VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

**Ing. Lukáš Povoda**

# DOLOVANIE ZNALOSTÍ Z TEXTOVÝCH DÁT POUŽITÍM METÓD UMELEJ INTELIGENCIE

TEXT MINING BASED ON
ARTIFICIAL INTELLIGENCE METHODS

*SKRÁTENÁ VERZIA DIZERTAČNEJ PRÁCE*

| | |
|---|---|
| Odbor: | Teleinformatika |
| Školiteľ: | doc. Ing. Radim Burget, Ph.D. |
| Oponenti: | |
| Dátum obhajoby: | |

## KĽÚČOVÉ SLOVÁ

Analýza sentimentu, dolovanie znalostí, hlboké učenie, klasifikácia emócií, klasifikácia textu, optimalizácia genetickým programovaním, spracovanie prirodzeného jazyka, textové dáta, umelá inteligencia

## KEYWORDS

Artificial intelligence, data mining, emotion classification, genetic programming optimization, natural language processing, sentiment analysis, text data, text mining

## MIESTO ULOŽENIA PRÁCE

Disertačná práca je k dispozícii na Vedeckom oddelení dekanátu FEKT VUT v Brne, Technická 10, Brno, 616 00



ISBN 80–214–
ISSN 1213–4198

# Obsah

# Úvod

Množstvo dát uložených v elektronickej podobe exponenciálne stúpa, pričom 80% týchto dát má textovú podobu a len 5% z týchto dát je možné považovať za hodnotné. [5] Vďaka súčasným trendom v *Internete vecí – Internet of Things* (IoT), stále lacnejšiemu úložisku a dostupnejším rýchlejším pripojeniam je možné očakávať, že tento trend bude naďalej pokračovať. Textové dáta nemajú jasne definovaný dátový model, ktorý by mohol byť použitý pri hľadaní požadovanej informácie. Spracovať veľký objem textových dát ľudskými silami je prakticky nemožné. V súčasnosti existujú metódy dolovania znalostí z textu (text-mining), ktoré sa snažia tento problém riešiť.

Znalosť jazyka a gramatiky konkrétneho textu predstavuje základný predpoklad k pochopeniu textu. Pokryť diverzitu všetkých gramatík a jazykov pomocou tradičných metód je obtiažne a časovo náročné. Predspracovanie vstupného textu sa v mnohých prípadoch spolieha na algoritmy pre určenie pôvodného (koreňového) tvaru slova, prípadne algoritmy pre opravu preklepov a gramatických chýb. Pre správnu funkčnosť potrebujú znalosť o jazyku, jeho štruktúre a tvarosloví.

Práca sa zaoberá niekoľkými experimentami: A) klasifikáciou 5 emócií z českého a anglického textu, ktorej riešenie bolo postavené na tradičnom prístupe – malé množstvo dát, použitie klasifikátorov postavených na metóde podporných vektorov (SVM), komplexné predspracovanie vstupného textu, rozšírenie o novo navrhnuté optimalizačné metódy publikované v článku [17] (IF pre rok 2016 = 0,945), ktoré zvýšili presnosť klasifikácie o 11,4%, B) abstrahovaním metódy pomocou Big Data prístupu, kde niektoré metódy optimalizácie nie je možné použiť (napr. metóda eliminácie vstupných parametrov), avšak bez jazykovo závislých častí systému dosahuje až o 11,0% vyššiu presnosť klasifikácie, než tradičný prístup, C) analýza textových dát pomocou metód postavených na hlbokom učení s využitím komplexných štruktúr, kde sa kládol dôraz na absolútnu jazykovú nezávislosť navrhnutej metódy, prekonanie úspešnosti metód súčasného stavu vedy o 0,5% (validované voči výsledkom výskumných tímov Google DeepMind [22] a Facebook Research [1] [7] na verejne dostupnej databáze Yelp reviews database zo súťaže z 1. 9. 2017).

Hlavným prínosom tejto práce je návrh a experimentálne overenie univerzálnej metódy pre automatickú analýzu textových dát, ako aj metodiky pre analýzu dát bez predošlej znalosti jazyka či gramatiky. Výsledky obsiahnuté v tejto práci boli publikované na medzinárodných konferenciách a vedeckých časopisoch s impakt faktorom. Validácia prebiehala na privátnych databázach, ako aj na databázach dostupných pre vedecké účely, pre možnosť porovnania so súčasným stavom vedy a techniky.

# 1 Ciele dizertačnej práce

Práca skúma súčasné metódy, ich možné zvýšenie presnosti vďaka optimalizačným metódam, ako aj nové metódy riešenia problému porozumenia textu s modelovaním kognitívneho správanie človeka pri spracovaní textových dát. Vďaka vývoju nových technológií a zdokonaľovaniu výrobných procesov dochádza k nárastu výpočtového výkonu, čo sa odzrkadľuje aj na smerovaní súčasných metód pre dolovanie znalostí z textových dát. Ich presnosť však stále nedosahuje presnosti človeka.

Hlavným cieľom dizertačnej práce je navrhnúť metódu pre strojové porozumenie neštruktúrovaným textovým dátam, teda bez ohľadu na charakter vstupných textových dát. Návrh metódy bol riadený niekoľkými požiadavkami: 1) dostatočná všeobecnosť, 2) znovu-použiteľnosť pre rôzne jazyky (prirodzené i strojové) bez nutnosti akéhokoľvek zásahu, z čoho vyplýva minimalizácia jazykovo závislých častí, v ideálnom prípade ich úplné odstránenie, a 3) dostatočnú presnosť validovanú na vybranom probléme.

Čiastkové ciele dizertačnej práce je možné rozdeliť do nasledujúcich bodov:

- vytvorenie databáz textových dát vhodných pre overenie a porovnanie úspešnosti súčasných metód,
- vytvorenie metódy pre klasifikáciu textu do viacerých emočných tried – návrh vychádza z tradičných metód, keďže k dispozícií môže byť len relatívne malé množstvo dát,
- návrh metódy trénovania viac-modelovej architektúry pre klasifikáciu viacerých tried s možnosťou implementácie optimalizačných metód,
- návrh optimalizačných metód s cieľom zvýšenia presnosti klasifikácie tradičných metód a minimalizácia možnosti pretrénovania,
- abstrahovanie metódy klasifikácie textov pomocou Big Data prístupu, ktorý zvyšuje pamäťovú a časovú náročnosť,
- vytvorenie optimalizačných metód použiteľných aj pre Big Data problém,
- výskum a vývoj metódy pre porozumenie textu založenej na hlbokom učení, so zameraním na všeobecnosť,
- experimentálne overenie navrhnutých metód na vytvorených databázach textových dát, ako aj experimentálne overenie navrhnutých optimalizačných metód a ich vplyv na úspešnosť klasifikačných metód.

Práca má za cieľ experimentálne overiť funkčnosť navrhnutých metód, pričom sa zameriava na extrakciu jednoduchých informácií prostredníctvom klasifikácie textových dát. V práci sú navrhnuté jazykovo nezávislé optimalizačné metódy a metódy na jazykovo nezávislé porozumenie textovým dátam validované na vybraných konkrétnych problémoch.

# 2 Navrhnuté metódy riešenia

Výskum a vývoj metódy pre porozumenie textu prebiehal vo viacerých krokov: návrh metódy pre klasifikáciu 5 emócií, návrh jazykovo nezávislých optimalizačných metód, zovšeobecnenie metódy prostredníctvom prístupu Big Data, čím došlo k odstráneniu jazykovo závislých častí, návrh optimalizačných metód v tomto systéme, a nakoniec návrh metódy pre porozumenie textu bez znalosti jazyka a gramatiky.

## 2.1 Návrh metódy klasifikácie emócií

Problém dolovania znalostí z textových dát je možné demonštrovať na probléme klasifikácie emócií autora textu. Emóciu autora je často možné pochopiť len z celého kontextu. Automatická extrakcia emócie je zložitá a jednoznačne určiť z pár slov zachytenú emóciu autora nie je možné, zvlášť v prípade sarkazmu a irónie.

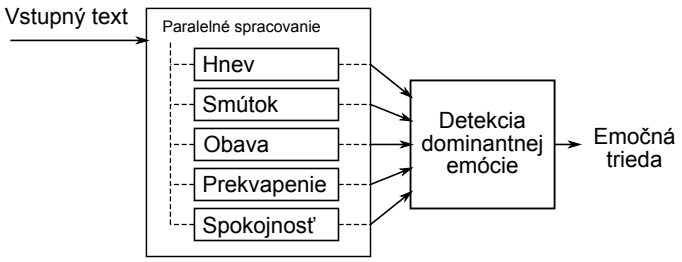
### 2.1.1 Definícia problému

Počas skúmania historických dát z podpory zákazníkov a pri tvorbe nových vzoriek do databáze textov bolo zistené, že najčastejšie sa vyskytujúce emócie v tomto systéme sú: hnev, smútok, spokojnosť, prekvapenie a obava. Tieto emócie sa javia ako dominantné z celkovej množiny. Problémovou emóciou v tomto návrhu je emócia „prekvapenie“, keďže do tejto emócie možno zaradiť aj emóciu „milo prekvapený“. Táto emócia v textoch evokuje šťastie a teda patrí do emočnej triedy „spokojnosť“. Texty s týmto emočným nábojom má však problém správne zaradiť aj človek.

Pre automatickú detekciu nespokojných zákazníkov je nutné navrhnúť metódu na detekciu emócie obsiahnutých vo vstupnom texte a na základe detekovanej emócie určiť prioritu správy. Metóda by mala mať na výstupe jednu z možných detekovaných emócií, resp. mieru istoty, že vstupný text spadá práve do danej triedy. Metóda SVM spadajúca pod tradičné metódy, vytvára nadrovinu, ktorá separuje dve triedy dát.

### 2.1.2 Popis navrhnutej štruktúry

Základnou myšlienkou, je natrénovať jeden model pre každú jednu emočnú triedu. Každý z týchto modelov bude rozhodovať, či vstupný text patrí do danej triedy (pozitívny výsledok) alebo nie (negatívny výsledok) – teda bude separovať dve triedy jednou nadrovinou. Výstupom takejto štruktúry je 5 reálnych čísel reprezentujúcich mieru istoty jednotlivých emočných tried (hnev, smútok, spokojnosť, prekvapenie a obava), z ktorých je následne potrebné určiť dominantnú emóciu vstupného textu. Popisovaná štruktúra sa nachádza na Obr. 2.1.




Obr. 2.1: Štruktúra systému klasifikácie emócií

Funkčnosť systému pozostáva zo spoločného predspracovania vstupného textu, aplikovaním piatich klasifikačných modelov emócií a následnej detekcie dominantnej emócie. Trénovanie celého systému predstavuje komplexný proces, ktorý svojou zložitosťou prináša problémy s nasadením bežných optimalizačných metód, ako sú napr. dopredný výber vstupných parametrov, spätná eliminácia apod.

