VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁVÁNÍ EMOCÍ Z TEXTU POMOCÍ UMĚLÉ INTELIGENCE

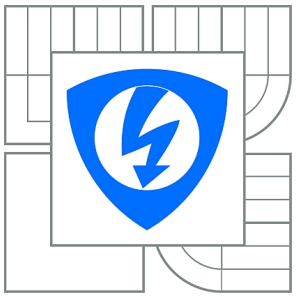
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE RADEK VYLÍČIL
AUTHOR

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁVÁNÍ EMOCÍ Z TEXTU POMOCÍ UMĚLÉ INTELIGENCE

RECOGNITION OF EMOTIONS IN TEXT USING ARTIFICIAL INTELLIGENCE

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

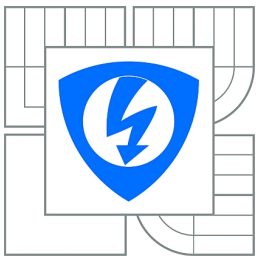
AUTOR PRÁCE — RADEK VYLÍČIL

AUTHOR

VEDOUCÍ PRÁCE — Ing. JAN MAŠEK

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor
Teleinformatika

Student:	Radek Vylíčil	***ID:***	130685
Ročník:	3	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Rozpoznávání emocí z textu pomocí umělé inteligence

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s algoritmy umělé inteligence (např. neuronové sítě) a nastudujte metody trénování algoritmů pro potřeby získávání znalostí z textů. Dle pokynů vedoucího vytvořte trénovací algoritmus v jazyce JAVA a demonstrujte jeho funkčnost na příkladě. Seznamte se také s problematikou ontologických bází a připravte trénovací množinu s ohodnocením. Natrénujte klasifikátor pro rozpoznávání emocí z textů a zhodnoťte dosažené výsledky.

DOPORUČENÁ LITERATURA:

[1] BURGET, R.; KARÁSEK, J.; SMÉKAL, Z. Classification and Detection of Emotions in Czech News Headlines. In The 33rd International Conference on Telecommunication and Signal Processing, TSP 2010. 2010.
[2] R. Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press

Termín zadání:	11.2.2013	***Termín odevzdání:***	5.6.2013

Vedoucí práce: Ing. Jan Mašek
Konzultanti bakalářské práce:

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Tato práce se zabývá rozpoznáváním emocí z textů pomocí strojového učení. V textu jsou popsány metody pro trénování a testování rozpoznávacích modelů. Hlavní přínos této práce spočívá ve vytvořeném algoritmu rozhodovacího stromu v jazyce Java. Vytvořený algoritmus byl integrován jako rozšíření do programu RapidMiner. V tomto programu vzniklo několik vzorových příkladů. Funkčnost byla ověřena na vytvořené databázi dat.

KLÍČOVÁ SLOVA

Rozhodovací strom, SVM, RapidMiner, Zpracování textu, Java

ABSTRACT

This thesis deals with the recognition of emotions from text using machine learning. The text describes methods how to train and test an recognition models. The main contribution of this thesis consists in creation decision tree in Java programming language. Created algorithm was integrated as plugin into the RapidMiner tool. The thesis contains some created examples for executing in RapidMiner. The functionality of decision tree was demonstrated on created database.

KEYWORDS

Decision tree, SVM, RapidMiner, Text processing, Java

VYLÍČIL, Radek *Rozpoznávání emocí z textu pomocí umělé inteligence*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2013. 42 s. Vedoucí práce byl Ing. Jan Mašek

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Rozpoznávání emocí z textu pomocí umělé inteligence“ jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu bakalářské práce panu Ing. Janu Maškovi za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Brno ....

(podpis autora)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této bakalářské práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno ....

(podpis autora)

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

V dnešní době je stav mediálních technologií takový, že umožňují uživateli používat velká množství dat a textů, které v sobě obsahují různé emoce, se kterými byly tvořeny. Bohužel je pro člověka analýza těchto textů velice omezená a časově náročná. Pro získání emocí z textu (Text Mining) jsou využívány různé techniky (např. strojové učení a umělá inteligence). Tato práce bude pojednávat o použití těchto i více metod, za účelem rozpoznání emocí z textu.

Emoce jsou psychické a sociálně konstruované procesy, zahrnující subjektivní zážitky libosti a nelibosti, provázené fyziologickými změnami (změna srdečního tepu, změna rychlosti dýchání), motorickými projevy (mimika, gestikulace), změnami pohotovosti a zaměřenosti. Hodnotí skutečnosti, události, situace a výsledky činností podle subjektivního stavu a vztahu k hodnocenému, vedou k zaujetí postoje k dané situaci [3].

Neuronové sítě jsou jen jednou z forem, jak získávat znalosti z dat (ať už s učitelem nebo bez něj). Jedna z dalších metod jsou například rozhodovací stromy. Využití procesu rozpoznávání emocí je možné najít i v jiných oblastech, například v bioinformatice, kde se zabývá výběrem důležitých informací z velkého a nepřehledného množství dat [7]. Rozpoznávání emocí nemusí být pouze u textu, ale také například u analýzy řeči [5], rozpoznávání emocí z pohybu, či z pořízeného obrazu obličeje, jako je uvedeno v pramenu [13]

Tato práce je rozdělena na několik částí. Teoretická část práce je zaměřená na poznatky, pomocí kterých je realizována praktická část. Popisuje teoretický návrh algoritmu pro rozhodovací strom. Dále je zde zobrazena architektura programu pro dolování znalostí a popis příkladů využití. V práci jsou popsány kroky pro zpracování textu (segmentace textu, filtrace tokenů, lemmatizace, ontologické báze).

Hlavním přínosem práce je návrh rozhodovacího stromu v jazyce Java. Vytvoření pluginu z tohoto programu pro prostředí RapidMiner. Do aplikace pro předzpracování textu bylo implementováno načítání a vyhledávání v databázi WordNet. Pomocí tohoto předzpracování byla vytvořena databáze pro klasifikaci do emočních tříd. Praktická část se zabývá návrhem a popisem algoritmu rozhodovacího stromu, vytvořením databází a následným trénováním a testováním v prostředí RapidMiner. Z těchto testování byly vytvořeny demonstrační příklady, spustitelné v programu RapidMiner.

V závěrečné části jsou pak prezentovány použité parametry pro nastavení klasifikátoru a dosažené procentuální výsledky úspěšnosti navrženého systému.

1 VYUŽITÍ ROZPOZNÁVÁNÍ EMOCÍ Z TEXTU

Využití rozpoznávání emocí z textu má uplatnění všude tam, kde se jakýmkoliv způsobem pracuje s velkým množstvím dat(textem), ve kterém je obsažen nějaký emoční náboj, se kterými byl daný dokument tvořen. Lidskými silami je velice náročné tyto texty analyzovat. Příkladů kde by se dolování emocí dalo využít je mnoho, zde budou uvedeny některé z nich.

1.1 Reklama

Reklama je jakákoliv forma propagace výrobku, služby, společnosti, mající za cíl zvýšení prodeje. Má několik forem: televizní, novinová, internetová, nebo plakátová. Reklamou se rozumí přesvědčovací proces, kterým jsou hledáni uživatelé zboží, služeb nebo myšlenek prostřednictvím komunikačních médií. Pro reklamu jsou emoce velice důležité. Systém pro získání emocí měří kvalitu reklamy a na základě získaných reakcí, se může reklama vylepšit. Čím lepší znalost emocí, tím je reklama více cílená. Což vede ke zvýšení prodeje, daného výrobku, služby, atp.

1.2 Projev veřejnosti

Využití rozpoznávání emocí se nabízí při zkoumání veřejného mínění. Kdy z různých dotazníků či anketních lístků je možné zjišťovat a posuzovat co si lidé myslí a s jakými emocemi byly dokumenty tvořeny. Tato možnost může být využita při pořádání různých propagačních akcí a posléze reagovat na negativní ohlasy.

1.3 Zákaznické centrum

Využití se také nabízí v případě technického centra, technických podpor nebo veškerých linek, které využívají přepisu telefonních rozhovorů do elektronické podoby. V případě negativních ohlasů je možné odstranit příčiny vzniku, které zlepší působení firmy na zákazníka.

1.4 Novinové články

Použití rozpoznávání emocí se dá využít u novinových titulků, případně článků. Je třeba nadefinovat vstupní databázi, která obsahuje všechna potřebná slova. V případě novinových článků by tato databáze měla obsahovat jak odborná slova, tak slova běžné komunikace. Na kvalitě databáze závisí kvalitní výsledek. Způsob jak

provézt určení emočního náboje v novinových článcích je provedením rozboru jednotlivých slov nebo vynecháním slov, která by mohla ovlivnit výsledek a poté jeho porovnání s použitím různých algoritmů [3].

1.5 Sociální sítě

Další využití se nabízí ve spojení se sociálními sítěmi. Během posledních let se sociální sítě staly oblíbeným místem pro projev lidských emocí. Její obliba stále vzrůstá, stejně tak jako potřeba sdělovat své pocity. Proto lze sociální sítě považovat za velmi obsáhlé co se emocí týče. Při použití algoritmu pro klasifikaci těchto dat je nutné aby vstupní databáze obsahovala jak slovník používaných slov, slovník zkratek, tak slovník obsahující emotikony. Hodnotící algoritmus by měl také počítat s nutnou úpravou textu kvůli nespisovnému jazyku, který se na těchto sítích používá.

