Využití neuronových sítí pro práci s filmovými daty

Semestrální projekt ENC-NS

Bc. Zuzana Lysová, Bc. Pavel Bulín

Obsah

1 Úvod 4
2 Teoretický základ 5
2.1 Výběr neuronové sítě 5
3 Data 6
4 Vypracování 8
4.1 Předzpracování a analýza dat 8
4.2 Upráva dat pro první model 12
4.3 Upráva dat pro druhý model 12
4.4 Příprava prvního modelu 15
4.5 První model 16
4.6 Příprava druhého modelu 22
4.7 Druhý model 22
5 Závěr 28
6 Literatura 29

Seznam obrázků

Obr. 1: Původní data po odstranění nepotřebných atributů 6
Obr. 2: Rozdělení žánrů ze seznamu 8
Obr. 3: Distribuce žánrů 9
Obr. 4: Vybrané žánry 10
Obr. 5: Častá slova 10
Obr. 6: Word cloud pro žánr Adventure 11
Obr. 7: Word cloud pro žánr Adventure po odstranění častých slov 11
Obr. 8: Upravená data pro první model 12
Obr. 9: Upravená data pro druhý model 13
Obr. 10: Analýza délky popisů 14
Obr. 11: Tokenizace a sekvencování textu 15
Obr. 12: Normalizace sekvencí 16
Obr. 13: Převedení dat do objektů 16
Obr. 14: Nastavení parametrů prvního modelu 17
Obr. 15: Vrstvy prvního modelu 17
Obr. 16: Jednotlivé iterace prvního modelu 18
Obr. 17: Přesnost prvního modelu 18
Obr. 18: Vizualizace přesnosti a ztráty prvního modelu 19
Obr. 19: Matice záměn prvního modelu 19
Obr. 20: Křivka ROC pro první model 20
Obr. 21: Ukázka predikce prvního modelu 21
Obr. 22: Kategorické kódování 22
Obr. 23: Nastavení parametrů druhého modelu 23
Obr. 24: Vrstvy druhého modelu 23
Obr. 25: Jednotlivé iterace druhého modelu 24
Obr. 26: Přesnost druhého modelu 24
Obr. 27: Vizualizace přesnosti a ztráty druhého modelu 25
Obr. 28: Matice záměn druhého modelu 26
Obr. 29: Ukázka predikce druhého modelu 27

1 Úvod

V rámci tohoto projektu jsme se věnovali rozvíjejícímu se oboru aplikace umělé inteligence v kontextu filmového průmyslu. Konkrétně se naše pozornost zaměřila na využití neuronových sítí pro analýzu a kategorizaci filmů z hlediska jejich žánrů. Cílem tohoto výzkumu je prozkoumat, jak mohou moderní metody strojového učení přispět k hlubšímu porozumění a kategorizaci filmového obsahu.

Struktura naší práce je rozdělena do dvou hlavních částí. První z nich je vytvoření a trénování neuronové sítě s cílem identifikovat, zda je daný film zařaditelný do žánru hororu na základě jeho popisu. Tento žánr byl vybrán pro svoji jedinečnost a výraznou odlišnost od ostatních filmových žánrů, což poskytuje jasný rámec pro klasifikační model. Druhou částí projektu je provedení experimentu s modelem, který se snaží kategorizovat filmy do širší škály žánrů na základě jejich popisu. I když tento druhý model nepřinesl zcela uspokojivé výsledky, nabídl cenné poznatky o omezeních a výzvách, které jsou spojeny s klasifikací do více žánrů.

V úvodu našeho projektu jsme nejprve podrobili důkladné analýze dostupná data a provedli příslušné úpravy a předzpracování datové sady. Následně jsme se zaměřili na vývoj a trénování modelů, při čemž jsme využili techniky jako rekurentní neuronová sít (RNN) a její varianta LSTM (Long Short-Term Memory), která je zvláště vhodná pro zpracování sekvenčních dat, jako je text.

Cílem tohoto projektu není pouze demonstrace aplikace neuronových sítí na konkrétním problému, ale také poskytnutí komplexního pohledu na celý proces práce s daty - od předzpracování dat, přes výběr vhodného modelu, až po analýzu výsledků. Výsledky našeho projektu by mohly posloužit jako inspirace pro další výzkum v oblasti strojového učení a jeho aplikací v kreativních odvětvích, včetně filmového průmyslu.

2 Teoretický základ

Neuronová síť je základním stavebním kamenem v oblasti umělé inteligence a strojového učení. Jedná se o model, který je inspirován strukturou a funkcí lidského mozku. Neuronové sítě jsou složeny z uzlů, tzv. neuronů, které jsou propojeny a komunikují mezi sebou. Tyto sítě jsou schopny učit se a identifikovat složité vzory v datech, a to prostřednictvím procesu trénování, kde síť upravuje své interní váhy na základě poskytnutých dat. Můžeme rozlišovat 4 hlavní druhy neuronových sítí (Sarker, 2021).

Perceptron představuje nejjednodušší typ neuronových sítí, které jsou vhodné pro jednoduché klasifikační úlohy. Skládá se z jedné vrstvy neuronů, které přijímají vstupní data a generují výstup (Geng et al., 2020).

Konvoluční neuronové sítě (CNN) jsou zvláště účinný typ neuronových sítí pro analýzu vizuálních dat, jako jsou obrázky a videa. CNN využívají konvoluční vrstvy k extrakci vzorů a charakteristik z obrazových dat (Taye, 2023).

Rekurentní neuronové sítě (RNN) jsou vhodné pro sekvenční data, jako je text nebo časové řady. Tyto sítě mají schopnost "pamatovat si" předchozí informace a používat je pro aktuální zpracování dat (Orojo et al., 2023).

Sítě s dlouhodobou pamětí (LSTM) jsou varianta RNN, které jsou navrženy tak, aby lépe zvládaly problémy s dlouhodobými závislostmi v sekvenčních datech. LSTM sítě jsou schopny uchovávat informace po delší dobu, což je činí ideálními pro složitější úlohy zpracování sekvenčních dat (Qiao et al., 2023).

2.1 Výběr neuronové sítě

Pro naše účely jsme si vybrali rekurentní neuronovou síť (RNN) a její variantu LSTM. Tento výběr byl motivován potřebou efektivně zpracovávat a analyzovat sekvenční data, konkrétně textové popisy filmů. LSTM sítě jsou ideální pro tyto účely, protože umožňují zachytávat složité vzory v datech a uchovávat důležité informace po delší časové období. Tento aspekt je klíčový pro úspěšné rozpoznání žánrů filmů z jejich popisů, kde může být důležitý kontext rozprostřený přes celý text.

3 Data

Dataset pro tento výzkum byl získán z platformy Kaggle, která je známá jako centrum pro datové vědce poskytující rozsáhlé množství datových sad a také komunitní prostředí pro sdílení poznatků a kódu. Kaggle je hojně využíván pro akademické i komerční účely a je ceněn pro svoji širokou nabídku otevřených datových sad pokrývající různé oblasti a tematiky.

Náš vybraný dataset obsahuje 10 178 záznamů a skládá se z několika atributů, které poskytují komplexní pohled na každý film. Mezi klíčové atributy, které byly použity pro naše analýzy, patří names (názvy filmů), genre (žánry), a overview (textové popisy). Kromě těchto hlavních atributů dataset obsahuje i další informace jako datum vydání (date_x), hodnocení (score), seznam tvůrců (crew), původní název (orig_title), status vydání (status), původní jazyk (orig_lang), rozpočet (budget_x), tržby (revenue) a země původu (country). Tato data poskytují široký základ pro analýzu a modelování, umožňují studium korelací mezi různými proměnnými a poskytují bohaté textové popisy, které jsou klíčové pro účely zpracování přirozeného jazyka pomocí neuronových sítí.

	sid	names	genre	overview
0	0	Creed III	[Drama, Action]	After dominating the boxing world, Adonis Cree...
1	1	Avatar: The Way of Water	[Science Fiction, Adventure, Action]	Set more than a decade after the events of the...
2	2	The Super Mario Bros. Movie	[Animation, Adventure, Family, Fantasy, Comedy]	While working underground to fix a water main,...
3	3	Mummies	[Animation, Comedy, Family, Adventure, Fantasy]	Through a series of unfortunate events, three ...
4	4	Supercell	[Action]	Good-hearted teenager William always lived in ...
...	...	...	...	...
10173	10173	20th Century Women	[Drama]	In 1979 Santa Barbara, California, Dorothea Fi...
10174	10174	Delta Force 2: The Colombian Connection	[Action]	When DEA agents are taken captive by a ruthles...
10175	10175	The Russia House	[Drama, Thriller, Romance]	Barley Scott Blair, a Lisbon-based editor of R...
10176	10176	Darkman II: The Return of Durant	[Action, Adventure, Science Fiction, Thriller,...	Darkman and Durant return and they hate each o...
10177	10177	The Swan Princess: A Royal Wedding	[Animation, Family, Fantasy]	Princess Odette and Prince Derek are going to ...