Algoritmus predspracovania transformuje text na vektor príznakov (parametrov), ktoré vstupujú do klasifikátorov. Text je rozdelený na slová (tokeny), následne sú skontrolované preklepy, doplnená diakritika, slová sú prevedené do pôvodného tvaru, synonymá sú nahradené hypernymom a výstupný text je prevedený pomocou metodiky hodnotenia relevancie TF–IDF na vektor reálnych čísel. [13]

Pokiaľ má byť minimalizovaná časová a pamäťová náročnosť, predspracovanie musí byť aplikované na vstupné dáta len raz. Ak by každý emočný model mal vlastný predspracovanie, náročnosť by vzrástla 5-násobne. Navrhnuté optimalizačné metódy musia byť dostatočne všeobecné a aplikovateľné na rôzne jazyky, aby nebolo potrebné pri tvorení systému pre iný jazyk meniť chod prípadne štruktúru tohto systému.

### 2.1.3 Popis trénovacích a testovacích dát

Dáta boli manuálne zozbierané z reálneho systému podpory zákazníkov, pre český a anglický jazyk. Texty zvyčajne zachytávajú diskusiu k produktom a službám, prípadne technicky zamerané otázky na funkčnosť produktu, možné nastavenia, otázky smerované k existujúcemu manuálu apod. Pre zvýšenie všeobecnosti použitia systému bola databáza rozšírená o frázy z verbálnej komunikácie s obsiahnutým emočným nábojom. Texty boli manuálne označkované jednou z emočných tried prostredníctvom skupiny ľudí, pričom priemerný počet značiek na jeden text bol 1,17. Tab. 2.1 zachytáva počty vzoriek jednotlivých emócií.

Pre trénovanie emočných klasifikátorov bolo potrebné dáta vhodne rozdeliť, zabezpečiť zastúpenie jednotlivých vzoriek pre pozitívnu a negatívnu množinu daného

Tab. 2.1: Veľkosti databáz testovaných jazykov

| Jazyk | Počet vzoriek emócie | | | | |
|---|---|---|---|---|---|
| | hnev | smútok | obava | prekvapenie | spokojnosť |
| Čeština | 351 | 441 | 241 | 56 | 992 |
| Angličtina | 114 | 106 | 104 | 110 | 112 |

klasifikátora – pomocou vyváženia množín. Vytvorené databázy majú rôzne veľkosti (ako zobrazuje Tab. 2.1) a rôzne zastúpenie množín. Je to spôsobené tým, že databáza českých textov vznikala zároveň s definíciou emočných tried. Databáza anglických textov bola vznikla s vedomím o zvolených triedach a potrebou vyvážených množín, s úlohou definovať aspoň 100 vzoriek od každej emócie.

Jednotlivé jazyky potrebujú ďalšie pomocné databázy textov (resp. slov) k predspracovaniu vstupného textu. Tieto databázy sú potrebné pri trénovaní modelu, tak aj pri použití natrénovaného modelu. Model založený na tradičnej metóde vyžaduje na vstupe predspracované dáta v rovnakej forme ako pri trénovaní. Pre rýchly beh sa tieto dáta držia v pamäti, čím sa rapídne zvyšuje pamäťová náročnosť tohto riešenia. Ide o nasledujúce množiny textových reťazcov:

1. databáza všetkých slov daného jazyka pre detekcia preklepov,
2. databáza synoným pre zmenšenie stavového priestoru,
3. databáza stop slov pre odfiltrovanie slov bez významu napr. spojky.

### 2.1.4 Popis predspracovania dát

Predspracovanie textu transformuje vstupný text do vektoru príznakov zastúpení jednotlivých slov. Proces počíta s výskytom slov vo vstupnom texte – je potreba vstupný reťazec znakov rozdeliť na slová (tokeny) pomocou tokenizácie, ktorá sama o sebe nám vnáša podmienku na syntax jazyka vstupného textu.

**Tokenizácia** zabezpečuje rozdelenie textu na jednotlivé slová (tokeny). Pred tokenizáciou je potreba detekovať špeciálne prípady ako e-mailové adresy alebo http adresy (môžu byť z textu odstránené – nesúvisia s emóciou textu), emotikony, čísela a prípadne vulgarizmy (tie nahradiť zastupujúcou značkou). Vyčistený text môže byť potom rozdelený na jednotlivé slová pomocou bielych znakov a znakov interpunkcie.

**Kontrola pravopisu** využíva databázu slov daného jazyka, ako aj všetkých možných tvarov. Každý token je porovnaný s touto databázou. Pokiaľ sa v databáze priamo nachádza, považuje sa za slovo bez chyby. Pokiaľ sa v databáze nenachádza, hľadajú sa slová ktoré majú najmenšiu Levensteinovú vzdialenosť k hľadanému slovu, aby sa chyba opravila, keďže sa môže prejaviť na ďalšom zmätení modelu.

**Lemmatizácia** transformuje slová do slovníkového tvaru. Aplikovaním dochádza k zmenšeniu stavového priestoru, keďže každý tvar slova je nahradený základným tvarom slova, s vždy malým začiatočným písmenom. Pre svoju funkciu využíva databázu tvarov a ich slovníkových tvarov, v ktorej prebieha vyhľadávanie. Často je lemmatizícia nahradzovaná efektívnejšou metódou – stemmingom, ktorý využíva súbor pravidiel pre dopracovanie sa ku koreňu slova.

**Nahradenie synoným** zabezpečuje ďalšie zmenšenie stavového priestoru a teda i dimenzionality vstupu trénovaného modelu. Slová ktoré predstavujú rovnaký význam a teda i emocionálne zafarbenie môžu byť nahradené zástupcom z danej skupiny synoným – hypernymom.

**Stop slová** slúžia na odstránenie slov, ktoré by mohli spôsobiť zmätenie modelu. To by malo za následok nižšiu úspešnosť klasifikácie. Najčastejšie ide o slová, ktoré sa v jazyku vyskytujú často. Samé o sebe však nenesú žiadnu významovú informáciu, zvyčajne v jazyku iba dotvárajú syntax. Do tejto kategórie spadajú aj slová ako predložky, spojky, zámená apod.

### 2.1.5 Popis experimentu

Najvhodnejší klasifikátor pre textové dáta bol nájdený pomocou experimentov na klasifikácií do 2 tried. V čase riešenia sa ako najvhodnejším zdal byť klasifikátor SVM. Vstupné dáta boli v experimente reprezentované váhami slov, teda hodnoty vypočítané pomocou TF–IDF. Experimenty dosahovali na vybranom klasifikátore 87,0 % úspešnosť. Riešenie pomocou viacerých emočných modelov nedosahuje výsledky, ktoré by sa mohli rovnať s človekom. Zvýšiť úspešnosť klasifikácie je možné pomocou optimalizačných metód, ktoré boli navrhnuté pre tento problém.

### 2.1.6 Navrhnuté optimalizačné metódy

Úspešnosť klasifikácie bola zvýšená pomocou navrhnutých optimalizačných metód, ktoré boli publikované v článku [17] (IF pre rok 2016 = 0,945). U každej zmeny v systéme je potreba dáta predspracovať znova. Proces je o to náročnejší, lebo počíta s existenciou viacerých modelov emócií. Navrhnuté metódy však ovplyvňujú aj ostatné modely emócií, nie len model, na ktorý bola daná zmena smerovaná. Toto prepojenie nie je nevyhnutné, ale oddelenie by spôsobilo omnoho vyššiu pamäťovú a výpočtovú náročnosť, keďže jednotlivé databázy by museli byť pre jednotlivé modely oddelené, predspracovanie by tiež bolo nutné spustiť 5x (pre každý emočný model).

### Sekvenčná eliminácia parametrov

Jednotlivé slová vstupného textu majú rôznu váhu v procese emočnej analýzy. Nie všetky parametre vstupu sú však pre umelú inteligenciu dôležité. Niektoré slová sú na vstupe prakticky úplne zbytočné – hlavne tie, ktoré neobsahujú žiaden emočný náboj. Naopak, niektoré slová spôsobujú len zmätenie modelu a znižujú výslednú úspešnosť modelu. Tieto slová môžu byť eliminovanú už pri predspracovaní.

Metóda sekvenčnej eliminácie parametrov je založená na známej a rozšírenej metóde spätnej eliminácie parametrov [19]. Metóda znižuje redundanciu modelu, kde na vstupe sú ponechané len najdôležitejšie parametre. V každom kroku sekvenčnej eliminácie je vybraný jeden parameter (jedno slovo), ktoré spôsobuje najväčšie zmätenie emočného modelu. Eliminácia tohto parametru môže zvýšiť presnosť daného modelu, ale zároveň znížiť presnosť iného emočného modelu. Pre tento prípad je každý eliminovaný parameter otestovaný výsledným modelom a stanovovania výslednej presnosti celého systému. Eliminované slovo je ponechané v databáze stop slov v prípade, ak je presnosť zvýšená o stanovenú hranicu. Tá je postupne znižovaná aby bolo slovo eliminované hneď v prvom cykle, je potrebné aby malo čo najväčší kladný dopad na celý systém.

### Zoskupovanie slov

Metóda používa definovaný slovník pre každú skupinu, teda slovníky sú jazykovo závislou časťou a musia byť manuálne definované pre daný jazyk. Skupina je definovaná pomocou niekoľkých slov, ktoré budú nahradené identifikačným tokenom skupiny. O slovách v skupine je možné povedať, že majú rovnaký význam – aspoň z emočného hľadiska.

Metóda znižuje dimenzionalitu vstupu vďaka nahradzovaniu slov s rovnakým emočným nábojom za jeden skupinový identifikátor. Podobný princíp je aplikovaný v metóde nahradzovania synoným, s tým rozdielom, že skupiny sú volené priamo pre konkrétny problém. Znížením počtu príznakov je možné predísť pretrénovaniu umelej inteligencie. Kompletný zoznam vytvorených skupín je v tab. 2.2.

### Rozširovanie o staticky definované vzorky

V praxi je nevyhnutné, aby bolo možné systém doladiť. Častokrát pri dosiahnutí vyššej úspešnosti bol skutočný pocit z klasifikovania rozdielny voči očakávanému. Frázy použité touto metódou sú založené na spätnej väzbe užívateľov, boli skontrolované človekom (zamestnancom zákazníckeho servisu) a následne vložené do „ladiacej databázy“ s označenou jednoznačne definovanou emóciou. Množiny fráz sú pri trénovaní spracovávané samostatne – mimo ostatné vzorky, aby bolo zabezpečené, že všetky fráze budú použité pri ďalšom trénovaní. Rovnako môžu byť definované aj

Tab. 2.2: Veľkosti definovaných skupín príznakov

| **Skupina** | **Počet príznakov** | |
|---|---|---|
| | Čeština | Angličtina |
| Smútok | 13 | 43 |
| Spokojnosť | 33 | 88 |
| Prekvapenie | 2 | 74 |
| Obava | 9 | 59 |
| Vulgárne 1 | 21 | 0 |
| Vulgárne 2 | 81 | 0 |
| Vulgárne 3 | 45 | 0 |

vety bez emočného náboja, teda tzv. neutrálne vzorky. Tieto sa použijú pri zostavovaní trénovacích množín len v negatívnej časti.