2 SOUČASNÉ PŘÍSTUPY

Zpracování velkého množství textu lidskými silami je založeno na lidské intuici. Hlavním klíčem k úspěchu je uspořádání dat tak, aby dávala smysl a bylo možné jim porozumět. Hlavní nevýhodou je časová náročnost, kdy analýza dat může trvat velice dlouhou dobu (dny, měsíce, roky). Výhodu oproti tomu má strojové učení, které rozhoduje na základě naučeného schématu [6]. Z časového hlediska se jedná o rychlejší způsob analýzy než pomocí lidských sil. K využití strojového řešení se využívají tzv. ontologické báze. Tyto báze obsahují člověkem vytvořená slovní spojení a jejich chápání pro strojové zpracování textů.

2.1 Detekce na základě klíčových slov

Jedná se o metodu, se kterou se pracuje na základě zadaných klíčových slov. Úspěšnost této metody závisí na zpracování textu před samotným použitím. Úpravou vět na rozdělená slova, a právě tvorbě klíčových slov, ze kterých je vytvořena databáze. Nevýhodou je, že tato metoda nepozná emoce ve větách, ve kterých nejsou klíčová slova obsažena. Způsob jak zpřesnit tuto metodu je vytvořením databáze tzv. „klíčových slov“ pomocí ontologické báze, ve které jsou obsaženy vzájemné spojitosti mezi těmito slovy. Díky tomu je možné vytvořit slovník [2].

2.2 Strojové učení

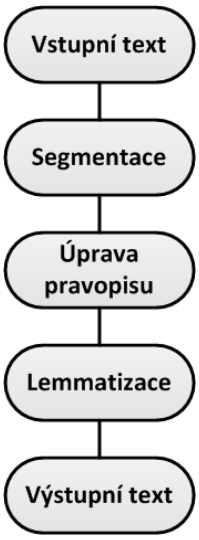
Strojové učení je souhrn moderních metod pro analýzu dat. Metody strojového učení jsou podoblastí umělé inteligence. Zabývají se algoritmy a technikami, pomocí kterých se počítač „učí“. Parametry jsou nastaveny už ve fázi trénování, v již předem ohodnoceném textu. Nejčastěji se využívá SVM algoritmus (Support Vector Machine) [3]. Úspěšnost této metody je založena na úspěšném zpracování trénovací množiny a na klíčových slovech.

2.3 Hybridní metody

Hybridní metody, jsou metody kombinující výhody z metody detekce na základě klíčového slova a metody strojového učení [17]. Tyto metody využívají ke zpracování textu tzv. lexikální databáze, ve kterých je obsaženo propojení mezi slovy a slovními spojeními. K získání těchto spojení se může využít například slovník Tezaurus, obsahující seznam synonym, někdy i antonym [1].

3 ZPRACOVÁNÍ TEXTU

Pro práci s textovým dokumentem je velice důležité jeho prvotní zpracování před použitím pro dolování znalostí. Pro počítač je text chápán jako směsice znaků, které se musí upravit do takové podoby, aby se s nimi mohlo dále pracovat. Touto úpravou dokumentu jsou získána data, ze kterých je snadnější rozpoznat význam textu. Na obrázku 3.1 je zjednodušené schéma procesu zpracování textu.



Obr. 3.1: Průběh zpracovaní textu

3.1 Segmentace textu

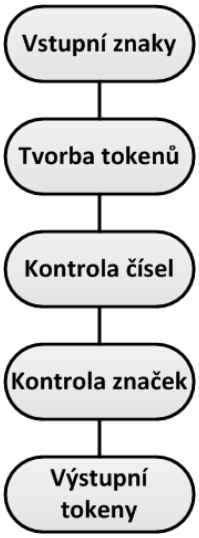
Segmentace je rozdělení textu do menších skupin. Segment je textový prvek, který aplikace při překladu považuje za nejmenší jednotku, určenou k přeložení. Pro správnou funkci je nutné, aby aplikace od sebe dokázala oddělit slova, rozpoznat začátky a konce vět atd. Segment neboli token je elementární nositel informace v daném jazyce. Tokeny nejsou vždy pouze jen slova, která vznikají rozdělením textu, např. podle mezer, nebo interpunkčních znamének. Mohou být také tvořeny různým skládáním znaků dohromady, tam kde by odděleně ztrácely smysl. Například různá registrační čísla, která jsou oddělena tečkou. U segmentace jsou i jisté problémy s určitými typy:

- *Čísla* - Různá spojení čísel oddělená čárkou nebo tečkou.
- *Emotikon* - Grafický symbol složený obvykle z interpunkčních a speciálních znaků (smajlík), který vyjadřuje autorovu náladu, postoj či emoce. Jedná se o speciální formu ASCII art[11] a její význam se interpretuje otočením znaků o 90° doprava. V tabulce 3.1 jsou uvedeny nejpoužívanější příklady Emotikonů.

V moderní době se tyto znaky používají pravidelně při tzv. „rychlé komunikaci“(chat, SMS)[15].

- *Zkratky* - Zkratky v sobě většinou obsahují tečku nebo tečky, které je nutné zachovat u sebe (a.s., s.r.o.), v tomto případě může nastat problém s rozeznáním konce věty. Tuto chybu můžeme vyřešit pomocí slovníku nejčastěji používaných zkratek při segmentaci.
- *Internetové tokeny* - Jedná se o identifikaci a následné zachování různých znakových posloupností(IP adres, URL, doménových jmen) a dalších významných řetězců, které se nesmí porušit.

Na obrázku 3.2 je zobrazena varianta tvorby tokenů [10].



Obr. 3.2: Tvorba tokenu

Tab. 3.1: Tabulka nejpoužívanějších emotikonů

Emotikon	**Emoční význam**
:-) :-D =)	Radost, úsměv
:-(:-/	Smutek, nespokojenost
:’(	Pláč, velký smutek
:-*	Láska, náklonnost
:-! :-x	Znechucení, hněv
:-o	Údiv
:-P	Vtipkování, laškování

3.2 Úprava pravopisu na základní tvar

Úprava pravopisu na základní tvar, je technika pro zpracování textu. Snaží se najít slova, jenž mají stejný význam, ale liší se spisovností a upravit je na stejný základní tvar daného slova (například slovo „levnějc“ se upraví na slovo „levněji“). Výhodou je, že sloučí slova se stejným základem na jeden tvar. Tato metoda, má ale určité problémy, jelikož slova, která se stejně píší, ale mají jiný význam převede na stejný základní tvar, čímž se daná informace ztratí. Dalším problémem je, že algoritmus vytvořený pro jeden jazyk nelze aplikovat na jazyky jiné.

3.3 Lemmatizátor

Lemmatizátor je nástroj (např. počítačový program), který převede dané slovo (vyhledá v databázi) do základního tvaru, tzv. „lemma“. Například slovo „počítačích“ je převedeno na slovo „počítač“. Umožňuje lepšímu porozumění strojovému textu a používá se při vyhledávání fulltextem. Lemmatizace je metoda vycházející ze stanovených pravidel, která jsou pro každý jazyk odlišná. Tento způsob je sice účinný, ale je velice časově náročný a pracný.

3.4 Ontologické báze

Ontologické báze jsou databáze, které jsou vytvořeny uživatelem. Tyto databáze obsahují vzájemné vztahy mezi pojmy. Účelem je zadefinovat chápání těchto pojmů pro strojové zpracování. Nejpoužívanější lexikální databází je databáze Word-Net. Tato databáze obsahuje podstatná jména, přídavná jména, slovesa a příslovce, které jsou vzájemně provázány do „sémantické síťe“. Word-Net obsahuje následující propojení [16]:

