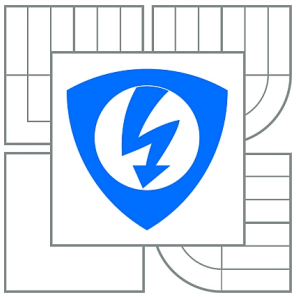
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁVÁNÍ EMOCÍ V ČESKY PSANÝCH TEXTECH

RECOGNITION OF EMOTIONS IN CZECH TEXTS

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

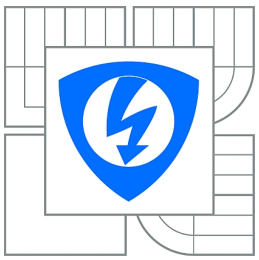
AUTOR PRÁCE — Bc. RADEK ČERVENEC

AUTHOR

VEDOUCÍ PRÁCE — Ing. RADIM BURGET, Ph.D.

SUPERVISOR

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student: Bc. Radek Červenec ***ID:*** 77712
Ročník: 2 ***Akademický rok:*** 2010/2011

NÁZEV TÉMATU:

Rozpoznávání emocí v česky psaných textech

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s problematikou dolování znalostí z textů a vhodně shrňte současný stav problematiky rozpoznávání emocí v česky psaných textech. Seznamte se také s problematikou ontologických bází a připravte trénovací množinu s ohodnocením. Navrhněte a vytvořte model pro natrénování klasifikátoru z pohledu rozpoznávání emocí 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: 7.2.2011 ***Termín odevzdání:*** 26.5.2011

Vedoucí práce: Ing. Radim Burget, Ph.D.

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

UPOZORNĚNÍ:

ABSTRAKT

Díky rozvoji informačních a komunikačních technologií v posledních letech došlo k velkému nárůstu množství informací, které denně vznikají ve formě elektronických dokumentů. Třídění a zpracování informací se stalo pro člověka velmi obtížné, a proto vzrůstá obliba systémů automatického dolování znalostí z textu. Zajímavou podoblastí jsou systémy pro analýzu sentimentu a automatického rozpoznání emocí v textech, které mají potencionálně široké uplatnění. V rámci této práce byl navržen a implementován systém využívající technik dolování znalostí z textu za účelem rozpoznávání emocí v česky psaných textech a bylo provedeno zhodnocení jeho úspěšnosti. Protože je systém postaven převážně na metodě strojového učení, byla navrhnuta a vytvořena trénovací množina, která byla posléze použita k vytvoření modelu klasifikátoru pomocí algoritmu podpůrných vektorů (SVM). Pro potřeby zpřesnění výsledků klasifikace textových dokumentů do předem definovaných emočních tříd, jsou do systému integrovány další prvky, jako např.: lexikální databáze, lemmatizátor a odvozený slovník klíčových slov. Součástí práce je také zhodnocení několika přístupů ke klasifikaci s různými modifikacemi navrženého systému.

KLÍČOVÁ SLOVA

dolování znalostí, genetický algoritmus, kategorizace, klasifikace, lemmatizace, lemmatizátor, lexikální databáze, rozpoznávání emocí, strojové učení, SVM, trénovací množina, WordNet

ABSTRACT

With advances in information and communication technologies over the past few years, the amount of information stored in the form of electronic text documents has been rapidly growing. Since the human abilities to effectively process and analyze large amounts of information are limited, there is an increasing demand for tools enabling to automatically analyze these documents and benefit from their emotional content. These kinds of systems have extensive applications. The purpose of this work is to design and implement a system for identifying expression of emotions in Czech texts. The proposed system is based mainly on machine learning methods and therefore design and creation of a training set is described as well. The training set is eventually utilized to create a model of classifier using the SVM. For the purpose of improving classification results, additional components were integrated into the system, such as lexical database, lemmatizer or derived keyword dictionary. The thesis also presents results of text documents classification into defined emotion classes and evaluates various approaches to categorization.

KEYWORDS

annotated corpus, categorization, classification, emotion detection, emotion recognition, feature selection, genetic algorithm, lemmatization, lemmatizer, lexical database, machine learning, SVM, text mining, WordNet

ČERVENEC, Radek *Rozpoznávání emocí v česky psaných textech*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2011. 67 s. Vedoucí práce byl Ing. Radim Burget, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Rozpoznávání emocí v česky psaných textech“ jsem vypracoval samostatně pod vedením vedoucího diplomové 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 diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucímu diplomové práce Ing. Radimovi Burgetovi, Ph.D. za vynikající vedení, cenné rady a neustálou motivaci při zpracování diplomové práce. Rád bych také poděkoval celé své rodině a přítelkyni Lence Martincové za neustálou podporu nejen při psaní této práce, ale i během studia.

V Brně dne ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Současný stav informačních a komunikačních technologií umožňuje uživatelům vytvářet a ukládat velké množství dat. Lidské schopnosti jejich analýzy jsou dosti omezené a tak nastává problémem dostupnosti relevantních informací v záplavě stále rostoucích objemů dat.

Dolování znalostí z textu (anglicky Text Mining) je vědní disciplína, která k extrakci užitečných informací z textových dokumentů využívá technik strojového učení, umělé inteligence, statistiky, zpracování přirozeného jazyka (NLP) a v neposlední řadě také dolování znalostí z dat (anglicky Data Mining).

Konkrétní typ a podoba informací, které mohou být z textu získávány jsou vždy závislé na povaze řešeného problému. Cílem této práce je použití metod dolování znalostí z textu za účelem rozpoznávání emocí v česky psaných textech. Analyzovány budou reálné příspěvky z blogů, technických podpor, sociálních sítí, elektronických diskuzí apod.

Emoce jsou obecně definovány jako subjektivní zážitky vztažené k povaze a momentálnímu rozpoložení jednotlivce. Vyskytují se v různých oblastech lidské komunikace a velmi často poskytují přídavnou informaci ve sdělení [7].

Na základě studií z neurologie a psychologie bylo zjištěno, že emoce hrají důležitou roli v rámci racionálního a inteligentního chování. Zároveň jsou také jedním z charakteristických znaků emoční inteligence, která je mnohými nadřazována nad inteligenci verbální či matematickou. Užitek v rozpoznávání emocí je možné najít při interakci mezi člověkem a strojem [26], ale i z hlediska analýzy lidského chování.

Rozpoznávání emocí není omezeno pouze na text. Byly zkoumány i techniky rozpoznávání emocí z akustického záznamu řeči, gest, snímků obličeje a dalších jedinečných biologických charakteristik [2], [27], [28].

Nicméně techniky detekce emocí v textu jsou stále nevyzrálou oblastí a před využitím v praktických aplikacích vyžadují podstatné zlepšení [23].

Hlavním přínosem této práce je návrh a implementace systému pro rozpoznávání emocí v česky psaných textech, včetně stanovení optimálních parametrů, zhodnocení jeho úspěšnosti a porovnání s ostatními systémy podobného typu, které byly vytvořeny pro jiné světové jazyky. Dále je to prezentace výsledků klasifikace textových dokumentů do předem definovaných emočních tříd a zhodnocení několika přístupů ke klasifikaci s různými modifikacemi navrženého systému. Mezi další přínosy patří návrh formátu a vytvoření početné trénovací množiny skládající se z reálných příspěvků a jejich manuální ohodnocení. Jako vedlejší produkt vznikl softwarový nástroj, který umožňuje hodnocení prvků trénovací množiny skrze grafické uživatelské rozhraní a umožňuje následný export do XML souborů v požadované podobě.

Práce má následující strukturu. První část je teoreticky zaměřená a snaží se

vybudovat základy, kterých je posléze využito v praktické části. Shrnuje současné přístupy k rozpoznávání emocí z textu a uvádí příklady konkrétního využití. Popisuje obecnou architekturu systému pro dolování znalostí v textu s detailním zaměřením na jednotlivé části, jako je předzpracování či klasifikace včetně problematiky lexikálních bází. Praktická část se zabývá návrhem a implementací systému pro rozpoznání emocí v textu s využitím technik dolování znalostí z textu. Popisuje návrh formátu a tvorbu trénovací množiny s ohodnocení, která byla základem pro implementaci metody založené na strojovém učení. V práci jsou detailně popsány jednotlivé kroky implementace předzpracování textu (segmentace textu, filtrace tokenů, lemmatizace), hledání hyperonymických vztahů pomocí lexikální databáze a optimalizační kroky v rámci selekce atributů. V závěrečné části jsou prezentovány nalezené optimální parametry pro modely klasifikátoru a dosažené výsledky klasifikace textu do definovaných emočních tříd. Dále je provedeno zhodnocení několika přístupů ke klasifikaci, analýza příčin neúspěšné klasifikace a porovnání vytvořeného systému s jinými systémy podobného charakteru.

1 MOŽNOSTI VYUŽITÍ ROZPOZNÁVÁNÍ EMOCÍ V TEXTU

Příkladů využití je mnoho, zde budou uvedeny jen ty nejzákladnější. Obecně systém pro automatické rozpoznávání emocí s využitím praktik dolování znalostí z textu, má potencionální uplatnění všude tam, kde je uchováváno velké množství textu a emoční náboj, který nese, je důležitý. A zároveň je časově náročné a nákladné tyto objemy textu zpracovávat lidskými silami.

1.1 Bezpečný Internet

Rozvoj elektronické komunikace a stále rostoucí počet uživatelů sociálních sítí přináší i nové on-line hrozby. Jednou z nich je např. kyberšikana, což je forma šikany, která k ubližování, obtěžování, vydírání či zastrašování využívá prostředků informačních technologií jako je e-mail, elektronická diskuze, chat, fóra, SMS zprávy nebo sociální sítě. Dle nedávné výzkumné studie [25], je téměř polovina českých dětí vystavena některé z forem kyberšikany (46,8 %). Potenciální aplikace je tedy v Centrech bezpečného Internetu a institucích, které se zabývají touto problematikou. Systémy mohou být využívány k automatické a z časového hlediska velmi efektivní filtraci blogů, s požadavkem na vyhledání těch, které vyvolávají negativní emoce (například strach, hněv, odpor, znechucení či úzkost) nebo mají vulgární obsah.

1.2 Průzkumy veřejného mínění

Možná aplikace se nabízí v souvislosti s průzkumy trhů a veřejného mínění. Z elektronických médií a publikací je možno zjišťovat s jakým emočním nábojem je psáno v souvislosti s danou firmou, politikem, institucí a poté patřičně na výsledky reagovat.

1.3 Reklamní průmysl

Reklama je velmi často založena na emocích. Proto systém pro detekci emocí může měřit kvalitu reklamy na základě analýzy reakcí a diskuzí s ní spojených. Zároveň s využitím znalosti systémem odvozených klíčových slov a slovních spojení, které jsou často spojovány s různými emocemi, umožňuje podávat náměty na obsah efektivní a cílené reklamy.

1.4 Péče o zákazníky

V rámci technických podpor, oddělení komunikace s veřejností nebo podpůrných linek, které využívají přepisy hovorů do elektronické podoby, je možno automaticky monitorovat reakce zákazníků. V případě vysoké koncentrace např. příliš negativních reakcí je možno analyzovat příčiny či učinit kroky, které zlepší obraz společnosti v očích zákazníka.

1.5 Interakce mezi člověkem a strojem

Jak je uvedeno v pramenu [26], automatické rozpoznávání emocí u systémů zvyšuje jejich schopnost zpracovávat znalosti a do jisté míry se samostatně rozhodovat, tzn. dělá stroj inteligentnější z pohledu uživatele.

2 SOUČASNÉ PŘÍSTUPY

Tradiční přístupy zpracování velkého množství dat lidskými silami jsou založeny na rozhodování s využitím přirozené inteligence, instinktu a intuice. Klíčem k úspěchu je uspořádání informací, tak aby bylo možné porozumět jejich kontextu. Velkou nevýhodou toho přístupu je časová náročnost, kdy analýza velkého množství dat může trvat týdny, měsíce i roky.

Naproti tomu strojové zpracování je charakteristické rozhodováním na základě faktů (nikoliv instinktu a intuice). Z hlediska časové náročnosti se jedná o několikanásobně rychlejší způsob analýzy a strukturování textových elektronických dokumentů. Donedávna velkou nevýhodou tohoto přístupu byla absence vzájemných vazeb a vztahů mezi jednotlivými analyzovanými informacemi (charakteristické pro člověka a jeho chápání světa). V této souvislosti bývají diskutovány tzv. ontologické báze, které obsahují člověkem vytvořené, formálně specifikované pojmy a vzájemné sémantické vztahy mezi nimi, za účelem společného chápání těchto pojmů pro potřeby strojového zpracování textu. Díky jejich vzniku je výše zmíněný nedostatek postupně eliminován a strojové zpracování přirozeného jazyka se stává velmi perspektivní nejen při dolování užitečných informací textu, ale i např. v boji s internetovou kriminalitou.

V obecné formulaci je možné problém rozpoznávání emocí v textu redukovat na hledání vztahu mezi specifickým textem a emocí, kterou autor textu prožíval, případně chtěl vyjádřit. Jinými slovy, je snahou efektivně klasifikovat vstupní text do předem určených emočních tříd.

Kromě níže uvedených metod, které převážně detekují emoce z textů na úrovni vět, existují i práce, zabývající se odvozením emocí na úrovni odstavců i kompletních článků, např. [10]. Podle pramenu [23], je možné rozčlenit současné přístupy k detekci emocí v textu do tří kategorií, které jsou zde rozebrány. Velká většina uvedených metod byla implementována pro anglický jazyk (případně čínštinu), tudíž nejsou přímo aplikovatelné pro češtinu. Důvody pro tuto skutečnost jsou zejména: značná rozdílnost charakteru jazyka (čeština vs. čínština) a závislost částí vytvořených systémů na jazyku (trénovací množina, slovníky klíčových slov, lexikální databáze, model klasifikátoru svázaný s trénovací množinou apod.).

2.1 Detekce na základě klíčových slov

Jedná se o nejpřímější a nejintuitivnější metodu. Úspěšnost je velmi závislá na fázi předzpracování textu a to především na separaci vstupních vět na slova, nalezení klíčových slov a tvorbě databáze klíčových slov příslušícím k jednotlivým emočním třídám. Mezi nevýhody této metody patří neschopnost detekce emocí ve větách, které

neobsahují klíčová slova. Další nepřesnosti způsobuje mnohoznačnost některých slov a fakt, že nebere v potaz lingvistickou strukturu vět. Jedna z možností jak zpřesnit klasifikaci prostřednictvím této metody spočívá ve tvorbě databáze klíčových slov za pomocí ontologické báze, která obsahuje vzájemné vazby mezi jednotlivými slovy [4]. Díky tomu je možné vytvořit komplexnější slovník.

2.2 Detekce s využitím strojového učení

Tento přístup využívá pro detekci emocí učících se algoritmů, jejichž parametry byly nastaveny ve fázi trénovaní na určité množině již předem ohodnoceného textu. V mnoha odborných pracích je často využíván SVM (algoritmus podpůrných vektorů – Support Vector Machine), např. v pramenu [7] je provedeno porovnání různých metod strojového učení v rámci klasifikace titulků novinových článků do definovaných emočních tříd a algoritmus SVM překonává ostatní metody. Stejný algoritmus byl také použit v pramenu [35] a [8]. Úspěch těchto metod je postaven na kvalitě trénovací množiny a často i na určitých klíčových slovech. Dalším problémem může být skutečnost, že většina přístupů založených jen na strojovém učení je schopna spolehlivě klasifikovat pouze do dvou tříd.