## 2.2 Klasifikácia pomocou „Big Data“

Zovšeobecnenie metódy pre automatickú analýzu textových dát je založené na odstránení jazykovo závislých častí predspracovania a nahradenie týchto častí iným inteligentným systémom, ktorý bude schopný pochopiť základné závislosti v textových dátach (napr. negáciu) ale aj komplexné štruktúry (napr. frazeologizmy, ióniu, sarkazmus), určiť váhu určitých slov a slovných spojení, a vo výsledku extrahovať znalosť – emóciu autora textu.

Omnoho komplexnejšiu analýzu je možné vykonať pomocou distribuovaných výpočtov a prístupu Big Data. Zozbieraním veľkého objemu dát je možné dosiahnuť podobných výsledkov bez nutnosti komplexného predspracovania. Rozdelením výpočtového problému na viac výpočtových staníc je možné znížiť celkový čas výpočtu na prijateľnú úroveň. Pre zvýšenie úspešnosti však tradičné optimalizačné metódy (napr. dopredný výber vstupných parametrov, postupná spätná eliminácia) u Big Data prístupu nie je možné použiť vzhľadom na časovú náročnosť a veľký počet vstupných parametrov.

### 2.2.1 Popis trénovacích a testovacích dát

Vytvorenie databázy s vysokým počtom záznamov je pri návrhu metódy založenej na „Big Data“ prístupe potrebné vytvoriť automaticky. Najjednoduchšou možnosťou je zozbierať dáta z dostupných webových diskusií, kde je text naviazaný na číselné hodnotenie priamo autorom textu (využité pri určení triedy). Takto vytvorené databázy sú tiež k dispozícií pre vedecké účely od spoločností ako sú Google, Yelp alebo

Amazon, a sú používané výskumnými tímami na celom svete.

Databáza s celkovým počtom presahujúcim 2,4 milióna vzoriek bola vytvorená so zameraním na 4 rôzne jazyky: český, anglický, nemecký a španielsky (počet v Tab. 2.3). V experimente bolo použité vyváženie množín dát, takže zo zozbieraných dát bolo nakoniec použitých len niečo cez 850-tisíc vzoriek. Práca s takýmito počtami dát však predstavuje obrovské nároky na operačnú pamäť.

Tab. 2.3: Veľkosti zozbieraných databáz testovaných jazykov

| **Jazyk** | **Celkový počet vzoriek** | **Počet použitých vzoriek** | |
|---|---|---|---|
| | | pozitívny | negatívny |
| Čeština | 675 325 | 99 222 | 99 222 |
| Angličtina | 1 176 316 | 235 334 | 235 334 |
| Nemčina | 224 911 | 50 418 | 50 418 |
| Španielčina | 359 770 | 40 460 | 40 460 |

### 2.2.2 Popis predspracovania dát

Predspracovanie dát je v porovnaní s predošlým experimentom omnoho jednoduchšie – pozostáva len z tokenizácie. Všetky jazykovo závislé časti predspracovania boli odstránené, čím bola minimalizovaná jazyková závislosť. Každý token (slovo) je vstupným parametrom SVM klasifikátoru, kde hodnota je daná metódou TF–IDF. Slová na vstupe klasifikátoru sú v tvare, v ktorom boli použité vo vstupnom texte. Vďaka tomuto prístupu je možné navrhnutý systém aplikovať na iné jazyky bez úpravy predspracovania. Týmto sa však stavový priestor značne zväčšil, keďže každý tvar slova predstavuje nový parameter. Tento problém je potreba riešiť dodatočne, pomocou optimalizačných metód.

Vstupné parametry boli obohatené o vytvorené $n$-gramy (metóda predstavená v roku 1992 v článku [3], prípadne [4]), ktoré do systému vnášajú informácie o zastúpení slovných spojení v jednotlivých triedach. Parametre $n$-gramov značne zvyšujú dimenzionalitu vstupu. Napr. ako znázorňuje Tab. 2.4 pre český jazyk, počet vstupných parametrov bol zvýšený zo 164 tisíc na 2,1 milónu použitím 2-gramov a na 6 miliónov pri použití 3-gramov. Z tohto dôvodu bola implementovaná minimálna frekvencia v dokumentoch (MDF). Určením hodnoty MDF je možné odfiltrovať $n$-gramy, ktoré sa v databáze textov nevyskytli aspoň v danom počte (skúmané boli hodnoty 12 až 32, na databáze 198 tisíc vzoriek). Správnou konfiguráciou je možné odstrániť aj slová s preklepmi a iné anomálie. Vďaka tejto metóde je možné počet parametrov dostať na nižšiu úroveň – desiatky až stovky tisíc parametrov aj u vyššej hodnoty $n$.

Tab. 2.4: Počet príznakov v závislostí na úrovni $n$-gramu pre český jazyk

| $n$ | **Počet unikátnych príznakov** | |
|---|---|---|
| | pozitívnych | negatívnych |
| 1 | 164 443 | 164 349 |
| 2 | 2 109 984 | 2 105 152 |
| 3 | 6 049 007 | 6 031 760 |

### 2.2.3 Popis experimentu

Výpočtová úloha je rozdelená na viacero výpočtových jednotiek, vďaka čomu je možné trénovať klasifikátor pomocou veľkých objemov dát (Big Data) v horizonte pár minút a tak validovať väčšie množstvo konfigurácií. Veľký objem dát je rozdelený medzi výpočtové jednotky, kde každá jednotka pracuje len s určitou podmnožinou. Tradičný klasifikátor SVM nie je možné v prípade distribuovaného trénovania použiť, je potrebné použiť distribuovanú variantu tohto klasifikátoru. Implementácií je v súčasnej dobe dostupných veľa, vhodným môže byť napr. kaskádové SVM. [18]

Vzhľadom na odstránenie krokov predspracovania, na dosiahnutie vyššej úspešnosti je potreba zvoliť vhodnú optimalizačnú metódu. Tradičné optimalizačné metódy nie je možné použiť kvôli vysokému počtu vstupných parametrov a časovej náročnosti trénovania. Samotný beh trénovania a optimalizácie takéhoto systému prostredníctvom napr. postupnej spätnej eliminácie by zabral niekoľko desiatok rokov. Je preto potreba navrhnúť sofistikovanejšiu metódu na voľbu vhodných parametrov.

### 2.2.4 Optimalizácia genetickým programovaním

Navrhnutá optimalizácia genetickým programovaním (GP) využíva chromozómy, ktorých štruktúra obsahuje množinu povolených parametrov (selekcia príznakov, ktoré môžu byť na vstupe klasifikátoru), úroveň $n$-gramu a hodnotu MDF. Prvá populácia je vygenerovaná náhodnou voľbou úrovne $n$-gramu v medziach 2 až 3 a konštantnej hodnoty parametru MDF. U každého génu je náhodne vybraných 99,5% povolených parametrov z celkovej množiny. Ohodnotenie prebieha na základe výsledného klasifikátoru s danou konfiguráciou, vhodnosť génu je reprezentovaná úspešnosťou klasifikácie na testovacej množine. GP využíva turnajovú selekciu.

Zmeny v obsahu chromozómov v danej populácií v behu GP boli dosiahnuté pomocou 6 navrhnutých operátorov. Operátory určené na zmenu povolených parametrov (slov) majú prístup ku kompletnej množine parametrov pre určenú konfiguráciu (množina je závislá na úrovni $n$-gramu), ktorú následne využívajú pri svo-

jom behu. Mutačný *operátor zmeny počtu povolených parametrov* môže meniť počet povolených parametrov. Rovnaký účel má mutačný *operátor nahradenia časti množiny*, ktorý nahradzuje časť množiny (percentuálne vyjadrenú) náhodne zvolenými parametrami z kompletnej množiny. Reprodukčný *operátor výmeny časti množiny* povolených parametrov využíva princíp vytvorenia nového chromozómu s časťami množín dvoch rôznych chromozómov v danej generácií. Nový chromozóm bude mať časť množiny povolených parametrov z prvého chromozómu a časť z druhého chromozómu. Pre zmenu parametru MDF bol navrhnutý *operátor zmeny hodnoty MDF*, ktorý mení túto hodnotu v nastavených medziach (1 až 32). *Operátor zmeny úrovne n-gramu* pracuje v nastavených medziach (2 až 3) pre zmenu úrovne $n$-gramu. Tento operátor má však vplyv na celkovú množinu povolených parametrov (slov), takže takouto mutáciou vznikajú takmer úplne nové chromozómy – s minimálnou podobnosťou k pôvodnému mutovanému chromozómu. *Operátor výmeny množín* zabezpečuje vzniku dvoch nových chromozómov, kde nové chromozómy majú vymenené množiny povolených parametrov medzi sebou. Táto výmena môže nastať len u operátorov s rovnakou úrovňou $n$-gramu.

## 2.3 Hlboké učenie pre klasifikáciu textu

Znalosť jazyka a gramatiky je základným predpokladom k pochopeniu textu a následnému extrahovaniu znalosti. Súčasné „state-of-the-art“ systémy využívajú modely zhlukovania slov (v angličtine označované ako „embedding models“, viac informácií v [9]), ktoré prevádzajú každé slovo z textu na vektor využívaný na vstupe hlbokej neurónovej siete (ďalej len DNN). Vytvoriť model zhlukovania slov pre každý jazyk predstavuje úlohu náročnú na dáta a výpočtové zdroje, navyše použitie takéhoto modelu v procese klasifikácie textu predstavuje obrovské nároky na operačnú pamäť (stovky MB).

Kognitívne správanie človeka pri čítaní textu sa však líši od súčasných metód automatického spracovania textu. Zvyčajne sú slová prevádzané na vektor príznakov a následne algoritmy strojového učenia extrahujú požadovanú znalosť. Prekážkou v tomto procese je nutnosť znalosti daného jazyka – či už vo forme komplexného predspracovania alebo vo forme modelu zhlukovania slov. Bez znalosti gramatiky a syntaxe daného jazyka nie je možné dosiahnuť pomocou súčasných systémov úspešnosť približujúcu sa človeku, a teda ani úplne jazykovo nezávislé riešenie.