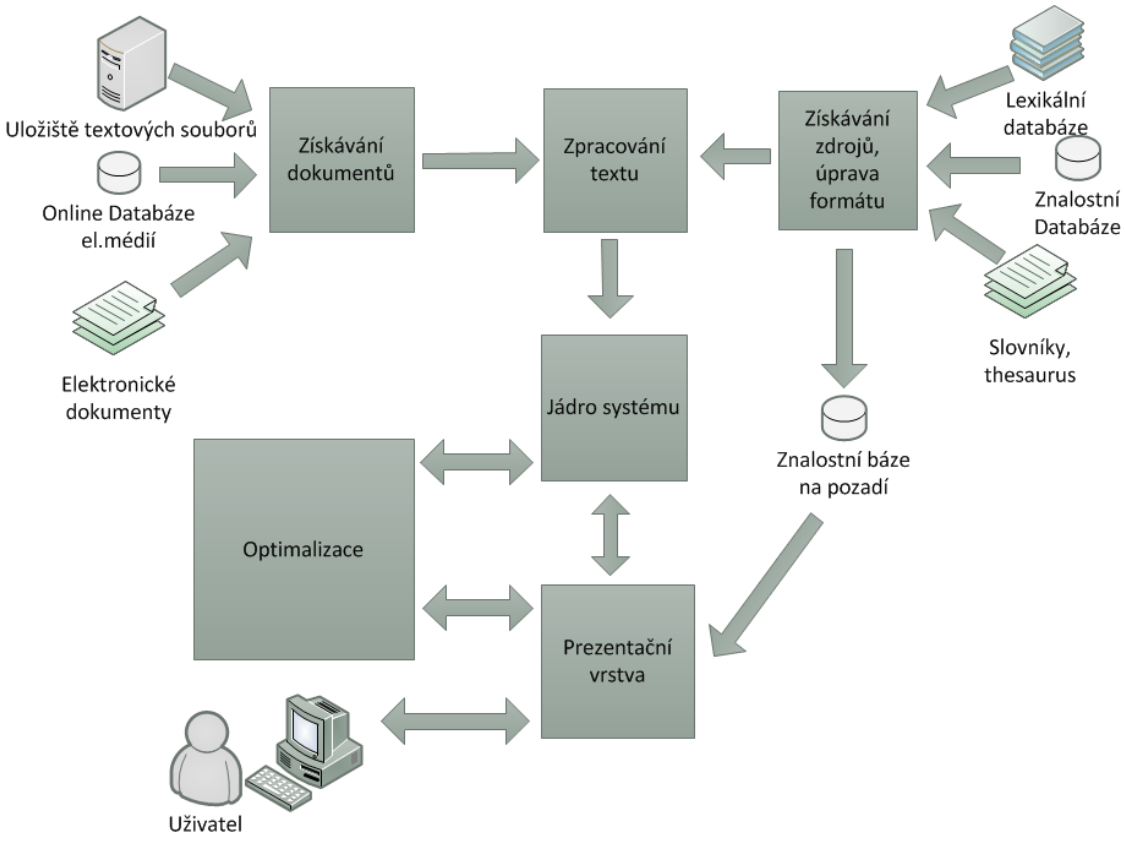
- *Synsety* - Obsahuje vzájemné vztahy mezi synonymy, tj. slovy, která mají stejný význam. Synsety dohromady tvoří lexikální databáze. Jednotlivé prvky synsetu se nazývají literály.
- *Hyperonymické vztahy* - Vztahy, které spojují určitý synset s jeho obecnějším významem (Ke slovu jablko je hyperonymický vztah slovo ovoce).
- *Hyponymické vztahy* - Vztahy, které spojují určitý synset s jeho konkrétnějším významem.
- *Meronymie* - Vytváří vztahy mezi částí a celkem.
- *Holonymie* - Vytváří vztahy mezi celkem a částí. (opačný smysl než meronymie)

4 ARCHITEKTURA SYSTÉMU

Aby se text mohl využít pro získání emocí, musí projít několika fázemi zpracování. Architektura systému pro získávání znalostí emocí z textu je zobrazena na obrázku 4.1 a může být rozdělena na tyto části [8]:

- *Zpracování textu* - V této fázi jsou obsaženy veškeré operace pro úpravu textu pro další zpracování. Jedná se o segmentaci textu (rozdělení na slova), úpravu pravopisu (do základního tvaru), lemmatizátor (nahrazení základním tvarem slova) a tvorbu klíčových slov. V této fázi mohou být obsaženy také metody, které připojí k dokumentům důležité informace (časové razítko, zdroj dokumentu, atd.) pro další zpracování.
- *Jádro systému* - Jedná se o hlavní část celého systému obsahující veškeré algoritmy, které se starají o dolování znalostí. Hledá propojení mezi dokumenty a entitami.
- *Prezentační vrstva* - Jedná se o rozhraní, přes které uživatel komunikuje s jádrem systému. V této vrstvě jsou obsaženy: Vizualizační programy, editory pro filtraci textu, GUI grafické rozhraní, atp.
- *Optimalizace výstupu* - Jedná se o nejrůznější vylepšení pro zobrazení prezentační vrstvy. Umožňuje zjednodušené zobrazení požadované akce. Filtraci zobrazených informací, hledání spojitostí v algoritmech a jejich následné zobrazení na prezentační vrstvě.

V případě, že systém pracuje s určitým daty, je dobré k analýze těchto dat využít znalostní databáze. Tyto databáze jsou k daným datům přidruženy a obsahují dodatečné informace, které dopomohou k lepšímu pochopení těchto dat. K analýze je také dobré použít Tezaurus, externí slovníky a lexikální databáze, které obsahují informace o datech, dokonce i význam slovních spojení a propojení mezi nimi.



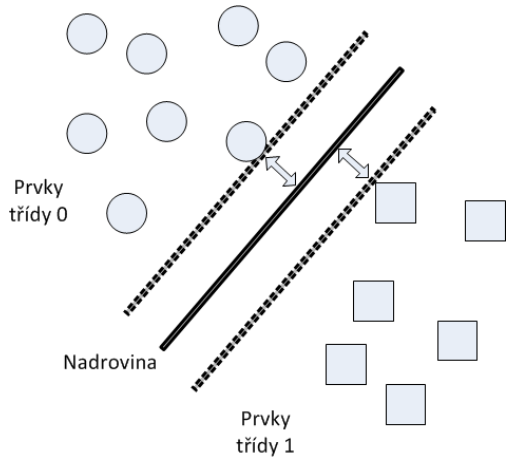
Obr. 4.1: Architektura systému dolování znalostí z textu.

5 KLASIFIKAČNÍ ALGORITMY

Umělá inteligence (AI) je obor informatiky, určen k tvorbě přístrojů vykazujících známky inteligentního chování. Při tvorbě umělé inteligence se využívají různé algoritmy. Algoritmus je přesný schématický návod či postup, pomocí něhož lze vyřešit daný typ určitého problému. Algoritmus má většinou přesný počet definovaných kroků, podle kterých se umělá inteligence chová.

5.1 SVM

Algoritmus SVM (Support Vector Machine) je metoda strojového učení. Jedná se o velice efektivní a rychlou metodu klasifikace [8]. Algoritmus slouží k lineární separaci dat, i těch která separovat nejdou. SVM algoritmus hledá nadrovinu, ve které se v prostoru příznaků optimálně rozdělují trénovací data. Požadavkem pro hledání nadroviny je vzdálenost mezi nadrovinou a nejbližším prvkem jednotlivých tříd viz. obrázek 5.1.



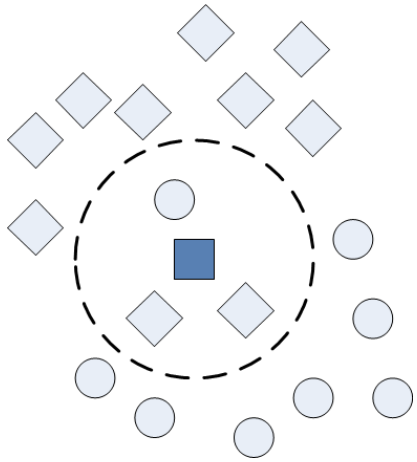
Obr. 5.1: Separace prvků pomocí SVM algoritmu

Jinými slovy, v okolí nadroviny je na obě strany co nejširší pruh, ve kterém se nenachází žádný bod. K popisu této nadroviny slouží pouze nejbližší body, kterých bývá velice málo. Tyto body se nazývají „podpůrné vektory“, odkud získala metoda své jméno. Metoda SVM je binární, data tedy rozděluje pouze do dvou tříd [14]. Nadrovina, která rozděluje obě strany je lineární funkcí v prostoru příznaků. Velice důležitou součástí metody Support Vector Machine je jádrová transformace (kernel transformation). Ta umožňuje převézt původně neseparovatelnou úlohu na úlohu separovatelnou, na kterou lze aplikovat další optimalizační algoritmus pro nalezení

rozdělující nadroviny. Výhoda této metody je ta, že se transformace dá použít i na různé typy předmětu např. Rozhodovací stromy a grafy [4].

5.2 K-nejbližších sousedů

Algoritmus k-nejbližších sousedů (K-NN) (K-Nearest Neighbours) je algoritmus sloužící ke strojovému učení pro rozpoznávání vzorů. Model klasifikátoru je vytvářen až při fázi klasifikace. Tato metoda vychází z tzv. „učení bez učitele“, kdy se část výstupních dat přivádí zpět na vstup. Ve fázi klasifikace se pak prvek umístí do nějakého místa v prostoru a najde si nejbližšího souseda. Objekt je pak klasifikován do té třídy, kam patří většina z těchto nejbližších sousedů, jako je na obrázku 5.2.

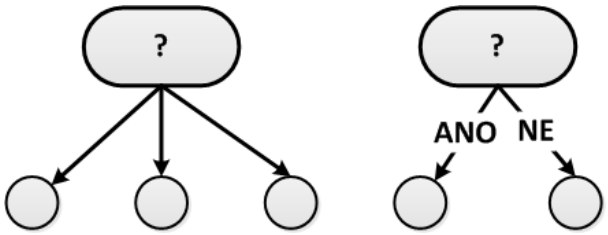


Obr. 5.2: Klasifikace pomocí algoritmu k-nejbližších sousedů.