10178 rows × 4 columns

Obr. 1: Původní data po odstranění nepotřebných atributů

Jak je vidět na přiloženém obrázku, který zobrazuje ukázku datasetu, data byla pečlivě vyčištěna a byly zachovány pouze relevantní atributy pro účel našeho výzkumu. Sloupec genre představuje žánry filmů, které jsou reprezentovány jako seznamy, umožňující snadnou manipulaci a analýzu. Sloupec overview obsahuje textové popisy, které poskytují

podrobný pohled na obsah a tématiku filmů, a jsou základem pro trénování našich modelů neuronových sítí.

Předzpracování dat zahrnovalo odstranění chybějících hodnot, normalizaci textových řetězců a v některých případech i transformaci dat pro zajištění konzistence a přesnosti následných analýz. Takovéto čištění a příprava dat jsou nezbytnými kroky pro zajištění, že modely neuronových sítí budou pracovat s kvalitními a relevantními informacemi.

4 Vypracování

4.1 Předzpracování a analýza dat

V této fázi projektu jsme se věnovali zásadnímu kroku v oblasti strojového učení – předzpracování dat. Vstupní data jsou často neúplná nebo nekonzistentní, což může vést k nepřesnostem při výsledcích modelů. Proto je klíčové data nejenom řádně vyčistit, ale také je připravit tak, aby odpovídala potřebám použitých algoritmů. Tento proces zahrnuje řadu kroků, včetně normalizace, transformace a výběru relevantních vlastností. Pro automatizaci a systematizaci procesu předzpracování jsme vyvinuli DataLoader, což je oddělená část v našem kódu, který nám umožňuje efektivně manipulovat s daty a připravovat je pro analýzu. Díky tomuto nástroji jsme mohli provést prvotní očištění dat, aby byla odstraněna jakákoliv nesrovnalost, která by mohla ovlivnit další práci s daty.

V rámci tohoto procesu bylo klíčové efektivně zpracovat a strukturovat žánry filmů. V původním datasetu byly žánry reprezentovány jako seznamy, což komplikovalo jejich analýzu a použití v neuronových sítích. Proto jsme se rozhodli implementovat metodu, která umožňuje rozdělit tyto seznamy a přeformátovat data tak, aby každý film byl reprezentován několika řádky odpovídajícími jednotlivým žánrům. Tento přístup nám umožnil rozšířit původní sloupec genre tak, aby každý žánr byl představen v samostatném řádku společně s odpovídajícím názvem filmu. Jak je patrné z přiloženého obrázku, každý film může nyní být spojen s více žánry, které jsou prezentovány v jedinečných řádcích. Tento krok zjednodušuje další analýzu a zpracování dat, jelikož umožňuje lépe pracovat s kategoriálními proměnnými a usnadňuje aplikaci metod strojového učení.

	names	genre
0	Creed III	Drama
0	Creed III	Action
1	Avatar: The Way of Water	Science Fiction
1	Avatar: The Way of Water	Adventure
1	Avatar: The Way of Water	Action
2	The Super Mario Bros. Movie	Animation
2	The Super Mario Bros. Movie	Adventure
2	The Super Mario Bros. Movie	Family
2	The Super Mario Bros. Movie	Fantasy
2	The Super Mario Bros. Movie	Comedy
3	Mummies	Animation
3	Mummies	Comedy
3	Mummies	Family
3	Mummies	Adventure
3	Mummies	Fantasy
4	Supercell	Action
5	Cocaine Bear	Thriller
5	Cocaine Bear	Comedy
5	Cocaine Bear	Crime
6	John Wick: Chapter 4	Action

Obr. 2: Rozdělení žánrů ze seznamu

Po rozdělení žánrů do jednotlivých řádků pro každý film se nám otevřely nové možnosti pro další analýzu. Byli jsme schopni provést podrobný průzkum distribuce žánrů v našem datasetu, což je klíčovým krokem pro pochopení skladby filmů a jejich charakteristik. Díky této metodě můžeme přesně určit, kolik filmů patří do každého žánru. Jak ilustruje přiložený obrázek, vytvořili jsme tabulku, ve které je každý žánr seřazený spolu s počtem filmů, které do něj spadají. Z této tabulky je patrné, že drama je nejčastějším žánrem v našem datasetu, následované komedií a akčními filmy. Tato data nám umožňují nejen lépe porozumět rozmanitosti filmů v datasetu, ale také identifikovat, které žánry jsou převládající a které jsou méně zastoupené.

	Genre	Count
0	Drama	3812
1	Comedy	2943
2	Action	2752
3	Thriller	2605
4	Adventure	1890
5	Romance	1576
6	Horror	1554
7	Animation	1468
8	Family	1407
9	Fantasy	1382
10	Crime	1272
11	Science Fiction	1261
12	Mystery	862
13	History	422
14	War	282
15	Music	277
16	Documentary	217
17	TV Movie	212
18	Western	131

Obr. 3: Distribuce žánrů

Po úvodní kvantitativní analýze distribuce žánrů jsme přistoupili k další fázi předzpracování dat, která zahrnovala výběr žánrů pro hloubkovou analýzu. Zatímco některé žánry jako drama, komedie nebo akce mají vysoký počet zastoupení a nabízejí robustní vzorek pro trénování našich modelů, jiné žánry mohou být méně vhodné kvůli své nejednoznačnosti nebo specifičnosti. V procesu selekce žánrů jsme se rozhodli zaměřit na ty, které byly nejčetněji zastoupené, a zároveň jsme hledali žánry s jasně definovanými charakteristikami. Žánry jako animace a rodinné filmy byly považovány za méně vhodné pro naše účely, protože často překrývají různé tematiky a mohou být subjektivně interpretovány.

Jako výsledek tohoto procesu jsme se zaměřili na žánry s významným počtem filmů, které jsou navíc jasné a rozlišitelné pro klasifikaci. To nám umožnilo vytvořit vyváženou a zaměřenou datovou sadu, která je představena v přiloženém obrázku a ukazuje vybrané žánry s rovnoměrným počtem zastoupení.

	Genre	Count
0	Drama	1000
1	Adventure	1000
2	Comedy	1000
3	Action	1000
4	Horror	1000

Obr. 4: Vybrané žánry

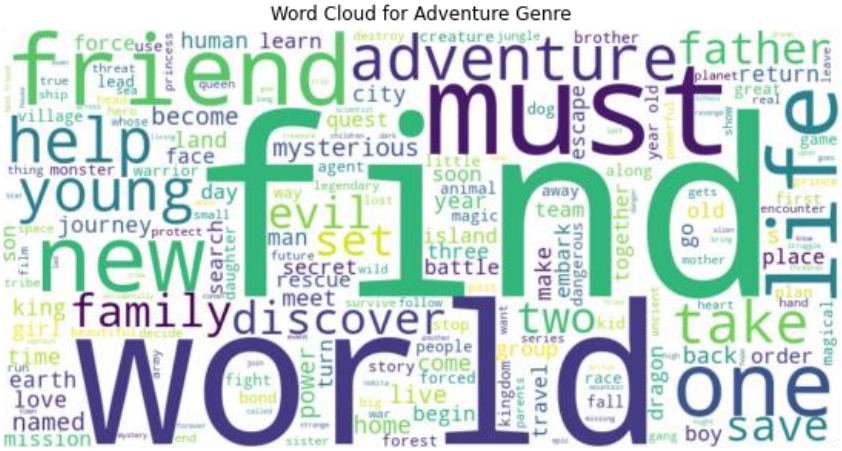
V rámci našeho projektu byl nadále kladen důraz na optimalizaci textových dat pro analýzu pomocí neuronových sítí. Jednou z nezbytných součástí předzpracování dat je i odstranění tzv. stop words. Tyto slova jsou běžná slova v jazyce, která však pro účely strojového učení a zpracování přirozeného jazyka obvykle nemají významovou váhu. Příkladem slov, která byla identifikována a odstraněna v tomto procesu, jsou běžně používané termíny bez významového zatížení pro žánrovou analýzu, jako jsou "the", "and" nebo "is". Je důležité poznamenat, že seznam těchto slov je často přizpůsoben konkrétním potřebám projektu a jazyka, ve kterém se data nacházejí. Naše přístupy k odstranění stop slov tedy byly pečlivě zváženy s ohledem na charakter a cíle našeho výzkumu.