Navrhnutá metóda postavená na hlbokom učení sa približuje k procesu porozumenia textových dát človekom – zo surových dát je pochopená typografia daného jazyka, sú detekované začiatky a konce slov, písmená a typografické príznaky sú formované do slabík a slov, slová do slovných spojení, následne do viet, z ktorých je potom pochopená hľadaná znalosť. Navrhnutá metóda teda simuluje celý proces

čítania a pochopenia textu človekom. Za týmto účelom bola navrhnutá štruktúra DNN, ktorá využíva niekoľko konvolučných vrstiev a komplexných štruktúr.

### 2.3.1 Popis trénovacích a testovacích dát

Pre overenie navrhnutej metódy bolo použitých 7 rôznych databáz textov. Prvých 5 databáz predstavujú dáta z predošlých experimentov – množiny textových dát zhromaždené z webových zdrojov, pre 5 rôznych jazykov: angličtina, nemčina, čeština, španielčina a čínština. Sumarizácia zozbieranej databázy je v Tab. 2.5. Ďalšie 2 databázy boli použité pre možnosť objektívneho porovnania so súčasnými metódami: 1) Databáza hodnotení Yelp[1] získaná z desiateho kola súťaže „Yelp challenge“ 1. Septembra 2017, 2) Databáza hodnotení Amazon[2] bola vytvorená projektom „Stanford Network Analysis Project (SNAP)“ zozbieraním hodnotení a užívateľských diskusií z webu Amazon.com do júla 2014, ako popisuje článok [11]. Ako znázorňuje Tab. 2.6, opäť ide o „Big Data“ problém. Navrhnutý systém však využíva masívny paralelizmus pomocou grafických kariet a dávkové spracovanie dát.

Tab. 2.5: Veľkosti zozbieraných databáz testovaných jazykov pre hlboké učenie

| **Jazyk** | **Celkový počet vzoriek** | **Počet použitých vzoriek** | |
|---|---|---|---|
| | | pozitívny | negatívny |
| Čeština | 675 325 | 99 222 | 99 222 |
| Angličtina | 1 176 316 | 235 334 | 235 334 |
| Nemčina | 224 911 | 50 418 | 50 418 |
| Španielčina | 359 770 | 40 460 | 40 460 |
| Čínština | 1 571 156 | 30 305 | 30 305 |

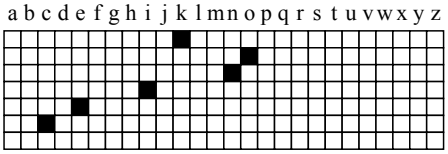
Tab. 2.6: Veľkosti testovaných verejne dostupných databáz

| **Databáza** | **Yelp** | **Amazon** |
|---|---|---|
| Celkový počet | 4 736 897 | 82 456 877 |
| 1 hviezdička | 639 849 | 6 702 809 |
| 5 hviezdičiek | 1 988 003 | 49 006 613 |
| **Použité vzorky** | **1 277 462** | **13 405 618** |

[1]Verejne dostupná pre vedecké účely na adrese: https://www.yelp.com/dataset
[2]Dostupná pre vedecké účely na adrese: http://jmcauley.ucsd.edu/data/amazon/

### 2.3.2 Prevod vstupných dát

Čítaním textu písmeno za písmenom a zakódovanie každého znaku do vektoru je inšpirované prácou [23], kde bola predstavená konverzia textového vstupu. Proces je postavený na kódovaní „one-hot“. Toto môže byť optimalizované vypustením bielych a špeciálnych znakov, a ich zakódovaním pomocou nulového vektoru. Príklad kódovania textu je znázornený na Obr. 2.2. Prístup je jazykovo nezávislý – jedinou jazykovo závislou časťou je abeceda, ktorá sa dá relatívne jednoducho extrahovať.




Obr. 2.2: Slovo „koniec“ reprezentované pomocou matice na vstupe

Obr. 2.2 zobrazuje slovo „koniec“ na vstupe DNN. Prvý riadok reprezentuje písmeno „k“, druhý písmeno „o“ apod. Posledný riadok reprezentuje prázdny vektor (prípade bielych znakov, špeciálnych znakov, znakov ktoré neboli v abecede definované, prípadne text nie je dostatočne dlhý). Prvé konvolučné vrstvy transformujú dáta do kanálov, kde dochádza k detekcií slabík, krátkych slov, začiatkov a koncov slov. Vďaka tomuto má neurónová sieť kontakt s väčšou koncentráciou informácií, než pri použití modelu zhlukovania slov.

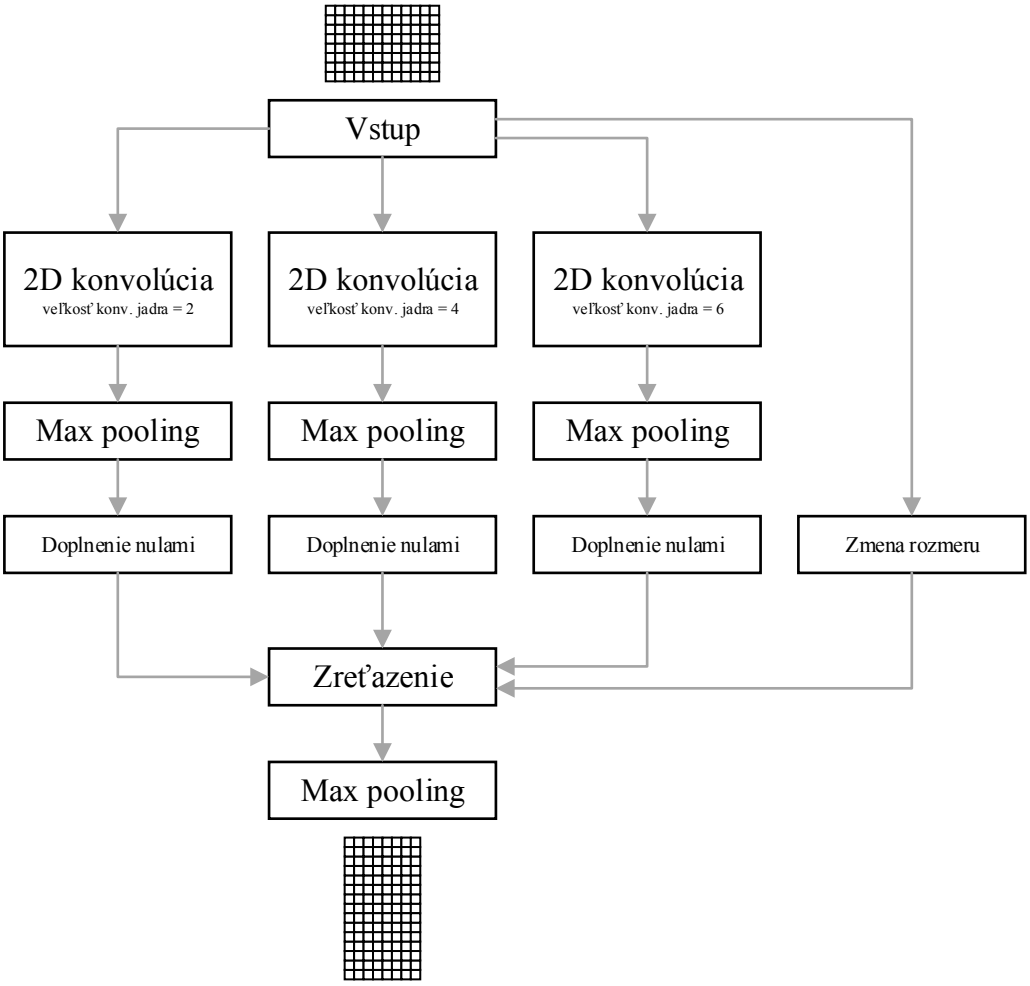
### 2.3.3 Návrh RCK jadra

Pri návrhu štruktúry DNN sa vychádzalo z existujúcich štruktúr používaných pri obrazovom spracovaní. Ide o siete VGG (článok [7]), Inception (článok [20]), viacstĺpcovú architektúru popísanú v článku [6] a reziduálne učenie popísané v článku [8]. V navrhnutej štruktúre DNN pre klasifikáciu textov boli identifikované opakujúce sa štruktúry, ktoré boli vyextrahované a označené ako „opakujúce sa jadro“ – *Recurring Kernel* (RCK), ktoré nesú rovnaké informácie rôznymi vetvami siete do hlbších vrstiev siete, čo umožňuje neurónovej sieti pohľad na text prostredníctvom rôznej miery abstrakcie.

Toto jadro je navrhnuté tak, aby získalo čo najviac informácií dostupných v danej hĺbke siete. Zapojením niekoľkých jadier za sebou je možné vytvoriť hlbokú architektúru, ktorá zároveň zabezpečuje vysokú informačnú priepustnosť. Dokonca najhlbšie vrstvy môžu mať prístup k dátam z prvých vrstiev. Takéto siete môžu byť trénované priamo pomocou algoritmu *Stochastický gradientný zostup – Stochastic*

*Gradient Descent* (SGD) [2]. Všetky dáta a parciálne informácie sú nesené do najnižších vrstiev, kde pomocou plne prepojených vrstiev môže dôjsť ku klasifikácií.

Použitie konvolučných vrstiev vytvára v jednotlivých úrovniach siete „odtlačky“ informácií v každom stĺpci tejto štruktúry. Tieto dátové odtlačky sú zozbierané a rozšírené o reziduálnu časť – o vstup daného jadra. Všetky výstupy sú považované za samostatné dátové kanály. Keďže ide o pamäťovo náročnú operáciu, pred výstupom dát z jadra dôjde k ich podvzorkovaniu. Túto operáciu je možné vykonať vďaka tomu, že niektoré dáta sú vďaka vytvoreným odtlačkom duplikované. Štruktúra RCK jadra je znázornená na Obr. 2.3.




Obr. 2.3: Štruktúra RCK jadra

# 3 Overenie navrhnutých metód

Funkčnosť navrhnutých metód bola overená na vybraných príkladoch: Optimalizačné metódy boli overené na klasifikácii 5 emočných tried, kde boli aplikované na tradičný prístup – relatívne malé množstvo dát, použitie jednoduchých klasifikátorov postavených na metóde SVM, komplexné predspracovanie vstupných dát. Abstrahovaná metóda založená na Big Data prístupe bola overená na probléme klasifikácia do dvoch tried, ako aj optimalizačná metóda pre tento prístup. Metóda postavená na hlbokom učení s novo navrhnutým jadrom pre účely analýzy textových dát bola overená na voľne dostupných databázach prístupných vedeckej verejnosti.

Všetky experimenty popísané v tejto kapitole prebiehali na osobnom počítači s procesorom Intel Core i7–2600K s frekvenciou 3,4GHz, 32GB operačnej pamäte, a grafickým akcelerátorom GeForce 1080Ti s 12GB operačnej pamäte. Pri experimente s Big Data prístupom bola využitá učebňa s 24 výpočtovými jednotkami pre distribuované výpočty – rôzne konfigurácie, prevažne však Intel Core i5 a 8GB operačnej pamäte.