5.3 Rozhodovací strom

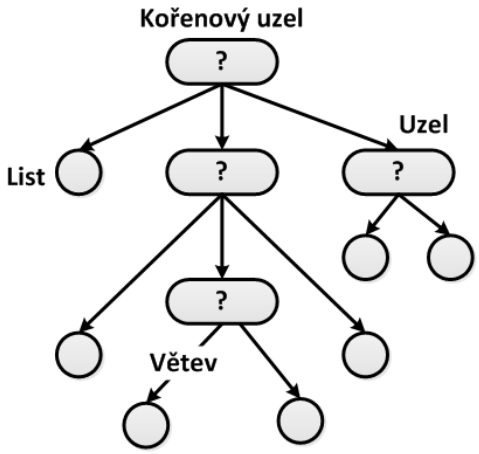
Rozhodovací strom (RS) je algoritmus umělé inteligence, určený k dolování znalostí z datového objektu. Cílem této metody je zjištění takových hodnot, které jsou schopny rozdělit data do příslušných tříd. Po rozdělení je možné získaná data použít pro analýzu dat nových nebo data uložit do struktury rozhodovacího stromu. Tato metoda je velice přehledná a jednoduchá. Umožňuje rychlé a efektní vyhodnocení získaných výsledků. Tvoření rozhodovacích stromů je dobře popsáno v algoritmech ID3 [9] a C4.5[18]. Více o rozhodovacích stromech je uvedeno v praktické části. Obecně existují dva druhy Rozhodujícího Stromu 5.3:

- *Více-cestné stromy* - z uzlu vystupuje n větví
- *Binární stromy* - z uzlu vystupují vždy 2 větve



Obr. 5.3: N-ární a Binární rozhodovací strom

Na obr 5.4 jsou zobrazeny části ze kterých se RS skládá, jejich popis je uveden v tabulce použitých termínů 5.1.



Obr. 5.4: Struktura rozhodovacího stromu

Tab. 5.1: Tabulka terminologie

Terminologie	**Popis**
Kořenový uzel (root node)	počáteční uzel.
Uzel (node)	dělí data na základě podmínky do větví.
Větev (branch)	Spojnice mezi dvěma uzly nebo uzlem a listem.
List (leaf, answer nodes)	Jeho dosažení vede ke klasifikaci objektu.

5.3.1 Významnost dělícího atributu

Pro správné rozdělení dat je nutné stanovit nejvýznamnější dělící atribut. Ten se určuje pomocí nejvyšší hodnoty Informačního zisku(I_{G}). Pro zjištění této hodnoty se využívá jedno z z následujících kritérii: Entropie (H), Gini index (G_{I}) a klasifikační chyba (M_{E}).

Informační zisk (I_{G})

Informační zisk je hodnotící kritérium, které je určeno pro správné rozdělení dat. Měří pokles v entropii.

$$I_{\mathrm{G}} = Entropy(S) - \sum_{v} \frac{S_v}{S} \cdot Entropy(S_v), \tag{5.1}$$

kde $Entropy(S)$ je celková entropie, S_v je počet jednotlivých rozdělených dat ,S je počet všech hodnot, $Entropy(S_v)$ je Entropie pro rozdělená data.

Entropie (H)

Entropie je číselná hodnota v intervalu <0;1> vyjadřující míru neuspořádanosti. Hodnota entropie je maximální, pokud jsou jednotlivé kategorie proměnných rovnoměrně zastoupeny v uzlech a minimální, pokud se v uzlu nacházejí pouze data jedné kategorie. Entropie se počítá pro každý dceřiný uzel.

$$H = \sum_{j} -p_j \log_2 p_j \tag{5.2}$$

GINI index (G_{I})

Gini index je další používaná statistika pro klasifikační stromy. Nabývá hodnot v intervalu <0;1>. Hodnota Gini indexu je minimální, pokud je v konečném uzlu zastoupena pouze jedna kategorie proměnných a maximální, pokud jsou v konečném uzlu proměnné rovnoměrně zastoupeny.

$$G_{\mathrm{I}} = 1 - \sum_{j} -p_j^2 \tag{5.3}$$

Klasifikační chyba (M_{E})

Klasifikační chyba je podíl chybně klasifikovaných pozorování. Klasifikační chyba je obvykle používaná k finálnímu měření přesnosti, proto je logické její použití jako kriteriální statistika. Celková klasifikační chyba je opět dána váženým součtem klasifikačních chyb v dceřiných uzlech.

$$M_{\mathrm{E}} = 1 - max(p_j) \tag{5.4}$$

kde p_j je počet vyskytujících se jednotlivých tříd pro dané hodnoty atributů.

6 PRAKTICKÁ ČÁST

V rámci praktické části bylo úkolem prezentovat možnosti analýzy emocí v textu. Jako řešení pro implementaci systému pro rozpoznávání emocí v textu byla použita metoda využívající princip strojového učení. Použité algoritmy pro klasifikaci textu jsou postaveny na algoritmech SVM (Support Vector Machine), jenž umožňují velmi rychlý a efektivní způsob klasifikace textu [8].

Mezi hlavními úkoly bylo navrhnout algoritmus rozhodovacího stromu pro rozdělení dat. Tento algoritmus byl navržen v jazyce Java. Z tohoto programu byl vytvořen plugin, který byl naimportován do prostředí RapidMiner(RM) [12].

K testování byla vytvořena databáze obsahující ohodnocená data. Pro vytvoření této databáze byla použita naprogramovaná metoda pro načtení a použití ontologických bází. Tato metoda je použita v části předzpracování textu. Proces předzpracování textu pak vypadá následovně: vstupní text se rozdělí pomocí tokenizace na jednotlivá slova. U těchto slov se poté opraví pravopis, aby slova byla spisovná. Dále se pomocí lemmatizátoru tato slova nahradí za slova v základním tvaru. Tyto slova se poté vyhledají v ontologických bázích (WordNet), pomocí kterých se nahradí za slova, jejichž význam je předem zadefinován právě v těchto bázích.

Poté následovalo trénování a testování natrénovaného modelu na databázi s implementací WordNetu. Pro správné trénování a testování modelu a vyhodnocení přesnosti bylo použito právě prostředí RapidMiner [12].

6.1 Rozhodovací strom

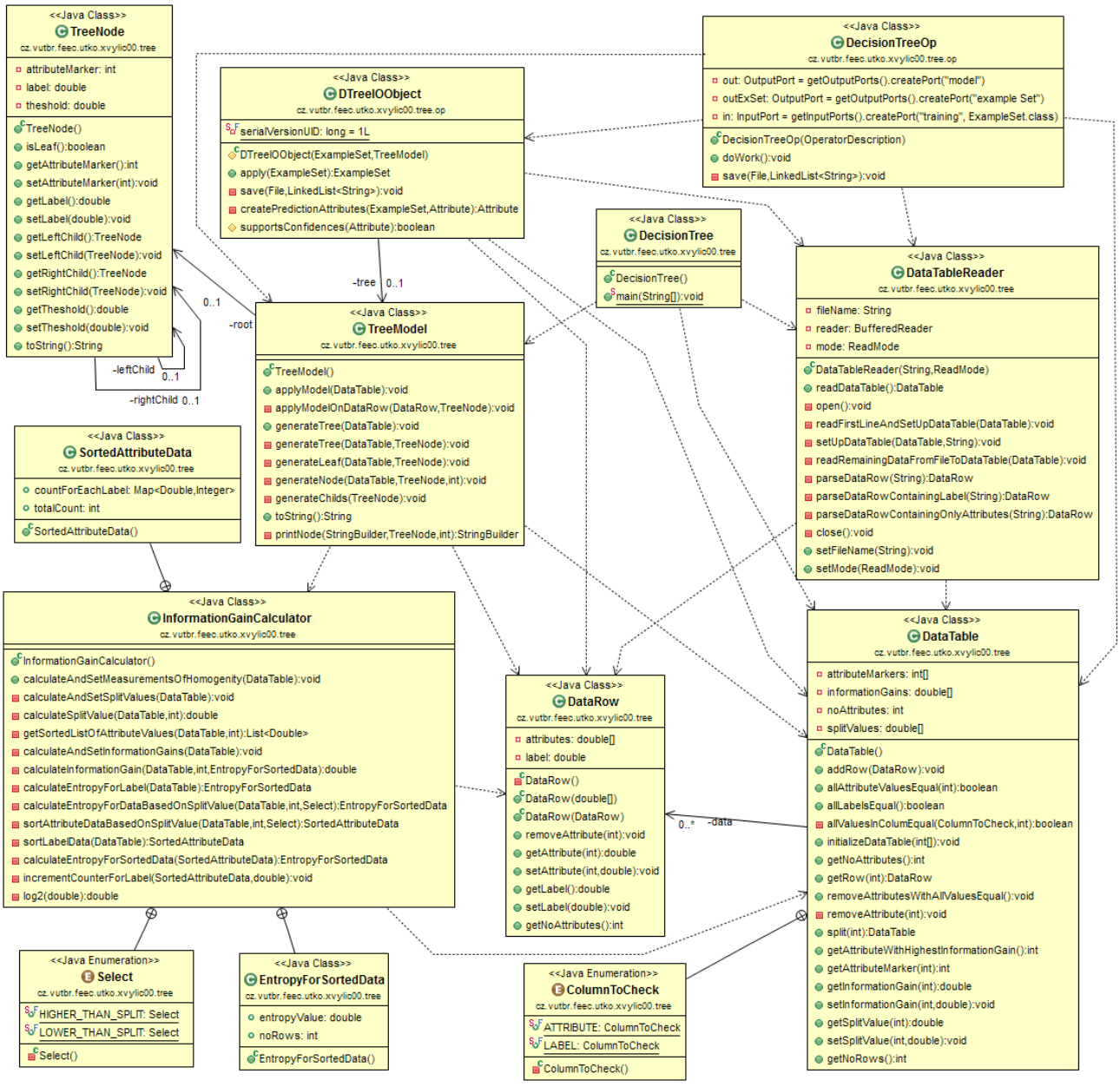
V rámci praktické části vznikl Java program pro rozhodovací strom. Typově se jedná se o binární strom, kdy se uzly dělí vždy jen na dvě větve. Z tohoto programu vznikl plugin, který byl naimportován do programu Rapid Miner. A pomocí něj bylo provedeno testování. Vytvořený plugin rapidminer-DecisionTree-1.0.0.jar je přiložen na CD.