2.3 Detekce pomocí hybridních metod

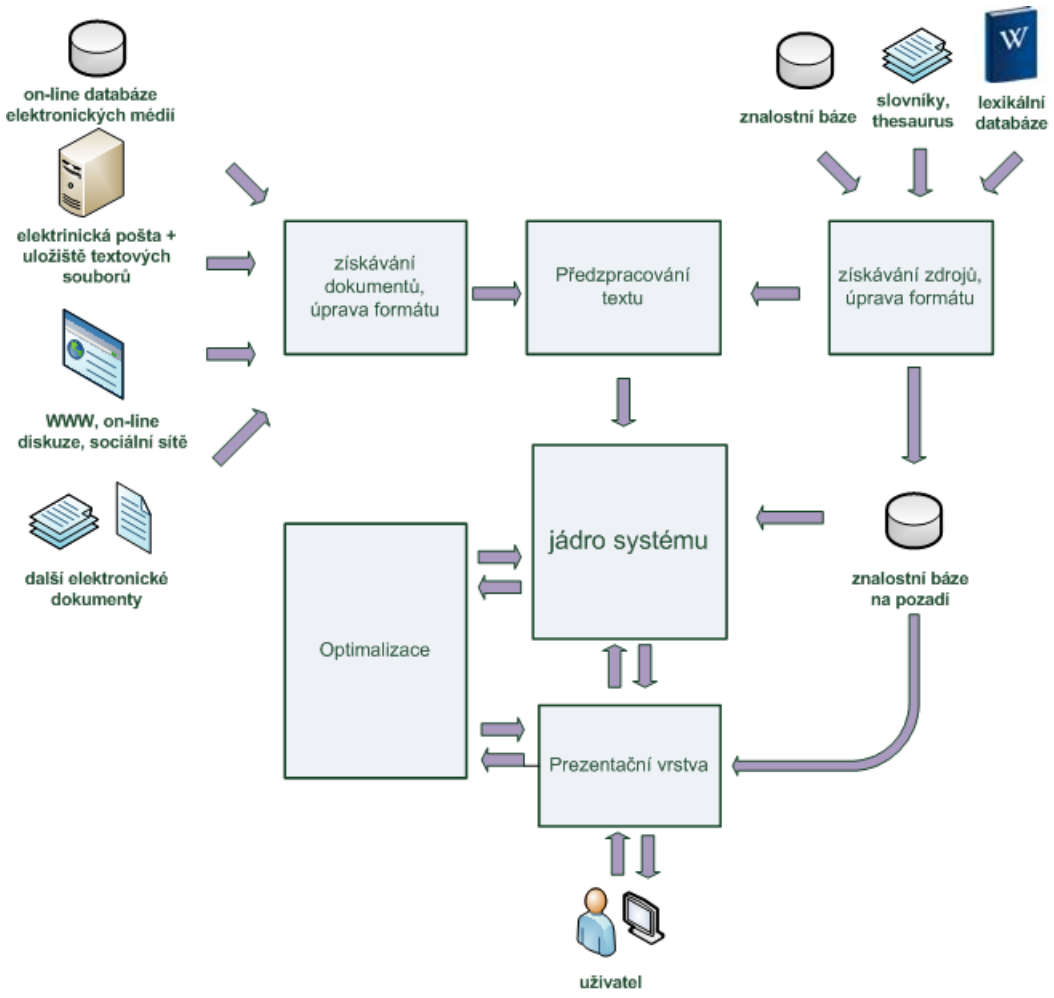
Hybridní metody pro zvýšení účinnosti kombinují výhody z detekce emocí na základě klíčových slov a metod vycházejících ze strojového učení [31], [38]. Tyto přístupy velmi často využívají pro vhodnou reprezentaci textu ve fázi předzpracování lexikální databáze, které obsahují vzájemné vazby mezi jednotlivými slovy a slovními spojeními. Společná idea těchto přístupů z hlediska analýzy, je použití velkého počtu slov sémanticky spojených s danou emocí. K získání takovýchto vazeb je možné použít např. tezaurus [1].

3 ARCHITEKTURA SYSTÉMU DOLOVÁNÍ ZNALOSTÍ Z TEXTU

Architektura obecného systému pro získávání znalostí z textu je zobrazena na obr. 3.1 a může být rozdělena na následující části [13]:

- *Předzpracování textu* – Zahrnuje všechny operace sloužící jako příprava textu pro další zpracování jádrem systému. Především se jedná o strukturování původní formy textu, extrakci klíčových slov a vytvoření nové reprezentace dokumentu. Může také obsahovat metody, které k dokumentu připojí pro další zpracování důležitou přídavnou informaci (časové razítko, zdroj dokumentu).
- *Jádro systému* – Je srdcem celého systému a zahrnuje algoritmy, které se bezprostředně podílejí na dolování znalostí. Jedná se tedy o hledání vztahů mezi dokumenty a jednotlivými entitami, určování druhu textu (klasifikace a shlukování dokumentů), určování sentimentu, shrnutí textu, analýza trendů, určení důležitosti dokumentu apod.
- *Prezentační vrstva* – Zahrnuje nástroje, které uživatelům prezentují znalosti získané z jádra systému a umožňují prostřednictvím ovládacích prvků ovlivňovat činnost systému. Do této kategorie řadíme: GUI (grafické uživatelské rozhraní – Graphical User Interface), vizualizační nástroje, editory pro zadávání příkazů a filtračních kritérií.
- *Optimalizace* – Jedná se o techniky vylepšení výstupu pro prezentační vrstvu (filtrace redundantních informací, shlukování, zobecnění), ale také o metody optimalizující činnost algoritmů pro dolování znalostí a hledání optimálních parametrů (např. genetické algoritmy).

V případě, že systém operuje převážně nad daty z určité domény (rybolov, medicína, automobily, IT atd.), je vhodné využít externích zdrojů v podobě znalostních bází, lexikálních databází, ontologických bází, tezaurů a slovníků. Tyto zdroje poskytují přídavnou informaci o dané doméně a umožňují identifikovat významné názvosloví, slovní spojení a všeobecně vzájemné vztahy mezi nimi. Z tohoto důvodu je vhodné jejich využití ve fázi předzpracování, v jádře systému i na úrovni prezentační vrstvy.



Obr. 3.1: Architektura obecného systému dolování znalostí z textu.

4 PŘEDZPRACOVÁNÍ TEXTU

Operace spojené s dolováním znalostí z textu jsou do velké míry závislé na předchozím předzpracování textu. Hlavní úkolem předzpracování je získat strukturovanou reprezentaci textu z původní nestrukturované podoby. Textový dokument ve své původní podobě je chápán jako abstraktní entita, která může nabývat různých významů. Strukturováním dokumentu je získána taková reprezentace, ze které je snadnější rozpoznat podstatu dokumentu. V následující části budou představeny nejpoužívanější kroky pro předzpracování textu.

4.1 Segmentace textu

Text v podobě vět a větných spojení představuje z hlediska dolování znalostí z textu pouze sekvenci znaků, bez explicitního vyjádření hranice mezi slovy a logických spojení. Jinými slovy je nutné tuto sekvenci znaků segmentovat do tzv. tokenů, které jsou elementárními nositeli informace v rámci daného jazyka. Tokeny nemusí být reprezentovány pouze slovy vzniklými rozdělením textu podle mezer a interpunkčních znamének. V závislosti na algoritmu segmentace, mohou být tokeny vytvořeny spojením znaků, které by odděleně ztrácely svůj význam. Například IP adresa je ve svém dekadickém vyjádření tvořena čísly oddělenými tečkami a tudíž pro zachování jejího významu nesmí být rozdělena na několik nezávislých částí.

I když se na první pohled může zdát, že segmentace představuje triviální operaci, jsou s ní spojené určité problémy:

- *Zkratky* – Jelikož zkratky obsahují tečky, může při segmentaci souvětí do vět vzniknout problém s rozpoznáním konce věty. Navíc je nutné zachovat soudržnost zkratek (například: př. n. l. - před naším letopočtem nebo t. č. – toho času). Řešení může spočívat v použití seznamu nejčastěji používaných zkratek během segmentace.
- *Internetové tokeny* – Spočívá v identifikaci a zachování IP adres, URL, doménových jmen, adres elektronické pošty a dalších významných řetězců, které splňují určitý, předem definovaný vzor. Vhodným nástrojem pro porovnávání vzoru s řetězcem jsou regulární výrazy.
- *Emotikony* – Jedná se o grafický symbol složený z různých znaků vyjadřujících emoce autora textu (smajlík). Z pohledu rozpoznávání emocí v textu je to užitečný prvek a neměl by být v rámci segmentace filtrován.
- *Čísla* - Identifikace a spojení čísel, která mohou být oddělena čárkou či tečkou. Mohou obsahovat i znaménko (+ nebo -).

Jedna z možných implementací systému pro tvorbu tokenů ze vstupního text je znázorněna v podobě blokového schéma na obr. 4.1 [24].

Obr. 4.1: Blokové schéma systému pro tvorbu tokenů.

4.2 Určení slovních druhů

Určování slovních druhů (anglicky POS tagging) má za úkol přiřadit jednotlivým slovům v textu slovní druh v závislosti na kontextu ve kterém se vyskytují. Znalost slovních druhů umožňuje rozlišit skutečný význam u mnohoznačných slov. Další velkou výhodou je identifikace vazeb při hledání v lexikálních bázích v rámci dalšího zpracování. Pro určování slovních druhů jsou často využívané statistické modely, např. HMM (skrytý Markovův model – Hidden Markov Model) [19].

4.3 Určení základního tvaru slova

Určení základního tvaru slova (anglicky stemming), je z hlediska předzpracování textu technikou, jenž se snaží využít lexikografických pravidel k získání kořene slov (anglicky stem), přičemž nemusí jít nutně o morfologický kořen slova. Výhodou je dosažení sjednocení slov se stejným základem, která se v textu vyskytují v různých tvarech. Jedním z problémů této metody je převedení slov s jiným významem na stejný základ, čímž dojde ke ztrátě informace. Další nevýhodou je fakt, že algoritmus vyvinutý pro jeden jazyk není univerzálně aplikovatelný na jiné jazyky. Mezi nejznámější algoritmy pro anglický jazyk patří Lovinsův stemmer a Porterův stemmer [17]. Strategii určování základního tvaru slova pro češtinu lze nalézt např. v pramenu [11].

4.4 Lemmatizátor

Lemmatizace je proces transformace slov na jejich normalizovaný neboli slovníkový tvar (lemma). Je to velmi užitečný postup zejména pro tzv. flektivní jazyky (např. čeština), v nichž se vyskytuje velké množství skloňovaných slovních tvarů. Normalizované (slovníkové) tvary se liší v závislosti na jazyku, pro češtinu to mohou být např.:

- *podstatná jména* – 1. pád jednotného čísla
- *přídavná jména* – 1. pád jednotného čísla, mužského rodu a prvního stupně v rámci skloňování
- *slovesa* – infinitiv
- *ostatní slovní druhy* – lemma je shodné s jediným tvarem

Tradiční metody lemmatizace vychází z pravidel, která jsou stanovena odborníky a jsou ušita na míru pro daný jazyk. Ačkoliv tento způsob může být účinný, je také velmi pracný a časově náročný. Ukazuje se, že techniky založené na strojovém učení mohou být levnější a rychlejší variantou z hlediska odvození pravidel pro lemmatizaci. Také umožňují lépe zpracovávat slova, která nejsou přímo v pravidlech stanovena a bývá snadnější udržovat, upravovat a rozšiřovat trénovací množinu nežli komplexně specifikovaná pravidla. Další výhodou tohoto přístupu je univerzálnost (jeden systém může být použit napříč více jazyky).

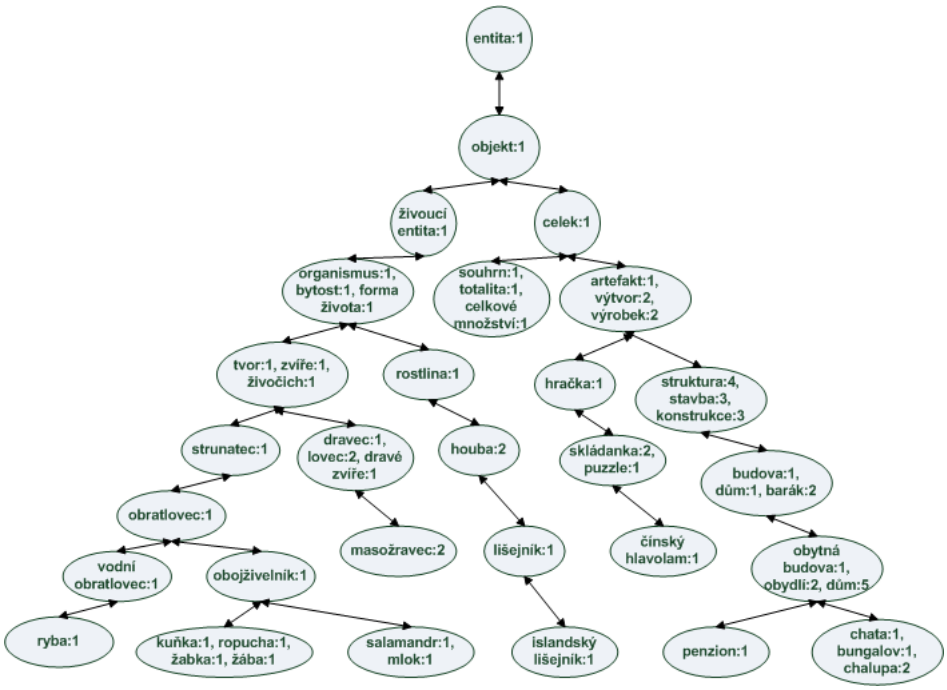
V současnosti jeden z nejefektivnějších lemmatizátorů pro slovanské jazyky (včetně češtiny) je založen na strojovém učení a prezentován v pramenu [22]. Tento systém je možné aplikovat až na 12 evropských jazyků.

4.5 Ontologické báze

V oblasti informatiky je ontologie chápána jako člověkem vytvořená explicitní, formální specifikace pojmů a vzájemných vztahů mezi nimi. Cílem je definovat společné chápání těchto pojmů v rámci strojového zpracování přirozeného jazyka [32]. V současnosti nejpopulárnější lexikální databází pro budování ontologie je WordNet. Jedná se o databázi ve které jsou podstatná jména, přídavná jména, slovesa a příslovce vzájemně provázány sémantickými vztahy. Zároveň platí, že tento lexikální zdroj obsahuje kromě jednoslovných záznamů i slovní spojení. Podle pramenu [37] a [32] jsou ve WordNet následující vazby:

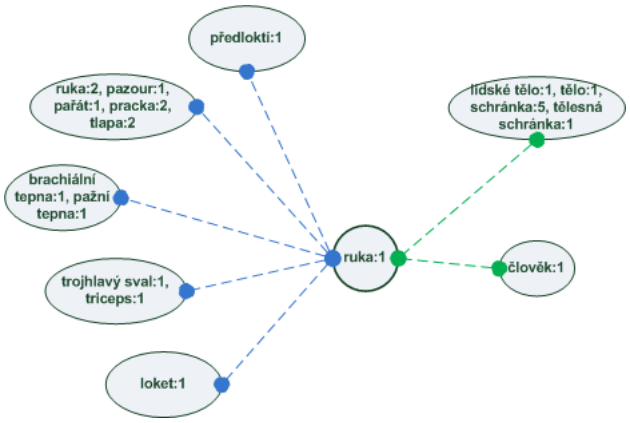
- *Synsety* – Jedná se o vzájemné vztahy mezi synonymy, tj. slovy s podobným či stejným významem. Synsety jsou základní stavební jednotkou lexikální databáze. Jednotlivé prvky synsetu jsou nazývány literály. Literály mají přidružený číselný identifikátor určující jejich význam.

- *Hyperonymické vztahy* – Jedná se o vztah, který spojuje určitý synset s jeho obecnějším významem, viz obr. 4.2 (vytvořeno pomocí nástroje [16]), kde obecnější význam od synsetu s literálem *ryba* je synset s literálem *vodní obratlovec*.
- *Hyponymické vztahy* – Jedná se o vztah, které spojuje určitý synset s jeho konkrétnějším významem, viz obr. 4.2, kde konkrétnější význam od synsetu s literály: obytná budova, obydlí, dům je synset s literálem *chata*.
- *Meronymie* – vytváří vztah mezi částí a celkem, viz obr. 4.3 (vytvořeno pomocí nástroje [16]), kde loket je část ruky.
- Holonymie – vztah mezi částí a celkem, ale v opačném smyslu než meronymie, viz obr. 4.3 (vytvořeno pomocí nástroje [16]) kde část člověka je ruka.



Obr. 4.2: Příklad hyperonymických a hyponymických vztahů mezi synsety.

Použití WordNet a lexikálních databází obecně, má v oblasti dolování znalostí z textu široké využití. Z hlediska detekce emocí můžeme najít aplikaci v podobě hledání hyperonymických vztahů za účelem dosažení redukce celkové dimenze všech analyzovaných slov, aniž by slova ztratila svůj sémantický význam. Tento přístup byl úspěšně aplikován v pramenu [38]. Další možností je využití vzájemných vztahů mezi synsety k tvorbě slovníku v kombinaci s metodami detekce emocí na základě klíčových slov [4].



Obr. 4.3: Příklad meronymie a holonymie pro synset s literálem ruka.

4.6 Vektorový model dokumentu

Běžné učící se algoritmy používané k vytvoření modelu klasifikátoru nemohou pracovat s textem v jeho původní podobě. Nevhodný je i formát, který vznikne po segmentaci textu (posloupnost tokenů). Proto je nutné v rámci předzpracovaní vytvořit jinou reprezentaci textu. Nejčastěji používaným přístupem pro tento účel je reprezentace dokumentu prostřednictvím vektorového modelu (anglicky vector space model). Dokument je v tomhle smyslu chápán jako obecný text tvořený slovy (případně posloupnostmi tokenů), větou nebo souvětím.

Mějme množinu všech dokumentů jenž obsahuje celkem n tokenů. Každý dokument z této množiny je možné reprezentovat jako vektor v n dimenzionálním prostoru R^n. Hodnoty jednotlivých složek vektoru mohou být různé. Nejjednodušším způsobem je binární reprezentace, kdy složka vektoru nabývá hodnoty 1 nebo 0 v závislosti na přítomnosti daného tokenu v dokumentu. Další možností je váhovat jednotlivé složky pomocí relativního vyjádření četnosti výskytu tokenů v dokumentu podle vztahu [14]:

$$TF_{rel}(t,d) = \frac{t_f(t,d)}{n(d)} \tag{4.1}$$

kde hodnota v čitateli $t_f(t,d)$ vyjadřuje počet výskytů tokenu t v dokumentu d. Hodnota ve jmenovateli $n(d)$ znamená počet všech tokenů v dokumentu d.