Zároveň jsme se zaměřili na identifikaci slov, která se často vyskytují napříč různými filmovými žánry. Tato slova, ačkoliv jsou běžnou součástí jazyka, nemusí přinášet podstatný přínos k rozlišení jednotlivých žánrů, protože jejich vysoká frekvence je činí méně diskriminačními při klasifikaci textů. Mezi tato často se vyskytující slova patří například "new", "young", "one", "must" a "find". Tato slova byla identifikována jako běžná napříč celým spektrem filmových popisů, a proto bylo rozhodnuto, že je nutné je v naší analýze zvlášť zvážit. I když nejsou tradičními stop slovy, jejich nadměrná přítomnost může ztížit odhalení charakteristických rysů, které jsou specifické pro daný žánr.

```
Slová, které jsou hodně používané napříč žánry:
new
young
one
must
find
```

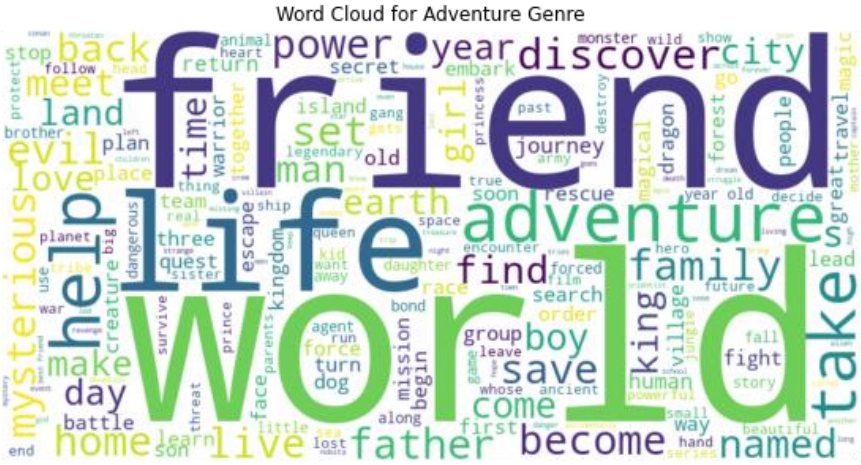
Obr. 5: Častá slova

Word clouds, neboli slovní mraky, jsou vizuální reprezentace textových dat, kde velikost slova odráží jeho frekvenci nebo důležitost v textu. Tyto mraky mohou poskytnout intuitivní přehled o klíčových tématech a konceptech, které se objevují v různých dokumentech nebo datových sadách, jako je například právě soubor filmových popisů. Vytvořili jsme tak word cloud pro každý zkoumaný žánr v našem datasetu. Níže ukázaná ukázka word cloudu, která obsahuje běžná slova jako "new", "young", "one", "must" a "find", poskytuje obecný přehled o jazyku, který je typicky používán v popisech dobrodružných filmů. Slova jako "quest", "journey", "mysterious", "world" a "adventure" jsou prominentně zastoupena a odrážejí klíčové prvky příběhu, které jsou pro žánr dobrodružství typické.



Obr. 6: Word cloud pro žánr Adventure

Druhý word cloud, ze kterého byla tato běžná slova odstraněna, nám ukazuje ještě specifičtější slovník, který je úzce spjat s tématy a příběhy dobrodružných filmů. Po odstranění těchto běžně používaných slov se objevují termíny, které jsou výraznější a poskytují hlubší porozumění pro to, co činí dobrodružství unikátním - například "discover", "treasure", "hero", "island" a "mystery". Tyto slova jsou klíčová pro dobrodružné žánry, protože odkazují na vzrušení z objevování nových světů, zápletky kolem hledání pokladů a charakteristické postavy hrdinů, kteří se vydávají na nebezpečné výpravy.



Obr. 7: Word cloud pro žánr Adventure po odstranění častých slov

4.2 Upráva dat pro první model

Při přípravě dat pro náš první model, jehož úkolem je klasifikace filmů na základě toho, zda patří do žánru hororu nebo ne, jsme přistoupili k významnému kroku v předzpracování dat. Kromě standardních úprav jako jsou normalizace textů a odstranění stop slov jsme dataset obohatili o nový sloupec is_horror. Sloupec is_horror byl přidán k naší datové sadě jako kritický prvek pro naše trénovací procesy. Tento sloupec obsahuje binární hodnoty, kde '0' indikuje, že film není klasifikován jako horor, zatímco '1' označuje, že film do této kategorie spadá. Použití binárních hodnot je v kontextu strojového učení preferováno, neboť modely jsou obvykle optimalizovány pro práci s numerickými daty a binární hodnoty jsou snadno interpretovatelné a efektivní pro trénování klasifikačních algoritmů.

	overview	sid	names	is_horror
0	dominating boxing world, adonis creed thriving...	0	Creed III	0
1	set decade events first film, learn story sull...	1	Avatar: The Way of Water	0
2	working underground fix water main, brooklyn p...	2	The Super Mario Bros. Movie	0
3	series unfortunate events, three mummies end p...	3	Mummies	0
4	good-hearted teenager william always lived hop...	4	Supercell	0
...	...	...	...	...
4995	haunted memory deceased mother, dana leaves ma...	10043	Blue Crush 2	0
4996	ronya lives happily father's castle comes acro...	10047	Ronia, The Robber's Daughter	0
4997	fred c. dobbs bob curtin, luck tampico, mexico...	10052	The Treasure of the Sierra Madre	0
4998	dinosaur families get trapped valley ice storm...	10058	The Land Before Time VIII: The Big Freeze	0
4999	polar bear norm three arctic lemming buddies f...	10064	Norm of the North	0

5000 rows × 4 columns

Obr. 8: Upravená data pro první model

4.3 Upráva dat pro druhý model

V rámci přípravy datové sady pro druhý model, který má za úkol klasifikovat filmy podle jejich žánrů, jsme implementovali klíčovou proměnnou genre_id. Tato proměnná nahrazuje textové názvy žánrů numerickými identifikátory, což představuje standardní praxi v strojovém učení pro efektivní zpracování kategoriálních dat. Díky genre_id

můžeme filmům přiřadit specifický žánr, což nám umožňuje lépe strukturovat a organizovat naše datové sady. Tato metodika je zvláště užitečná v situacích, kde je potřeba provádět klasifikaci na základě více tříd, a umožňuje modelu rozpoznat a naučit se vzorce, které jsou specifické pro každý žánr.

	sid	names	overview	genre	genre_id
0	0	Creed III	dominating boxing world, adonis creed thriving...	Drama	0
1	1	Avatar: The Way of Water	set decade events first film, learn story sull...	Adventure	1
2	2	The Super Mario Bros. Movie	working underground fix water main, brooklyn p...	Adventure	1
3	3	Mummies	series unfortunate events, three mummies end p...	Comedy	2
4	4	Supercell	good-hearted teenager william always lived hop...	Action	3
...	...	...	...	...	...
4995	10043	Blue Crush 2	haunted memory deceased mother, dana leaves ma...	Adventure	1
4996	10047	Ronia, The Robber's Daughter	ronya lives happily father's castle comes acro...	Adventure	1
4997	10052	The Treasure of the Sierra Madre	fred c. dobbs bob curtin, luck tampico, mexico...	Adventure	1
4998	10058	The Land Before Time VIII: The Big Freeze	dinosaur families get trapped valley ice storm...	Adventure	1
4999	10064	Norm of the North	polar bear norm three arctic lemming buddies f...	Adventure	1

5000 rows × 5 columns

Obr. 9: Upravená data pro druhý model

Nakonec jsme se zaměřili i na analýzu délky textových popisů, což nám poskytlo další vhled do našeho datasetu. Tato analýza je důležitá, neboť délka textu může ovlivnit jak kvalitu dat pro trénování modelů, tak výsledky klasifikace. Statistický souhrn délky popisů, jak je znázorněno v přiloženém obrázku, odhalil průměrnou délku popisu filmu okolo 200 slov s odchylkou přibližně 103 slov. To znamená, že ačkoliv některé popisy jsou velmi stručné (s minimální délkou 17 slov), jiné jsou poměrně podrobné (s maximální délkou 738 slov). Tento rozsah délek nám ukazuje diverzitu v tom, jak jsou filmy prezentovány v textové formě. Kratší popisy mohou poskytnout méně kontextu pro klasifikaci, zatímco delší popisy mohou obsahovat více podrobností a mohou být tak pro modely přínosnější. Zjištění, že většina popisů se drží kolem průměru, nám také naznačuje, že pro trénování modelů máme k dispozici dostatečně bohatá data.

```
✓ final_df['overview_len'].describe()
count    5000.000000
mean      200.998000
std       103.759458
min        17.000000
25%       121.000000
50%       179.000000
75%       259.000000
max       738.000000
```

Obr. 10: Analýza délky popisů

Všechny tyto poznatky jsou důležité pro další kroky při návrhu modelů, protože nám umožňují zvolit nejvhodnější techniky pro zpracování přirozeného jazyka, jako je například tokenizace a normalizace, aby byly popisy filmů co nejvíce informativní pro naše účely.