## 3.1 Tradičné metódy a ich optimalizácie

Prvým krokom pri návrhu metódy klasifikácie emócií vola voľba vhodného existujúceho klasifikátoru. Experimenty prebiehali na klasifikácií dvoch tried pre porovnanie rôznych klasifikátorov. Úspešnosť klasifikácie bola relatívne vysoká pre klasifikátor SVM – s hodnotou 87,00 % ako naznačuje tab. 3.1. Experiment prebiehal na databáze so 7000 vzorkami (5000 vzoriek určených pre trénovanie, 2000 vzoriek pre testovanie). Databáza bola vytvorená automaticky z textov hodnotenia produktov a komentárov s hodnotením, na základe ktorého bolo rozhodnuté, či ide o pozitívny alebo negatívny text.

Na základe týchto výsledkov bol pre ďalšie experimenty zvolený práve klasifikátor SVM. Základnou myšlienkou bolo natrénovanie jednoduchého modelu pre každú jednu emočnú triedu tak, že každý bude rozhodovať či vstupný text patrí do danej triedy alebo nie, resp. s akou pravdepodobnosťou. Pre overenie bola využitá druhá polovica databázy – 1673 vzoriek pre český jazyk, 279 vzoriek pre anglický jazyk. Niektoré metódy boli testované prevažne len pre český jazyk z dôvodu časovej náročnosti tejto operácie. Všetky uvedené výsledky boli prijaté vedeckou komunitou a publikované v časopise RadioEngineering (IF pre rok 2016 = 0,945) [17] a na medzinárodných konferenciách – články [13] a [14], prípadne v článku [16] v slovenskom jazyku. Tab. 3.2 zobrazuje počiatočnú úspešnosť navrhnutého riešenia pred aplikovaním optimalizačných metód.

Tab. 3.1: Porovnanie tradičných metód na klasifikáciu textových dátach

| | Skutočný neg. | Skutočný pos. | Predikcia triedy |
|---|---|---|---|
| **FLM klasifikátor** | | | |
| Predikovaný neg. | 799 | 133 | 85.73% |
| Predikovaný pos. | 201 | 867 | 81.18% |
| Vyťaženosť triedy | 79.90% | 86.70% | **83.30%** |
| **Random Forest klasifikátor** | | | |
| Predikovaný neg. | 663 | 696 | 48.79% |
| Predikovaný pos. | 337 | 304 | 47.43% |
| Vyťaženosť triedy | 66.30% | 30.40% | **48.35%** |
| ***k*-NN klasifikátor** | | | |
| Predikovaný neg. | 993 | 976 | 50.43% |
| Predikovaný pos. | 7 | 24 | 77.42% |
| Vyťaženosť triedy | 99.30% | 2.40% | **50.85%** |
| **SVM klasifikátor** | | | |
| Predikovaný neg. | 867 | 127 | 87.22% |
| Predikovaný pos. | 113 | 873 | 86.78% |
| Vyťaženosť triedy | 86.70% | 87.30% | **87.00%** |

Tab. 3.2: Presnosť klasifikácie pred aplikovaním optimalizačných metód

| **Presnosť modelu emócie** | | | | | **Presnosť** |
|---|---|---|---|---|---|
| obava | smútok | spokojnosť | hnev | prekvapenie | **klasifikácie** |
| **Český jazyk** | | | | | |
| 91,11% | 73,03% | 71,54% | 70,00% | 71,74% | **75,49%** |
| **Anglický jazyk** | | | | | |
| 79,81% | 75,44% | 56,73% | 69,09% | 76,85% | **71,58%** |

### 3.1.1 Optimalizácia sekvenčnou eliminációu parametrov

Niektoré kroky sekvenčnej eliminácie pre český jazyk sú zachytené v Tab. 3.3, v poslednom stĺpci je presnosť v danom kroku. Tabuľka popisuje správanie pri jednotlivých elimináciách parametru a ich prípadné zapísanie do databázy eliminovaných parametrov. V tabuľke je táto akcia zobrazená pomocou zaškrtnutia pri danom slove.

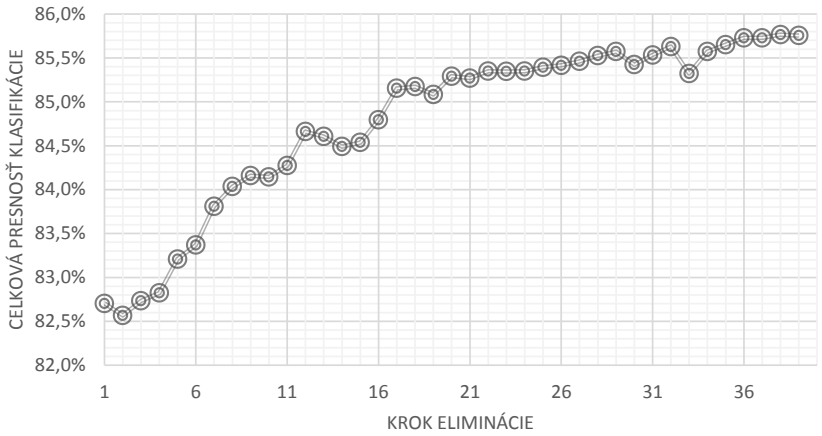
Eliminované parametre ovplyvňujú úspešnosť klasifikácie všetkých modelov, keďže je medzi nimi vzťah[1]. Eliminácia parametru môže zvýšiť presnosť jedného emočného modelu, ale zároveň znížiť presnosť iného. Toto správanie je vidno napr. u 3. kroku, kde došlo k zvýšeniu presnosti u modelov emócií hnev a spokojnosť, ale k zníženiu úspešnosti klasifikácie emócie obava. V 5. kroku (teda po eliminovaní ďalších dvoch parametrov) sa však tento emočný model dostal späť na svoju pôvodnú hodnotu.

[1]Závislosť medzi modelmi emócií bola vysvetlená v kap. 2.1.6 – ide o optimalizáciu pamäťovej a výpočtovej náročnosti, keďže jednotlivé databázy by museli byť upravované pre emočné modely oddelene, predspracovanie by tiež bolo nutné spustiť 5x (pre každý emočný model).

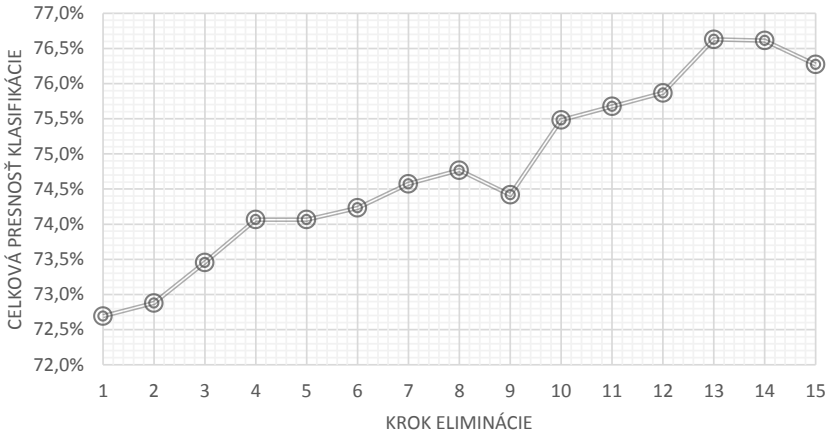
Tab. 3.3: Vývoj hodnôt presnoti klasifikácie českého jazyka v priebehu eliminácie

| **It.** | **Presnosť modelu emócie** obava smútok spokojnosť hnev prekvapenie | **Eliminovaný parameter** | **Presnosť klasifikácie** |
|---|---|---|---|
| 1 | 71.43% 92.08% 79.77% 81.66% 88.57% | | 82.70% |
| 2 | 71.43% 92.08% 79.09% 81.66% 88.57% | efekt | 82.57% |
| 3 | 69.64% 92.92% 80.23% 82.30% 88.57% | sem ✓ | 82.73% |
| 4 | 69.64% 92.92% 80.68% 82.30% 88.57% | akční ✓ | 82.82% |
| 5 | 71.43% 92.92% 80.68% 82.73% 88.29% | dát ✓ | 83.21% |
| ... | ... | ... | ... |
| 20 | 76.79% 93.75% 82.73% 82.62% 90.57% | pletka ✓ | 85.29% |
| 21 | 76.79% 93.75% 82.73% 82.52% 90.57% | esej | 85.27% |
| 22 | 76.79% 93.75% 82.73% 82.62% 90.86% | pokračování ✓ | 85.35% |
| 23 | 76.79% 93.75% 82.50% 82.84% 90.86% | dávno | 85.35% |
| 24 | 76.79% 93.75% 82.73% 82.62% 90.86% | dement | 85.35% |
| 25 | 76.79% 93.75% 82.95% 82.62% 90.86% | celosvětový ✓ | 85.39% |
| ... | ... | ... | ... |
| 36 | 76.79% 94.17% 83.18% 83.37% 91.14% | dojít ✓ | 85.73% |
| 37 | 76.79% 94.17% 83.18% 83.37% 91.14% | exaktní | 85.73% |
| 38 | 76.79% 94.17% 83.18% 83.26% 91.43% | drobek ✓ | 85.77% |
| 39 | 76.79% 94.17% 82.95% 83.16% 91.71% | inteligentní | 85.76% |

Na ukážku bola vybraná sledovaná eliminácia ktorá popisujú možné situácie – elimináciu z etapy 4, u ktorej došlo k zvýšeniu presnosti o 3,07% pre český jazyk. Priemerné zlepšenie úspešnosti pomocou sekvenčnej eliminácie bolo stanovené na 1,69%, ktoré bolo určené zo 6 meraných etáp pri nasadzovaní systému klasifikácie emócií. Pre anglický jazyk došlo k zlepšeniu úspešnosti klasifikácie o 3,58% na sledovanej elimináciíi, pričom výsledná úspešnosť dosahuje 76,63%. Priebehy elimináciíi pre oba jazyky, so všetkými krokmi eliminácie, sú znázornené na Obr. 3.1 pre český jazyk a na Obr. 3.2 pre anglický jazyk.




Obr. 3.1: Priebeh sekvenčnej eliminácie pre klasifikáciu českého jazyka




Obr. 3.2: Priebeh sekvenčnej eliminácie pre klasifikáciu anglického jazyka

### 3.1.2 Optimalizácia metódou zoskupovania slov

Druhou meranou optimalizačnou metódou je metóda zoskupovania slov. Táto kompresia vstupných parametrov pomáha zabraňovať pretrénovaniu modelu, vďaka čomu je možné dosiahnuť vyššiu úspešnosť klasifikácie v reálnej prevádzke. Počty slov v jednotlivých definovaných skupinách sú znázornené v Tab. 2.2.