Jedním z nejpoužívanějších způsobů je TF-IDF (Term Frequency-Inverse Document Frequency), který k určení váhy pro jednotlivé složky vektoru využívá jednak četnosti výskytu tokenu v rámci jednoho dokumentu i v rámci celého korpusu [29]:

$$TF\text{-}IDF(t,d) = t_f(t,d) \cdot \log \frac{N}{d_f(t)} \tag{4.2}$$

kde $t_f(t, d)$ vyjadřuje počet výskytů tokenu t v dokumentu d, N je počet všech dokumentů v korpusu a $d_f(t)$ vyjadřuje počet dokumentů obsahující token t.

5 KLASIFIKACE TEXTU

Úkolem klasifikace je rozčlenit analyzované dokumenty do předem definovaných skupin (tříd). Existuje mnoho aplikací využívající různá kritéria pro klasifikaci: podle pohlaví autora, klasifikace nevyžádané pošty, podle obsahu webových stránek, podle autora dokumentu, na základě předem definovaných emočních tříd apod.

Současné přístupy ke klasifikaci textu je možné rozdělit na dvě kategorie [13]:

- *Na základě znalostního inženýrství* – Pravidla pro kategorizaci jsou stanovena experty v dané oblasti, ta jsou poté integrována do systému. Tento přístup je drahý a časově náročný, na druhou stranu bývá velmi účinný.
- *Na základě strojového učení* – Model klasifikátoru je postupně odvozován z již klasifikovaných příkladů (trénovací množina). Jedná se o levnější variantu, která je i výhodnější z hlediska časové náročnosti. Úspěšnost klasifikátoru je velmi závislá na kvalitě trénovacích dat. Tento proces je znám jako kontrolované učení (anglicky supervised learning). Existuje také nekontrolované učení pomocí shlukování dat (anglicky data clustering).

V další části budou představeny nejčastěji používané algoritmy pro vytvoření modelu klasifikátoru.

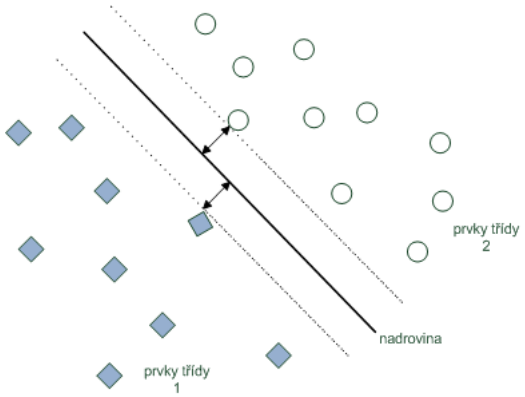
5.1 Algoritmy podpůrných vektorů

Algoritmus podpůrných vektorů SVM je velmi rychlý, efektivní a často používaný způsob klasifikace [13]. Algoritmus dokáže lineárně separovat i data, která původně lineárně separabilní nejsou. Prostřednictvím nelineárního mapování převede trénovací data (vyjádřená pomocí vektorového modelu) do vyšší dimenze, kde hledá optimální nadrovinu (rozhodovací hranici), která by ideálně oddělovala jednotlivé třídy. Jak ilustruje obrázek 5.1, požadavkem při hledání nadroviny je maximální vzdálenost mezi nadrovinou a nejbližším prvkem jednotlivých tříd [14]. Samotná rozhodovací hranice je určena pouze relativně malým množstvím prvků z trénovací množiny (podpůrnými vektory). Ostatní prvky nemají na utváření modelu klasifikátoru žádný vliv [13].

Detailní popis transformace z původního prostoru na lineárně separabilní úlohu je možné najít v pramenu [6].

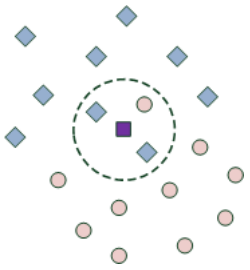
5.2 K-nejbližších sousedů

Algoritmus k-nejbližších sousedů k-NN (k-Nearest Neighbours) je řazen mezi algoritmy líného učení (anglicky lazy learner). Model klasifikátoru je vytvářen až ve fázi



Obr. 5.1: Ukázka separace prvků tříd nadrovinou nalenzenou pomocí SVM.

klasifikace. Zjednodušený příklad klasifikace ve dvourozměrném prostoru ilustruje obrázek 5.2, v prostoru se nachází prvky trénovací množiny zařazené do dvou tříd. Při určovaní třídy neznámého prvku je vyhledáno k nejbližších sousedů, třída která je nejvíce zastoupena v těchto prvcích je poté přiřazena i neznámému prvku. Je vhodné volit za k liché číslo a předejít tak nejednoznačnostem při klasifikaci.



Obr. 5.2: Ukázka klasifikace pomocí algoritmu k-nejbližších sousedů.

Často používanou metrikou pro určování vzdálenosti je Euklidovská metrika, kde obecně vzdálenost dvou bodů x_1 a x_2 v n dimenzionálním prostoru lze určit podle vztahu [14]:

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{1i})^2} \tag{5.1}$$

6 TVORBA VLASTNÍHO ŘEŠENÍ

Jako vhodné řešení pro implementaci systému pro rozpoznávání emocí v textu byla zvolena metoda, která využívá principů strojového učení, kdy model klasifikátoru je postupně odvozován na základě předem ohodnocených příkladů z trénovací množiny. Algoritmy pro klasifikaci textu, jsou postaveny na algoritmech podpůrných vektorů (SVM), jenž umožňují velmi rychlý a efektivní způsob klasifikace textu [13]. Fáze předzpracování textu, společně s využitím lexikální báze Český WordNet a vytvoření nové reprezentace textu pro další zpracování byla implementována jako samostatný program v programovacím jazyku Java. Pro účely optimalizace, klasifikace a vyhodnocení přesnosti klasifikace bylo použito prostředí RapidMiner [30].

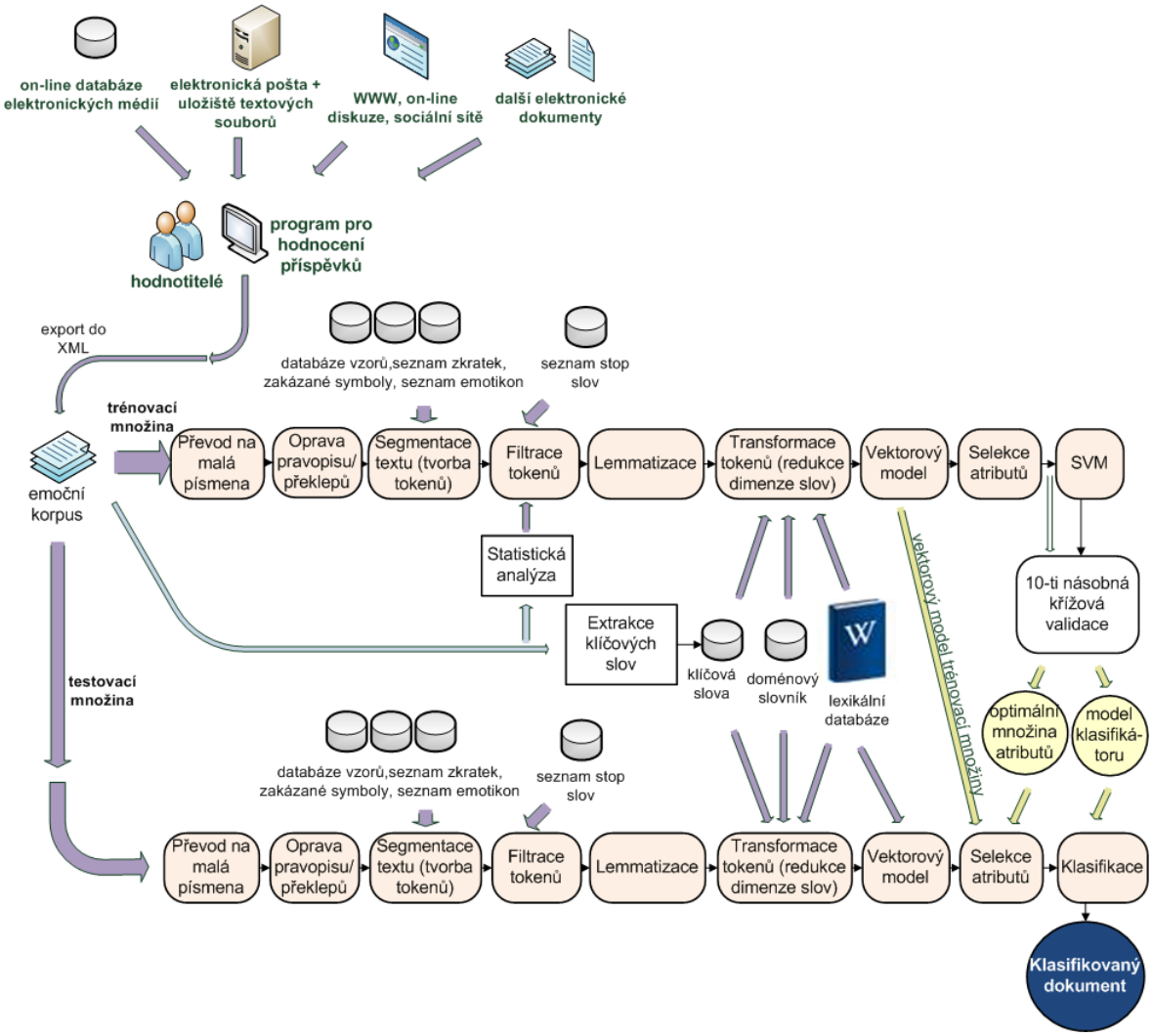
6.1 Blokové schéma systému

Celkové řešení je možno rozdělit na tří části: tvorba trénovací množiny, trénovací část a část zahrnující validaci. Blokové schéma na obr. 6.1 ilustruje jednotlivé kroky, jenž byly v rámci implementace použity a jejich vzájemnou návaznost.

Tvorba trénovací množiny zahrnuje výběr vhodných příspěvků z elektronických diskuzí a dokumentů. Ohodnocení těchto příspěvků hodnotiteli prostřednictvím vytvořeného softwarového nástroje a následný export ohodnocených příspěvků v požadované struktuře ve formátu XML.

Takto ohodnocená množina příkladů různých textů je základem pro trénovací fázi. V rámci lexikální analýzy je proveden převod všech písmen na malá písmena abecedy. Poté je provedena automatická kontrola a případná oprava pravopisu a překlepů. Následuje rozčlenění vět do tzv. tokenů, jakožto elementárních nositelů informace v rámci daného jazyka. Dalším krokem je filtrace tokenů využívající seznam slov, která sama o sobě nenesou žádný sémantický význam. Následuje využití lexikální databáze Český WordNet pro vytvoření nové reprezentace tokenů na základě nalezených hypernymů. Při hledání hyperonymických vztahů je využíván doménový slovník klíčových slov. Reprezentace textu je poté znovu změněna v rámci tvorby vektorového modelu s určením vah pro jednotlivé tokeny. S ohledem na přesnost klasifikace je provedena selekce atributů. Posledním krokem trénovací fáze je aplikace učícího se algoritmu podpůrných vektorů pro vytvoření modelu klasifikátoru. Ověření klasifikátoru je realizováno pomocí 10-ti násobné křížové validace. Výstupem trénovací části jsou: optimální množina atributů podílející se na tvorbě klasifikátoru, nastavení vah pro jednotlivé atributy v rámci vektorového modelu a optimální model klasifikátoru.

Prvky testovací množiny jsou podrobeny stejným operacím předzpracování textu jako prvky sloužící pro trénovací fázi. Rozdíl je v tvorbě vektorového modelu doku-



Obr. 6.1: Blokové schéma řešení systému pro detekci emocí v textu.

mentu kdy je brán ohled na vektorový dokument vzniklý po testovací fázi a s pomocí lexikální databáze je provedeno odpovídající mapování jednotlivých atributů. V posledním kroku se již prvky nepodílejí na tvorbě modelu klasifikátoru, ale tento model je na ně aplikován za účelem klasifikace do předem definovaných tříd.

Jednotlivé bloky budou v následujících části podrobně rozebrány.

6.2 Návrh a tvorba trénovací množiny

Trénovací množina hraje klíčovou roli v úspěšnosti řešení založených na kontrolovaném strojovém učení. Proto návrhu a manuální tvorbě množiny s ohodnocením byla věnována značná pozornost. Na návrh formátu trénovací množiny byly kladeny následující požadavky:

- Vytvořit takový systém ohodnocení příspěvků, který bude zajišťovat nezávislost dat na budoucí použité technice, algoritmu či prostředí jejich analýzy.
- Umožnit hodnocení a posléze analýzu na úrovni odstavců, vět, klíčových slov či slovních spojení.
- Vytvořit přenositelný a na platformě nezávislý formát.
- Umožnit hodnotit příspěvky teoreticky neomezenému počtu hodnotitelů.
- Možnost zahrnout identifikaci zdroje příspěvku včetně kategorie.
- Umožnit bližší identifikaci smyslu příspěvků (dotaz od zákazníka, odpověď na zákazníkův dotaz, neformální rozhovor na fóru, apod.).
- V rámci hodnocení umožnit kombinování různých emočních tříd.

Výsledkem je formát XML (Extensible Markup Language), jehož struktura bude posléze popsána. Tento formát jednoznačně splňuje požadavek na přenositelnost dat. Všechny výše uvedené cíle byly do návrhu úspěšně zahrnuty. Navíc formát XML umožňuje další případnou rozšiřitelnost.

6.2.1 Třídy emocí

Jednotlivé příspěvky byly v rámci tvorby trénovací množiny manuálně hodnoceny (kategorizovány) do jednotlivých tříd vzhledem k dané skupině emocí, která je zastřešuje. Hodnocení probíhalo na úrovni vět i větných spojení a je nutné zdůraznit, že bylo subjektivní. Tudíž dva různí hodnotitelé mohou mít na daný příspěvek zcela odlišný názor.

V rámci této práce byly zvoleny dvě skupiny tříd. Hodnocení probíhalo výhradně z pohledu autora, tzn. snahou bylo přiřadit k danému textu takovou emoci, kterou autor prožíval či chtěl vyjádřit. První s názvem *1DET* je zaměřena výhradně na třídy emocí ve vztahu k pozitivním a negativním emocím (podobný způsob kategorizace textu je možné najít např. v pramenu [33]). Popis jednotlivých tříd včetně názvosloví pro XML formát je uveden v tabulce 6.1. Druhá skupina tříd emocí s názvem *2DET*, je popsána v tabulce 6.2. Snahou této skupiny je komplexnější definice tříd emocí a hodnocení textu. Tato skupina emočních tříd bývá často používána v odborných publikacích [7].

6.2.2 Popis formátu

Seznam všech elementů, které jsou součástí formátu trénovací množiny včetně jejich povinných a nepovinných vnořených elementů je uveden v tabulce 6.3. Konkrétní příklad jednoho prvku trénovací množiny ve formátu XML je uveden na obr. 6.2.

Tab. 6.1: Popis názvosloví a významu pro skupinu emočních tříd *1DET*

XML název třídy	**Charakteristika třídy**
vulgar	Text byl vytvořen s negativní emocí a navíc má i vulgární charakter. Např.: obsahuje vulgarismy, má urážlivý význam, může obsahovat tabuizovaná témata, nevhodné pro danou situaci apod.
negativeL3	Autor vyjádřil negativní emoce subjektivně nejvyššího stupně bez vulgarismů (např.: zloba, hněv, vztek, znechucení, pomluva, odpor, strach, žal, smutek apod.).
negativeL2	Autor vyjádřil negativní emoce subjektivně středního stupně stupně bez vulgarismů (např.: zloba, hněv, vztek, znechucení, pomluva, odpor, strach, žal, smutek apod.).
negativeL1	Autor vyjádřil negativní emoce subjektivně nejnižšího stupně stupně bez vulgarismů (např.: zloba, hněv, vztek, znechucení, pomluva, odpor, strach, žal, smutek apod.).
neutral	Text byl vytvořen s neutrální emocí (nespadá do žádné s dalších skupin).
positive	Autor vyjádřil pozitivní emoce (např.: pochvalu, radost, případně positivní překvapení atd.).