4.4 Příprava prvního modelu

Na základě připraveného kódu můžeme popsat první model, který byl navržen pro klasifikaci filmů jako hororových nebo nehororových. Model je postaven na rekurentní neuronové síti s architekturou LSTM (Long Short-Term Memory), což je vhodné pro sekvenční data, jako jsou texty, jelikož dokáže zachytit dlouhodobé závislosti v textových sekvencích.

První krok zahrnoval načtení a zpracování dat pomocí DataLoader, který byl speciálně vytvořen pro tento účel. Po načtení byla data očištěna a připravena do finální podoby, včetně přidání sloupce is_horror, který indikuje, zda je film horor či nikoli, přičemž hodnoty byly převedeny na binární formát pro lepší zpracování modelu.

Další fáze zahrnovala rozdělení dat na trénovací a testovací množinu, což je nezbytné pro ověření schopnosti modelu generalizovat na zatím neviděných datech. S využitím train_test_split z knihovny sklearn byla data rozdělena v poměru 70% pro trénování a 30% pro testování. Testovací sada slouží jako náhrada za skutečný svět a pomáhá ověřit, zda je model schopen efektivně předpovídat nebo klasifikovat nové příklady, které nebyly součástí jeho trénování. Tímto způsobem můžeme získat realistický odhad toho, jak by se model mohl chovat v reálných podmínkách a zabránit případnému nadměrnému přetrénování. Zároveň byly také použity váhy tříd, aby se vyrovnaly případné nerovnováhy v zastoupení tříd hororových a nehororových filmů, což pomáhá předcházet zkreslení modelu ve prospěch dominantnější třídy.

Následně byla provedena tokenizace a sekvencování textů pomocí knihovny tensorflow.keras, kde tokenizér převádí texty na sekvence tokenů, které model může zpracovat. Poté byly sekvence doplněny do stejné délky pomocí paddingu, aby byly kompatibilní s modelem LSTM.

```
oov_token = "<OOV>"
tokenizer = Tokenizer(oov_token=oov_token)
tokenizer.fit_on_texts(X_train)

word_index = tokenizer.word_index

X_train_sequences = tokenizer.texts_to_sequences(X_train)
X_test_sequences = tokenizer.texts_to_sequences(X_test)
```

Obr. 11: Tokenizace a sekvencování textu

Další část je součástí procesu zpracování textových dat pro výstupní vrstvu LSTM neuronové sítě. LSTM sítě vyžadují, aby vstupní data byla ve formě sekvencí stejné délky, proto je potřeba sekvence normalizovat. Nejprve kód určuje délku každé sekvence (seznam tokenů) v trénovací a testovací sadě. Délka sekvence je počet tokenů v každém popisu filmu po tokenizaci. Následně kód spojuje délky sekvencí z trénovací a testovací sady do jednoho seznamu a najde maximální délku sekvence v celém datasetu. Tato maximální délka se pak používá jako referenční hodnota pro všechny sekvence při jejich doplňování (padding) nebo ořezávání (truncating).

Poté pad_sequences funkce z knihovny tensorflow.keras doplní (nebo ořízne) všechny sekvence tak, aby odpovídaly této maximální délce. V této fázi je možné sekvence buď doplnit na konec (padding="post") nebo oříznout na konci (truncating="post"), v závislosti na zvolené konfiguraci.

```
padding_type = "post"
trunction_type= "post"

# Find the length of each sequence
sequence_lengths_train = [len(seq) for seq in X_train_sequences]
sequence_lengths_test = [len(seq) for seq in X_test_sequences]
sequence_lengths = sequence_lengths_train + sequence_lengths_test
# Determine the maximum length
max_length = max(sequence_lengths)
# max_length = 512

X_train_padded = pad_sequences(X_train_sequences,maxlen=max_length, padding=padding_type, truncating=trunction_type)

X_test_padded = pad_sequences(X_test_sequences,maxlen=max_length, padding=padding_type, truncating=trunction_type)
```

Obr. 12: Normalizace sekvencí

Nakonec náš kód inicializuje pseudonáhodný generátor tensorflow s pevnou hodnotou seedu, aby zajistil reprodukovatelnost výsledků při trénování. Trénovací a validační data jsou poté převedena do objektů, které efektivně umožňují iteraci přes data během trénovacího procesu. Tyto objekty jsou pak použity pro trénování a validaci modelu.

```
tf.random.set_seed(0)

training_data = tf.data.Dataset.from_tensor_slices((X_train_padded, y_train))
validation_data = tf.data.Dataset.from_tensor_slices((X_test_padded, y_test))
```

Obr. 13: Převedení dat do objektů

4.5 První model

Pro efektivní a strukturované trénování neuronové sítě je nezbytné definovat parametry, které určují, jak bude trénování probíhat. V našem případě jsme nastavili batch_size na 64, což znamená, že model bude zpracovávat 64 popisů najednou během jedné iterace (tzv. epochy) trénovacího procesu. Toto číslo je kompromisem mezi paměťovými požadavky a rychlostí učení modelu a bylo vybráno s ohledem na optimální využití zdrojů a efektivitu trénování. Počet epoch přes celou datovou sadu byl nastaven na 5. Toto číslo reprezentuje počet celkových průchodů trénovacími daty, které model vykoná před ukončením trénování.

Pro zamezení přetrénování modelu byla zavedena zpětná vazba v podobě EarlyStopping, která monitoruje 'val_loss', tedy ztrátu na validační sadě. Pokud se během tří po sobě jdoucích epoch nezlepší, trénování se předčasně zastaví. Tento mechanismus slouží jako pojistka proti přizpůsobení se modelu pouze na trénovací data bez schopnosti generalizace.

Velikost slovníku vocab_size určuje počet unikátních tokenů, které model může zpracovat, a byla vypočítána jako počet unikátních slov v indexu tokenizeru plus jeden pro padding.

```
batch_size = 64
training_data = training_data.batch(batch_size)
validation_data = validation_data.batch(batch_size)
epochs = 5
callback = tf.keras.callbacks.EarlyStopping(monitor='val_loss', patience=3)
vocab_size = len(tokenizer.word_index) + 1
```

Obr. 14: Nastavení parametrů prvního modelu

Tento model je sekvenční sestávající z několika vrstev, navržený pro práci s textovými daty. Začíná vrstvou Embedding, která převádí celočíselné indexy slov na husté vektory nižší dimenze a umožňuje modelu zachytit a naučit se bohaté reprezentace slov.

Následuje dvoucestná LSTM vrstva, která zpracovává sekvence z obou směrů, což poskytuje modelu kontext z minulosti i budoucnosti v sekvenci. S dropoutem 0.5 a rekurentním dropoutem 0.5 je model robustnější proti přizpůsobení na trénovací data. Regularizace L1 a L2 pomáhá omezit složitost modelu a předcházet přetrénování.

Dále model obsahuje plně propojené vrstvy s relu aktivační funkcí a dalším dropoutem pro další regulaci. Výstupní vrstva s aktivační funkcí sigmoid je vhodná pro binární klasifikaci, kde model predikuje pravděpodobnost příslušnosti filmu k žánru hororu.

Model je kompilován s optimizátorem Adam s nižší učící rychlostí 0.001, což umožňuje jemnější nastavení váh a často vede k lepší konvergenci při trénování. Metriky jako přesnost (acc) a binární crossentropy ztráta (binary_crossentropy) jsou použity k monitorování a hodnocení výkonu modelu.