6.1.1 Popis funkce programu rozhodovacího stromu

Program, poté co se spustí, načte data pro zpracování. Data v jednotlivých sloupcích atributů se seřadí podle velikosti od nejmenších po největší. Z těchto seřazených dat se určí střední hodnota, která data rozdělí na dvě části, $\leq$ než střední hodnota a $>$ než střední hodnota. Pro tyto části je vypočítána entropie, ze které se určí informační zisk. Podle informačního zisku se určí nejvýznamnější atribut a podle něj se vytvoří dělící uzel, který data rozdělí. Tento postup se opakuje tak dlouho, dokud nejsou všechna data rozdělena do jednotlivých tříd.

6.1.2 Popis jednotlivých tříd programu

Na obrázku 6.1 je zobrazen UML diagram programu. V podkapitole jsou popsány vybrané třídy programu.



Obr. 6.1: UML diagram programu

DecisionTreeOp.java

Tato třída popisuje implementování algoritmu rozhodovacího stromu, jako operátoru do RM.

DataTable.java

Tato třída reprezentuje tabulku dat. Může uchovávat jak ohodnocená (trénovací), tak i neohodnocená data. Samotná tabulka dat je uložena za použití kolekce ArrayList, která uchovává záznamy typu DataRow (pro jednotlivé řádky tabulky). DataTable uchovává také dělící hodnoty (split values) a informační zisk (information gain) pro jednotlivé atributy. DataTable provádí rozdělení v sobě obsažených dat na dvě tabulky na základě dělící hodnoty předaného atributu.

DataRow.java

Třída reprezentující jeden řádek v tabulce s daty. Uchovává hodnoty jednotlivých atributů (dat v tomto řádku) a jejich ohodnocení (label/třídu).

DataTableReader.java

Slouží pro načítání dat ze souboru do třídy DataTable.

TreeNode.java

Reprezentuje jeden uzel stromu. Uchovává dělící hodnotu, index atributu, který budeme porovnávat s dělící hodnotou a ohodnocení dat (label/třídu). Obsahuje i odkazy na svůj levý a pravý podstrom (uzel). V případě, že je tento uzel list, je nastavena pouze hodnota uzlu (label). Pokud se jedná o uzel musí byt nastavena dělící hodnota a index atributu, se kterým budeme porovnávat a odkazy na potomky.

TreeModel.java

Reprezentuje strom. Uchovává referenci na kořenový uzel stromu. Obsahuje metodu, která pro DataTable obsahující ohodnocená data vygeneruje rozhodovací strom. A metodu, která na neohodnocená data aplikuje vytvořený model.

InformationGainCalculator.java

Na základě DataTable obsahující ohodnocená data vypočte dělící hodnoty a informační zisk pro jednotlivé atributy.

6.1.3 Vzorové příklady výpočtu

Veškeré výpočty jsou tvořeny pro data TrainingSet.txt, která jsou přiložena na CD. Mají pouze informativní účely pro lepší pochopení funkce programu.

1.Střední hodnota

Pro určení střední hodnoty je použit medián. Medián se u daných hodnot určí tak, že se hodnoty seřadí podle velikosti a nalezne se hodnota, která se nalézá uprostřed seznamu. V případě, že má soubor sudý počet prvků, tak se za medián označuje aritmetický průměr hodnot na místech n/2 a n/2+1, kde n je počet prvků. Pro atribut A0 je střední hodnota = 5,8 ; pro A1 = 3,0; pro A2 = 4,35; pro A3 = 1,3

2.Entropie

Pro výpočet entropie se vychází ze vzorce 5.2. Entropie se pro každý atribut vypočítává dvakrát, jednou pro data $\leq$ střední hodnotě, a pro data $>$ než střední hodnota. Pro Atribut A0 $\leq$ než 5,8:

$$H = \sum_j -p_j \log_2 p_j = -\frac{50}{80}\log_2\left(\frac{50}{80}\right) - \frac{24}{80}\log_2\left(\frac{24}{80}\right) - \frac{6}{80}\log_2\left(\frac{6}{80}\right) = 1,225$$

Pro Atribut A0 $>$ než 5,8:

$$H = -\frac{26}{70}\log_2\left(\frac{26}{70}\right) - \frac{44}{70}\log_2\left(\frac{44}{70}\right) = 0,9517$$

Pro Atribut A1 $\leq$ než 3,0:

$$H = -\frac{8}{83}\log_2\left(\frac{8}{83}\right) - \frac{42}{83}\log_2\left(\frac{42}{83}\right) - \frac{33}{83}\log_2\left(\frac{33}{83}\right) = 1,3516$$

Pro Atribut A1 $>$ než 3,0:

$$H = -\frac{42}{67}\log_2\left(\frac{42}{67}\right) - \frac{8}{67}\log_2\left(\frac{8}{67}\right) - \frac{17}{67}\log_2\left(\frac{17}{67}\right) = 1,2905$$

Pro Atribut A2 $\leq$ než 4,35:

$$H = -\frac{50}{75}\log_2\left(\frac{50}{75}\right) - \frac{25}{75}\log_2\left(\frac{25}{75}\right) = 0,91829$$

Pro Atribut A2 $>$ než 4,35:

$$H = -\frac{25}{75}\log_2\left(\frac{25}{75}\right) - \frac{50}{75}\log_2\left(\frac{50}{75}\right) = 0,91829$$

Pro Atribut A3 $\leq$ než 1,3:

$$H = -\frac{50}{78}\log_2\left(\frac{50}{78}\right) - \frac{28}{78}\log_2\left(\frac{28}{78}\right) = 0,9418$$

Pro Atribut A3 > než 1,3:

$$H = -\frac{22}{72}\log_2\left(\frac{22}{72}\right) - \frac{50}{72}\log_2\left(\frac{50}{72}\right) = 0,8879$$

Pro Atribut tříd(celková Entropie):

$$H = -\frac{50}{150}\log_2\left(\frac{50}{150}\right) - \frac{50}{150}\log_2\left(\frac{50}{150}\right) - \frac{50}{150}\log_2\left(\frac{50}{150}\right) = 1,5849$$

3.Informační zisk

Výpočty jsou pro určení prvního dělícího atributu. Je použit vzorec 5.1.

I_{G} pro A0:

$$I_{\mathrm{G}} = Entropy(S) - \sum_v \frac{S_v}{S} Entropy(S_v) = 1,5849 - \left(\frac{80}{150}\cdot 1,225 + \frac{70}{150}\cdot 0,9517\right) = 0,48744$$

I_{G} pro A1:

$$I_{\mathrm{G}} = 1,5849 - \left(\frac{83}{150}\cdot 1,3516 + \frac{67}{150}\cdot 1,2905\right) = 0,26059$$