Tab. 6.2: Popis názvosloví a významu pro skupinu emočních tříd *2DET*

XML název třídy	**Charakteristika třídy**
anger	Autor textu vyjádřil hněv, zlost nebo vztek.
disgust	Autor textu vyjádřil odpor či znechucení.
fear	Autor textu vyjádřil strach či obavu.
sadness	Autor textu vyjádřil smutek nebo žal.
surprise	Autor textu vyjádřil překvapení, úžas, údiv.
joy	Autor textu vyjádřil radost či štěstí.
none	Autor textu nevyjádřil žádnou z předchozích emocí (neutrální obsah).

```
<?xml version="1.0" encoding="UTF-8"?>
<emotionsTrainData xml:lang="en">
  <trainUnit>
    <emotionEvalParagraph>
        <class7>anger</class7>
        <class1d>negativeL2</class1d>
        <term>nic nefunguje</term>
        <term>přicházim</term>
        <term>není možný</term>
    </emotionEvalParagraph>
    <emotionEvalSentence>
        <class7>anger</class7>
        <class1d>negativeL2</class1d>
        <ending>nefunguje.</ending>
    </emotionEvalSentence>
    <emotionEvalSentence>
        <class7>anger</class7>
        <class1d>negativeL1</class1d>
        <ending>aukce.</ending>
    </emotionEvalSentence>
    <emotionEvalSentence>
        <class7>anger</class7>
        <class1d>negativeL2</class1d>
        <ending>kdo?</ending>
    </emotionEvalSentence>
    <author>customer</author>
    <text xml:lang="cs">To snad není možný, tady nikdo nic nedělá,nic nefunguje.
    Tři týdny nejdou vystavovat aukce. Ty peníze o který přicházim mi jako
    zaplatí kdo?
    </text>
    <source>
        <url>http://im.aukro.cz/phorum/</url>
        <category>auction</category>
    </source>
  </trainUnit>
</emotionsTrainData>
```

Obr. 6.2: Příklad jednoho prvku trénovací množiny ve formátu XML.

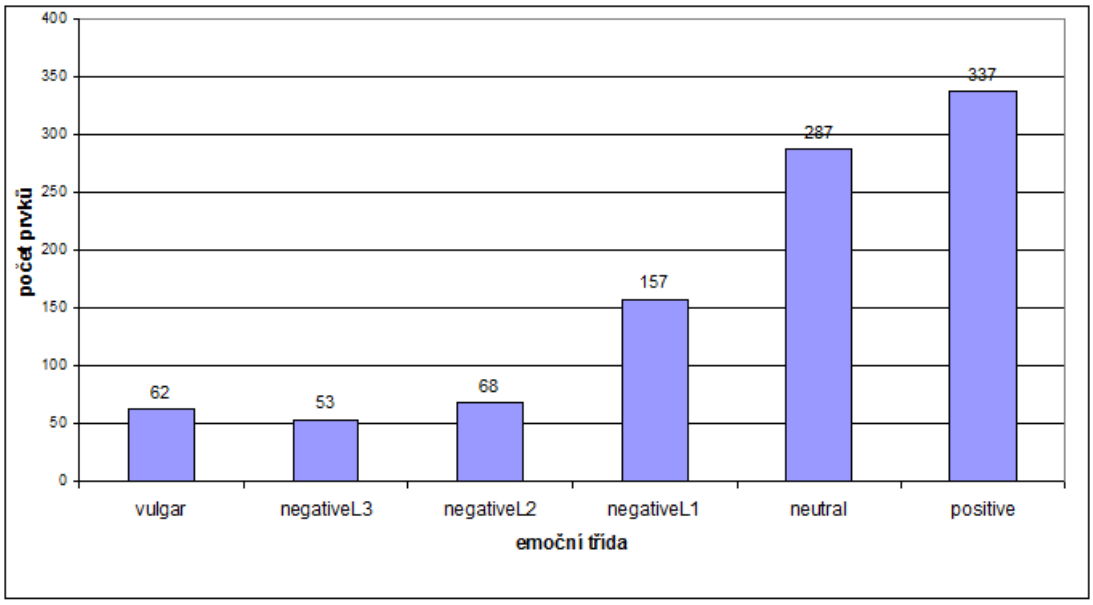
6.2.3 Tvorba nástroje pro hodnocení příspěvků

Vzhledem k relativní komplexnosti struktury prvků trénovací množiny, vznikla potřeba softwarového nástroje, který by umožňoval snadnější hodnocení prvků pro hodnotitele skrze GUI (grafické uživatelské rozhraní – Graphical User Interface) a také automaticky kontroloval formální správnost při hodnocení příspěvků. Dalším požadavkem bylo umožnit export ohodnocených příspěvků do XML souborů v požadované struktuře. Z těchto důvodů byl vytvořen program v programovacím jazyku Java s názvem *Annotation Editor* (viz. příloha B), který v sobě zahrnuje všechny

výše zmíněné vlastnosti.

6.2.4 Distribuce emočních tříd

Při volbě zdrojů prvků trénovací množiny byl kladen důraz na to, aby byl obsah příspěvků emočně bohatý a obsahoval hojný počet různých druhů písemných vyjádření emocí z různých oblastí lidské činnosti. Důležitým faktorem byla i různorodost stylu psaní (formálních i neformálních). Jako vhodné zdroje příspěvků byly zvoleny blogy, formální i neformální elektronické diskuze a dotazy na technické podpory. Přehled o rozložení prvků napříč emočními třídami je možné získat z obrázku 6.3.



Obr. 6.3: Zastoupení prvků emočních tříd v trénovací množině.

6.3 Automatická kontrola pravopisu a překlepů

Elektronická textová komunikace je charakteristická zvýšeným procentem chybovosti, překlepů, chybějících nebo přeházených písmen ve slovech, včetně slov vyskytujících se bez diakritiky. Takovéto tvary se mohou negativním způsobem podílet na výsledcích klasifikace a proto bylo snahou jejich počet zredukovat.

Pro implementaci byl využit nástroj, který je popsán v pramenu [5]. Vychází z knihovny *Jazzy - The Java Open Source Spell Checker* [21]. Tato knihovna umožňuje kontrolu pro několik světových jazyků v závislosti na dostupném slovníku.

Tyto slovníky mohou být volně uživateli upravovány a rozšiřovány o vlastní výrazy. V případě identifikace překlepu či pravopisné chyby je automaticky nabídnuta i správná varianta.

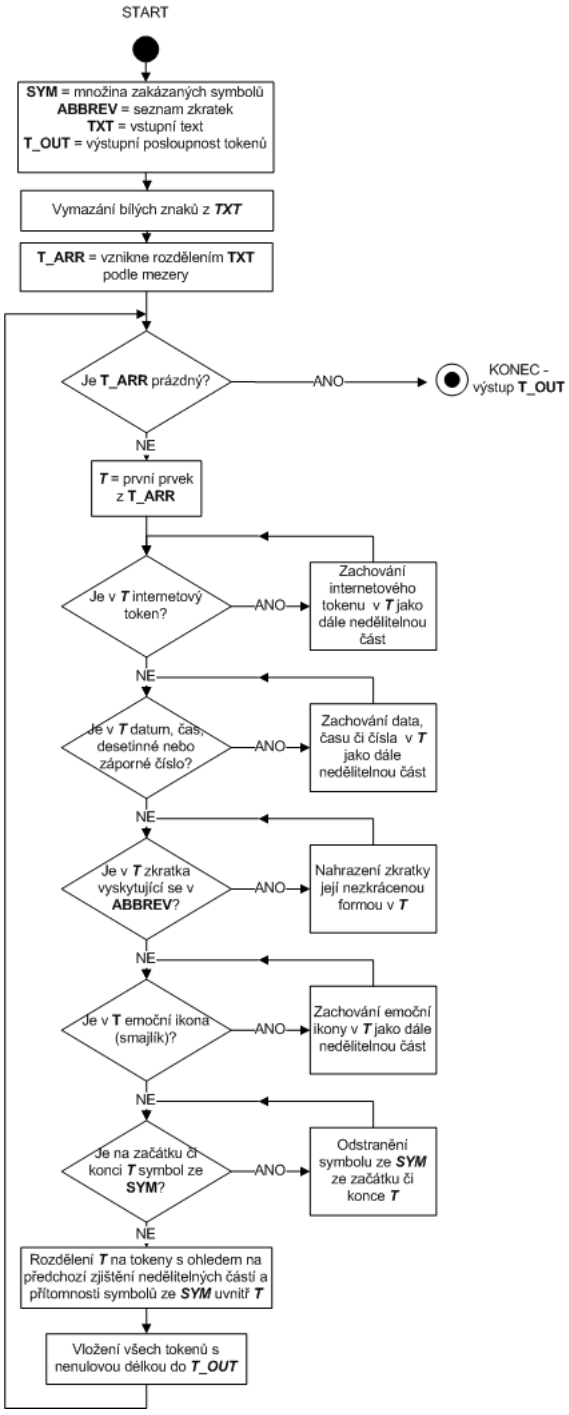
6.4 Implementace segmentace textu

Úkolem segmentace textu je rozčlenit věty na tokeny, které jsou chápány jako základní nositelé informace. Velmi často bývají tvořeny pouze slovy vzniklými rozdělením věty podle mezer a interpunkčních znamének. Nicméně, je nutné také zachovat posloupnosti znaků, které by odděleně ztrácely svůj význam. Z tohoto důvodu byly při implementaci použity seznamy nejčastěji používaných zkratek a emočních ikon. Dále bylo snahou zachovat datum a čas v ucelené podobě a tzv. Internetové tokeny, které zahrnují IP adresy, doménové adresy, adresy elektronické pošty a URL.

Obrázek 6.4 prezentuje vývojový diagram průběhu rozčlenění vstupního textu na tokeny. Nejdříve jsou ze vstupního textu ořezány bílé znaky, což jsou znaky odřádkování, odsazení či přebytečné mezery na začátku a konci řetězce. Poté je řetězec rozdělen dle mezer. Každý takovýto vzniklý podřetězec je kontrolován na přítomnost speciální posloupnosti symbolů, ať už pomocí slovníku (zkratky, emoční ikony) nebo na základě porovnávání se vzorem za pomocí regulárních výrazů (Internetové tokeny, datum, čas, desetinné nebo záporné číslo). V případě nalezení speciální posloupnosti, je tato zachována a označena za nedělitelný token v rámci dalšího zpracování. Podřetězec je dále kontrolován na přítomnost zakázaných symbolů (převážně interpunkční znaménka) na jeho začátku a konci. Pokud to předchozí zpracování umožňuje, jsou takovéto znaky odstraněny. Výstupní podoba tokenů vznikne z podřetězce rozdělením dle nalezených tokenů v předešlém zpracování, případně jeho dalším dělením na základě přítomnosti některého ze zakázaných symbolů uvnitř řetězce (například čárka mezi slovy bez mezery). Při implementaci byl kladen důraz i na zachování pořadí výstupní posloupnosti tokenů vzhledem k podobě vstupního textu.

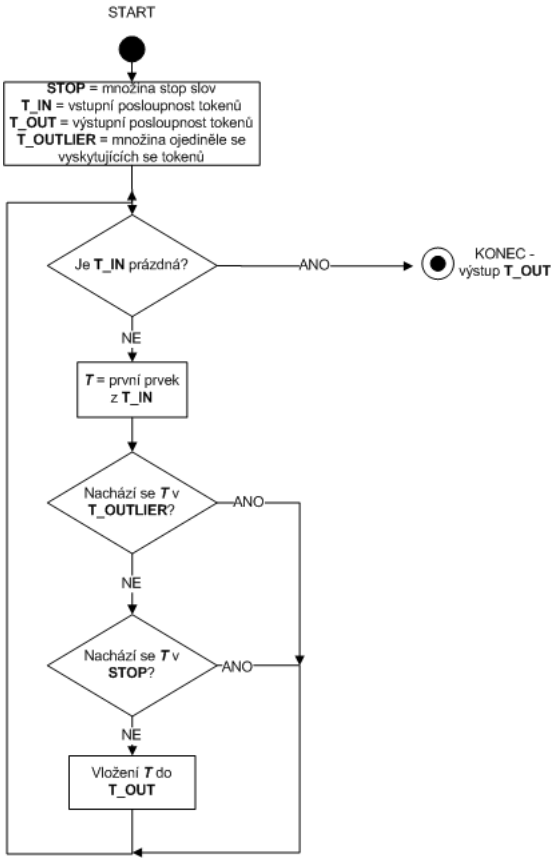
6.5 Implementace filtrace tokenů

Filtrace tokenů slouží k odstranění prvků, které nemají z hlediska dalšího zpracování sémantický význam nebo jsou statisticky nevýznamné pro další zpracování. Tento krok může podstatně zlepšit výsledky klasifikace, jelikož se jedná o jistou formu odstranění šumu. Průběh filtrace tokenů ilustruje vývojový diagram na obr. 6.5. Pro účel filtrace byla použita množina stop slov obsahující 1763 záznamů. Pro slova obsahující diakritiku byla uvedena i jejich varianta bez diakritiky, jelikož v rámci



Obr. 6.4: Vývojový diagram rozčlenění textu na tokeny.

elektronické komunikace se často tato varianta vyskytuje. Množina ojediněle vyskytujících se tokenů byla získána na základě statistické analýzy. Tato množina obsahuje tokeny jenž se nevyskytují více jak jednou napříč všemi prvky trénovací množiny. Tento postup si kladl za cíl eliminovat překlepy a další statisticky nevýznamné tvary.



Obr. 6.5: Vývojový diagram filtrace tokenů.

6.6 Lemmatizace

Lemmatizace spočívá v transformaci tokenů na jejich normalizovaný neboli slovníkový tvar (lemma). Tento krok předzpracování byl do systému zařazen z následujících důvodů:

- *Redukce počtu zpracovávaných tokenů* – Čeština je charakteristická velkým množstvím skloňovaných slovních tvarů (jedná se o tzv. flektivní jazyk). Z hlediska strojového zpracování textu je obměna tvarů slovní jednotky beze změny významu nevýhodná, jelikož jsou všechny tvary zpracovány samostatně bez přímé vazby. Nalezení normalizovaného tvaru umožňuje podstatnou redukci zpracovávaných tokenů. Tato operace významově sjednotí skloňované tvary a tím zjednoduší další fáze předzpracování společně s potencionálním zlepšením výsledků klasifikace.
- *Efektivita vyhledávání v lexikální databázi* – V lexikálních databázích (např. Český WordNet) jsou slova a slovní spojení převážně uvedena v normalizovaném (slovníkovém) tvaru. Tento tvar je nutné znát v případě, že chceme

využít sémantické vazby mezi slovy a slovními spojeními, které se v takovéto databázi vyskytují.

Pro implementaci lemmatizátoru byl využit nástroj, který je popsán v pramenu [34]. Vychází ze systému LemmaGen [22]. Tento systém využívá pro automatickou lemmatizaci technik strojového učení. Předpokládá přítomnost značného souboru slov v normalizovaném tvaru (trénovací množina). Určení lemma neznámého slova je realizováno zařazením do jedné ze tříd normalizovaných tvarů jenž byly odvozeny na základě trénovací množiny. Z toho vyplývá, že problém lemmatizace je transformován na problém klasifikace.

Pravidla pro lemmatizaci odvozená systémem LemmaGen, jsou uložena v čitelné podobě v textovém souboru, což umožnilo jejich začlenění do systému pro rozpoznávání emocí. Podle autorů v pramenu [22] je LemmaGen v současnosti jeden z nejúčinnějších veřejně dostupných lemmatizátorů použitelný napříč několika evropskými jazyky.

6.7 Využití lexikální databáze Český WordNet

Hlavním cílem transformace tokenů byla celková redukce počtu analyzovaných slov. Za tímto účelem byla využita lexikální databáze Český WordNet. Vývojový diagram na obr. 6.6 ilustruje posloupnost kroků použitých pro transformaci vstupního tokenu na výstupní.

Lexikální databáze byla použita zejména pro hledání hyperonymický vztahů, jenž přiřazují k jednotlivým synsetům synsety s obecnějším významem. Synsety jsou základní stavební jednotkou lexikální databáze. Jednotlivé prvky synsetu jsou nazývány literály (viz. kap. 4.5). Zároveň bylo využito znalosti klíčových slov příslušících k jednotlivým prvkům trénovací množiny při její tvorbě a také znalosti domény ze které prvek pochází. Nahrazení klíčových slov, jenž hrála podstatnou roli při určování výsledné emoce příspěvku, obecnějším významem bylo chápáno jako nežádoucí z pohledu klasifikace. Z toho důvodu byl vytvořen doménový slovník klíčových slov. Například klíčové slovo *Škoda* v doméně automobilů má s velkou pravděpodobností jiný význam, než stejné slovo v jiné doméně. V závislosti na doméně byly tedy určité tokeny zachovány v původní podobě.

V rámci dalšího zpracování jsou pro vstupní token vyhledávány všechny synsety v lexikální databázi, jejichž literály obsahují řetězec obsažený v tokenu. Pokud alespoň jeden synset existuje, je proveden výběr nejvhodnějšího literálu. Kritéria pro výběr byla: maximální podobnost mezi tokenem a literálem, společně s minimální hodnotou významu přiřazenou jednotlivým literálům (v lexikální databázi Český

WordNet má každý literál přiřazenou číselnou hodnotu významu: tzv. sense number).

