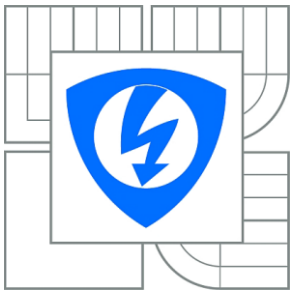
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

AUTOMATICKÉ ZJIŠŤOVÁNÍ VÝZNAMU TEXTU

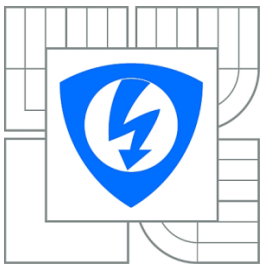
AUTOMATIC RECOGNITION OF MEANING IN TEXTS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE JIŘÍ JELEČEK
AUTHOR

VEDOUCÍ PRÁCE Ing. LUKÁŠ POVODA
SUPERVISOR

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor

Teleinformatika

Student: Jiří Jeleček ***ID:*** 145437

Ročník: 3 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Automatické zjišťování významu textu

POKYNY PRO VYPRACOVÁNÍ:

Vytvořte databázi textů s různým významem pro klasifikaci do dvou tříd. S pomocí metod umělé inteligence (poskytne školitel) natrénujte klasifikátor pro zjišťování pozitivního a negativního náboje v textu. Statisticky zhodnoťte dosažené výsledky pro různé klasifikátory. Vyberte vhodný klasifikátor a natrénujte modely pro rozpoznávání emočného náboje pro 3 různe jazyky - český, anglický, německý, případně jiné.

DOPORUČENÁ LITERATURA:

[1] Ronen Feldman and James Sanger. 2006. Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY, USA.

[2] SCHLEGEL, Katja; GRANDJEAN, Didier; SCHERER, Klaus R. Emotion recognition: Unidimensional ability or a set of modality-and emotion-specific skills?. Personality and Individual Differences, 2012, 53.1: 16-21.

[3] FRANCISCO, Virginia; GERVÁS, Pablo. EmoTag: An Approach to Automated Markup of Emotions in Texts. Computational Intelligence, 2013, 29.4: 680-721.

Termín zadání: 9.2.2015 ***Termín odevzdání:*** 2.6.2015

Vedoucí práce: Ing. Lukáš Povoda

Konzultanti semestrální práce:

doc. Ing. Jiří Mišurec, CSc.

Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

V rámci této práce byl navržen a implementován systém využívající technik dolování znalostí z textu za účelem rozpoznávání emocí v česky, anglicky a německy 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 vybraných algoritmů.

Klíčová slova

klasifikace, dolování znalostí, rozpoznávání emoce, SVM, negativní text, pozitivní text, strojové učení

Abstract

As part of this work it was designed and implemented a system using data mining techniques from the text in order to detect emotions in Czech, English and German language texts. Because the system is built mostly on machine learning techniques, was designed and created training set, which was later used to build the model classifier using the selected algorithms.

Keywords

classification, text mining, emotion detection, SVM, negative text, positive text, machine learning

JELEČEK, J. *Automatické zjišťování významu textu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 38 s. Vedoucí bakalářské práce Ing. Lukáš Povoda.

Prohlášení

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

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

V Brně dne ....

podpis autora

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Lukáši Povodovi za odborné vedení, konzultace, trpělivost a cenné rady při zpracování práce.

V Brně dne ....

podpis autora

Faculty of Electrical Engineering
and Communication

Brno University of Technology
Technicka 12, CZ-61600 Brno, Czechia

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

Výzkum popsaný v této bakalářské práci byl realizovaný v laboratořích podpořených projektem Centrum senzorických, informačních a komunikačních systémů (SIX); registrační číslo CZ.1.05/2.1.00/03.0072, operačního programu Výzkum a vývoj pro inovace.

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

Obsah

Úvod

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).

Podoba informací a jejich konkrétní typ, které mohou být z textu získávány, jsou závislé na povaze řešeného problému. Cílem této práce je použití metod dolování znalostí z textu za účelem rozpoznávání emocí v psaných textech. Analyzovány budou reálné příspěvky elektronických diskuzí. Jednou z těchto informací, které nás mohou zajímat, jsou emoce.

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 [4].