Tab. 3.4: Dopad optimalizácie zoskupovania slov na presnosť klasifikácie

| **Jazyk** | **Presnosť modelu emócie**<br>obava smútok spokojnosť hnev prekvapenie | **Presnosť klasifikácie** |
|---|---|---|
| Bez optimalizácie | | |
| Čeština | 75,00% 95,83% 80,91% 81,98% 89,71% | 84,69% |
| Angličtina | 79,81% 75,44% 56,73% 69,09% 76,85% | 71,58% |
| S optimalizáciou | | |
| Čeština | 82,14% 96,25% 82,05% 83,16% 90,86% | 86,89% |
| Angličtina | 79,81% 75,44% 57,69% 70,00% 76,85% | 71,96% |

### 3.1.3 Optimalizácia rozširovaním trénovacej množiny

Výsledky optimalizácie je možné rozdeliť do 6 testov (rozširovaná v 6 etapách). V každej etape pribudli nové vzorky, čo popisuje aj Tab. 3.5. Po definícií novej množiny textov je spustený proces trénovania a sekvenčnej eliminácie. Na začiatku eliminácie u každej etapy teda dôjde k zníženiu nameranej presnosti klasifikácie, pričom sa očakáva zvýšenie presnosti na konci. Pri jednotlivých testoch je zároveň

uvedená aktuálna veľkosť množiny statických textov. Celkový priebeh eliminácie a úspešnosť celého modelu, ako aj jeho čiastkových častí je uvedený na Obr. 3.3.

Tab. 3.5: Vývoj presností v jednotlivých etapách rozširovania trénovacej množiny

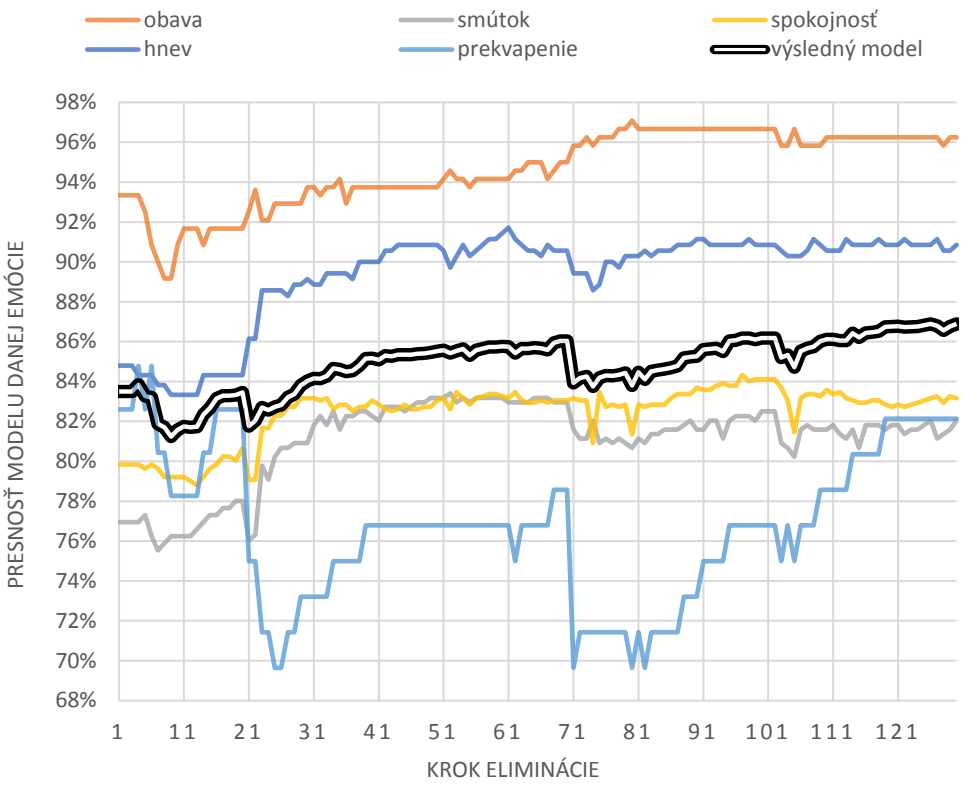
| **Etapa** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Emócia** | **Počet vzoriek rozširujúcej množiny** | | | | | |
| Obava | 46 | 50 | 50 | 50 | 81 | 95 |
| Hnev | 61 | 61 | 63 | 63 | 92 | 103 |
| Smútok | 42 | 42 | 42 | 42 | 61 | 68 |
| Spokojnosť | 29 | 29 | 29 | 29 | 34 | 37 |
| Prekvapenie | 1 | 1 | 4 | 4 | 13 | 16 |
| Neutrál | 1 | 1 | 1 | 1 | 1 | 1 |
| Suma | 180 | 184 | 189 | 189 | 282 | 320 |
| **Presnosť klasifikácie** | | | | | | |
| na začiatku | 75,49% | 78,97% | 81,70% | 82,70% | 83,93% | 85,20% |
| na konci | 76,44% | 80,37% | 83,30% | 85,77% | 86,18% | **86,89%** |
| zlepš. eliminácie | 0,95% | 1,40% | 1,60% | **3,06%** | 2,25% | 1,69% |
| zlepš. v etape | — | 3,93% | 2,93% | 2,47% | 0,41% | 0,71% |

Pre porovnanie bol prevedený experiment, kedy mal človek určiť správnu emóciu vopred označeného textu. Na 200 vybraných textoch z testovacej množiny dosiahol človek úspešnosť 88,50%. Test prebiehal na 5 ľuďoch, pričom každý zúčastnený mal určiť emóciu u 40 textov. Človek teda správne určil emočnú triedu v 177 prípadoch. Určovanie emócií textu človekom je ovplyvňované rôznymi faktormi ako sú únava, emočné a psychické rozpoloženie, kvôli ktorým môže emóciu v danom texte pochopiť inak.

## 3.2 Abstrahovaný systém pre klasifikáciu textu

Tak ako v predošlom prípade, testy boli prevedené pomocou druhej polovice zozbieranej databázy (veľkosť databáz zachytáva Tab. 2.3). Hodnotenie správnosti klasifikácie do 2 tried (pozitívny a negatívny), prah klasifikácie nastavený na minimálnu istotu rovnú 0,5, pričom nebola použitá žiadna zakázaná oblasť.

Pomocou navrhnutého distribuovaného riešenia bolo validovaných 720 rôznych konfigurácií. Každá z konfigurácií bola trénovaná na vytvorenej databáze s veľkým objemom dát (napr. 235 tisíc vzoriek u anglického jazyka). Keďže jedna iterácie môže byť časovo náročná, výpočet bol distribuovaný na 112 výpočtových jednotiek (jadier). Experiment porovnáva rôzne úrovne $n$-gramov, ktoré zvyšujú dimenzionalitu




Obr. 3.3: Priebeh sekvenčnej eliminácie a metódy rozširovania trénovacej množiny pre klasifikáciu emócií českého jazyka

vstupu, ako aj rôzne hodnoty MDF, vďaka ktorým sa dimenzionalita znižuje odfiltrovaním parametrov, ktoré sa v trénovacej množine nevyskytujú v dostatočnom množstve.

Tab. 3.6: Porovnanie najlepších konfigurácií pre klasifikáciu Big Data

| **Jazyk** | **Iterácií** | **MDF** | **Úroveň *n*-gramu** | **Presnosť** |
|---|---|---|---|---|
| Čeština | 10 | 12 | 2 | 89.05% |
| Angličtina | 100 | 32 | 2 | **95.31%** |
| Nemčina | 10 | 12 | 2 | 88.34% |
| Španielčina | 50 | 12 | 2 | 93.23% |

Z výsledkov v Tab. 3.6 vyplýva, že vo všeobecnosti pre klasifikáciu textových dát trénované veľkými objemami dát postačujú 2-gramy. Táto konfigurácia sa zdá byť vhodná pre „Big Data“ prístup, keďže dokáže vyriešiť problém negovania. Úspešnosť klasifikácie dosiahnutá pomocou tri-gramov bola napr. u angličtiny o 1,15% nižšia, u 4-gramov dokonca o 1,61% nižšia. Vzhľadom na tieto výsledky boli v ďalších

experimentoch použité len hodnoty $n$-gramu rovné 2 alebo 3. Pre české texty by bol počet príznakov cez 2,1 milióna, po odfiltrovaní pomocou MDF=12 je počet príznakov už len 67 250. U angličtiny je možné pozorovať iný jav – napriek tomu, že má omnoho nižší počet slov než čeština (len 55 tisíc voči 800 tisíc v českom jazyku), výskyt neodfiltrovaných slovných spojení je omnoho vyšší. Práve z tohto dôvodu vykazuje vyššiu presnosť klasifikácie pri vyšších hodnotách MDF.

Navrhnutý klasifikátor dosahuje najlepšiu úspešnosť klasifikácie na anglickom jazyku, kde presnosť 95,31% predstavuje zvýšenie presnosti až o 11% voči tradičnému prístupu bez použitia „Big Data“. Navrhnuté riešenie je navyše jazykovo nezávislé, predspracovanie je časovo menej náročné a neobsahuje žiadne časti, ktoré by potrebovali úpravu pri nasadení klasifikátoru na iný jazyk.