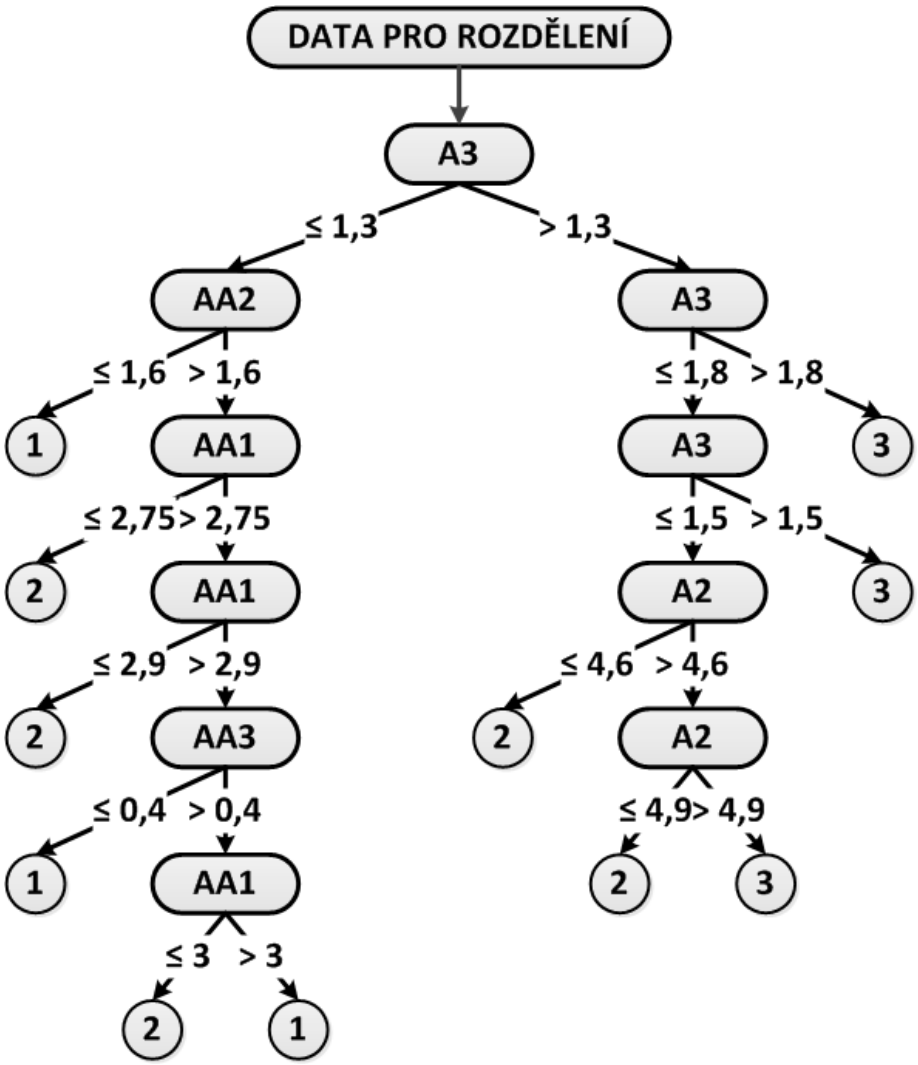
I_{G} pro A2:

$$I_{\mathrm{G}} = 1,5849 - \left(\frac{75}{150}\cdot 0,91829 + \frac{75}{150}\cdot 0,91829\right) = 0,66661$$

I_{G} pro A3:

$$I_{\mathrm{G}} = 1,5849 - \left(\frac{78}{150}\cdot 0,9418 + \frac{72}{150}\cdot 0,8879\right) = 0,668972$$

Jako dělící atribut se použije atribut A3, jelikož je jeho informační zisk nejvyšší. Prvotní rozdělení dat bude podle střední hodnoty 1,3. Tím vzniknou dvě oddělené tabulky dat, pro atributy A3 $\leq$ 1,3 a pro atributy A3 > než 1,3. U těchto tabulek se znovu provedou výpočty a data se znovu rozdělí. Na obrázku 6.2 je názorně zobrazen finální rozhodovací strom. Kde symboly A,AA jsou jednotlivé dělící atributy a čísla 1,2,3 reprezentují jednotlivé třídy.



Obr. 6.2: Rozhodovací strom

6.2 Databáze dat

V rámci práce vznikly databáze obsahující ohodnocená data pro 5 emočních tříd, tabulka 6.1.

Databáze obsahuje okolo 1850 příspěvků určených k ohodnocení. Tyto příspěvky byly stahovány z diskuzí na stránkách Novinky.cz [1] a IHned.cz [2]. Jejich ohodnocení bylo prováděno v aplikaci vytvořené na ústavu telekomunikací výzkumnou skupinou SPLAB [3], do které jsem implementoval načítání a vyhledávání v databázi WordNet. Ohodnoceno bylo okolo 390 příspěvků.

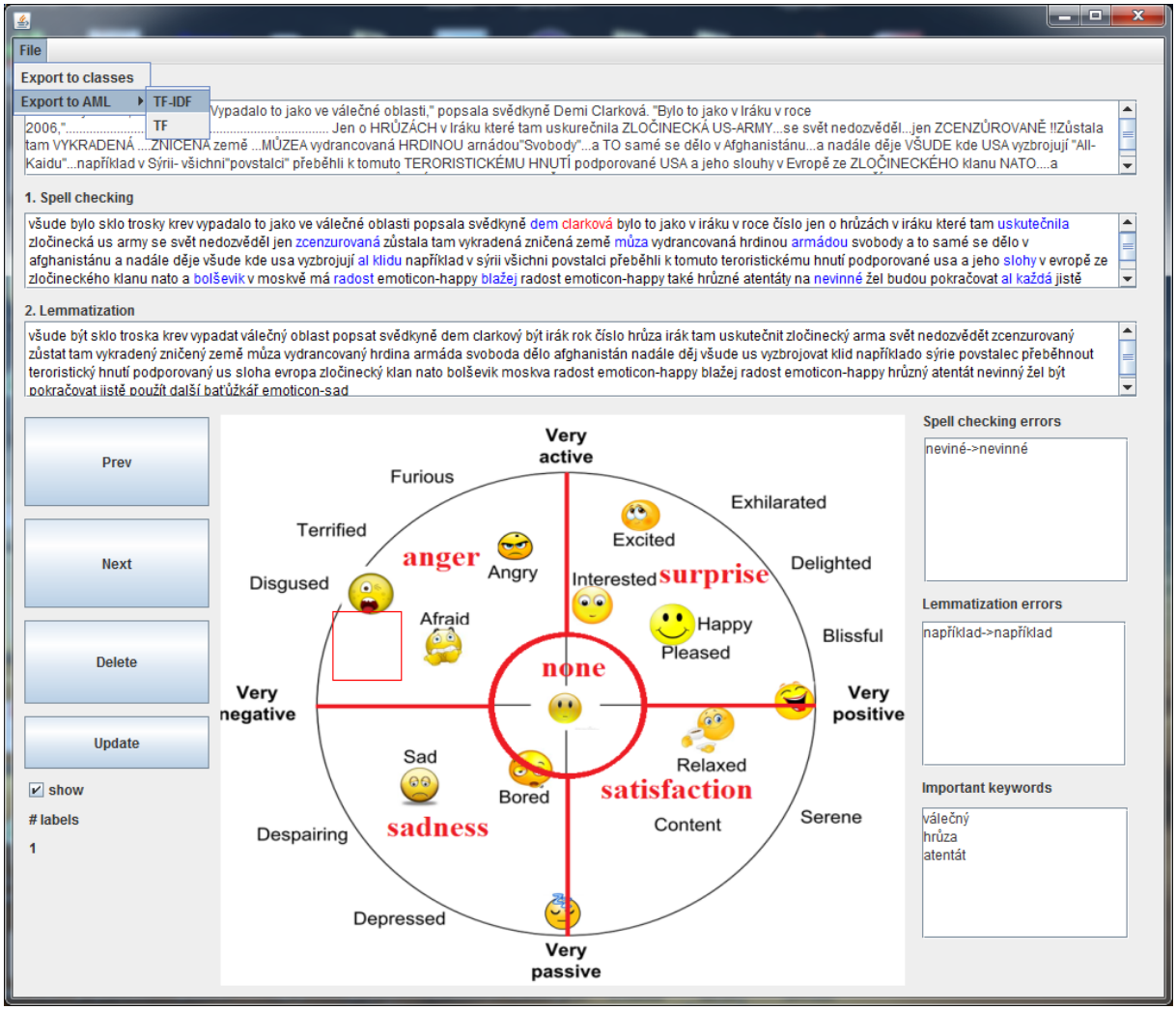
Tato aplikace je zobrazena na obrázku 6.3 a skládá se z těchto částí:

[1]Dostupný z URL: http://www.novinky.cz/
[2]Dostupný z URL: http://ihned.cz/
[3]Dostupný z URL: http://splab.cz/

Tab. 6.1: Popis názvosloví a významu pro skupinu emočních tříd

XML název třídy	**Charakteristika třídy**
Anger	Text byl vytvořen s vulgární emocí (sprostá slova, urážlivý význam, nevhodné pro danou situaci apod.)
Sadness	Text vyjadřuje smutek nebo žal (zoufalost, deprimovanost)
Neutral	Text nevyjadřuje žádnou emoci (nespadá do žádné s dalších skupin).
Satisfaction	Text byl vytvořen se spokojenou emocí (absence negativně ovlivňujících podmětů; vyrovnanost, spokojenost)
Surprise	Text byl vytvořen s pozitivní emocí (radost, nadšení, pochvala, překvapení).

- Input - zobrazení/načtení věty z databáze
- Spell checking - zobrazení opravené věty
- Lemmatization - zobrazení věty v základním tvaru slov
- Spell checking errors - okno pro opravu pravopisu
- Lemmatization errors - okno pro opravu Lemmatizace
- Important keywords - okno pro výběr klíčových slov
- Souřadnicová mapa - okno pro přiřazení labelovacích souřadnic
- NEXT - Posun na další větu v databázi
- PREV - Posun na předchozí větu v databázi
- DELETE - Smazání věty
- UPDATE - Uložení všech přiřazených nastavení.
- Export - Vyexportování knihovny dat do souboru .aml , nebo do jednotlivých klasifikačních tříd, soubor.txt.