Hlavním přínosem této práce je analýza metod umělé inteligence pro zjišťování pozitivního a negativního náboje v textu z vytvořené databáze reálných příspěvků internetové diskuze. Nalezení vhodného klasifikátoru a natrénování modelu pro český, anglický a německý jazyk.

Práce má následující strukturu. První část je teoretická, zaměřená na Text Mining, zde jsou vysvětleny jeho základy a stručně popsána historie. Dále je vysvětlen rozdíl mezi text miningem, data miningem a vyhledáváním textu. Následuje využití text miningu a kde lze text mining uplatit v praxi. V další části jsou popsány tři základní metody na zpracování dat: hledání na základě klíčových slov, pomocí strojového učení a hybridní metoda. Obecné schéma systému získávání znalostí z textu a popis jeho základních částí. Poslední část teorie je věnována klasifikátorům použitých v praktické části: algoritmu podpůrných vektorů, algoritmu nejbližších sousedů a algoritmu náhodných lesů. Praktická část se zabývá vlastním vytvořením trénovací a testovací databáze, měřením a zpracováním výsledků simulací.

1 Teoretická část

V této části jsou popsány základy Text Miningu a jeho využití. Dále potom metody zpracování dat, jednotlivé metodiky pro vyhodnocování úspěšnosti.

1.1 Podstata text miningu

Spolu s masivní elektronizací dokumentů ve všech sférách lidských činností začalo v letech nedávno minulých docházet k masivnímu nárůstu množství dat uložených v elektronické podobě. Přibližně 80 procent dat uložených v nejrůznějších databázích má podobu textu, je tedy v podobě nestrukturovaných dat [5]. Všechny tyto data jsou výsledkem práce různých webových vyhledávačů, anket, blogů, výzkumů, sociálních sítí, internetových diskuzí, reakcí zákazníků. Jsou tedy výsledkem jak přímé lidské činnosti, tak i činnosti různých strojů a robotů, kteří generují v reálném čase velké množství textových dat. Přečíst a manuálně provést analýze tak obrovského kvanta dat je nemožné, avšak tyto textové informace v sobě obsahují zpravidla údaje, jejichž využití pomůže získat komplexní obraz
o daném bodu zájmu a zvýšit v tomto směru také efektivitu rozhodování.

Pro uživatele je proto nezbytné, aby ve všech nestrukturovaných textových informacích uložených na svých serverech neztratily přehled, průběžně je systematizovaly a samozřejmě také vytěžili prospěch z včasného získání dalších informací analýzou stávajících dat. Právě zde nastupuje na scénu text mining.

"Text mining je metoda, která umí nestrukturovaná textová data zpracovat a poskytne nám stěžejní informaci obsaženou v textu dokumentu, setřídí dokumenty podle podobnosti bez toho, aby je musel někdo číst. Celý tento automatický proces bez potřeby lidských zdrojů je v současné době velmi žádoucí" [5].

Historie metody text miningu je poměrně krátká a logicky souvisí se samotnou existencí dat v digitální formě. Přibližně před čtyřiceti lety inženýři začali hledat způsob, jak propojit sbírky textových dokumentů pomocí počítačových technologií. Položili tak základy vědecké disciplíně, která je známa jako počítačová lingvistika a je v současné době populární na mnohých univerzitách a různých výzkumných ústavech celého světa.

V současné době je tento vědecký obor základním zdrojem informací a metod pro text mining, jako souhrnné označení systému dolování informací v digitální textové formě, které se skládá ze složitých lingvistických metod a kompletní sady nástrojů pokročilé analytiky a statistiky. Text mining se stal nejrozšířenější technologií při řešení úloh reálného světa, počínaje analýzou malých záznamů až k organizaci inteligentního vyhledávání a interpretaci tržních zpráv. Obor text miningu obecně spadá pod soubor data miningových metod, kde vznikl jako další odvětví data miningu, pokrývající požadavky po zpracování textů za souběžného vyhledání informací v nich obsažených. Důvodem separace text mining od data miningu je především skutečnost, že data mining má obecnější záběr, vyhledává a zpracovává informace i v číslech, nominálních a ordinárních proměnných, naopak text mining se specializuje výhradně na práci s nestrukturovaným textem [5] [6].

