model_imdb.py` – súbor, ktorý načíta model zo zložky `best_trained_roberta` a vyhodnotí ho na validačnej testovacej sade.
  - `compare_plots.py` – súbor, ktorý porovnáva dva zhrnutia deja pomocou metrických podobnosti.
  - `predict_plot.py` – súbor, ktorý určí kľúčové dejové línie a aj predpovede pre zadané zhrnutie deja.

- `thesis.pdf` – súbor, obsahujú technickú správu vo formáte PDF určenú na čítanie.
- `thesis_print.pdf` – súbor, obsahujú technickú správu vo formáte PDF určenú na tlač.
- `plagat.pdf` – súbor, obsahujúci plagát popisujúci túto prácu.
- `requirements.txt` – súbor, obsahujúci informáciu o potrebných knižniciach pre spustenie skriptov.
- `README` – súbor, obsahujúci obsah pamäťového média a manuál.

## Príloha B

# Zoznam kľúčových dejových línií

- Doppelganger
- Memory Loss
- Manipulation
- Near Death Experience
- Mind Control
- Rise To Power
- Rise And Fall
- Saving The World
- Loss Of Father
- Quest
- Pursuit
- Rescue
- Revenge
- Riddle
- Rivalry
- Underdog
- Temptation
- Guilt
- Secret
- Justice
- Prejudice
- Racism
- Abuse Of Power
- Power Struggle
- Time Travel
- Supernatural Power
- Time Loop
- Alternate Reality
- Good Versus Evil
- One Against Many
- One Man Army
- Alternate History
- Romantic Rivalry
- Race Against Time

- Detention
- Shot To Death
- Gunfight
- Flashback
- Escape
- Beating
- Betrayal
- Drunkenness
- Fear
- Drinking
- Deception
- Competition
- Dream
- Lie
- Vandalism
- Conspiracy
- Corruption
- Courage
- Ambush
- Paranoia
- Faith
- Trauma
- Self Sacrifice
- Childhood Memory
- Moral Dilemma
- Erased Memory
- Brainwashing
- Friendship
- Family Relationships
- Love
- Marriage
- Pregnancy
- Infidelity
- Jealousy
- Divorce
- Love Triangle
- Marriage Proposal
- Unrequited Love
- Love At First Sight
- Death
- Murder
- Violence
- Fight
- Kidnapping
- Chase
- Suicide
- Battle
- Investigation
- Blackmail
- Metamorphosis
- Transformation
- Forbidden Love
- Sacrifice
- Discovery

## Príloha C

# Zhrnutia dejov

Zhrnutie deja knihy od Erich Maria Remarque - Na západe nič nové:

*”The book tells the story of Paul Bäumer, a German soldier who—urged on by his school teacher—joins the German army shortly after the start of World War I. Bäumer arrives at the Western Front with his friends and schoolmates (Tjaden, Müller, Kropp and a number of other characters). There they meet Stanislaus Katczinsky, an older soldier, nicknamed Kat, who becomes Paul’s mentor. While fighting at the front, Bäumer and his comrades have to engage in frequent battles and endure the dangerous and often dirty conditions of warfare. At the very beginning of the book Erich Maria Remarque says ‘This book is to be neither an accusation nor a confession, and least of all an adventure, for death is not an adventure to those who stand face to face with it. It will try simply to tell of a generation of men who, even though they may have escaped shells, were destroyed by the war.’ The book does not focus on heroic stories of bravery, but rather gives a view of the conditions in which the soldiers find themselves. The monotony between battles, the constant threat of artillery fire and bombardments, the struggle to find food, the lack of training of young recruits (meaning lower chances of survival), and the overarching role of random chance in the lives and deaths of the soldiers are described in detail. The battles fought here have no names and seem to have little overall significance, except for the impending possibility of injury or death for Bäumer and his comrades. Only pitifully small pieces of land are gained, about the size of a football field, which are often lost again later. Remarque often refers to the living soldiers as old and dead, emotionally drained and shaken. ‘We are not youth any longer. We don’t want to take the world by storm. We are fleeing from ourselves, from our life. We were eighteen and had begun to love life and the world; and we had to shoot it to pieces.’ Paul’s visit on leave to his home highlights the cost of the war on his psyche. The town has not changed since he went off to war; however, he finds that he does ‘not belong here anymore, it is a foreign world.’ He feels disconnected from most of the townspeople. His father asks him ‘stupid and distressing’ questions about his war experiences, not understanding ‘that a man cannot talk of such things.’ An old schoolmaster lectures him about strategy and advancing to Paris, while insisting that Paul and his friends know only their ‘own little sector’ of the war but nothing of the big picture. Indeed, the only person he remains connected to is his dying mother, with whom he shares a tender, yet restrained relationship. The night before he is to return from leave, he stays up with her, exchanging small expressions of love and concern for each other. He thinks to himself, ‘Ah! Mother, Mother! How can it be that I must part from you? Here I sit and there you are lying; we have so much to say, and we shall never say it.’ In the end, he concludes that he ‘ought never to have come home*