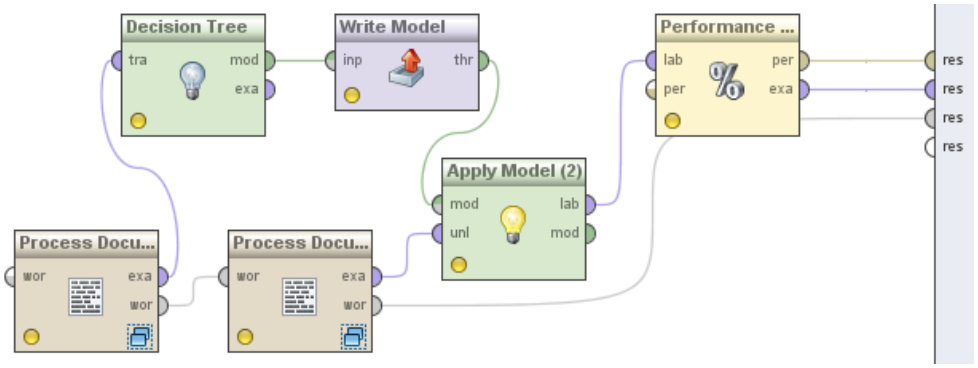
Obr. 6.3: Labelovací aplikace

6.3 Proces trénování v RapidMineru

RapidMiner je volně dostupné prostředí, které může být použito např. pro klasifikaci textů.

6.3.1 Trénování pomocí rozhodovacího stromu

Na obrázku 6.4 je zobrazeno blokové schéma procesu trénování pomocí navrženého rozhodovacího stromu. Toto schéma obsahuje několik bloků: operátory pro načtení trénovacích a testovacích dat, operátor rozhodovacího stromu, operátor zhodnocení úspěšnosti natrénovaného modelu a operátor pro testování natrénovaného modelu. Za pomoci těchto operátorů byl vytvořen demonstrační příklad pro trénování a testování, spustitelný v programu RapidMiner.



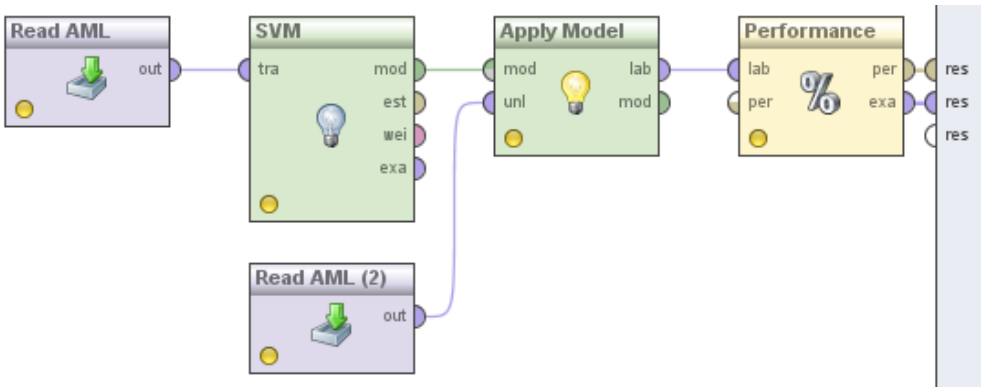
Obr. 6.4: Proces trénování pomocí rozhodovacího stromu

6.3.2 Trénování pomocí algoritmu SVM

Obrázek 6.5 zobrazuje blokové schéma procesu trénování pomocí SVM. Toto schéma obsahuje tyto bloky: operátory pro načtení trénovacích a testovacích dat, operátor SVM, operátor zhodnocení úspěšnosti natrénovaného modelu a operátor pro testování natrénovaného modelu. S použitím těchto operátorů byl vytvořen demonstrační příklad pro trénování a testování, spustitelný v programu RapidMiner

Operátory pro načtení trénovacích a testovacích dat

Jedná se o bloky, do kterých se nahrávají data určená k trénování a testování systému. Nahrávají se sem např. vulgarní a non-vulgarní data, která se poté pomocí křížové validace dále zpracovávají.



Obr. 6.5: SVM Proces trénování

Operátor rozhodovacího stromu

Operátor reprezentující plugin navrženého algoritmu rozhodovacího stromu.

Operátor SVM

Operátor algoritmu SVM sloužící k lineární separaci dat. Algoritmus slouží k lineární separaci dat, i těch která separovat nejdou. SVM algoritmus hledá nadrovinu, ve které se v prostoru příznaků optimálně rozdělují trénovací data.

Operátor pro testování natrénovaného modelu

Tento operátor aplikuje natrénovaný model na testovací data. Obsahuje informace o datech, která byla natrénována. Tyto informace mohou být použity k předpovězení výsledné hodnoty neznámé proměnné. Reprodukuje přetvoření jak během tréninku, tak při jiných změnách. Veškeré potřebné parametry jsou uložené uvnitř tohoto objektu.

Operátor zhodnocení úspěšnosti natrénovaného modelu

Operátor zhodnocení je funkce, na které se nastavuje jaké výsledky chceme zobrazit (% úspěšnost natrénovaného modelu na testovacích datech, chybovost, atd.). Pro zjištění úspěšnosti u rozhodovacího stromu se využívá vzorce:

$$R_{\text{acc}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}} + N_{\text{TN}}}[\%], \tag{6.1}$$

kde N_{TP} vyjadřuje počet správně zařazených do kategorie (true positive), N_{TN} vyjadřuje počet správně zařazených do kategorie (true negative), N_{FP} počet chybně zařazených do dané kategorie (false positive), N_{FN} počet chybně zařazených do dané kategorie (false negative).

7 VÝSLEDKY

Pro navržený program bylo nutné zjistit jeho procentuální úspěšnost. Toto testování bylo provedeno v prostředí RapidMiner, kde byl navržený systém také porovnán s jinými způsoby klasifikace 7.1.

Tab. 7.1: Definice použitých způsobů

Způsob	**Popis**
Rozhodovací strom	využívá se implementovaného algoritmu rozhodovacího stromu.
SVM	využívá se SVM algoritmu.
Rozhodovací strom RM	Výchozí rozhodovací strom RM

7.1 Testování pomocí navrženého rozhodovacího stromu

Testování prováděné pomocí implementovaného pluginu rozhodovacího stromu. Pro vyhodnocení procentuální úspěšnosti, byl použit vzorec 6.1. Názorný příklad výpočtu úspěšnosti pro třídu anger:

$$R_{\text{acc}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}} + N_{\text{TN}}} = \frac{53 + 156}{53 + 19 + 105 + 156} = 0,6276 = 62,76\%$$

V tabulce 7.2 jsou uvedeny úspěšnosti systému rozhodovacího stromu pro jednotlivé třídy, i pro testování všech pěti tříd zároveň.

7.2 Testování pomocí SVM

Pro otestování pomocí SVM algoritmu, bylo třeba nastavit parametry C a ϵ. Tato nastavení jsou uvedena v tabulce 7.3.

Jako hodnotící kritérium bylo zvoleno **Root Mean Square Error**, které vyjadřuje Střední kvadratickou chybu. Tato chyba určuje rozdíl mezi hodnotami předpovědi a hodnotami skutečnými. Testování bylo prováděno na labelech X a Y. Rozmezí hodnot těchto labelů je od -1 do 1.

V tabulce 7.4 jsou uvedeny výsledky testování pomocí SVM algoritmu.

Tab. 7.2: Tabulka úspěšnosti vytvořeného algoritmu

typ třídy	**úspěšnost programu**
anger	62,76%
neutral	90,74%
sadness	84,81%
satisfaction	79,77%
surprise	64,19%
Celkový	46,32%

Tab. 7.3: Tabulka nastavení parametrů SVM

Typ klasifikace	**C**	ϵ
detekce negativních emocí	0,001953	0,5
detekce pozitivních emocí	9	0,001953
detekce neutrálních emocí	32	0,001953

7.3 Testování pomocí rozhodovacího stromu RM

Testování prováděné pomocí výchozího rozhodovacího stromu obsaženém v RM. Pro vyhodnocení procentuální úspěšnosti byl taktéž použit vzorec 6.1.

$$R_{\text{acc}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}} + N_{\text{TN}}} = \frac{91 + 136}{91 + 39 + 67 + 136} = 0,6789 = 67,89\%$$

V tabulce 7.5 jsou uvedeny procentuální úspěšnosti jednotlivých systémů.

Tab. 7.4: Tabulka úspěšnosti systému

Kernel Type	**C**	ϵ	**X souřadnice RMSE**	**Y souřadnice RMSE**
dot	0	0	0.508	0.491
dot	0,001953	0,5	0.495	0.430
dot	9	0,001953	0.508	0.499
dot	32	0,001953	0.508	0.499
radial	0	0	0.493	0.439
radial	0,001953	0,5	0.494	0.444
radial	9	0,001953	0.494	0.438
radial	32	0,001953	0.494	0.438

Tab. 7.5: Tabulka úspěšnosti rozhodovacího stromu v RM

typ třídy	**úspěšnost programu**
anger	68,17%
neutral	93,52%
sadness	89,45%
satisfaction	80,92%
surprise	66,92%
Celkový	67,89%

7.4 Zhodnocení výsledků

Účelem těchto testování bylo zjištění procentuální úspěšnosti navrženého algoritmu a jeho porovnání s výchozím rozhodovacím stromem v RapidMineru. Veškeré testování bylo prováděno za pomocí databáze obsahující ohodnocená data. Z dosažených výsledků je zřejmé, že nejvyšší úspěšnost měl výchozí rozhodovací strom RM. Důvodem pro tuto úspěšnost je odlišný výpočet dělícího atributu, než u navrženého algoritmu programu, kde je tato úspěšnost o něco menší. Výsledky jsou ale v určitých třídách velice zkreslené, příčinou toho je malý počet trénovacích a testovacích dat. Pro určení lepších a kvalitnějších výsledků by bylo dobré použít větší trénovací databázi dat.