Formálněji lze text mining popsat následujícím způsobem. Text mining (textová analýza) nebo někdy může být alternativně nazýván data miningem je metoda netriviální automatické extrakce skryté, implicitní, předem neznámé a potenciálně užitečné a důležité informace z velkého množství „nestrukturovaných" a částečně strukturovaných textových dat pomocí kombinaci strojového učení, pokročilých statistických analýz, různých algoritmu, identifikace jádrových konceptů, postojů a trendů a následného použití této informace [6].

Další možnou definicí je popis text miningové metody jako proces objevování respektive získávání znalostí, který má za cíl identifikovat a analyzovat užitečné informace v textech, jež jsou důležité pro uživatele používajícího text miningový software [6]. Ten odhaluje propojení a vztahy ne pouze v rámci jednoho dokumentu, ale napříč celým spektrem dokumentů, se kterými v daný okamžik pracuje. Dokumentem pak může být například článek v odborném časopisu, nebo volné textové odpovědi v dotazníku s otevřenými otázkami, různé záznamy databáze, emailová korespondence i běžné články v novinách. Prvořadou úlohou text miningu je převést nestrukturovaná textová data do strukturované podoby, co nejblíže tomu, jak by to udělal člověk, který by dokumenty četl. Tento softwarově strukturovaný výstup pak lze třídit a vybírat pomocí standardních data miningových metod.

Častou mylnou představou je to, že text mining je prakticky to samé, co vyhledávání v textu. Vyhledávací softwary postupují tak, že hledají informace v textovém materiálu chronologicky. To má za následek fakt, že abychom dospěli k požadovanému výsledku, musíme přesně vědět, co hledáme a také přesně formulovat otázku. Textová analýza používá přesně opačný postup. Logicky pak není ani potřeba, abychom přesně znali hledaný termín, naopak, text miningem se odkrývají slova (předměty) a slovní spojení (koncept) obsažené v těle dokumentů a následně se mapují vztahy mezi nimi. Tento rozdíl vyplývá už ze samotné podstaty vyhledávání, například na webu. Tam vyhledáváme věci, které známe, ale chceme si o nich zjistit další informace. Cílem text miningu je naopak získání informace nové, doposud neznámé [7].

Další s text miningem zaměňovanou metodou je data mining. Rozdíl je ten, že text mining zpravidla vychází ze přirozeného jazyka, tedy volného textu a data mining ze strukturovaných dat. To ovšem nevylučuje kombinované použití text miningu a data miningu, kdy si data miningem vypomáháme při samotné analýze strukturovaných dat. Příkladem může být zpracování dotazníků s uzavřenými a otevřenými otázkami, kde je na dotazy s uzavřenou odpovědí použit data mining, a na dotazy s otevřenou odpovědí text mining.

Na text mining se nahlíží jako na činnost, která se skládá ze tří částí. První část procesu lze popsat jako před přípravu textových dokumentů. Vstupní dokument je převeden do standardizované podoby, na které se dále pracuje. Znalosti jsou získávány ve druhé části, které jsou odvozovány z první části procesu a následně se analyzují. Třetí fáze je už jen export získaných dat z druhé části do srozumitelné formy, jako je například tabulka, graf.

1.2 Využití – uplatnění

Využití a uplatnění text miningu je velmi široké. Své uplatnění nachází při nejrůznějších analýzách zákaznických dat, například záznamů z call center, dále pak při organizaci a inteligentním vyhledávání v klíčových tržních zprávách, reportech atd. Dalším využitím je analýza odpovědí otevřeného průzkumu. Pomocí text miningu lze v

odpovědích objevit svazky slov nebo fráze používané respondenty při hodnocení kladů a záporů daného produktu, služby nebo značky.

Náročnější na zpracování jsou otevřené odpovědi, ale ve výsledku dávají přesnější a kvalitnější výsledky. Hlavním důvodem kvalitnějších výsledků je prostor pro vyjádření své odpovědi, kde není omezen hranicemi ani možnostmi jak vyjádřit jsou odpověď.

Další uplatnění je hledání souvislostí v různých historických dokumentech, ať už se jedná o reakce na regionální informace, politickou situaci, marketingovou kampaň, nebo i sledovaní konkurence [6]. Text mining našel své využití také ve spam filtrech. Emaily jsou automaticky kontrolovány, zpracovávány, filtrovány a třízeny.