on leave.' Paul feels glad to be reunited with his comrades. Soon after, he volunteers to go on a patrol and kills a man for the first time in hand-to-hand combat. He watches the man die, in pain for hours. He feels remorse and asks forgiveness from the man's corpse. He is devastated and later confesses to Kat and Albert, who try to comfort him and reassure him that it is only part of the war. They are then sent on what Paul calls a 'good job'. They must guard a village that is being shelled too heavily. The men enjoy themselves but while evacuating the villagers, Paul and Albert are wounded. They recuperate in a Catholic hospital and Paul returns to active duty. By now, the war is nearing its end and the German Army is retreating. In despair, Paul watches as his friends fall one by one. It is the death of Kat that eventually makes Paul careless about living. In the final chapter, he comments that peace is coming soon, but he does not see the future as bright and shining with hope. Paul feels that he has no aims left in life and that their generation will be different and misunderstood. When he finally dies at the end of the novel, the situation report from the frontline states, 'All is Quiet on the Western Front,' symbolizing the cheapness of human life in war. "

Zhrnutie deja hry od William Shakespeare - Rómeo and Júlia:

"The play, set in Verona, begins with a street brawl between Montague and Capulet supporters who are sworn enemies. The Prince of Verona intervenes and declares that further breach of the peace will be punishable by death. Later, Count Paris talks to Capulet about marrying his daughter, but Capulet asks Paris to wait another two years (then he later orders Juliet to marry Paris) and invites him to attend a planned Capulet ball. Lady Capulet and Juliet's nurse try to persuade Juliet to accept Paris's courtship. Meanwhile, Benvolio talks with his cousin Romeo, Montague's son, about Romeo's recent depression. Benvolio discovers that it stems from unrequited infatuation for a girl named Rosaline, one of Capulet's nieces. Persuaded by Benvolio and Mercutio, Romeo attends the ball at the Capulet house in hopes of meeting Rosaline. However, Romeo instead meets and falls in love with Juliet. After the ball, in what is now called the 'balcony scene', Romeo sneaks into the Capulet orchard and overhears Juliet at her window vowing her love to him in spite of her family's hatred of the Montagues. Romeo makes himself known to her and they agree to be married. With the help of Friar Laurence, who hopes to reconcile the two families through their children's union, they are secretly married the next day. Juliet's cousin Tybalt, incensed that Romeo had sneaked into the Capulet ball, challenges him to a duel. Romeo, now considering Tybalt his kinsman, refuses to fight. Mercutio is offended by Tybalt's insolence, as well as Romeo's 'vile submission,' and accepts the duel on Romeo's behalf. Mercutio is fatally wounded when Romeo attempts to break up the fight. Grief-stricken and wracked with guilt, Romeo confronts and slays Tybalt. Montague argues that Romeo has justly executed Tybalt for the murder of Mercutio. The Prince, now having lost a kinsman in the warring families' feud, exiles Romeo from Verona, with threat of execution upon return. Romeo secretly spends the night in Juliet's chamber, where they consummate their marriage. Capulet, misinterpreting Juliet's grief, agrees to marry her to Count Paris and threatens to disown her when she refuses to become Paris's 'joyful bride.' When she then pleads for the marriage to be delayed, her mother rejects her. Juliet visits Friar Laurence for help, and he offers her a drug that will put her into a deathlike coma for 'two and forty hours.' The Friar promises to send a messenger to inform Romeo of the plan, so that he can rejoin her when she awakens. On the night before the wedding, she takes the drug and, when discovered apparently dead, she is laid in the family crypt. The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken,

*Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two 'star-cross'd lovers'. The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: 'For never was a story of more woe Than this of Juliet and her Romeo.' "*