V závislosti na slovním druhu je poté vyhledán hypernym, tj. synset s obecnějším významem. Pro přídavná jména, příslovce a slovesa je vyhledán kořenový hypernym stromu jenž je tvořen heperonymickými vazbami. U podstatných jmen není možné nahradit token kořenovým hypernymem, jelikož ten je pro všechna podstatná jména společný (*entita*). Tím by došlo k nežádoucímu přílišnému zobecnění. V rovnici

$$C_{\mathrm{NODE}} = \frac{H_{\mathrm{SYN}} - 3}{2} \tag{6.1}$$

představuje C_{NODE} počet uzlů o které je možné se posunout v rámci stromu tvořeného hyperonymickými a hyponomickými vazbami u podstatných jmen. H_{SYN} je aktuální hloubka uzlu (synsetu) v rámci stromu, tj. délka cesty od kořenového uzlu.

6.8 Tvorba vektorového modelu

Pro potřeby dalšího zpracování učícím se algoritmem podpůrných vektorů (SVM) bylo nutné změnit vyjádření textu pomocí posloupností tokenů do jiné podoby. Byla zvolena reprezentace textu pomocí vektorového modelu, tzn. vyjádření dokumentu (zde posloupnost tokenů) v n dimenzionálním prostoru R^n, kde n je počet tokenů v rámci celého korpusu (respektive jeho části určené pro trénovací množinu). Pro určení váhy jednotlivých složek vektoru byla nejdříve vypočítána četnost výskytu tokenu v rámci dokumentu podle vztahu [29]:

$$TF(t,d) = \begin{cases} 0, & t_f(t,d) = 0 \\ 1 + \log(1 + \log(t_f(t,d))), & jinak \end{cases} \tag{6.2}$$

kde $t_f(t,d)$ vyjadřuje počet výskytů tokenu t v dokumentu d.

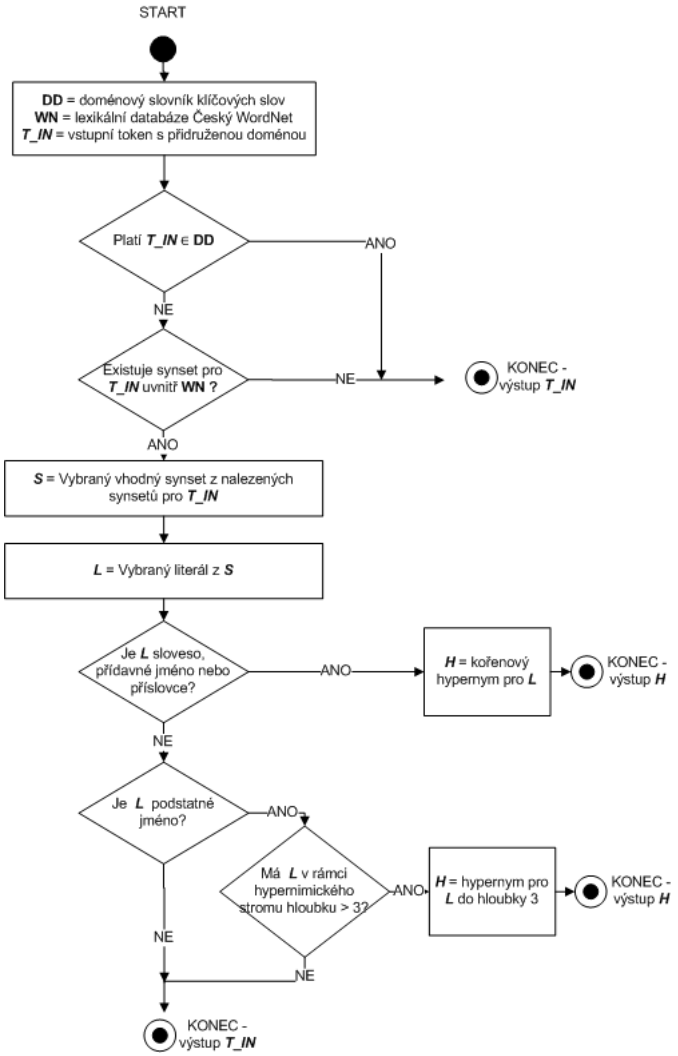
Dále podle následujícího vztahu byla vypočtena důležitost tokenu v rámci celého korpusu [29]:

$$IDF(t,d) = \log \frac{1+N}{d_f(t)} \tag{6.3}$$

kde N je počet všech dokumentů v korpusu a $d_f(t)$ vyjadřuje počet dokumentů obsahující token t.

Násobením 6.2 a 6.4 dostáváme vztah pro výpočet výsledné váhy tokenu v dokumentu (příslušné složky vektoru v modelu dokumentu) [29]:

$$TF\text{-}IDF(t,d) = TF(t,d) \cdot IDF(t,d) \tag{6.4}$$



Obr. 6.6: Vývojový diagram transformace tokenů s využitím lexikální databáze.

6.9 Selekce atributů

Selekce atributů má za úkol vybrat z vytvořeného vektorového modelu optimální množinu prvků (zde atributů), které budou mít vliv na tvorbu modelu klasifikátoru. I přes předchozí kroky v předzpracování (filtrace tokenů), obsahuje podle pramenu [13] vektorový model dokumentu stále mnoho irelevantních atributů, jejichž odstraněním snížíme výpočetní a paměťovou náročnost. Vhodnou volbou selekčních mechanizmů dokonce můžeme zlepšit výsledky klasifikace. Atribut je v této souvislosti chápán jako jeden token z emočního korpusu, který má pro jednotlivé dokumenty přiřazenou různou hodnotu na základě algoritmu TF-IDF.

Selekce atributů byla realizována ve dvou krocích, které budou následně popsány.

6.9.1 Korelační koeficient

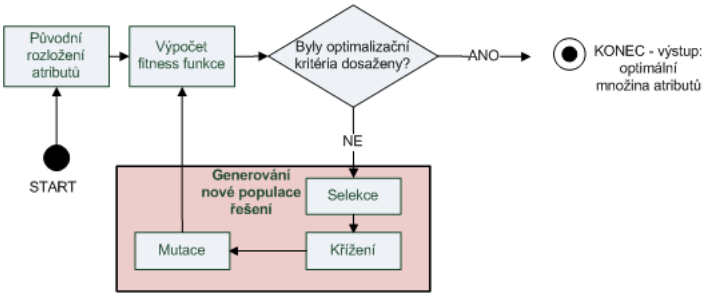
První krok selekce atributů je realizován pomocí výpočtu Pearsonova korelačního koeficientu. Pro atributy X a Y se jedná o podíl kovariance s násobkem směrodatných odchylek obou atributů podle vztahu [3], [36]:

$$r_{PCC}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{6.5}$$

k odstranění jednoho z dvojice atributů dojde v případě, že absolutní hodnota korelačního koeficientu je větší než *0,95*.

6.9.2 Evoluční optimalizace

Druhým krokem selekce atributů je použití genetických algoritmů [9], [18]. Tyto algoritmy bývají zejména využívány k řešení optimalizačních úloh v případě, že není možné explicitně stanovit postup řešení dané úlohy. Jsou založeny na přírodních zákonech evoluce a přirozeného výběru. Hledání optimálního řešení probíhá formou soutěže v rámci populace. Obrázek 6.7 ilustruje princip činnosti genetického algoritmu pro optimální výběr množiny atributů. Každý jedinec v dané populaci vyjadřuje jedno konkrétní řešení problému (množina vybraných atributů). Je vypočtena fitness funkce (zdatnost) každého jedince prostřednictvím aplikace klasifikátoru a 10-ti násobné křížové validace. Výstupem fitness funkce je tedy hodnota F-skóre (vyjádřeno vztahem 7.1) pro danou množinu atributů. V případě, že nějaký jedinec v populaci řešení dosáhl maximální možné hodnoty fitness funkce nebo se již jedná o hraniční (předem stanovenou) generaci řešení, algoritmus skončí. V případě, že tomu tak není, dojde ke generaci nové populace řešení prostřednictvím selekce a genetických operátorů křížení a mutace.



Obr. 6.7: Princip činnosti genetického algoritmu při výběru atributů.

Parametry genetického algoritmu, které byly použity při selekci atributů jsou uvedeny v tabulce 6.4.

6.10 Volba algoritmu pro klasifikaci

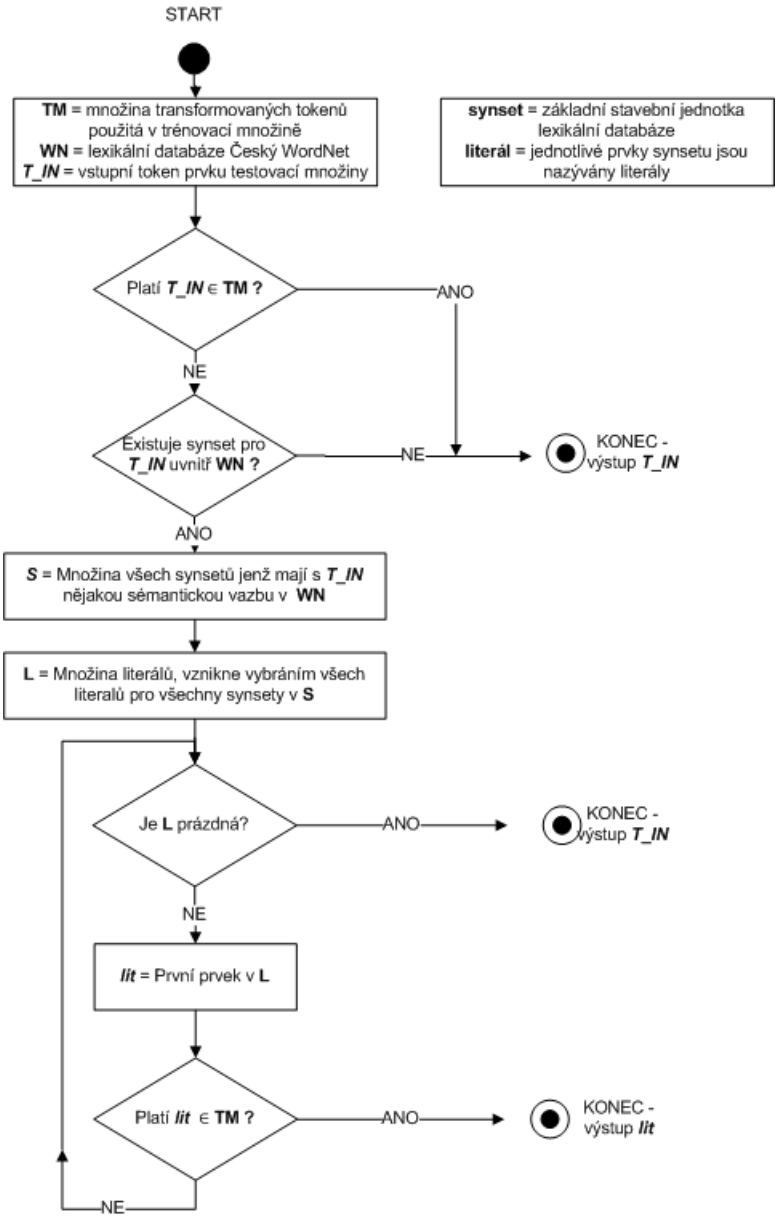
Algoritmus podpůrných vektorů (SVM) byl použit pro vytvoření modelu klasifikátoru jelikož se již mnohokrát ukázal jako vhodný způsob klasifikace textových dokumentů [13], [7]. Implementace SVM byla založena na knihovně LIBLINEAR, jenž byla prezentována v pramenu [12]. Její výhodou je optimalizace pro klasifikaci prvků s velkým množstvím atributů. V souladu s [20] bylo použito RBF (Radial Basis Function) jádro. Tento typ využívá nelineárního mapování vzorků do prostoru s vyšší dimenzí. Na rozdíl tedy od lineárního jádra umožňuje řešit i případy kdy je vztah mezi třídami pro klasifikaci a vzorky nelineární. Hledání optimálních parametrů C a ϵ potřebných pro vytvoření modelu klasifikátoru je uveden v části 7.2.

6.11 Tvorba vektorového modelu pro prvky testovací množiny

Prvky testovací množiny prochází stejnými kroky předzpracování textu jako tomu bylo u prvků trénovací množiny. Problém nastává při tvorbě vektorového modelu dokumentu dle TF-IDF (Term Frequency-Inverse Document Frequency). Tento model pro určení vah bere v úvahu četnost výskytu daného tokenu jednak v rámci dokumentu, ale také v rámci celého souboru dokumentů (v tomto případě testovací množině). V případě, že by byl tímto způsoben pro testovací množinu vytvořen nový vektorový model, obsažené tokeny a k nim odpovídající váhy by byly zcela odlišné od vektorového modelu, který vznikl po trénovací fázi. Takovýto stav je nežádoucí, jelikož model klasifikátoru je vytvořen a nastaven na základě prvků trénovací množiny a tudíž by byla přesnost klasifikace podstatně snížena.

Snahou bylo tedy mapovat tokeny prvků testovací množiny na prvky trénovací množiny a k nim odpovídající váhy. Vzhledem k bohatosti jazyka není možné zaručit, že daný token z prvku v testovací množině je také obsažen v některém z prvků testovací množiny. Z tohoto důvodu byla pro tento účel využita lexikální databáze Český WordNet, která obsahuje sémantické vazby mezi slovy a slovními spojeními. Díky tomu je umožněno provést odpovídající mapování i v případě, že daný token není explicitně obsažen v testovací množině. Vývojový diagram na obr. 6.8 ilustruje

posloupnost kroků, které byly v rámci procesu mapování tokenu z testovací do trénovací množiny použity.



Obr. 6.8: Vývojový diagram mapování tokenů testovací množiny na tokeny trénovací množiny.

Tab. 6.3: Popis XML elementů formátu trénovací množiny

Název elementu	**Povinné**	**Nepovinné**	**Popis**
emotionsTrainData	trainUnit	–	Kořenový element, může obsahovat libovolný počet elementů *trainUnit*.
trainUnit	emotionEvaluation, autor, text	source	Reprezentuje prvek z trénovací množiny.
emotionEvalParagraph	class7, class1d	term	Reprezentuje ohodnocení přidruženého textu jedním hodnotitelem.
class7	–	–	Musí obsahovat jednu emoční třídu ze skupiny tříd *1DET*.
class1d	–	–	Musí obsahovat jednu emoční třídu ze skupiny tříd *2DET*.
term	–	–	Musí obsahovat klíčové slovo nebo více slov (fráze), které hrálo podstatnou roli při výběru dané emoční třídy.
emotionEvalSentence	class7, class1d, ending	–	Reprezentuje ohodnocení jedné věty jedním hodnotitelem.
ending	–	–	Musí obsahovat řetězec, který jednoznačně identifikuje ukončení věty v rámci celého přidruženého textu.
author	–	–	Možné hodnoty: *customer*, *business*, *forum*. Blíže specifikuje původ příspěvku.
text	–	–	Příspěvek (v původní podobě, tzn. bez opravy pravopisu a překlepů).
source	url, category	–	Element reprezentující zdroj příspěvku.
url	–	–	URL (jednotný lokátor zdrojů – Uniform Resource Locator).
category	–	–	Kategorie (doména) příspěvku. Je možné definovat v podstatě libovolné, ale musí být později dodržovány napříč prvky.

Tab. 6.4: Parametry genetického algoritmu

Parametr	**Nastavení**
pevně stanovený počet atributů pro jedno řešení	NE
minimální počet atributů pro jedno řešení	1
velikost populace	30
maximální počet generací řešení	30
selekční mechanizmus	turnajová selekce
část populace podílející se v turnaji	0,25
zachování nejlepšího jedince v populaci	NE
typ křížení	uniformní
pravděpodobnost mutace	$(počet\_atributů)^{-1}$
pravděpodobnost křížení	0,5

7 VÝSLEDKY

7.1 Metodika hodnocení úspěšnosti klasifikace

Pro vyhodnocení úspěšnosti klasifikace byly použity následující metriky a principy.