8 ZÁVĚR

Cílem této práce bylo prozkoumat současné přístupy k rozpoznávání emocí v textu, společně s teoretickým rozborem základních technik používaných při dolování znalostí z textu.

Hlavním přínosem této práce je naprogramování algoritmu pro rozhodovací strom v jazyce Java. Z tohoto programu byl vytvořen plugin, který byl importován do prostředí RapidMiner.

V rámci práce byly vytvořeny databáze obsahující data pro pět skupin emočních tříd. K vytvoření těchto databází bylo využito systému na předzpracování textu s implementací načítání a vyhledávání v databázi WordNet. Následně bylo na těchto databázích provedeno trénování a testování v prostředí RapidMiner.

Dalším přínosem je prezentace výsledků klasifikace textových dokumentů do definovaných emočních tříd a zhodnocení klasifikace s různými modifikacemi navrženého systému. Zjišťování procentuální úspěšnosti bylo prováděno několika způsoby 7.1. Pro vytvořený algoritmus byla na vytvořených databázích ověřena účinnost s úspěšností 62,76% pro třídu Anger; 90,74% pro třídu Neutral; 84,81% pro třídu Sadness; 79,77% pro třídu Satisfaction; 64,19% pro třídu Surprise; a 46,32% pro testování všech pěti tříd zároveň.

Při použití SVM algoritmu bylo testování prováděno na labelech X a Y. Rozmezí těchto labelů je od -1 do 1. Dále bylo třeba nastavit parametry C a ϵ podle tabulky 7.3. Výsledky testování byly prováděny pro dva typy jader „dot“ a „radial“. Výsledky tohoto testování jsou uvedeny v tabulce 7.4. Za pomoci operátorů byly vytvořeny demonstrační příklady pro trénování a testování, spustitelné v programu RapidMiner.

Na vytvořených databázích byla také ověřena úspěšnost pomocí výchozího rozhodovacího stromu v prostředí RM. Dosažené hodnoty byly následující: 68,17% pro třídu Anger; 93,52% pro třídu Neutral; 89,45% pro třídu Sadness; 80,92% pro třídu Satisfaction; 66,92% pro třídu Surprise; a 67,89% pro testování všech pěti tříd zároveň.

Důvodem pro větší procentuální úspěšnost při použití stromu z RM je odlišný vzorec pro výpočet dělícího atributu.

LITERATURA

[1] AMAN, S.; SZPAKOWICZ, S. *Using Roget's Thesaurus for Fine-grained Emotion Recognition. Affective Text: Semeval Task at the 4th International Workshop on Semantic Evaluations* [online] 2007,[cit. 2010-11-20]. Dostupné z WWW: `http://aclweb.org/anthology/I/I08/I08-1041.pdf`.

[2] BRACEWELL, D.B. *Semi-Automatic WordNet Based Emotion Dictionary Construction, Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on* s.629-634, 12-14 Dec. 2010. Dostupné z WWW: `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5708896&isnumber=5708804`

[3] BURGET, R.; KARÁSEK, J.; SMÉKAL, Z. *Classification and Detection of Emotions in Czech News Headlines. In The 33rd International Conference on Telecommunication and Signal Processing, TSP 2010.* 2010.

[4] CRISTIANINI,N; SHAWE-TAYLOR, J. *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004, ISBN 0-521-81397-2

[5] ČERNÝ, L. *Rozpoznávání a klasifikace emocí na základě analýzy řeči: Emotional State Recognition and Classification Based on Speech Signal Analysis.* Brno: Vysoké učení technické, Fakulta elektrotechniky a komunikačních technologií, 2010.

[6] ESPARZA, J. *Automatic emotion classification vs. human perception: Comparing machine performance to the human benchmark, Information Science, Signal Processing and their Applications (ISSPA) 2012*,11th International Conference on , s. 1253-1258, 2-5 July 2012. Dostupné z WWW: `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6310484&isnumber=6310432`

[7] FAIIAZEE, H.; ABDULLAH, R.; AL-HADDAD, S.A.R.; SAMSUDIN, K. *Text mining in bioinformatics: Past, present and future* Information Retrieval and Knowledge Management (CAMP), 2012 International Conference on , s .327-330, 13-15 March 2012. Dostupné z WWW: `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6205000&isnumber=6204970`

[8] FELDMAN, R.; SANGER, J. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data.* Cambridge : Cambridge University Press,2007.
410 s. ISBN 978-0-521-83657-9.

[9] CHEN JIN; LUO DE-LIN; MU FEN-XIANG. *An improved ID3 decision tree algorithm, Computer Science Education, 2009. ICCSE '09. 4th International Conference on* s.127,130, 25-28 July 2009. Dostupné z WWW: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5228509&isnumber=5228118

[10] KONCHADY, M. *Text Mining Application Programming.* Boston (Mass.) Charles River Media, 412 s., 2006. ISBN 978-1-58450-460-3.

[11] O'GRADY, P.D.; RICKARD, S.T. *Automatic ASCII Art conversion of binary images using non-negative constraints* Signals and Systems Conference, 208. (ISSC 2008). IET Irish , s.186-191, 18-19 June 2008. Dostupné z WWW: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4780951&isnumber=4780919

[12] Rapid - I - *RapidMiner* [online]. 2001 [cit. 2010-11-07]. RapidMiner. Dostupné z WWW: http://rapid-i.com/content/view/181/190/

[13] ROSŮLEK, L.; CHALOUPKA, J. *Rozpoznávání lidských emocí na základě pořízeného obrazu obličeje* Liberec : Technická univerzita v Liberci, 66 s., 2009.

[14] STEINWART, I.; CHRISTMANN, A. *Support Vector Machines*, Springer, New York, 2008. 602 s. ISBN 978-0-387-77241-7

[15] WALTHER, J. B.; D'ADDARIO, K. P. *The impacts of emoticons on message interpretation in computer-mediated communication* Social Science Computer Review vol.19 no.3 s.323–345 2001. Dostupné z WWW: http://ssc.sagepub.com/content/19/3/324

[16] *The WordNet Home Page* [online]. 1998 [cit. 2010-11-09]. The WordNet Reference Manual. Dostupné z WWW: http://wordnet.princeton.edu/wordnet/man/wngloss.7WN.html

[17] YE WU; FUJI REN. *Improving emotion recognition from text with fractionation training, Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on* ,s.1-7,2010. Dostupné z WWW: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5587800&isnumber=5587762

[18] ZHAO, HONG YAN. *The Analysis and Application of the C4. 5 Algorithm in Decision Tree Technology. Advanced Materials Research* s.754-757, 2012

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

a.s. Akciová společnost

G_{I} Gini Index

GUI Graphical User Interface - Grafické uživatelské rozhraní

H Entropie

I_{G} Information Gain - Informační zisk

IP Internet Protocol

k-NN k-Nearest Neighbours - algoritmus k-nejbližších sousedů

M_{E} Classification Error - Klasifikační chyba

N_{FN} počet chybně zařazených dat do dané kategorie (false negative)

N_{FP} počet chybně zařazených dat do dané kategorie (false positive)

N_{TN} počet správně zařazených dat do kategorie (true negative)

N_{TP} počet správně zařazených dat do kategorie (true positive)

R_{acc} accuracy - procentuální úspěšnost

RM Rapid Miner

RMSE Root Mean Square Error - Střední kvadratická chyba

RS Rozhodovací strom

s.r.o. Společnost s ručením omezeným

SMS Short message service - Služba krátkých textových zpráv

SVM Support Vector Machine - algoritmus podpůrných vektorů

TXT Textový soubor

URL Uniform Resource Locator - jednotný lokátor zdrojů

XML Extensible Markup Language

SEZNAM PŘÍLOH

A OBSAH CD

- **xvylic00-BP.pdf** – Elektronická verze této práce ve formátu PDF.
- **Data.zip** – Archiv obsahující zdrojové kódy a databázi trénovacích a testovacích dat.
- **rapidminer-DecisionTree-1.0.0.jar** - Knihovna obsahující vytvořený algoritmus rozhodovacího stromu prom RM.
- **rapidminer-Text Processing-5.3.001.jar** - Knihovna pro práci s textem spustitelná v RM.
- **rapidminer-5.2.008x32-install.exe** – Instalační soubor pro operační systém Windows – 32 bitová verze.
- **rapidminer-5.2.008x64-install.exe** – Instalační soubor pro operační systém Windows – 64 bitová verze.
- **TrainingSet.txt** - Trénovací set pro názorný příklad RS
- **ukazkove-procesy.zip** - Spustitelné příklady pro zpracování dat v RM.
- **wncze20** - Databáze WordNet.