Zhrnutie deja knihy od Anthony Burgess - Mechanický pomaranč:

*"Alex, a teenager living in near-future England, leads his gang on nightly orgies of opportunistic, random 'ultra-violence'. Alex's friends ('droogs' in the novel's Anglo-Russian slang, Nadsat) are: Dim, a slow-witted bruiser who is the gang's muscle; Georgie, an ambitious second-in-command; and Pete, who mostly plays along as the droogs indulge their taste for ultra-violence. Characterized as a sociopath and a hardened juvenile delinquent, Alex is also intelligent and quick-witted, with sophisticated taste in music, being particularly fond of Beethoven, or 'Lovely Ludwig Van'. The novel begins with the droogs sitting in their favorite hangout (the Korova Milkbar), drinking milk-drug cocktails, called 'milk-plus', to hype themselves for the night's mayhem. They assault a scholar walking home from the public library, rob a store leaving the owner and his wife bloodied and unconscious, stomp a panhandling derelict, then scuffle with a rival gang. Joyriding through the countryside in a stolen car, they break into an isolated cottage and maul the young couple living there, beating the husband and raping his wife. In a metafictional touch, the husband is a writer working on a manuscript called 'A Clockwork Orange', and Alex contemptuously reads out a paragraph that states the novel's main theme before shredding the manuscript. Back at the milk bar, Alex punishes Dim for some crude behaviour, and strains within the gang become apparent. At home in his dreary flat, Alex plays classical music at top volume while fantasizing of even more orgiastic violence. Alex skips school the next day. Following an unexpected visit from P.R. Deltoid, his 'post-corrective advisor', Alex meets a pair of ten-year-old girls and takes them back to his parents' flat, where he administers hard drugs and then rapes them. That evening, Alex finds his droogs in a mutinous mood. Georgie challenges Alex for leadership of the gang, demanding that they pull a 'man-sized' job. Alex quells the rebellion by slashing Dim's hand and fighting with Georgie, then in a show of generosity takes them to a bar, where Alex insists on following through on Georgie's idea to burgle the home of a wealthy old woman. The break-in starts as farce and ends in tragic pathos, as Alex's attack kills the elderly woman. His escape is blocked by Dim, who attacks Alex, leaving him incapacitated on the front step as the police arrive. Sentenced to prison for murder, Alex gets a job at the Wing chapel playing religious music on the stereo before and after services as well as during the singing of hymns. The prison chaplain mistakes Alex's Bible studies for stirrings of faith (Alex is actually reading Scripture for the violent passages). After Alex's fellow cellmates blame him for beating a troublesome cellmate to death, he agrees to undergo an experimental behaviour-modification treatment called the Ludovico Technique. The technique is a form of aversion therapy in which Alex receives an injection that makes him feel sick while watching graphically violent films, eventually conditioning him to suffer crippling bouts of nausea at the mere thought of violence. As an unintended consequence, the soundtrack to one of the films—Beethoven's Fifth Symphony—renders Alex unable to listen to his beloved classical music. The effectiveness of the technique is demonstrated to a group of VIPs, who watch*



as Alex collapses before a walloping bully, and abases himself before a scantily-clad young woman whose presence has aroused his predatory sexual inclinations. Though the prison chaplain accuses the state of stripping Alex of free will, the government officials on the scene are pleased with the results and Alex is released into society. Since his parents are now renting his room to a lodger, Alex wanders the streets and enters a public library where he hopes to learn a painless way to commit suicide. There, he accidentally encounters the old scholar he assaulted earlier in the book, who, keen on revenge, beats Alex with the help of his friends. The policemen who come to Alex's rescue turn out to be none other than Dim and former gang rival Billyboy. The two policemen take Alex outside of town and beat him up. Dazed and bloodied, Alex collapses at the door of an isolated cottage, realizing too late that it is the house he and his droogs invaded in the first half of the story. Because the gang wore masks during the assault, the writer does not recognize Alex. The writer, whose name is revealed as F. Alexander, shelters Alex and questions him about the conditioning. During this sequence, it is revealed that Mrs. Alexander died from the injuries inflicted during the gang-rape, and her husband has decided to continue living 'where her fragrant memory persists' despite the horrid memories. Alexander, a critic of the government, hopes to use Alex as a symbol of state brutality and thereby prevent the incumbent government from being re-elected. Eventually, he begins to realize Alex's role in the happenings of the night two years ago. One of Alexander's radical associates manages to extract a confession from Alex after removing him from F. Alexander's home and then locks him in a flatblock near his former home. Alex is then subjected to a relentless barrage of classical music, prompting him to attempt suicide by leaping from a high window. Alex wakes up in hospital, where he is courted by government officials anxious to counter the bad publicity created by his suicide attempt. With Alexander safely packed off to a mental institution, Alex is offered a well-paying job if he agrees to side with the government. As photographers snap pictures, Alex daydreams of orgiastic violence and realizes the Ludovico conditioning has been reversed: 'I was cured all right'. In the final chapter, Alex has a new trio of droogs, but he finds he is beginning to outgrow his taste for violence. A chance encounter with Pete, now married and settled down, inspires Alex to seek a wife and family of his own. He contemplates the likelihood of his future son being a delinquent as he was, a prospect Alex views fatalistically. "