7.1.1 Přesnost (Precision)

Přesnost při klasifikaci je poměr počtu správně zařazených dokumentů do dané kategorie ku počtu všech zařazených dokumentů do dané kategorie [14]:

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \qquad [-] \tag{7.1}$$

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

7.1.2 Výtěžnost (Recall)

Výtěžnost při klasifikaci je poměr počtu správně zařazených dokumentů do dané kategorie ku počtu všech relevantních dokumentů v rámci dané kategorie [14]:

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \qquad [-] \tag{7.2}$$

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

7.1.3 F-skóre (F-score)

Kompromisní hodnotou mezi přesností a výtěžností je F-skóre vyjádřené vztahem [14]:

$$F_{\mathrm{score}} = 2 \cdot \frac{P \cdot R}{P + R} \qquad [-] \tag{7.3}$$

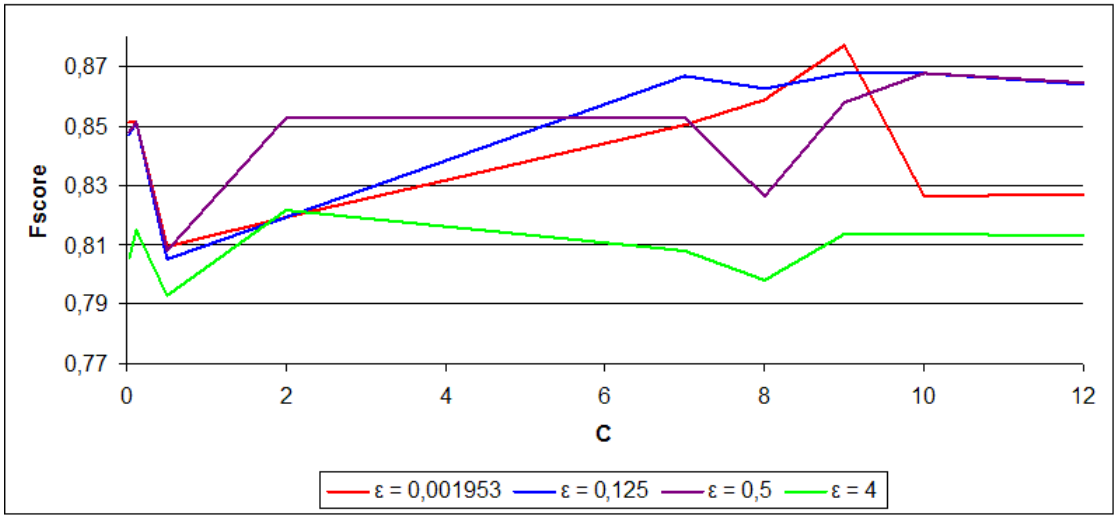
kde P vyjadřuje přesnost (vztah 7.1) a R výtěžnost (vztah 7.2).

7.1.4 10-ti násobná křížová validace

Jako vhodný způsob validace byla zvolena 10-ti násobná křížová validace, kdy v deseti krocích byla vždy použita jedna desetina prvků z trénovací množiny pro účely validace. Zbytek byl použit jako trénovací množina. Výsledky z jednotlivých kol (přesnost 7.1 a výtěžnost 7.2) byly pro určení výsledné hodnoty zprůměrovány.

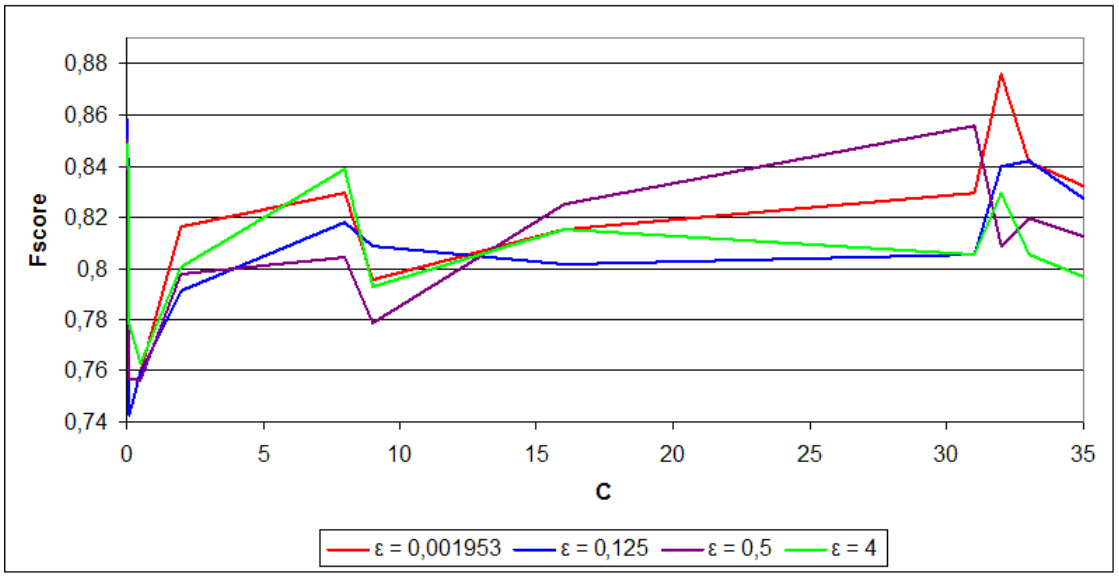
7.2 Hledání optimálních parametrů pro SVM

Implementace SVM s RBF jádrem na základě knihovny LIBLINEAR [12] má dva parametry: C a ϵ. Tyto parametry ovlivňují tvorbu modelu klasifikátoru a není možné určit předem jaké hodnoty parametrů budou ideální pro danou trénovací množinu a řešený problém. Z tohoto důvodu musely být tyto parametry stanoveny experimentálně. Optimální parametry byly hledány pro tři modely klasifikátorů textových dokumentů: detekce negativních + vulgárních emocí (sdružuje třídy *vulgar* a *negativeL1-negativeL3* ze skupiny *1DET*), detekce pozitivních emocí (odpovídá třídě *positive* ve skupině *1DET*) a detekce neutrálních emocí (odpovídá třídě *neutral* ve skupině *1DET*). Jako vhodný nástroj k validaci jednotlivých nastavení byla zvolena 10-ti násobná křížová validace, jejímž výstupem byla hodnota F-skóre (viz. vztah 7.3). Volba parametru ϵ i rozsahu hodnot C proběhla v souladu s pramenem [20]. Závislosti F-skóre na hodnotě C a parametru ϵ jsou uvedeny v grafech: 7.1 pro model klasifikátoru pro detekci negativních emocí + vulgárních příspěvků, 7.2 pro model klasifikátoru pro detekci pozitivních emocí a 7.3 pro model klasifikátoru pro detekci neutrálních emocí. V případě neutrálních emocí (graf 7.3), bylo nejvyššího F-skóre dosaženo při nízkých hodnotách C. Kvůli lepší rozlišitelnosti je proto graf zobrazen pouze v rozsahu 0-1,2 na ose x.

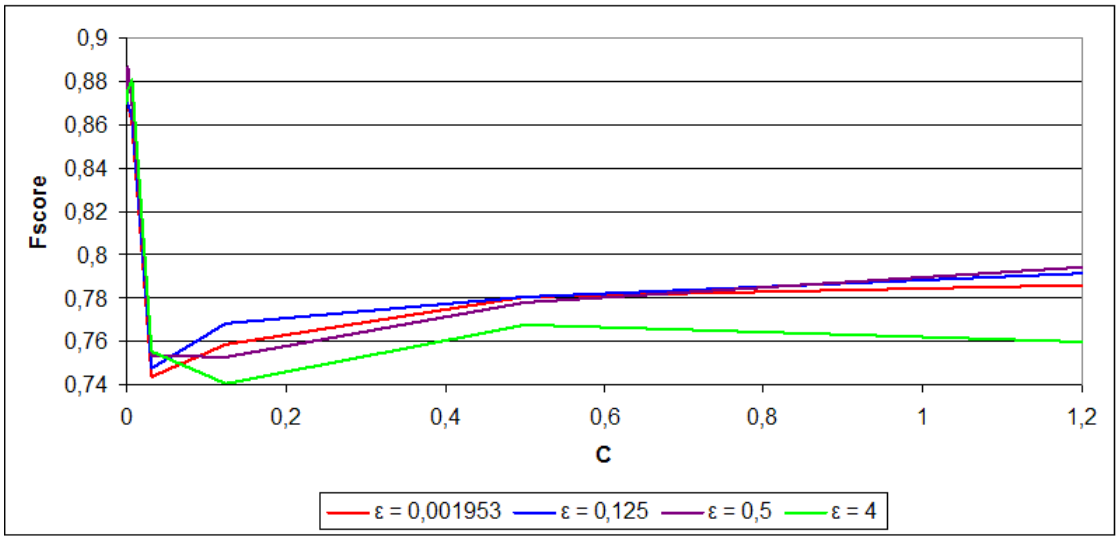


Obr. 7.1: Závislost F-skóre na hodnotě C a parametru ϵ při detekci negativních emocí a vulgárních příspěvků.

Optimální nalezené parametry pro jednotlivé typy modelu klasifikátoru jsou uvedeny v tabulce 7.1.



Obr. 7.2: Závislost F-skóre na hodnotě C a parametru ϵ při detekci pozitivních emocí.



Obr. 7.3: Závislost F-skóre na hodnotě C a parametru ϵ při detekci neutrálních emocí.

7.3 Úspěšnost klasifikace

Vytvořený systém klasifikoval textové dokumenty do tří emočních tříd: negativní + vulgární (sdružuje třídy *vulgar* a *negativeL1-negativeL3* ze skupiny *1DET*), positivní (odpovídá třídě *positive* ve skupině *1DET*) a neutrální (odpovídá třídě *neutral*

Tab. 7.1: Optimální parametry pro tvorbu modelu klasifikátoru

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

ve skupině *1DET*). Klasifikace probíhala nezávisle, tzn. že vytvořený model klasifikátoru s ideálními parametry (viz. část 7.2) byl použit na celou trénovací množinu pro každou z tříd zvlášť. Byly vytvořeny různé scénáře klasifikace, které jsou uvedené v tabulce 7.2. Tyto scénáře modifikují systém tak, že z něj odebírají vždy nějakou část. Záměrem bylo porovnání vytvořeného systému s různými modifikacemi, tak aby bylo patrné, jestli případná chyba v návrhu negativně neovlivňuje výsledky klasifikace.

Tab. 7.2: Definice scénářů pro klasifikaci

Název scénáře	**Popis**
kompletní	Představuje použití kompletního vytvořeného systému se všemi jeho částmi.
bez selekce atributů	Představuje použití systému bez všech kroků v rámci selekce atributů (evoluční optimalizace a selekce na základě korelačního koeficientu).
bez WordNet	Představuje použití systému bez lexikální databáze Český WordNet pro hledání hyperonimických vztahů a redukci dimenze slov.
bez lemmatizace	Představuje použití systému bez lemmatizátoru.
bez filtrace tokenů	Představuje použití systému bez filtrace tokenů na základě seznamu stop slova a statistické analýzy.
bez aut. opravy	Představuje použití systému bez automatické opravy překlepů, pravopisu a slov vyskytujících se bez diakritiky.

Výsledky klasifikace do jednotlivých emočních tříd v závislosti na použitém scénáři zobrazuje tabulka 7.3, kde jsou uvedeny hodnoty přesnosti P, výtěžnosti R a F-skóre F_{score}. Graf 7.4 ilustruje hodnoty F-skóre uvedené v tabulce 7.3.

Výsledky ukazují, že vytvořený systém (scénář *kompletní*) dosahuje nejvyšších hodnot F-skóre napříč všemi třídami. To poukazuje na fakt, že zvolené kroky v před-

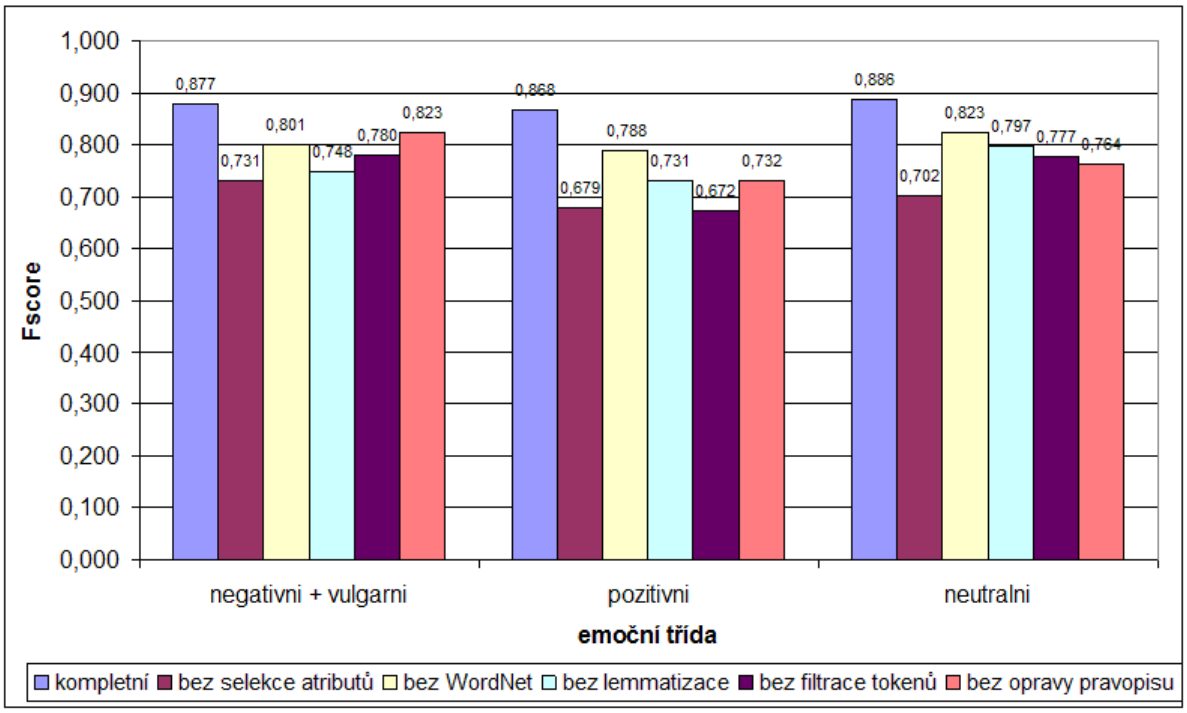
Tab. 7.3: Výsledky klasifikace v závislosti na použitém scénáři a emoční třídě

Scénář	**negat.+vulg.**			**pozitivní**			**neutrální**		
	P	R	F_{score}	P	R	F_{score}	P	R	F_{score}
kompletní	0,893	0,862	**0,877**	0,838	0,838	**0,868**	0,886	0,886	**0,886**
bez selekce a.	0,717	0,745	0,731	0,690	0,668	0,679	0,691	0,713	0,702
bez WordNet	0,745	0,865	0,801	0,757	0,821	0,787	0,805	0,842	0,823
bez lemma.	0,691	0,814	0,748	0,668	0,807	0,731	0,756	0,842	0,797
bez filtrace	0,780	0,780	0,780	0,683	0,662	0,672	0,795	0,760	0,777
bez opravy	0,867	0,782	0,823	0,677	0,797	0,732	0,7719	0,757	0,764

zpracování textu a při selekci atributů negativně neovlivnili dosažené výsledky. Lze vysledovat, že selekce atributů byl postup, jenž měl velký vliv na přesnost klasifikace. Tento fakt se projevil u všech emočních tříd, kdy úspěšnost klasifikace výrazně poklesla při odstranění selekce atributů (scénář *bez selekce atributů*). Zajímavým ukazatelem je také úspěšnost nalezených slov v lexikální databázi Český WordNet (viz. tabulka 7.4), jenž má pozitivní vliv na další zpracování. Zde je patrný velký vliv zařazení lemmatizátoru do systému, který převede skloňované tvary slov do normalizovaného (slovníkového) tvaru. Lemmatizátor zvyšuje procentuální úspěšnost nalezení slov v databázi z 27 % na 64 %. Důvody pro neúspěšné vyhledávání slov v databázi jsou následující: slova nepřevedená do normalizovaného tvaru (chyby lemmatizátoru), slova bez diakritiky a s překlepy (chyby automatické opravy), nepřítomnost určitých slov v databázi (nedostatky lexikální databáze).

Tab. 7.4: Úspěšnost vyhledávání v lexikální databázi v závislosti na použitém scénáři

Scénář	**Úspěšnost nalezených slov** [%]
kompletní	64
bez lemmatizace	27
bez filtrace tokenů	41
bez aut. opravy	58



Obr. 7.4: Závislost F-skóre na použité metodě klasifikace.

7.4 Příčiny pro chybnou klasifikaci

Byla provedena analýza výstupů klasifikace a na jejím základě byly zjištěny následující možné příčiny nepřesností při klasifikaci.

První soubor příkladů je uveden v tabulce 7.5. Jedná se o ironické texty, které jsou charakteristické tím, že vyjadřují něco jiného než je v nich ve skutečnosti napsáno. Člověk dokáže na základě přirozené inteligence, případně i díky kontextu pochopit jejich skutečný význam. Nicméně z hlediska strojového zpracování je pochopení skutečného významu obtížným úkolem.

Tab. 7.5: Chybná klasifikace - ironie

Číslo	**Věta**
1	To hodnotí situaci ten pravej.
2	To snad není ani pravda, zcela neuvěřitelný, neskutečný!
3	Jak je krásně u koryta, že.

Elektronická textová komunikace je charakteristická zvýšeným procentem slov vyskytujících se bez diakritiky (případně s překlepy). Toto použitý systém automa-

tické opravy nedokázal vždy dostatečně eliminovat. Tabulka 7.6 ilustruje původní podobu vět a stav po automatizované opravě. Je patrné, že ne vždy došlo k opravě slov bez diakritiky. Takováto slova nebylo možné dále zpracovávat pomocí lexikální databáze a tím docházelo k negativnímu ovlivnění výsledků klasifikace. Navíc občas docházelo i ke změně významu slova a tím i celé věty (viz. příklad 3 v tabulce).