Celkově tato konfigurace představuje promyšlenou kombinaci hyperparametrů, architektury modelu a procesu trénování. Níže je možné vidět i průběh jednotlivých iterací

```
model = Sequential([
    Embedding(vocab_size, 64, input_length=max_length),
    Bidirectional(LSTM(32, return_sequences=False, dropout=0.5,
                       recurrent_dropout=0.5, kernel_regularizer=l1_l2(l1=1e-5, l2=1e-4))),
    Dropout(0.5),
    Dense(32, activation='relu', kernel_regularizer=l1_l2(l1=1e-5, l2=1e-4)),
    Dropout(0.5),
    Dense(1, activation='sigmoid')
])

# Using a smaller learning rate
optimizer = Adam(learning_rate=0.001)

model.compile(optimizer=optimizer, loss='binary_crossentropy', metrics=['acc'])

history = model.fit(training_data, epochs=epochs, verbose=1, validation_data=validation_data, callbacks=[callback])
```

Obr. 15: Vrstvy prvního modelu

```
55/55 [==============================] - 18s 192ms/step - loss: 1.1463 - acc: 0.6620 - val_loss: 0.7184 - val_acc: 0.4567
Epoch 2/5
55/55 [==============================] - 10s 179ms/step - loss: 1.1249 - acc: 0.6146 - val_loss: 0.7244 - val_acc: 0.3920
Epoch 3/5
55/55 [==============================] - 10s 190ms/step - loss: 0.7980 - acc: 0.8063 - val_loss: 0.3770 - val_acc: 0.8440
Epoch 4/5
55/55 [==============================] - 10s 185ms/step - loss: 0.3013 - acc: 0.9411 - val_loss: 0.5809 - val_acc: 0.8427
Epoch 5/5
55/55 [==============================] - 10s 185ms/step - loss: 0.1783 - acc: 0.9734 - val_loss: 0.5576 - val_acc: 0.8660
```

Obr. 16: Jednotlivé iterace prvního modelu

Po dokončení fáze trénování byl výkon modelu vyhodnocen na trénovacích a validačních datech. Tento krok je zásadní, protože nám poskytuje přehled o tom, jak dobře model funguje a jak dobře se může přizpůsobit neviděným datům.

Na trénovacích datech dosáhl model přesnosti téměř 98,89 %, což ukazuje, že se model velmi dobře naučil rozpoznávat vzory a rozlišovat hororové filmy od nehororových v rámci trénovací sady. Tento vysoký stupeň přesnosti naznačuje, že model dobře zapadá do trénovacích dat a je schopen správně klasifikovat filmy na základě poskytnutých popisů.

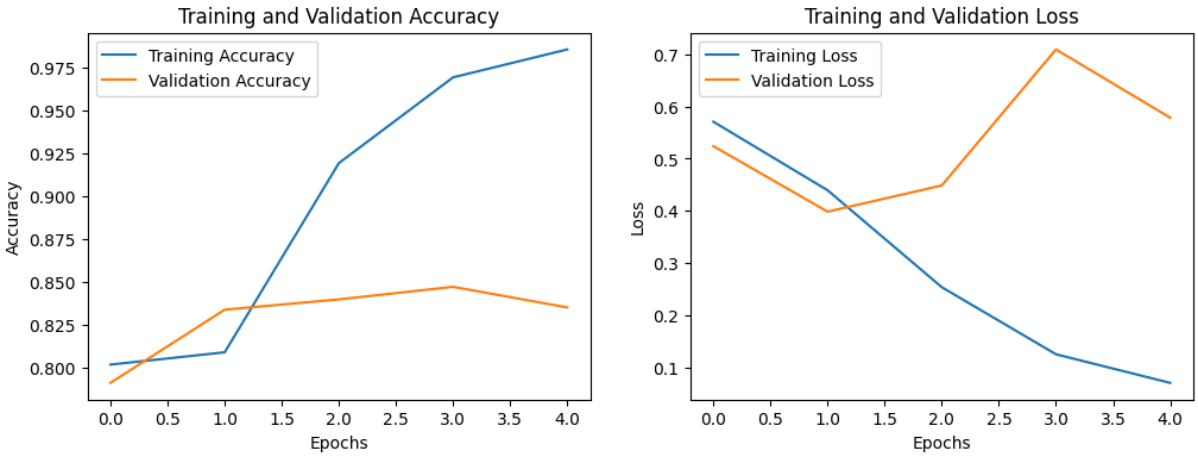
Při vyhodnocení modelu na validační sadě, která obsahuje data, na kterých model nebyl trénován, byla přesnost zhruba 86,6 %. Tento pokles oproti trénovací sadě je očekávaný, neboť testovací data mohou obsahovat vzory nebo příklady, které model během trénování neviděl. Přesto je tato přesnost poměrně vysoká a naznačuje, že model má dobrou generalizační schopnost.

```
✓ loss, accuracy = model.evaluate(training_data) ···

55/55 [==============================] - 1s 23ms/step - loss: 0.0601 - acc: 0.9889
Training Accuracy is 98.88571500778198


✓ loss, accuracy = model.evaluate(validation_data) ···

24/24 [==============================] - 1s 32ms/step - loss: 0.5576 - acc: 0.8660
Testing Accuracy is 86.59999966621399
```

Obr. 17: Přesnost prvního modelu

Grafy přesnosti a ztráty představují vizualizaci výkonu modelu v průběhu jeho trénování a validace. Na grafu přesnosti vidíme, že trénovací přesnost (modrá čára) konzistentně roste s každou epochou, což signalizuje, že model se s každým průchodem daty zlepšuje ve své schopnosti správně klasifikovat trénovací příklady. Naopak validační přesnost (oranžová čára) se zvyšuje mnohem pomaleji a po určitém bodě se stabilizuje, což naznačuje, že model může dosahovat svého výkonnostního stropu na neviděných datech.

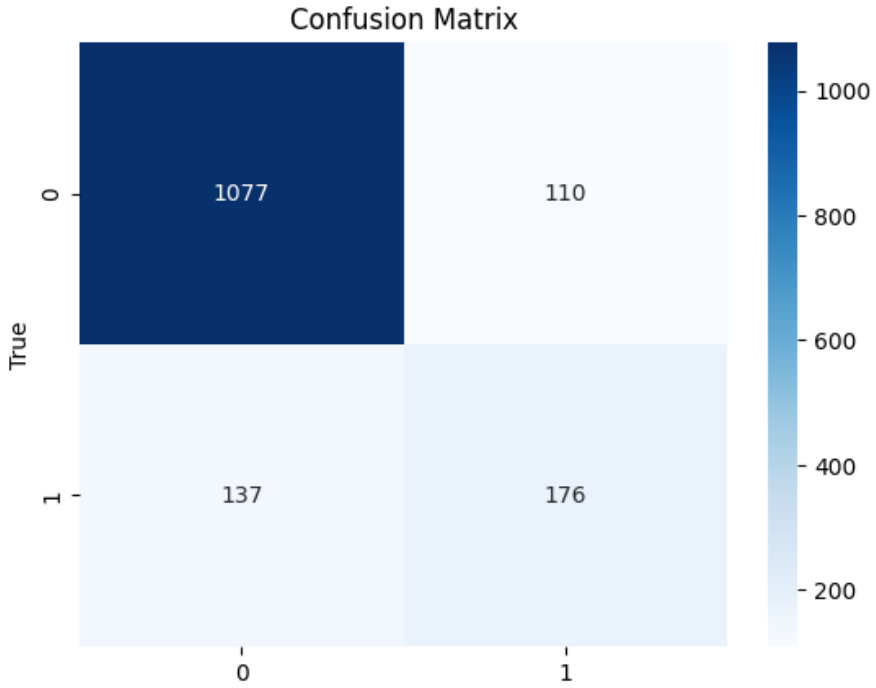
Na druhém grafu ztráty pozorujeme klesající trend trénovací ztráty (modrá čára), což ukazuje, že model se stává stále jistějším ve svých predikcích. Validační ztráta (oranžová čára) však po počátečním poklesu vykazuje nárůst, což může být signálem přetrénování nebo nesouladu mezi trénovacími a validačními daty.



Obr. 18: Vizualizace přesnosti a ztráty prvního modelu

Další důležitou částí vyhodnocení je matice záměn (confusion matrix). Jedná se o vizuální nástroj používaný pro vyhodnocení výkonu klasifikačních modelů. V poskytnuté matici záměn vidíme počty případů, kde model správně predikoval (diagonální hodnoty) ve srovnání s počty, kde se model mýlil (nediagonální hodnoty).