Zhrnutie deja knihy od Fyodor Dostoyevsky - Zločin a trest:

"Raskolnikov, a conflicted former student, lives in a tiny, rented room in Saint Petersburg. He refuses all help, even from his friend Razumikhin, and devises a plan to murder and to rob an unpleasant elderly pawn-broker and money-lender, Alyona Ivanovna. His motivation comes from the overwhelming sense that he is predetermined to kill the old woman by some power outside of himself. While still considering the plan, Raskolnikov makes the acquaintance of Semyon Zakharovich Marmeladov, a drunkard who recently squandered his family's little wealth. He also receives a letter from his sister and mother, speaking of their coming visit to Saint Petersburg, and his sister's sudden marriage plans which they plan on discussing upon their arrival. After much deliberation, Raskolnikov sneaks into Alyona Ivanovna's apartment where he murders her with an axe. He also kills her half-sister, Lizaveta, who happens to stumble upon the scene of the crime. Shaken by his actions, Raskolnikov manages to only steal a handful of items and a small purse, leaving much of the pawn-broker's wealth untouched. Raskolnikov then flees and, due to a series of coincidences, manages to leave unseen and undetected. After the bungled murder, Raskolnikov falls into a feverish state and begins to worry obsessively over the murder. He hides the stolen items and purse

under a rock, and tries desperately to clean his clothing of any blood or evidence. He falls into a fever later that day, though not before calling briefly on his old friend Razumikhin. As the fever comes and goes in the following days, Raskolnikov behaves as though he wishes to betray himself. He shows strange reactions to whoever mentions the murder of the pawnbroker, which is now known about and talked of in the city. In his delirium, Raskolnikov wanders Saint Petersburg, drawing more and more attention to himself and his relation to the crime. In one of his walks through the city, he sees Marmeladov, who has been struck mortally by a carriage in the streets. Rushing to help him, Raskolnikov gives the remainder of his money to the man's family, which includes his teenage daughter, Sonya, who has been forced to become a prostitute to support her family. In the meantime, Raskolnikov's mother, Pulkheria Alexandrovna, and his sister, Avdotya Romanovna (or Dounia) have arrived in the city. Avdotya had been working as a governess for the Svidrigailov family until this point, but was forced out of the position by the head of the family, Arkady Ivanovich Svidrigailov. Svidrigailov, a married man, was attracted to Avdotya's physical beauty and her feminine qualities, and offered her riches and elopement. Avdotya, having none of this, fled the family and lost her source of income, only to meet Pyotr Petrovich Luzhin, a man of modest income and rank. Luzhin proposes to marry Avdotya, thereby securing her and her mother's financial safety, provided she accept him quickly and without question. It is for these very reasons that the two of them come to Saint Petersburg, both to meet Luzhin there and to attain Raskolnikov's approval. Luzhin, however, calls on Raskolnikov while he is in a delirious state and presents himself as a foolish, self-righteous and presuming man. Raskolnikov dismisses him immediately as a potential husband for his sister, and realizes that she only accepted him to help her family. As the novel progresses, Raskolnikov is introduced to the detective Porfiry, who begins to suspect him for the murder purely on psychological grounds. At the same time, a chaste relationship develops between Raskolnikov and Sonya. Sonya, though a prostitute, is full of Christian virtue and is only driven into the profession by her family's poverty. Meanwhile, Razumikhin and Raskolnikov manage to keep Avdotya from continuing her relationship with Luzhin, whose true character is exposed to be conniving and base. At this point, Svidrigailov appears on the scene, having come from the province to Petersburg, almost solely to seek out Avdotya. He reveals that his wife is dead, and that he is willing to pay Avdotya a vast sum of money in exchange for nothing. She, upon hearing the news, refuses flat out, suspecting him of treachery. As Raskolnikov and Porfiry continue to meet, Raskolnikov's motives for the crime become exposed. Porfiry becomes increasingly certain of the man's guilt, but has no concrete evidence or witnesses with which to back up this suspicion. Furthermore, another man admits to committing the crime under questioning and arrest. However, Raskolnikov's nerves continue to wear thin, and he is constantly struggling with the idea of confessing, though he knows that he can never be truly convicted. He turns to Sonya for support and confesses his crime to her. By coincidence, Svidrigailov has taken up residence in a room next to Sonya's and overhears the entire confession. When the two men meet face to face, Svidrigailov acknowledges this fact, and suggests that he may use it against him, should he need to. Svidrigailov also speaks of his own past, and Raskolnikov grows to suspect that the rumors about his having committed several murders are true. In a later conversation with Dounia, Svidrigailov denies that he had a hand in the death of his wife. Raskolnikov is at this point completely torn; he is urged by Sonya to confess, and Svidrigailov's testimony could potentially convict him. Furthermore, Porfiry confronts Raskolnikov with his suspicions and assures him confession would substantially lighten his sentence. Meantime, Svidrigailov attempts to seduce Avdotya, but when he realizes that she will never love him, he lets her go. He then spends a night in confusion and