Tab. 7.6: Chybná klasifikace - chyby automatické opravy

Označení	**Věta**
původní 1	Auta pozvolna vytesnuji chodce s tim jak kazda socka (viz ten imbecil nize) si nejakou tu ohyzdnou smrdutou plechovku (na lyzink) koupi
po opravě 1	auta pozvolna vytesanější chodce s tm jak kazaa socka viz ten imbecil knize si nijakou tu ohyzdnou smrdutou plechovku na lyzin koupi
původní 2	a pak cela stastna ze se na ten pekac konecne zmohla machruje a jezdi co nejvice aby vsichni videli ze i on na to ma
po opravě 2	a pak cela sytostní ze se na ten pekař kankán zmohla machruje a jezd co nejvíce aby všichni viděli ze i on na to ma
původní 3	To hodnotí situaci ten pravej.
po opravě 3	to hodnotí situaci ten plavej

Dalším charakteristickým znakem elektronické textové komunikace je, že lidé často nevystupují pod svými skutečnými jmény, ale pod různými přezdívkami (především u diskuzí, sociálních sítích apod.). V případě, že přezdívka je slovníkovým výrazem a navíc nese sama o sobě určitý emoční náboj, může dojít při jejím použití v textu chybně ke změně výsledná emoce věty (viz. příklad 1 v tabulce 7.7). Toto nebyl častý jev, nicméně se vyskytoval. Dalším příčinou byla nevhodná filtrace tokenů na základě množiny stop slov (viz. příklad 2 a 3 v tabulce 7.7), kdy při odstranění stop slov docházelo v některých případech ke změně významu (a tím pádem i výsledné emoce). Také chyby lemmatizátoru byly příčinou nesprávné klasifikace. Toto je ilustrováno na příkladu 4 v tabulce 7.7, kde kromě chybného nahrazení slova slovníkovým tvarem došlo také k nežádoucí modifikaci klíčového slova. Nepřesnosti klasifikace byly také způsobené chybami při hledání hyperonimických vztahů v lexikální databázi Český WordNet, protože některé záznamy v databázi mají přiřazeno několik významů. V případě, že byl pro další zpracování zvolen špatný význam, došlo k volbě nesprávné větve hyperonymického stromu a tím pádem i k chybnému nahrazení slova obecnějším významem (viz. příklad 5 v tabulce 7.7). Dalším fakto-

rem může být nekonzistentní hodnocení prvků trénovací množiny, jelikož ta vznikala průběžně přibližně osm měsíců. Různé výrazy mohou vyvolávat rozdílné emoce v závislosti na čase. To platí zejména u názvu států, institucí, firem, výrobků apod.

Tab. 7.7: Chybná klasifikace - ostatní

Označení	**Věta**
1	Lotře já jsem to říkala.
2.1 původní	Není snad nic co by se mi u vás nelíbilo.
2.2 po filtraci	není snad nelíbilo
3.1 původní	je to kariérní úředník bez páteře
3.2 po filtraci	kariérní úředník páteře
4.1 původní	ale většinou jsou to šmejdi kteří se nikde jinde neuživí
4.2 po filtraci	většinou jsou šmejdi nikde jinde neuživí
4.3 po lemma.	většina jsa šmejít nikde jinde neuživý
5.1 původní	milí rybářští přátelé velmi rád všem oznamuji že naše diskuse nese ovoce
5.2 po filtraci	milí rybářští přátelé velmi rád oznamuji diskuse nese ovoce
5.3 po lemma.	milý rybařit přítel velmi rád oznamovat diskuse nést ovoce
5.4 po WordNet	milý vzít kamarád velmi rád oznamovat akt hnout prostředek reprodukce

8 MOŽNOSTI ROZŠÍŘENÍ SYSTÉMU

Jednotlivé části vytvořeného systému mohou být dále rozšířeny. Zde jsou uvedeny některé návrhy, které mohou potencionálně zvýšit úspěšnost klasifikace:

- *Hodnocení trénovací množiny více hodnotiteli* – Jelikož jsou prvky trénovací množiny hodnoceny lidmi podle jejich subjektivního dojmu, můžeme předpokládat, že napříč prvky trénovací množiny nalezneme jistou nekonzistentnost v hodnocení. Při hodnocení trénovací množiny např. deseti hodnotiteli a následném průměrování jejich hodnocení, by mohlo být dosaženo kvalitnější (objektivnější) trénovací množiny, která je základem u metod založených na strojovém učení.
- *Automatická aktualizace trénovací množiny a slovníků klíčových slov* – Existuje i časová platnost pro hodnocené prvky trénovací množiny, případně klíčová slova a slovní spojení, která jsou mapována na danou emoci. Jinými slovy, různé výrazy mohou vyvolávat rozdílné emoce v závislosti na čase. To platí zejména u názvu států, institucí, firem, výrobků apod. Pro získání časové nezávislosti by tedy systém měl obsahovat i část, která by dokázala odpovídající zdroje aktualizovat a to nejlépe automatizovaným způsobem s využitím stávajícího sytému pro klasifikaci textových dokumentů.
- *Optimalizace automatické opravy překlepů a slov bez diakritiky* – Ačkoliv použitá knihovna pro opravu překlepů, pravopisných chyba a slov bez diakritiky (viz. pramen [21]) umožňuje detekovat chybná slova, ne vždy je vybrána optimální varianta opravy z nabízených možných. To je dáno zejména faktem, že knihovna je primárně určená pro grafické uživatelské rozhraní, kdy uživatel případné chyby dokáže sám korigovat. Jeden z možných způsobů optimalizace je vytvoření databáze dvojic či trojic slov, které se v textu běžně společně vyskytují. Na základě takto vzniklých databází by posléze bylo možné vybírat nejpravděpodobnější variantu k opravě z nabízených možných.

9 POROVNÁNÍ SYSTÉMU S JINÝMI PRACEMI

Pro porovnání navrženého a vytvořeného systému se systémy podobného druhu bylo použito několik prací, které se zabývají problematikou rozpoznávaní emocí z textu.

Práce v rámci SemEval-2007 [33] zkoumala detekci emocí z anglicky psaných textů, metoda vycházela ze strojového učení a nejlepších výsledků bylo dosaženo při použití algoritmu SVM. Navíc pro získání sémantických vztahů mezi jednotlivými slovy bylo využito WordNet-Affect, což je rozšíření anglické lexikální databáze WordNet. Ačkoliv byl návrh systému velmi podobný systému navrženém v rámci této diplomové práce, systém pro anglický jazyk dosahoval přibližně o 10-15 % nižší přesnosti při klasifikaci textových dokumentů. Jedním z důvodů může být fakt, že ve výše zmíněné práci bylo provedeno jemnější dělení emočních tříd.

Další systém podobného typu byl navržen pro čínštinu (viz. pramen [35]), jenž mimo jiné využívá také algoritmus SVM. Rozdělení emočních tříd bylo velmi podobné. Úspěšnost detekce negativních a pozitivních emocí byla přibližně stejná jako úspěšnost prezentovaná v této práci, pouze v detekci neutrálních emocí předčil systém určený pro čínský jazyk zde prezentovaný systém přibližně o 4 %.

Jiný způsob detekce emocí v čínsky psaných textech je možné nalézt v pramenu [38]. Nejdříve byla využita lexikální databáze pro redukci počtu vyskytujících se slov a reprezentaci textu pomocí sémantických značek a atributů. Poté byly vyhledávány podobnosti mezi vhodně transformovanou větou určenou ke klasifikaci a odpovídajícím vzorem jenž byl namapován na danou emoci. Při stejném rozdělení emočních tříd v článku prezentovaný systém vykazoval o 15-20 % nižší hodnoty F-skóre.

Další způsob detekce emocí v textu pro čínský jazyk je uveden v pramenu [39]. Tento systém je naopak postaven převážně na metodě strojového učení a porovnává různé algoritmy pro klasifikaci, včetně SVM. Systém vytvořený v rámci této diplomové práce dosahuje přibližně o 30-40 % lepších výsledků při klasifikaci.

Zde uvedené srovnání bere v úvahu pouze výsledky úspěšnosti klasifikace, případně základní přístup k návrhu a implementaci systémů. Nezahrnuje odlišnosti lingvistické struktury jazyků a tudíž se jedná pouze o zjednodušené srovnání. Z výše uvedeného je patrné, že navržený a vytvořený systém předčí z hlediska úspěšnosti klasifikace systémy, jenž pro určování emocí využívají pouze metod strojového učení. Zároveň dosahuje podobných nebo lepších výsledků při klasifikaci textových dokumentů jako systémy, které pro určování emocí využívají kromě strojového učení také vzájemných sémantických vztahů mezi slovy (lexikální databáze).

10 ZÁVĚR

V rámci této práce byl proveden průzkum současných přístupů k rozpoznávání emocí v textu společně s teoretickým rozborem základních technik používaných při dolování znalostí z textu.

Hlavním přínosem této práce je návrh a implementace systému pro rozpoznávání emocí v česky psaných textech, včetně stanovení optimálních parametrů pro jednotlivé modely klasifikátoru, zhodnocení jejich úspěšnosti a porovnání se systémy, které byly vytvořeny pro jiné světové jazyky. Dalším přínosem je prezentace výsledků klasifikace textových dokumentů do předem definovaných emočních tříd a zhodnocení několika přístupů ke klasifikaci s různými modifikacemi navrženého systému. Mezi další přínosy patří návrh formátu a vytvoření početné trénovací množiny skládající se z reálných příspěvků a jejich manuální ohodnocení. Jako vedlejší produkt vznikl softwarový nástroj, který umožňuje hodnocení prvků trénovací množiny skrze grafické uživatelské rozhraní a umožňuje následný export do XML souborů v požadované podobě.

Dosažené hodnoty F-skóre, jenž je jedním z měřítek úspěšnosti klasifikace do předem definovaných emočních tříd, byly následující: 0,877 pro třídu negativní+vulgární, 0,868 pro třídu positivní a 0,886 pro třídu neutrální. Bylo vytvořeno několik scénářů klasifikace: kompletní, bez selekce atributů, bez WordNet, bez lemmatizace, bez filtrace tokenů a bez automatické opravy pravopisu. Vyjma scénáře kompletní, každý odstraňoval z navrženého systému určitou část pro zjištění, jestli negativně neovlivňuje výsledky klasifikace. Bylo ověřeno, že scénář kompletní dosahoval nejlepších výsledků napříč všemi emočními třídami.

Bylo provedeno srovnání navrženého a vytvořeného systému se systémy určenými pro jiné světové jazyky. Z nich vyplývá, že systém předčí z hlediska úspěšnosti klasifikace systémy, jenž pro určování emocí využívají pouze metod strojového učení. Zároveň dosahuje podobných nebo lepších výsledků při klasifikaci textových dokumentů jako systémy, které pro určování emocí využívají kromě strojového učení také vzájemných sémantických vztahů mezi slovy. V této práci prezentovaný systém dosahuje při hrubším rozdělení emočních tříd, přibližně o 10-15 % vyšší přesnosti při klasifikaci textových dokumentů, nežli systém prezentovaný v rámci SemEval-2007 [33].

Byla provedena analýza výstupů klasifikace a na jejím základě byly zjištěny možné příčiny nepřesností při klasifikaci (viz. kap. 7.4). Negativní vliv má velké procento slov vyskytujících se bez diakritiky (což je charakteristické pro elektronickou komunikaci). Toto použitý systém automatické opravy nedokázal vždy dostatečně eliminovat. Takováto slova nebylo možné dále zpracovávat pomocí lexikální databáze. Dalším faktorem je přítomnost ironických textů. Ty člověk dokáže z kontextu pochopit, nicméně u strojového zpracování je obtížné rozlišit, že text vyjadřuje něco

jiného než je v něm ve skutečnosti napsáno. Mezi další příčiny například patří: chyby lemmatizátoru, nevhodná filtrace stop slov a chyby při hledání hyperonymických vztahů s pomocí lexikální databáze.

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, 1, [cit. 2010-11-20]. Dostupný z WWW: <http://aclweb.org/anthology/I/I08/I08-1041.pdf>.

[2] ATASSI, H.; SMÉKAL, Z. Real-Time Model for Automatic Vocal Emotion Recognition. *In Proceedings of the 31st International Conference on Telecommunications and Signal Processing*, September 3-4, 2008, Parádfürdö, Hungary, pp.21-25. ISBN 978-963-06-5487-6

[3] BIESIADA, J.; DUCH, W. Feature Selection for High-Dimensional Data : A Pearson Redundancy Based Filter . *Computer Recognition Systems 2* [online]. 2007, Volume 45, [cit. 2010-11-17]. Dostupný z WWW: <http://www.springerlink.com/content/38v2855878037k1q/>.

[4] BRACEWELL , D.B. Semi-automatic creation of an emotion dictionary using WordNet and its evaluation. *IEEE Conference on Cybernetics and Intelligent Systems* [online]. 2008, 1, [cit. 2010-10-30]. s. 1385 - 1389 . Dostupný z WWW: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4670735>. eclanek doi:10.1109/ICCIS.2008.4670735.

[5] BUREŠ, S. Kontrola pravopisu v českých textech. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 40 s. Vedoucí diplomové práce Ing. Lucie Fojtová.

[6] BURGES Christopher J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121-167,1998.

[7] BURGET, R.; SMÉKAL, Z.; KARÁSEK, J. Classification and Detection of Emotions in Czech News Headlines. *The 33rd International Conference on Telecommunication and Signal Processing, TSP 2010.* 2010, s. 1-5.

[8] CHANGQIN, Q.; REN, F. Recognizing sentence emotions based on polynomial kernel method using Ren-CECps. *International Conference on Natural Language Processing and Knowledge Engineering* [online]. 2009, 1, pp. 1 - 7. Dostupný z WWW: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5313834>.

[9] DAVIS L. *Handbook of Genetic Algorithms.* Thomson Publishing Group / Van Nostrand Reinhold Company, New York, USA, January 1991. ISBN: 0-4420-0173-8, 978-0-44200-173-5.

[10] DEVILLERS L., VASILESCU I., LAMEL L. Annotation and detection of emotion in a task-oriented human-human dialog corpus. *Proc. ISLE Workshop on Dialogue Tagging for Multi-Modal Human-Computer Interaction*, [online]. 2002, [cit. 2010-10-30]. Dostupný z WWW: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.5524&rep=rep1&type=pdf>.

[11] DOLAMIC, L.; SAVOY, J. Indexing and stemming approaches for the Czech language. *Information Processing and Management*. 2009, 45, s. 714–720.

[12] FAN, Rong-En, et al. LIBLINEAR : A Library for Large Linear Classification. *Journal of Machine Learning Research*. 2008, 9, s. 1871-1874.

[13] 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.

[14] HAN, J.; KAMBER, M. *Data Mining : Concepts and Techniques*. San Francisco : Morgan Kaufmann, 2006. 770 s. ISBN 978-1-55860-901-3.

[15] HOFMANN, T.; PUZICHA, J. Statistical Models for Co-occurrence Data. *Massachusetts Institute of Technology* [online]. 1998 , AIM-1625, CBCL-159, [cit. 2010-10-29]. Dostupný z WWW: <http://hdl.handle.net/1721.1/7253>.

[16] HORÁK, A.; PALA K.; RAMBOUSEK A.; RYCHLÝ P. New clients for dictionary writing on the DEB platform. *In DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems*. 2006. vyd. Torino, Italy : Lexical Computing Ltd., U.K., 2006. od s. 17-23. ISBN 0-620-36890-X.

[17] HULL, D.A., GREFENSTETTE, G.: A Detailed Analysis of English Stemming Algorithms[online], *Xerox Research Center* [online]. 1996, [cit. 2010-11-20]. Dostupný z WWW: <http://www.xrce.xerox.com/content/download/6676/51464/file/DHull-GGrefenstette-Technical-report-MLTT96.pdf>.

[18] HYNEK, Josef. *Genetické algoritmy a genetické programování*. 1. vyd. Praha : Grada, 2008. 200 s. ISBN 978-80-247-2695-3.

[19] CHARNIAK, Eugene. Statistical Techniques for Natural Language Parsing. AI Magazine : Volume 18 [online]. 1997, 4, [cit. 2010-10-29]. Dostupný z WWW: <http://www.aaai.org/ojs/index.php/aimagazine/article/view/1320/1221>.

[20] CHIH-WEI, H.; CHIH-CHUNG, Ch.; CHIH-JEN, L. A Practical Guide to Support Vector Classification. *Department of Computer Science National Taiwan University.* 2003, s. 1-16.

[21] Jazzy [online]. 2006 [cit. 2011-05-14]. The Java Open Source Spell Checker. Dostupné z WWW: <http://jazzy.sourceforge.net/>. [webová stránka]

[22] JURŠIČ, M., et al. LemmaGen : Multilingual Lemmatisation with Induced Ripple-Down Rules. *Journal of Universal Computer Science* [online]. 2010, vol. 16, no. 9, [cit. 2010-11-21]. Dostupný z WWW: <http://www.jucs.org/jucs_16_9/lemma_gen_multilingual_lemmatisation/jucs_16_09_1190_1214_jursic.pdf>.