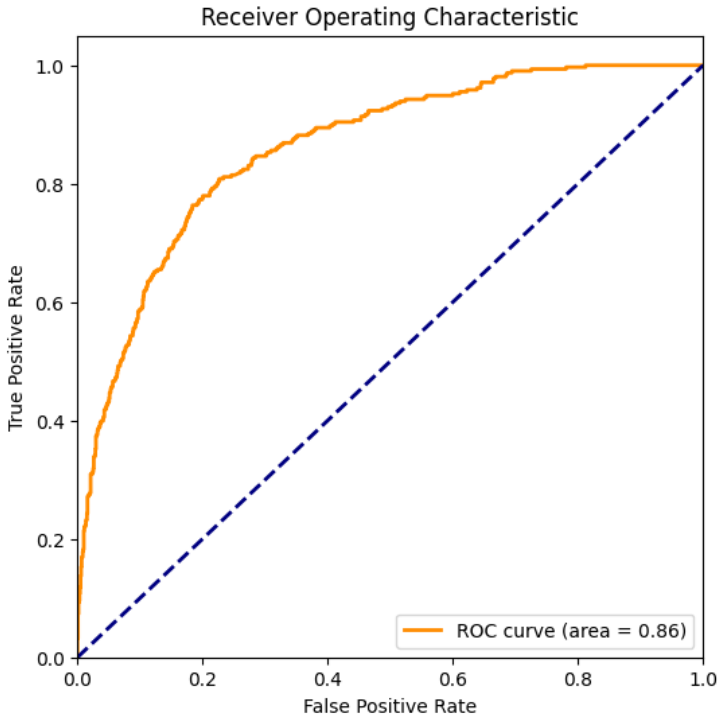
V levém horním kvadrantu (0,0) vidíme 1077 správných predikcí pro třídu "není horor", zatímco v pravém dolním kvadrantu (1,1) je 176 správných predikcí pro třídu "horor". Levý dolní kvadrant (1,0) ukazuje 137 falešně negativních výsledků, kde model nesprávně predikoval, že film není horor, a pravý horní kvadrant (0,1) obsahuje 110 falešně pozitivních výsledků, kde model nesprávně uvedl, že film je horor.



Obr. 19: Matice záměn prvního modelu

ROC křivka (Receiver Operating Characteristic curve) je grafické vyjádření výkonu klasifikačního modelu. Na obrázku vidíme ROC křivku pro náš model, která znázorňuje vztah mezi mírou pravdivě pozitivních výsledků (na vertikální ose) a mírou falešně pozitivních výsledků (na horizontální ose) při různých prahových hodnotách.

Křivka ukazuje, jak dobře model dokáže rozlišit mezi třídami – v tomto případě mezi hororovými a nehororovými filmy. Ideální model by měl křivku, která se táhne k hornímu levému rohu grafu, což by znamenalo vysokou míru pravdivě pozitivních výsledků při nízké míře falešně pozitivních výsledků. Plocha pod křivkou (AUC - Area Under the Curve) hodnocená jako 0.86 naznačuje vysokou prediktivní schopnost modelu - čím blíže je AUC k hodnotě 1, tím lépe. Tato ROC křivka a přidružené skóre poskytují silné náznaky, že model je schopný efektivně rozlišovat mezi pozitivními a negativními příklady v našem datasetu.



Obr. 20: Křivka ROC pro první model

Ukázka predikcí náhodně vybraných filmů nám poskytuje konkrétní příklady výkonu našeho modelu v reálném provozu. V tomto náhodném výběru model úspěšně rozpoznal nehororové filmy, jako jsou "Predator", "Fighting Spirit - Mashiba vs. Kimura", "Dumb and Dumberer: When Harry Met Lloyd" a "Aladdin", což naznačuje, že se model naučil rozlišovat typické vlastnosti a klíčová slova, která nejsou přítomna v hororovém žánru. Naproti tomu, u filmu "Maid in Sweden" došlo k nesprávné klasifikaci, kde model chybně označil film za horor. To nám ukazuje, že ačkoliv je model výkonný, stále existuje prostor pro jeho další vylepšení v případě filmů, jejichž popisy mohou obsahovat klamavé nebo méně jednoznačné klíčové slova.

```
Title: Predator
Actual Genre: Not Horror
Predicted Genre: Not Horror
Result: Correct ✓
----------------------------------------
Title: Fighting Spirit - Mashiba vs. Kimura
Actual Genre: Not Horror
Predicted Genre: Not Horror
Result: Correct ✓
----------------------------------------
Title: Dumb and Dumberer: When Harry Met Lloyd
Actual Genre: Not Horror
Predicted Genre: Not Horror
Result: Correct ✓
----------------------------------------
Title: Aladdin
Actual Genre: Not Horror
Predicted Genre: Not Horror
Result: Correct ✓
----------------------------------------
Title: Maid in Sweden
Actual Genre: Not Horror
Predicted Genre: Horror
Result: Incorrect ✗
----------------------------------------
```

Obr. 21: Ukázka predikce prvního modelu

Celkově tento model předvedl schopnost efektivně zpracovat a analyzovat velké množství textových dat a poskytnout přesné predikce. Výsledky, kterých jsme dosáhli, potvrzují, že model by mohl mít s určitými vylepšeními potenciál být užitečným nástrojem například pro streamovací platformy, doporučovací systémy nebo archivy obsahu pro automatické tagování a kategorizaci filmů. Budoucí práce by tak mohla zahrnovat zkoumání metod pro zlepšení jeho přesnosti, snížení míry falešně pozitivních a falešně negativních výsledků a zvýšení jeho schopnosti generalizace napříč rozmanitějšími a komplexnějšími datovými sadami.

4.6 Příprava druhého modelu

Postup vývoje druhého modelu byl podobný prvnímu, avšak s jedním klíčovým rozdílem: zatímco první model se zaměřoval na binární klasifikaci (horor/nehoror), druhý model byl určen pro kategorickou klasifikaci, kde každý film byl přiřazen do jednoho ze širšího spektra žánrů. To vyžadovalo použití kategorického kódování výstupních tříd, což bylo realizováno pomocí funkce to_categorical, která převedla cílové proměnné na binární matici tříd pro použití v kategorické cross-entropy funkci ztráty. Zároveň byla data rozdělena v poměru 80 % pro trénování a 20 % pro testování.

```
unique_genres = final_df['genre_id'].unique()
y = to_categorical(y, num_classes=len(unique_genres))
X_train, X_test , y_train, y_test = train_test_split(X, y , test_size = 0.20)
```

Obr. 22: Kategorické kódování

Druhý model byl tedy koncipován s cílem rozpoznat a klasifikovat žánr filmu na základě jeho textového popisu. K dosažení tohoto cíle bylo třeba zpracovat popisy filmů do podoby, kterou je možné efektivně využít pro trénování modelu schopného rozpoznávat jednotlivé žánrové charakteristiky.

V rámci zpracování dat bylo každému filmu v datové sadě přiřazen unikátní identifikátor žánru, který odpovídal široké škále filmových žánrů. Tento krok byl klíčový pro zajištění, že model bude schopen provádět více třídní klasifikaci, což je významně složitější úkol než binární klasifikace, jelikož vyžaduje, aby model rozpoznal a odlišil více vzorů a souvislostí v datech.

Proces přípravy dat, trénování a validace modelu byl proveden s pečlivým důrazem na vyvážení tříd a zajištění, že model nebude mít předpojatost k některému žánru kvůli nerovnoměrnému zastoupení ve vstupních datech. Tokenizace, sekvencování a normalizace délky sekvencí byly provedeny podobně jako u prvního modelu, přičemž se zohlednila potřeba zachovat informace důležité pro rozlišení mezi různými žánry.

4.7 Druhý model

Pro druhý model bylo nastavení parametrů trénování upraveno v porovnání s prvním modelem, čím reflektujeme kategorickou povahu úlohy. Velikost dávky (batch_size) byla snížena na 32, což může vést ke stabilnějšímu a podrobnějšímu učení, jelikož menší dávky umožňují modelu činit jemnější úpravy váh na základě menšího počtu vzorků. Toto může být obzvláště prospěšné při práci s více třídami, kde může být každá třída reprezentována různým počtem vzorků a kde se hledají jemnější vzory v datech.

Počet epoch byl nastaven na 8, což indikuje, že model by měl projít datovou sadou celkem osmkrát během procesu trénování. Tento počet epoch značí potřebu zabránit přetrénování, které může být víc pravděpodobné při kategorické klasifikaci, jelikož modely se snaží naučit složitější vzory a mohou být náchylnější k zapamatování konkrétních detailů trénovací sady.