*in the morning shoots himself. This same morning, Raskolnikov goes again to Sonya, who again urges him to confess and to clear his conscience. He makes his way to the police station, where he is met by the news of Svidrigailov's suicide. He hesitates a moment, thinking again that he might get away with a perfect crime, but is persuaded by Sonya to confess. The epilogue tells of how Raskolnikov is sentenced to penal servitude in Siberia, where Sonya follows him. Avdotya and Razumikhin marry and are left in a happy position by the end of the novel, while Pulkheria, Raskolnikov's mother, falls ill and dies, unable to cope with her son's situation. Raskolnikov himself struggles in Siberia. It is only after some time in prison that his redemption and moral regeneration begin under Sonya's loving influence. "*

# Príloha D

## Plagát

### Analyza dějových linií na základě shrnutí obsahu knih a uživatelských recenzí

#### Abstrakt

Cieľom tejto práce je vytvoriť systém pre analýzu a klasifikáciu kľúčových dejových línií zo zhrnutých dejových zápletiiek a užívateľských recenzií v anglickom jazyku. Zvolený problém je riešený pomocou techniky strojového učenia založenej na transformeroch. Vo vytvorenom riešení je implementované aj sťahovanie dát a bol vytvorený dataset užívateľských recenzií a informácií o knihách prevyšujúci 23 miliónov recenzií a takmer 900 tisíc informácií o knihách. Systém dokáže predikovať aké typy dejových zápletiiek sa v dátach nachádzajú.

#### Príklad: Andy Weir – The Martian

Popis knihy Marťan z Goodreads:

„Six days ago, astronaut Mark Watney became one of the first people to walk on Mars. Now, he’s sure he’ll be the first person to die there. After a dust storm nearly kills him and forces his crew to evacuate while thinking him dead, Mark finds himself stranded and completely alone with no way to even signal Earth that he’s alive—and even if he could get word out, his supplies would be gone long before a rescue could arrive. Chances are, though, he won’t have time to starve to death. The damaged machinery, unforgiving environment, or plain-old “human error” are much more likely to kill him first. But Mark isn’t ready to give up yet. Drawing on his ingenuity, his engineering skills — and a relentless, dogged refusal to quit — he steadfastly confronts one seemingly insurmountable obstacle after the next. Will his resourcefulness be enough to overcome the impossible odds against him?“

Výsledné dejové línie:

- Rescue 92 %
- Race against time 86 %
- Near death experience 86 %
- Fear 71 %
- Courage 59 %
- Death 58 %

---

- Self sacrifice 48 %
- Flashback 35 %
- Escape 35 %
- ...

Autor: Peter Růček

E-mail: xrucek00@stud.fit.vutbr.cz

Vedúci: Smrž Pavel, doc. RNDr., Ph.D.

Brno 2022



Obr. D.1: Náhľad vytvoreného plagátu.