Textová analýza může pomoci odhalit slabé a silné stránky produktu. Tomuto účelu slouží analýza reklamací nebo pojistných škod, obecně analýza otevřeného textu z komerčních sfér. Aplikací text miningového algoritmu jsou příslušná data zpracovány a výstupem mohou být třeba nejčastější závady, stížnosti nebo důvody vrácení zboží.
Text mining našel využití i ve spam filtrech. Emaily jsou automaticky zpracovávány a filtrovány třízeny. Třídění nemusí být pouze na skupiny nevyžádaná pošta (SPAM) a běžná emailová komunikace

Své uplatnění také nachází při zachytávání šikany na internetu, tj. kyberšikana. To je forma šikany, která využívá prostředků informační technologie, jako jsou např. SMS zprávy, sociální sítě, email, chat nebo různé elektronické diskuze. Může být využit k automatickému a velmi rychlému prohledávání webů za účelem hledání vulgárních výrazů, urážlivých textů a negativních emocí např. úzkost, strach.

1.3 Metody zpracování dat

První způsob zpracování velkého množství dat je za pomoci lidských sil. Toto zpracování dat je založeno na využití přirozené inteligence, intuice a instinktu jedince. Obrovskou nevýhodou tohoto způsobu je časová náročnost, kde může analýza většího množství trvat několik týdnů, měsíců, někdy i roky. Oproti tomu strojové zpracování je úplný opak. Charakteristické je rozhodováním na základě faktů, nikoliv instinktu. Časově náročné je strojové zpracování o mnohonásobně kratší.

Hledání na základě klíčových slov je nejpříjemnější metoda. Úspěšnost je velmi závislá na předzpracování textu, především na rozdělení vstupních vět na slova, dále na

nalezení klíčových slov a tvorbě databáze klíčových slov patřícím k jednotlivým emočním třídám. Mezi hlavní nevýhody patří nemožnost detekce emocí ve větách, ve kterých není obsaženo klíčové slovo.

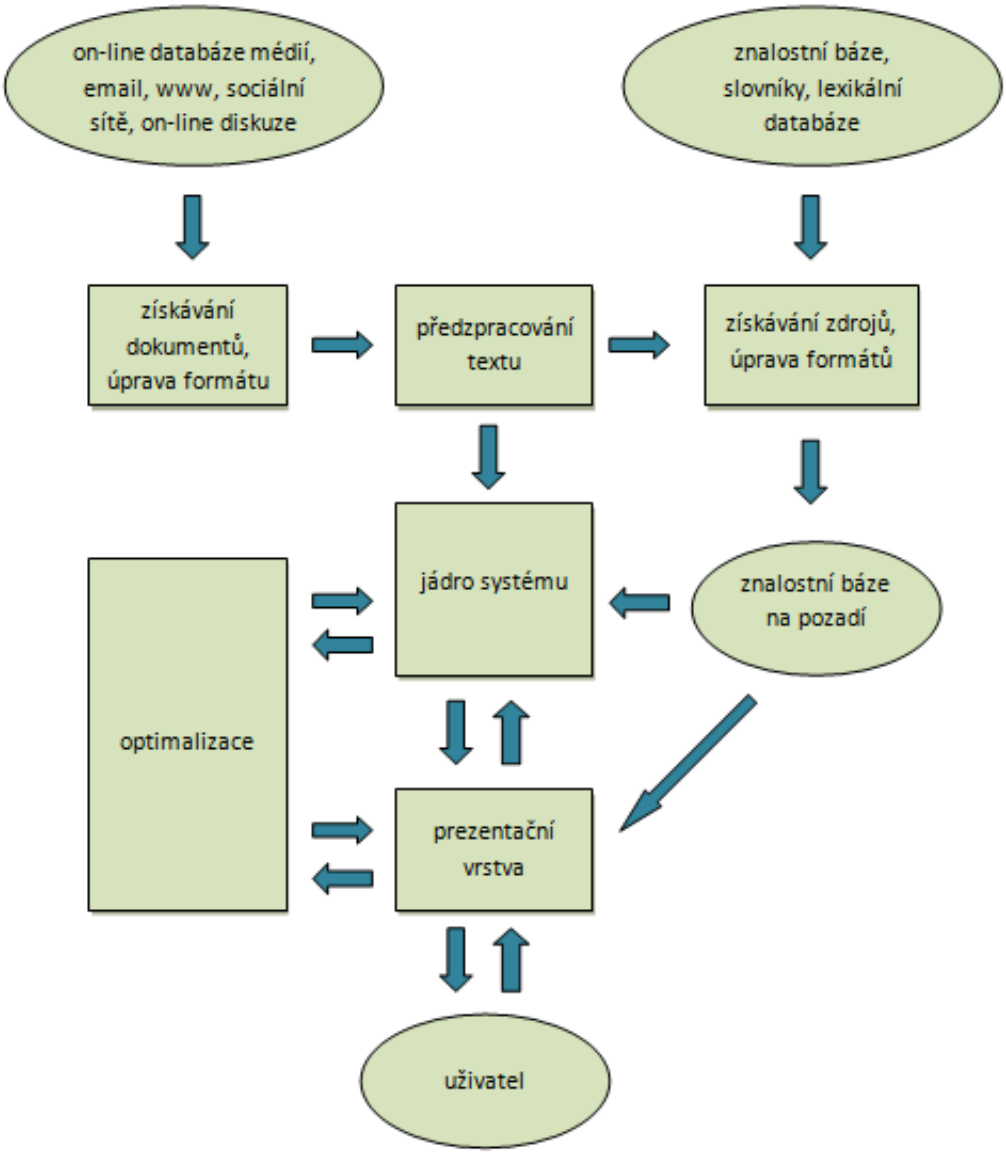
Další metodou je hledání pomocí strojového učení. Tato metoda využívá při detekci emocí učící algoritmy, jejichž parametry byly nastaveny ve fázi trénování na určité množině již předem ohodnoceného textu. Úspěch této metody je založen na kvalitě a velikosti trénovací databáze.