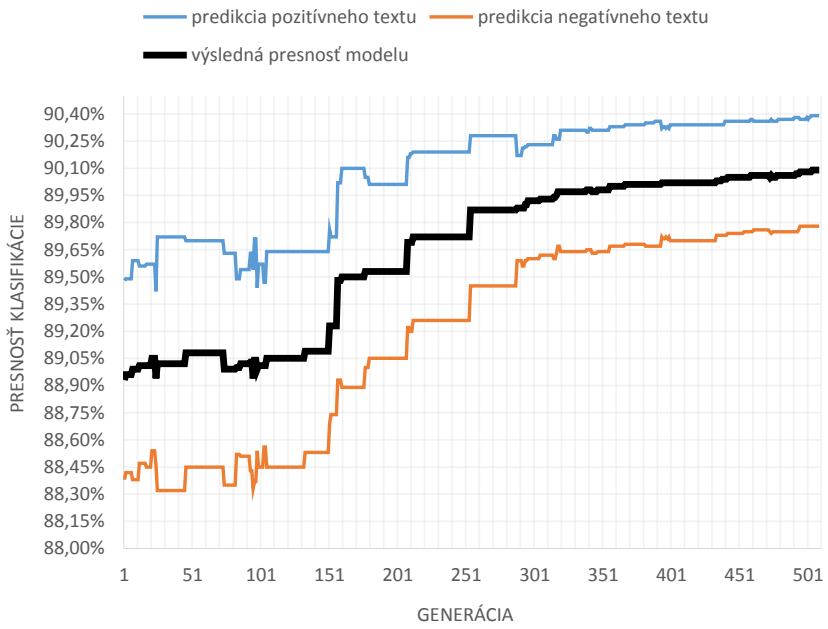
### 3.2.1 Optimalizácia genetickým programovaním

Metóda ma za úlohu nájsť optimálnu konfiguráciu spomínaných dvoch parametrov (úroveň $n$-gramu a hodnota MDF) a podmnožinu povolených vstupných parametrov (slov a slovných spojení), ktoré budú dosahovať na testovacej množine čo najvyššiu presnosť klasifikácie. Ohodnotenie prebieha pomocou druhej polovice vytvorenej databázy textov. Výsledky tejto optimalizačnej metódy boli publikované v článku [15] a demonštrované na medzinárodnej konferencií v Indii. Konfigurácia využívajúca 2-gramy a MDF rovné 12 pre český jazyk a 32 pre anglický jazyk dosahovali najvyššiu úspešnosť v predošlých experimentoch. Zbytok chromozómov do populácie bol vygenerovaný náhodne v rozsahoch 2 až 3 pre úroveň $n$-gramu a 1 až 32 pre hodnotu MDF. Vyššie hodnoty u týchto parametrov nepredstavovali možnosť zlepšenia úspešnosti klasifikácie.

Obr. 3.4 znázorňuje vývoj úspešnosti klasifikácie českých textov v jednotlivých generáciách. Prvá generácia vykazuje úspešnosť 88,93%, čo je o 0,12% nižšie než systém bez selekcie príznakov, pokles je spôsobený znížením počtu vstupných parametrov. Posledná dosiahnutá generácia má poradové číslo 509, kde bola dosiahnutá úspešnosť 90,09%, čo predstavuje zlepšenie presnosti klasifikácie o 1,04% voči predošlému riešeniu.

## 3.3 Hlboká neurónová sieť založená na RCK jadre

Odstránenie všetkých jazykovo závislých častí bolo možné vďaka návrhu založenom na hlbokom učení a novo navrhnutých RCK jadrách. Tieto jadrá disponujú veľkou schopnosťou učenia a vysokou informačnou priepustnosťou. Takéto riešenie nie je vhodné pre jazyky ako je napr. čínština, kde slová nie sú oddelené prostredníctvom medzery. Prvý experiment dokazuje, že navrhnuté riešenie je vhodné pre akýkoľvek




Obr. 3.4: Priebeh úspešnosti klasifikácie pri optimalizácií GP

druh jazyka: **angličtina** – ako zástupca anglo-frízskeho jazyka, množina testovacích textov bola cielená na bežnú komunikáciu, **nemčina** – ako zástupca germánskych jazykov, testuje schopnosť spracovať dlhšie slová a zloženiny, komplexnú gramatiku, **čeština** – ako zástupca slovanských jazykov, zameriava sa na slová s diakritikou a bez, ako aj na časté chyby v pravopise a prípadné preklepy, **španielčina** – ako zástupca románskych jazykov, množina textov je zameraná prevažne na krátke texty, **čínština** – pre dokázanie, že metóda je schopná pracovať s rôznymi abecedami syntaxami, jazyk nespolieha na jednoznačné oddelenie slov. Druhý experiment porovnáva úspešnosť navrhnutého riešenia so súčasnými metódami použitými v iných publikáciach. K týmto účelom boli využité verejné databázy textov – databáza hodnotení Yelp a databáza hodnotení Amazon, ktoré sú používané širokou vedeckou komunitou pri porovnávaní nových navrhnutých metód zameraných na dolovanie znalostí z textových dát. Tab. 3.7 zhrňuje štruktúry navrhnutých sietí. Z vytvorených štruktúr bola hľadaná tá, ktorá dosiahne najvyššiu úspešnosti klasifikácie, kde k porovnaniu slúžila práve testovacia množina dát.

Tab. 3.8 zobrazuje dosiahnuté úspešnosti klasifikácie u rôznych jazykov. Rozptyl presností 12,65% a priemerná úspešnosť klasifikácie 87,71% dokazujú, že metóda je aplikovateľná na rôznych jazykoch. Ako u každej metódy založenej na hlbokom učení, aj tu rozhoduje kvalita vytvorenej databázy a priamo ovplyvňuje možné dosiahnuteľné úspešnosti. Pre španielský a čínsky jazyk boli dosiahnuté najvyššie úspešnosti – 92,46% a 91,22%.

Tab. 3.7: Súhrn navrhnutých štruktúr hlbokých neurónových sietí

| **Č.** | **Parametre RCK jadra** | | **Počet neurónov** | **Trénovateľných** |
|---|---|---|---|---|
| | hĺbka | počet konv. jadier | **plne prepojených** | **parametrov** |
| 1 | 6 | 64, 64, 64, 32, 32, 32 | 256, 128 | 3 311 970 |
| 2 | 4 | 64, 64, 64, 32 | 256, 128 | 6 779 298 |
| 3 | 5 | 64, 64, 64, 32, 32 | 128, 128 | 2 783 490 |
| 4 | 4 | 64, 64, 64, 32 | 128, 128 | 3 793 186 |
| 5 | 5 | 64, 64, 128, 256, 256 | 512, 256 | 27 746 818 |
| 6 | 4 | 64, 64, 64, 64 | 256, 128 | 7 807 362 |
| 7 | 7 | 64, 64, 64, 64, 64, 32, 32 | 256, 128 | 3 761 410 |
| 8 | 7 | 64, 64, 64, 64, 64, 32, 32 | 128, 128 | 3 127 938 |

Tab. 3.8: Výsledky klasifikácie pomocou hlbokej neurónovej siete

| **Jazyk** | **pozitívna trieda** | **negatívna trieda** |
|---|---|---|
| **Čeština** | model 1 s 3 493 602 trénovateľných parametrov | |
| precíznosť | 89,90% | 84,07% |
| senzitivita | 82,81% | 90,69% |
| presnosť klasifikácie | | **86,75%** |
| **Angličtina** | model 8 s 3 296 898 trénovateľných parametrov | |
| precíznosť | 84,08% | 85,45% |
| senzitivita | 85,73% | 83,77% |
| presnosť klasifikácie | | **84,75%** |
| **Nemčina** | model 5 s 28 321 282 trénovateľných parametrov | |
| precíznosť | 85,95% | 81,17% |
| senzitivita | 79,83% | 86,95% |
| presnosť klasifikácie | | **83,39%** |
| **Španielčina** | model 8 s 3 296 898 trénovateľných parametrov | |
| precíznosť | 92,90% | 92,04% |
| senzitivita | 91,96% | 92,97% |
| presnosť klasifikácie | | **92,46%** |
| **Čínština** | model 7 s 3 947 266 trénovateľných parametrov | |
| precíznosť | 92,14% | 90,34% |
| senzitivita | 90,13% | 92,31% |
| presnosť klasifikácie | | **91,22%** |

Výsledky z verejných databáz textov sú zhrnuté v Tab. 3.9, kde sieť č. 8 dosahuje najlepšie výsledky. Sieť pozostáva zo 7 RCK jadier a počet konvolučných jadier nepresahuje číslo 64. Natrénovaný model má 3,3 milióna trénovateľných parametrov, ktoré je možné reprezentovať v pamäti pomocou 24MB dát (o 85% nižšia než u najjednoduchšieho modelu GloVe [12]). Z toho vyplýva, že model je možné použiť aj na jednoduchých vstavaných systémoch (tzv. „embedded“ systémy), ktoré

Tab. 3.9: Výsledky klasifikácie na verejných databázach

| **Databáza** | **Štruktúra č.** | **Úspešnosť** | |
|---|---|---|---|
| | | validačná | testovacia |
| Yelp | 8 | 96.14% | **96.15%** |
| | 2 | 96.10% | 96.05% |
| | 1 | 96.02% | 96.03% |
| Amazon | 3 | 94.74% | **94.59%** |
| | 1 | 94.62% | 94.57% |
| | 2 | 94.19% | 94.02% |

figurujú nízkymi parametrami technického vybavenia – teda aj malou operačnou pamäťou. Modely zhlukovania slov vyžadujú stovky MB – napr. najjednoduchší pred-trénovaný model Glove vytvorený Stanfordskou univerzitou [12] má 171MB.

Metóda je kompletne jazykovo nezávislá, nevyužíva žiadnu pripravenú znalosť o analyzovanom jazyku textu, má porovnateľnú úspešnosť než súčasné „state-of-the-art“ metódy (znázorňuje Tab. 3.10) a je 10-násobne pamäťovo efektívnejšia než riešenia postavené na modeloch zhlukovania slov. U databáze hodnotení Yelp dosahuje dokonca o 0,5% vyššiu úspešnosť klasifikácie, než súčasné metódy. Metóda je tiež vhodná na strojovo generované texty, u ktorých nie je známa gramatika.

Tab. 3.10: Porovnanie navrhnutej metódy so súčasnými riešeniami na databáze Yelp

| **Model** | **Rok, publikácia** | **Úspešnosť** |
|---|---|---|
| Bag of Words | 2015, [23] | 92.2% |
| n-grams | 2015, [23] | 95.6% |
| Char-CNN | 2015, [23] | 94.7% |
| Char-CRNN | 2016, [21] | 94.5% |
| Very deep CNN | 2016, [7] | 95.7% |
| FastText | 2016, [1] | 93.8% |
| FastText with bigrams | 2016, [1] | 95.7% |
| Discriminative LSTM | 2017, [22] | 92.6% |
| Generative LSTM | 2017, [22] | 90.0% |
| RCK č. 8 | 2018, v recenzií | **96.2%** |

# 4 Záver

Práca sa zaoberá problémom dolovania znalostí z textových dát, ktorý je stále aktuálnejší vzhľadom na exponenciálny rast množstva uložených dát v elektronickej podobe. Vďaka súčasným trendom je možné predpokladať zvýšený záujem o spracovanie týchto dát a ich analýzu. Vývoj nového výpočtového technického vybavenia so zameraním na masívny paralelizmus umožňuje túto analýzu značne urýchliť. Hlavným problémom súčasných metód je závislosť na konkrétnom jazyku textu a ich presnosť, ktorá nedosahuje úspešnosti človeka.