[23] KAO, E.C.-C., et al. Towards Text-based Emotion Detection : A Survey and Possible Improvements. *2009 International Conference on Information Management and Engineering* [online]. 2009, 10720278, [cit. 2010-10-29]. s. 70 - 74 . Dostupný z WWW: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5077000&tag=1>. eclanek doi:10.1109/ICIME.2009.113.

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

[25] KREJČÍ, V.; KOPECKÝ, K. NEBEZPEČÍ ELEKTRONICKÉ KOMUNIKACE. www.e-bezpeci.cz [online]. 2010, 406/08/P106), [cit. 2011-04-17]. Dostupný z WWW: <http://prvok.upol.cz/index.php/vyzkum/37-kyberikana-u-eskych-dti-zavry-z-vyzkumneho-eteni-projektu-e-bezpei-a-centra-prvok-zai-listopad-2009>.

[26] PICARD, R.W. ; VYZAS, E.; HEALEY, J. Toward machine emotional intelligence : analysis of affective physiological state. *IEEE Transaction on Pattern Analysis and Machine Intelligence* [online]. 2001, 10, [cit. 2010-10-29]. s. 1175 - 1191 . Dostupný z WWW: <http://www.computer.org/portal/web/csdl/doi/10.1109/34.954607> ISSN 0162-8828.

[27] PŘINOSIL, J., SMÉKAL, Z. Robust Real Time Face Tracking System. *In Proceedings of the 32nd International Conference on Telcommunications and Signal Processing- TSP2009*, August 26-27, 2009, Dunakiliti, Hungary, pp.101-104. ISBN 978-963-06-77169-5h

[28] PŘINOSIL, J.; SMÉKAL, Z.; ESPOSITO, A. Combining Features for Recognizing Emotional Facial Expressions in Static Images. *In Proceedings of Conference Information: International Conference on Verbal and Nonverbal Features of Human and Human-Machine Interaction*, Lecture Notes in Artificial Intelligence, 2007,Vol. 5042, s. 56-69 .

[29] RAMOS, Juan. Using TF-IDF to Determine Word Relevance in Document Queries. *Rutgers University,Department of Computer Science* [online]. 2003, 1, [cit. 2010-11-14]. Dostupný z WWW: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.1424>.

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

[31] REN , Ye Wu Fuji . Improving emotion recognition from text with fractionation training. *International Conference on Natural Language Processing and Knowledge Engineering.* 2010, s. 1 - 7 .

[32] SMRŽ P.; PITNER T. Sémantický web a jeho technologie (3). *Zpravodaj ÚVT MU.* ISSN 1212-0901, 2004, roč. XIV, č. 5, s. 14-16.

[33] STRAPPARAVA, C.; MIHALCEA, R. SemEval-2007 Task 14: Affective Text. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* 2007, 4, s. 70–74.

[34] ŠANDA, Pavel Určení základního tvaru slova: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2010. 67 s. Vedoucí práce byl Ing. Jan Karásek.

[35] TENG, Z.; REN, F.; KUROIWA, S. Emotion Recognition from Text based on the Rough Set Theory and the Support Vector Machines. *International Conference on Natural Language Processing and Knowledge Engineering* [online]. 2007, 1, [cit. 2010-10-30]. s. 36 - 41. Dostupný z WWW: <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4368008>. eclanek doi:10.1109/NLPKE.2007.4368008 .

[36] WESTON, J., et al. Feature Selection for SVMs. *NIPS* [online]. 2000, 13, [cit. 2010-11-17]. Dostupný z WWW: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20.825&rep=rep1-&type=pdf>.

[37] *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>.

[38] WU , Chung-Hsien; CHUANG, Ze-Jing; LIN, Yu-Chung . Emotion recognition from text using Semantic Labels and Separable Mixture Models. *ACM Transactions on Asian Language Information Processing (TALIP)* [online]. 2006, 5, [cit. 2010-10-30]. s. 165 - 183. Dostupný z WWW: <http://portal.acm.org/citation.cfm?id=1165255.1165259>. ISSN 1530-0226.

[39] YANG, Ch.; LIN, K.; CHEN, H. Emotion Classification Using Web Blog Corpora. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* [online]. 2007, 1, [cit. 2011-05-17]. Dostupný z WWW: <http://portal.acm.org/citation.cfm?id=1331857>.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

GUI grafické uživatelské rozhraní – Graphical User Interface

HMM skrytý Markovův model – Hidden Markov Model

IP Internet Protocol

k-NN k-nejbližších sousedů – k-Nearest Neighbours

LemmaGen generátor lematizátoru – Lemmatiser Generator

NLP zpracování přirozeného jazyka – Natural Language Processing

POS Part of Speech

RBF Radial Basis Function

SVM algoritmus podpůrných vektorů – Support Vector Machine

TF Term Frequency

TF-IDF Term Frequency-Inverse Document Frequency

URL jednotný lokátor zdrojů – Uniform Resource Locator

XML Extensible Markup Language

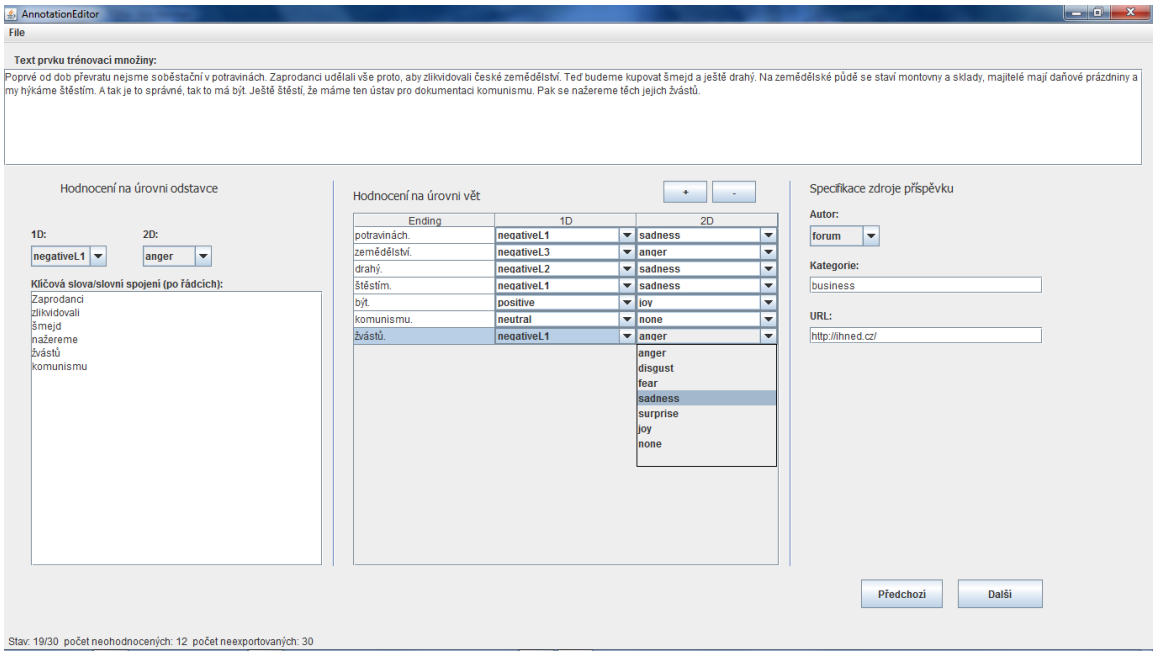
SEZNAM PŘÍLOH

A OBSAH PŘILOŽENÉHO MÉDIA

Přiložené médium má následující strukturu:

- *Elektronická verze práce ve formátu PDF* – Nachází se v kořenovém adresáři.
- Adresář *SW Emotion Recognition* – Obsahuje program vytvořený v jazyku Java pro předzpracování textu včetně transformace s využití lexikální Český WordNet (zahrnuje také zdrojové kódy).
- Adresář *SW Annotation Editor* – Obsahuje program vytvořený v jazyku Java pro hodnocení prvků trénovací a testovací množiny skrze grafické uživatelské rozhraní a následný export do formátu XML v definované struktuře. Adresář zahrnuje také zdrojové kódy.
- Adresář *Emotion Corpus* – Obsahuje ohodnocené prvky trénovací a testovací množiny ve formátu XML s definovanou strukturou.
- Adresář *Data* – Obsahuje množinu stop slov a pravidla použitá pro lemmatizátor.
- Adresář *Training Set* – Obsahuje předzpracované prvky trénovací množiny.
- Adresář *Testing Set* – Obsahuje předzpracované prvky testovací množiny.
- Adresář *Attributes* – Obsahuje optimální množinu atributů.

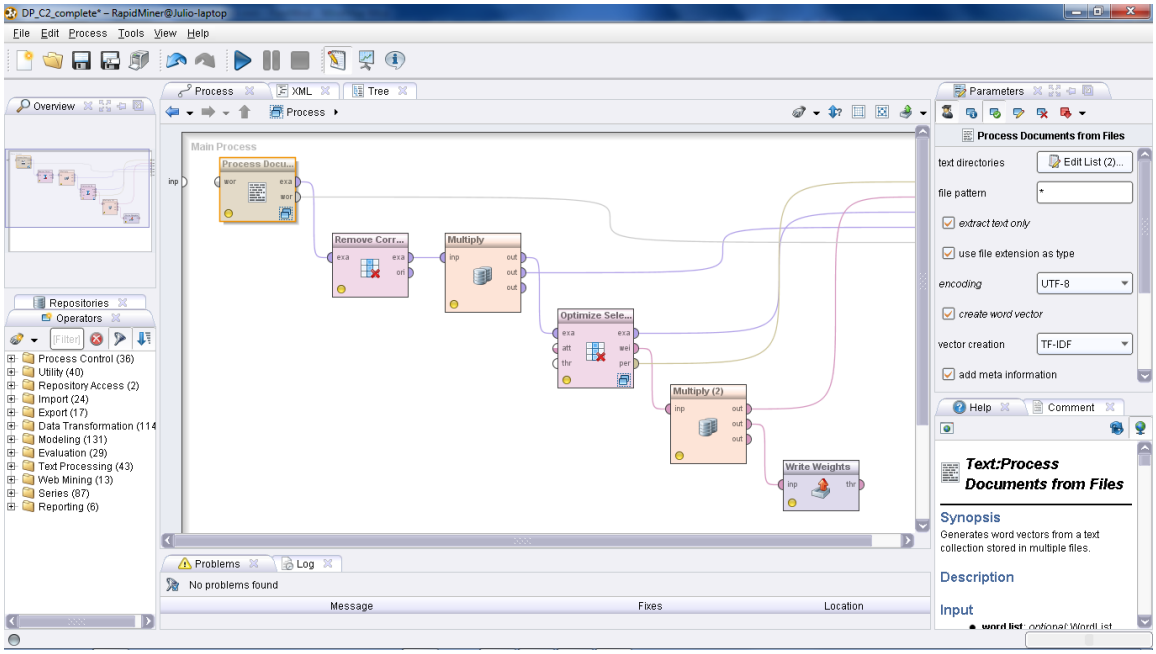
B UKÁZKA PROGRAMU PRO HODNOCENÍ A TVORBU TRÉNOVACÍ MNOŽINY



Obr. B.1: Ukázka vytvořeného programu pro hodnocení a tvorbu trénovací množiny.

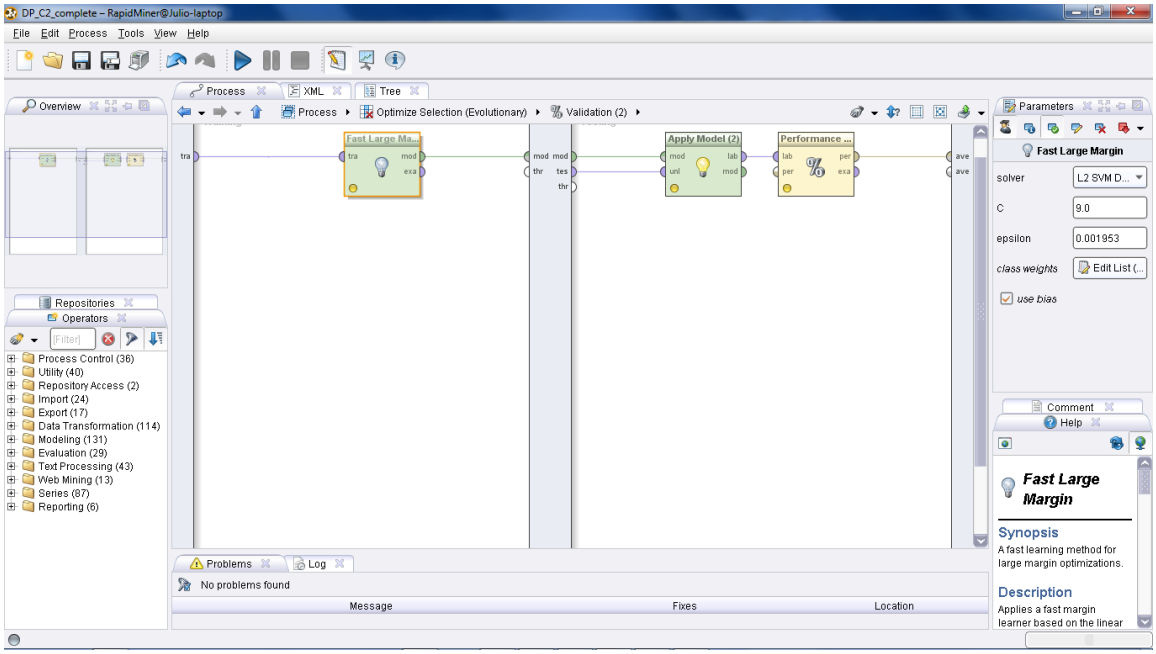
C UKÁZKY ROZLOŽENÍ OPERÁTORŮ V PROGRAMU RAPIDMINER

C.1 Trénovací fáze - hlavní proces



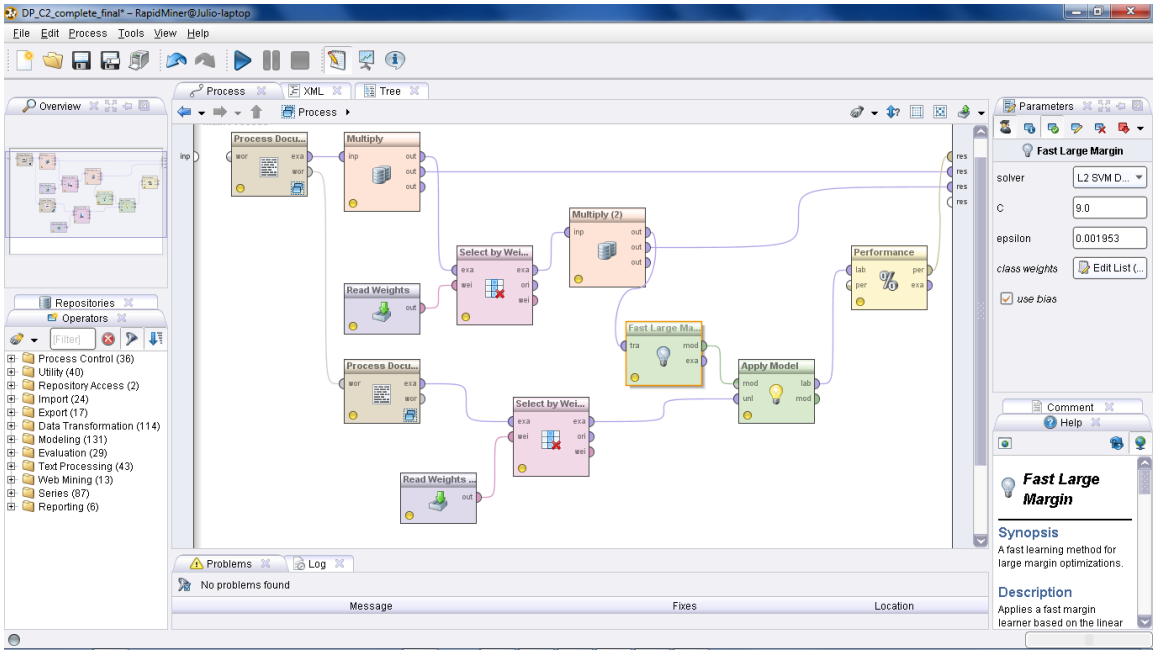
Obr. C.1: Rozložení operátorů v programu RapidMiner pro trénovací fázi - hlavní proces.

C.2 Trénovací fáze - evoluční operátor



Obr. C.2: Rozložení operátorů v programu RapidMiner pro trénovací fázi - vnitřní proces evolučního operátoru.

C.3 Testovací fáze - hlavní proces



Obr. C.3: Rozložení operátorů v programu RapidMiner pro testovací fázi - hlavní proces.