Stejně jako u prvního modelu, byla použita zpětná vazba EarlyStopping pro monitorování ztráty na validační sadě. Patience byla nastavena na 3, což znamená, že trénování se zastaví, pokud se během tří epoch nezlepší hodnota ztráty na validační sadě. Toto nastavení je klíčové pro udržení schopnosti generalizace modelu a zabránění jeho nadměrného přizpůsobení se trénovacím datům.

```
batch_size = 32
training_data = training_data.batch(batch_size)
validation_data = validation_data.batch(batch_size)
epochs = 8
callback = tf.keras.callbacks.EarlyStopping(monitor='loss', patience=3)
```

Obr. 23: Nastavení parametrů druhého modelu

Architektura druhého modelu byla navržena s ohledem na složitější úkol kategorické klasifikace, kde model musí rozpoznat a přiřadit správný žánr filmu z více možných kategorií. Jako základ byla použita vrstva Embedding. Dále byly implementovány dvě LSTM vrstvy, první s 64 jednotkami s možností vrácení sekvencí, což je důležité pro propojení následujících vrstev, a druhá s 32 jednotkami. Obě vrstvy byly nastaveny s dropoutem a rekurentním dropoutem pro prevenci přetrénování. Dropout pomáhá modelu zobecnit lépe tím, že redukuje závislosti na konkrétních vzorcích v trénovacích datech, čímž se zvyšuje jeho robustnost.

Výstupní vrstva využívá aktivační funkci softmax, což je standardní volba pro kategorickou klasifikaci. V tomto případě je počet neuronů v této vrstvě roven počtu unikátních žánrů. Regularizace je použita i v této vrstvě k omezení složitosti modelu a snížení rizika přetrénování.

Celkově byl druhý model kompilován s adam optimizátorem a funkcí ztráty categorical_crossentropy, což je opět typické pro úlohy kategorické klasifikace. Metrika přesnosti (acc) byla použita pro sledování výkonu modelu během trénování.

```
vocab_size = len(tokenizer.word_index) + 1
# vocab_size = 5000
model = Sequential([
    Embedding(vocab_size, 64, input_length=max_length),
    Bidirectional(LSTM(64, return_sequences=True, dropout=0.5, recurrent_dropout=0.25)),
    Bidirectional(LSTM(32, dropout=0.5, recurrent_dropout=0.25)),
    Dense(len(unique_genres), activation='softmax', kernel_regularizer=regularizers.l2(0.01),
          activity_regularizer=regularizers.l1(0.01))
])

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['acc'])
history = model.fit(training_data, epochs=epochs, verbose=1,validation_data = validation_data, callbacks = [callback])
```

Obr. 24: Vrstvy druhého modelu

```
125/125 [==============================] - 86s 461ms/step - loss: 1.6307 - acc: 0.2548 - val_loss: 1.5119 - val_acc: 0.3310
Epoch 2/8
125/125 [==============================] - 52s 411ms/step - loss: 1.2538 - acc: 0.4465 - val_loss: 1.4022 - val_acc: 0.4080
Epoch 3/8
125/125 [==============================] - 61s 486ms/step - loss: 0.9038 - acc: 0.6125 - val_loss: 1.5439 - val_acc: 0.3800
Epoch 4/8
125/125 [==============================] - 60s 477ms/step - loss: 0.6915 - acc: 0.7275 - val_loss: 1.7098 - val_acc: 0.3970
Epoch 5/8
125/125 [==============================] - 59s 469ms/step - loss: 0.5323 - acc: 0.8215 - val_loss: 1.8803 - val_acc: 0.4160
Epoch 6/8
125/125 [==============================] - 60s 482ms/step - loss: 0.3820 - acc: 0.8970 - val_loss: 2.0159 - val_acc: 0.4460
Epoch 7/8
125/125 [==============================] - 69s 553ms/step - loss: 0.3525 - acc: 0.9097 - val_loss: 1.8921 - val_acc: 0.4740
Epoch 8/8
125/125 [==============================] - 47s 378ms/step - loss: 0.2298 - acc: 0.9550 - val_loss: 2.0199 - val_acc: 0.4820
```

Obr. 25: Jednotlivé iterace druhého modelu

Vyhodnocení druhého modelu odhalilo značný rozdíl mezi přesností na trénovacích a validačních datech. Na trénovací sadě dosáhl model vysoké přesnosti přibližně 98,32 %, což ukazuje, že se model byl schopen naučit rozlišovat mezi různými žánry na základě poskytnutých textových popisů. Tato vysoká přesnost na trénovacích datech naznačuje, že model má dobrou schopnost zapamatovat si charakteristiky každého žánru, jak jsou prezentovány v trénovací sadě.

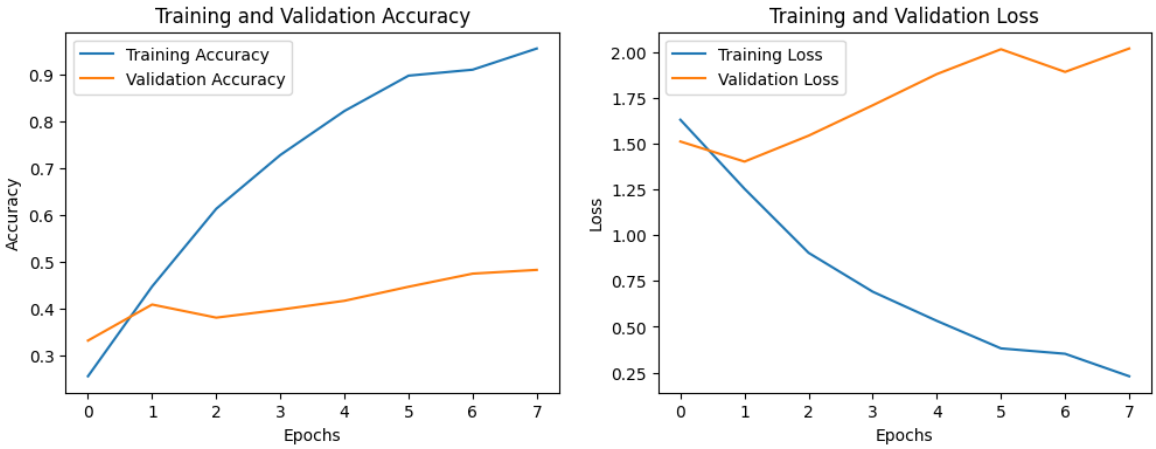
Avšak při testování na validační sadě došlo k výraznému poklesu přesnosti na 48,20 %. Tento výsledek může signalizovat, že model má potíže s generalizací na neviděná data a může být náchylný k přetrénování. Tento rozdíl v přesnosti mezi trénovacími a validačními daty naznačuje, že model může být příliš uzpůsobený specifikům trénovací sady a nezachytává dostatečně obecné vzory, které jsou aplikovatelné na širší spektrum filmů

```
✓ loss, accuracy = model.evaluate(training_data) ···

··· 125/125 [==============================] - 4s 31ms/step - loss: 0.1609 - acc: 0.9833
Training Accuracy is 98.32500219345093


✓ loss, accuracy = model.evaluate(validation_data) ···

··· 32/32 [==============================] - 1s 30ms/step - loss: 2.0199 - acc: 0.4820
Testing Accuracy is 48.19999933242798
```

Obr. 26: Přesnost druhého modelu

Grafy přesnosti a ztrát pro druhý model ukazují výrazný rozdíl mezi výkonem modelu na trénovacích a validačních datech. Z grafu přesnosti je patrné, že zatímco přesnost na trénovací sadě se postupně zvyšuje a dosahuje hodnoty nad 90%, přesnost na validační sadě stagnuje na nižší úrovni okolo 40 %. Tento jev může naznačovat přetrénování modelu, kde model vyniká na datech, na kterých byl trénován, ale nepředvádí dobrou generalizaci na nových datech.

V porovnání s prvním modelem, kde validační ztráta mírně stoupala, zde vidíme výraznější rozkol mezi trénovací a validační ztrátou. Tato data naznačují, že druhý model by mohl mít příliš komplexní strukturu nebo by nebyl dostatečně regularizován, aby předešel nadměrnému přizpůsobení se specifickým charakteristikám trénovací sady.