V práci bolo navrhnuté riešenie pozostávajúce z tradičnej metódy a jej optimalizácie, abstrahovanie tejto metódy a jej optimalizácií, a vývoja novej metódy pre strojové porozumenie textu bez znalosti jazyka a gramatiky. U tradičných metód bola navrhnutá štruktúra pre detekciu 5 emočných tried, ktorá pozostávala z 5 samostatných klasifikátorov a jedného spoločného predspracovania textu. Pre tento systém boli navrhnuté jazykovo nezávislé optimalizačné metódy s cieľom zvýšiť úspešnosť klasifikácie. Abstrahovaním tejto metódy prostredníctvom odstránenia jazykovo závislejších častí a použitím prístupu „Big Data“ bol dosiahnutý návrh všeobecnejšieho systému, ktorý bol následne optimalizovaný prostredníctvom navrhnutého algoritmu genetického programovania. U tradičných metód nebolo možné dosiahnuť vyššiu mieru jazykovej nezávislosti riešenia (riešenie nebolo vhodné napr. pre čínsky jazyk, ktorého syntax nevyužíva slová), preto ďalší výskum smeroval na nové metódy založené na hlbokom učení. Po návrhu prevodu vstupných textových dát do maticovej podoby bola navrhnutá štruktúra hlbokej neurónovej siete so zameraním na textové dáta. Pri návrhu sa vychádzalo z existujúcich štruktúr často používaných pri obrazovom spracovaní. V navrhnutých štruktúrach boli identifikované opakujúce sa vzory, ktoré boli izolované do novej štruktúry nazvanej „RCK jadro“. Táto bola použitá pri návrhu komplexnejších hlbokých neurónových sietí vďaka jej schopnosti extrakcie informácií, vysokej informačnej priepustnosti a jej kapacite.

Hlavným prínosom tejto práce je navrhnutá univerzálna metóda pre dolovanie znalostí z textových dát, ktorá bola experimentálne overená na vybranom prípade klasifikácie textu, kde bola dosiahnutá vyššia presnosť v porovnaní so súčasnými metódami (prekonanie o 0,5%) a zníženie pamäťovej náročnosti o 85% v porovnaní so súčasnými metódami postavenými na modeloch zhlukovania slov. Výsledky obsiahnuté v tejto práci boli publikované v časopisoch s impakt faktorom [17] (IF pre rok 2016 = 0,945), riešenie založené na novo-navrhnutom RCK jadre bolo v dobe písania práce v recenznom konaní k publikácií v časopise Cognitive Computation (IF pre rok 2016 = 3,441). Čiastkové výsledky boli publikované na medzinárodných konferenciách [13, 14, 10, 15].

# Literatúra


[1] Bojanowski, P.; Grave, E.; Joulin, A.; aj.: Enriching Word Vectors with Subword Information. *Facebook AI Research, ArXiv e-prints*, Červenec 2016, 1607.04606.

[2] Bottou, L.: Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, Springer, 2010, s. 177–186.

[3] Brown, P. F.; Desouza, P. V.; Mercer, R. L.; aj.: Class-based n-gram models of natural language. *Computational linguistics*, ročník 18, č. 4, 1992: s. 467–479.

[4] Cavnar, W. B.; Trenkle, J. M.; aj.: N-gram-based text categorization. *Ann arbor mi*, ročník 48113, č. 2, 1994: s. 161–175.

[5] Chakraborty, G.; Krishna, M.: Analysis of unstructured data: Applications of text analytics and sentiment mining. In *SAS global forum*, 2014, s. 1288–2014.

[6] Cireşan, D.; Meier, U.; Schmidhuber, J.: Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[7] Conneau, A.; Schwenk, H.; Barrault, L.; aj.: Very Deep Convolutional Networks for Text Classification. *Facebook AI Research, ArXiv e-prints*, Červen 2016, 1606.01781.

[8] He, K.; Zhang, X.; Ren, S.; aj.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, s. 770–778.

[9] Lai, S.; Liu, K.; He, S.; aj.: How to generate a good word embedding. *IEEE Intelligent Systems*, ročník 31, č. 6, 2016: s. 5–14.

[10] Masek, J.; Burget, R.; Povoda, L.; aj.: Multi–GPU Implementation of Machine Learning Algorithm using CUDA and OpenCL. *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, ročník 5, č. 2, 2016: s. 101–107.

[11] McAuley, J.; Yang, A.: Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, s. 625–635.

[12] Pennington, J.; Socher, R.; Manning, C.: Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, s. 1532–1543.

[13] Povoda, L.; Arora, A.; Singh, S.; aj.: Emotion Recognition from Helpdesk Messages. In *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, IEEE, 2015, s. 310–313.

[14] Povoda, L.; Burget, R.; Dutta, M. K.: Sentiment analysis based on Support Vector Machine and Big Data. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, 2016, s. 543–545.

[15] Povoda, L.; Burget, R.; Dutta, M. K.; aj.: Genetic Optimization of Big Data Sentiment Analysis. In *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, 2017, s. 141–144.

[16] Povoda, L.; Burget, R.; Mašek, J.; aj.: Automatické rozpoznávanie emócií z českého textu pomocou umelej inteligencie. In *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz)*, 17.1, 2015, s. 15–18.

[17] Povoda, L.; Burget, R.; Masek, J.; aj.: Optimization Methods in Emotion Recognition System. *Radioengineering*, ročník 25, č. 3, 2016: s. 565–572.

[18] Stolpe, M.; Bhaduri, K.; Das, K.: *Distributed Support Vector Machines: An Overview*. Cham: Springer International Publishing, 2016, ISBN 978-3-319-41706-6, s. 109–138, doi:10.1007/978-3-319-41706-6_5.

[19] Sutter, J. M.; Kalivas, J. H.: Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical journal*, ročník 47, č. 1-2, 1993: s. 60–66.

[20] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; aj.: Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, s. 2818–2826.

[21] Xiao, Y.; Cho, K.: Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*, 2016.

[22] Yogatama, D.; Dyer, C.; Ling, W.; aj.: Generative and Discriminative Text Classification with Recurrent Neural Networks. *Google DeepMind, ArXiv e-prints*, Březen 2017, `1703.01898`.

[23] Zhang, X.; Zhao, J.; LeCun, Y.: Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 2015, s. 649–657.

# Curriculum Vitæ

Lukáš Povoda

---

### Osobné informácie


| | |
|---|---|
| Dátum narodenia: | 18. 1. 1990 |
| Miesto narodenia: | Bojnice |
| Telefón: | +420 732 252 572 |
| E-mail: | lupo112@gmail.com |


### Vzdelanie

| | |
|---|---|
| 2014 – 2018 | Doktorské štúdium – Teleinformatika<br>Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno<br>titul: Ph.D. (predpokladaný rok ukončenia: 2018) |
| 2012 – 2014 | Magisterské štúdium – Telekomunikační a informační technika<br>Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno<br>titul: Ing. |
| 2009 – 2012 | Bakalárske štúdium – Teleinformatika<br>Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno<br>titul: Bc. |
| 2005 – 2009 | Úplné stredné odborné vzdelanie – Elektrotechnika<br>Stredná odborná škola Handlová, Lipová 8, 972 51 Handlová |

### Zamestnanie

| | |
|---|---|
| 2013 – súčasnosť | Vedecký pracovník Centra senzorických informačních a komunikačních systémů<br>Vysoké učení technické v Brně, Technická 3058/10, 616 00 Brno |
| 2013 – súčasnosť | Senior PHP Developer<br>a-net group, a.s., Národních hrdinů 3, 190 00 Praha 9 |

## Spolupráca na projektoch

| | |
|---|---|
| 2017 – 2020 | Expertní systém pro automatickou analýzu a řízení big data úložišť výrobních společností<br>MPO reg. č. FV20044 |
| 2017 – 2019 | Hardwarově akcelerovaná identifikace osob z videozáznamů s použitím hlubokého strojového učení<br>MVČR reg. č. VI2VS/554 |
| 2016 – 2016 | Connected homes - thermostat datamining<br>hospodárska zmluva (750 669 CZK), referenčná osoba Ondřej Lipták - Leader of electronic design group at Honeywell, Brno Hardware Center of Excellence |
| 2016 – 2016 | Klasifikace hutních materiálů pomocí spektrometrie laserem indukovaného plazmatu za použití hlubokých neuronových sítí<br>IGA reg. č. FEKT/FSI-J-16-3564 |
| 2015 – 2016 | Pokročilé meteorologické informace pro letectví<br>TAČR reg. č. TH01010503 |
| 2014 – 2015 | Výzkum a vývoj technologie pro detekci emocí v nestrukturovaných datech<br>MPO reg. č. FR-TI4/151 |

## Publikačná aktivita

- Publikácie v časopisoch s impakt faktorom: 1 + 1 (v recenznom riadení)
- Publikácie v časopisoch bez impakt faktoru: 2
- Publikácie v konferenčných zborníkoch: 10
- Software: 3
- Príspevky indexované v databáze WoS: 9
- Príspevky indexované v databáze Scopus: 9
- H-index podľa databáze WoS: 2
- H-index podľa databáze Scopus: 3

## ABSTRAKT

Práca sa zaoberá problémom dolovania znalostí z textových dát, ktorý je stále aktuálnejší vzhľadom na exponenciálny rast množstva uložených dát v elektronickej podobe, kde 80% týchto dát je v textovej podobe. Práca skúma súčasné metódy, ich možné zvýšenie presnosti vďaka optimalizačným metódam, ako aj nové metódy riešenia problému porozumenia textu s modelovaním kognitívneho správanie človeka pri spracovaní textových dát. Problém súčasných metód, ktorým je závislosť na konkrétnom jazyku textu, ako aj ich presnosť, ktorá nedosahuje úspešnosti človeka, rieši prostredníctvom troch smerov: tradičnými metódami a ich optimalizáciami, prístupom Big Data a abstrahovaním prostredníctvom minimalizácie jazykovo závislých častí, a prístupom hlbokého učenia. Hlavným cieľom dizertačnej práce bolo navrhnúť metódu pre strojové porozumenie neštruktúrovaným textovým dátam. Metóda bola experimentálne overená na probléme extrakcie jednoduchých informácií prostredníctvom klasifikácie textových dát v 5 jazykoch – čeština, angličtina, nemčina, španielčina a čínština, čím bola dokázaná možnosť aplikácie na rôzne rodiny jazykov. Pri validácií na databáze hodnotení Yelp bola dosiahnutá presnosť vyššia o 0,5% než poskytujú súčasné metódy.



## ABSTRACT

This work deals with the problem of text mining which is becoming more popular due to exponential growth of the data in electronic form. The work explores contemporary methods and their improvement using optimization methods, as well as the problem of text data understanding in general. The work addresses the problem in three ways: using traditional methods and their optimizations, using Big Data in train phase and abstraction through the minimization of language-dependent parts, and introduction of the new method based on the deep learning which is closer to how human reads and understands text data. The main aim of the dissertation was to propose a method for machine understanding of unstructured text data. The method was experimentally verified by classification of text data on 5 different languages – Czech, English, German, Spanish and Chinese. This demonstrates possible application to different languages families. Validation on the Yelp evaluation database achieve accuracy higher by 0.5% than current methods.