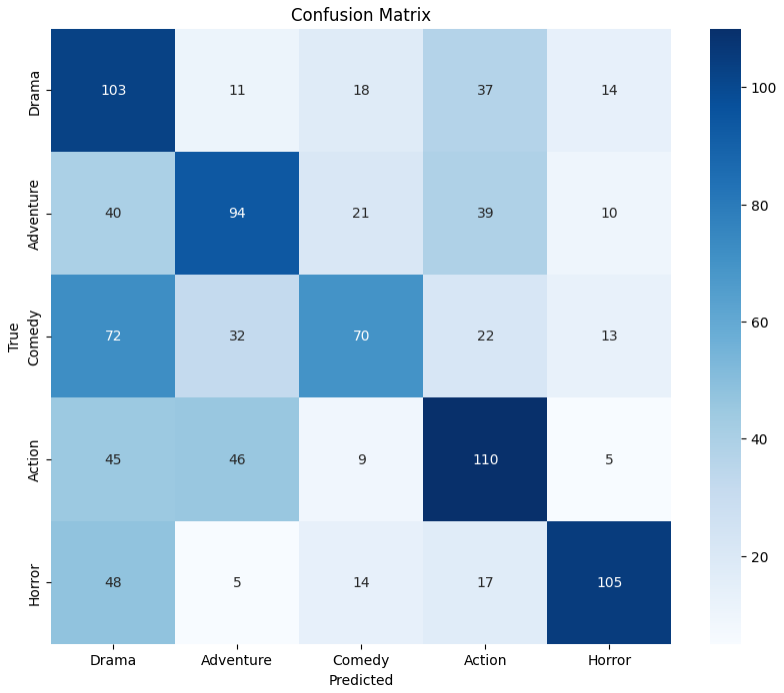


Obr. 27: Vizualizace přesnosti a ztráty druhého modelu

Matice záměn pro druhý model poskytuje užitečný vhled do jeho schopnosti klasifikace více žánrů. Každý řádek matice představuje skutečné třídy, zatímco každý sloupec představuje predikované třídy. Diagonální hodnoty ukazují počet správně klasifikovaných příkladů pro každý žánr, což je ideální situace. Například model správně identifikoval 103 dramat, 94 dobrodružných filmů, 70 komedií, 110 akčních filmů a 105 hororů.

Nediagonální hodnoty ukazují případy, kde model klasifikoval filmy do špatného žánru. Tyto hodnoty ukazují, jaké typy chyb model dělá. Například 48 filmů bylo nesprávně klasifikováno jako drama, když ve skutečnosti to byly horory, což naznačuje, že model může zaměňovat určité narativní prvky mezi těmito žánry.

Tato matice záměn odhaluje silné a slabé stránky modelu v rozlišování mezi jednotlivými žánry a může pomoci v identifikaci specifických žánrů, které modelu dělají problémy, což by mohlo vést k dalším úpravám v předzpracování dat nebo v architektuře modelu, aby se tyto chyby minimalizovaly. V porovnání s prvním modelem, který klasifikoval pouze mezi hororovými a nehororovými filmy, druhý model čelí výzvě klasifikovat filmy do širšího spektra žánrů, což je úloha s mnohem vyšší složitostí a vyšším rizikem chyb.



Obr. 28: Matice záměn druhého modelu

Ukázka predikcí druhého modelu v rámci mezí demonstruje jeho schopnost klasifikovat filmy do příslušných žánrů. Model úspěšně identifikoval "Vertical Limit" jako dobrodružný film a "Thor: Love and Thunder" jako akční, což ukazuje, že má schopnost správně rozpoznávat určité žánrové charakteristiky. Na druhou stranu, u filmů "Johnny English Reborn" a "Shin Godzilla" model udělal chybu, když je nesprávně klasifikoval jako akční, respektive dobrodružný, což naznačuje, že některé aspekty žánrových popisů mohou být pro model matoucí a vedou k nesprávným predikcím.

Obr. 29: Ukázka predikce druhého modelu

Jako závěrečný postřeh ohledně druhého modelu lze říci, že i když model prokázal schopnost naučit se a predikovat žánry filmů, celková úspěšnost není tak vysoká, jak bylo zamýšleno. Nízká přesnost na validační sadě a chyby v klasifikaci poukazují na potenciální problémy s přetrénováním a generalizací. Tento výsledek může být způsoben mnoha faktory, včetně možné potřeby dalšího ladění modelu, revize předzpracování dat, nebo přepracování architektury modelu pro lepší zachycení rozmanitosti žánrových charakteristik. V budoucí práci by se mělo zaměřit na zdokonalení těchto oblastí s cílem zvýšit přesnost a spolehlivost modelu v rozpoznávání žánrů filmů. Na základě provedených analýz v našem projektu jsme dospěli k závěru, že popis filmů není dostačně relevantní vůči žánru filmu.

Druhý model představoval náš prvotní pokus o vytvoření neuronové sítě, který zde uvádíme i přes to, že nelze považovat za úplně úspěšný. Díky zjištěním týkajících se kategorické klasifikace jsme si uvědomili daná omezení v souvislosti s filmovými daty a byli jsme schopni vypracovat jiný a úspěšnější model.

5 Závěr

V průběhu našeho výzkumu jsme se věnovali propojení pokročilých technik strojového učení s analýzou filmového obsahu a jeho žánrové klasifikace. Prostřednictvím dvou hlavních modelů jsme zkoumali, jak může umělá inteligence přispět k hlubšímu porozumění a strukturované kategorizaci filmového obsahu. Naše práce zdůrazňuje důležitost pečlivého předzpracování dat a výběru správné architektury neuronových sítí pro zadanou úlohu.

První model, zaměřující se na binární klasifikaci filmů do žánru hororu nebo nehororu, prokázal schopnost s vysokou přesností identifikovat a klasifikovat filmy na základě textových popisů. Přesnost dosažená na trénovacích i validačních datech, stejně jako vizualizace procesu trénování a validační metriky, naznačují, že model by mohl být cenným nástrojem pro automatické kategorizace filmů a může sloužit jako základ pro další výzkum a vývoj aplikací v oblasti doporučovacích systémů a obsahových archivů.

Druhý model, jehož cílem bylo rozpoznávat a kategorizovat filmy do širšího spektra žánrů, se setkal s určitými výzvami. I přes vysokou přesnost na trénovací sadě, pokles přesnosti na validační sadě a chyby v klasifikaci ukázaly, že model potřebuje další vylepšení pro zlepšení své schopnosti generalizace. Závěrem lze říci, že i když tento model nepřinesl zcela uspokojivé výsledky, poskytl cenné poznatky o komplexnosti kategorické klasifikace a představuje důležitý krok k porozumění, omezením a možnostem aplikace neuronových sítí v této doméně.

Na základě zkušeností získaných během tohoto výzkumu a analýzy výsledků obou modelů jsme přesvědčeni o potenciálu umělé inteligence ve filmovém průmyslu. Je jasné, že aplikace strojového učení může přinést nové možnosti pro analýzu a pochopení filmového obsahu, ale také zdůrazňuje potřebu pokračujícího vývoje a zdokonalování modelů pro dosažení lepších výsledků.

Celý zdrojový kód tohoto projektu je k dispozici na odkazu zde:
https://github.com/xlysova/NS_projectZP

6 Literatura

GENG, Chao, Qingji SUN a Shigetoshi NAKATAKE. Implementation of Analog Perceptron as an Essential Element of Configurable Neural Networks. *Sensors (Basel, Switzerland)*. 2020, roč. 20, č. 15, s. 4222. ISSN 1424-8220. DOI: 10.3390/s20154222

OROJO, Oluwatamilore et al. The Multi-Recurrent Neural Network for State-Of-The-Art Time-Series Processing. *Procedia Computer Science*. 2023, roč. 222, s. 488–498. ISSN 1877-0509. DOI: 10.1016/j.procs.2023.08.187

QIAO, Yi, Kaiwen XU a Kai ZHOU. *Research on time series based on improved LSTM*. 2023. DOI: 10.1109/ICPECA56706.2023.10076103

SARKER, Iqbal H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*. 2021, roč. 2, č. 6, s. 420. ISSN 2661-8907. DOI: 10.1007/s42979-021-00815-1

TAYE, Mohammad Mustafa. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*. Multidisciplinary Digital Publishing Institute, 2023, roč. 11, č. 3, s. 52. ISSN 2079-3197. DOI: 10.3390/computation11030052

The Movies Dataset. In: [cit. 23.01.2024]. Dostupné z: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

xlysova/NS_projectZP [online]. 2024 [cit. 23.01.2024]. Dostupné z: https://github.com/xlysova/NS_projectZP