Poslední je hledání pomocí hybridních metod. Tato metoda kombinuje výhody z vyhledávání emocí pomocí metody hledání na základě klíčových slov a strojového učení [7].

1.4 Architektura systému získávání znalostí z textu

Základní architektura obecného systému pro dolování znalostí z textu je zobrazena na obr. 1 a je rozdělena na tyto následující části [1]:

- Předzpracování textu – Tato část zahrnuje všechny operace, které slouží jako příprava textu pro další práci s jádrem systému. Hlavně se jedná o převedení původní formy textu, výběr klíčových slov, vytvoření nového dokumentu. Může ale také připojit z dokumentu důležitou informaci jako je např. zdroj dokumentu nebo časové razítko.
- Jádro systému – Jádro je hlavní část celého systému. Zahrnuje všechny algoritmy, které se přímo podílejí na získávání znalostí z dokumentu. Jedná se tedy o hledání vztahů mezi jednotlivými dokumenty, určování druhu textu (negativní, pozitivní), analýza a shrnutí textu, určení důležitosti dokumentu.
- Optimalizace – Zajišťuje vylepšení výstupu, optimalizaci algoritmů pro získávání znalostí a vyhledávání optimálních parametrů.
- Prezentační vrstva – Obsahuje grafické uživatelské rozhraní, editor pro zadávání příkazů a filtrů. Dále se podílí na prezentaci znalostí získaných z jádra systému a umožňuje ovlivňovat činnost systému pomocí ovládacích prvků.



Obr. 1: Architektura obecného systému dolování znalostí z textu, převzato z [7]

Jestliže systém pracuje s daty určitého druhu zájmu (IT, architektura, lékařství atd.) je možné využít externích zdrojů například lexikální databáze, slovníky a znalostní báze. Tyto zdroje mohou identifikovat známé slovní spojení nebo odborné názvosloví.

1.5 Klasifikace textu

Úkolem klasifikátorů je rozdělit text do předem připravených skupin (tříd). Existuje mnoho možností, do jakých tříd lze text rozložit, ať už podle definovaných emočních skupin, podle obsahu webových stránek, podle autora dokumentu nebo třídění nevyžádané pošty.

Klasifikaci textu lze rozdělit na dvě základní skupiny [1]:

- Na základě strojového učení – Model klasifikátoru je postupně odvozován z již klasifikovaných příkladů (trénovací množina). Tato metoda je jednoduší, levnější a bývá i z hlediska časové náročnosti výhodnější. Úspěch metody je velmi závislý na kvalitě a obsahu trénovací množiny.
- Na základě znalostního inženýrství – Pravidla pro rozdělení jsou stanovena experty v dané oblasti a poté vložena do systému. Tato metoda je oproti strojovému učení drahá a bývá i časově náročná. Výhoda je v tom, že klasifikace na základě znalostního inženýrství bývá velmi přesná a účinná.

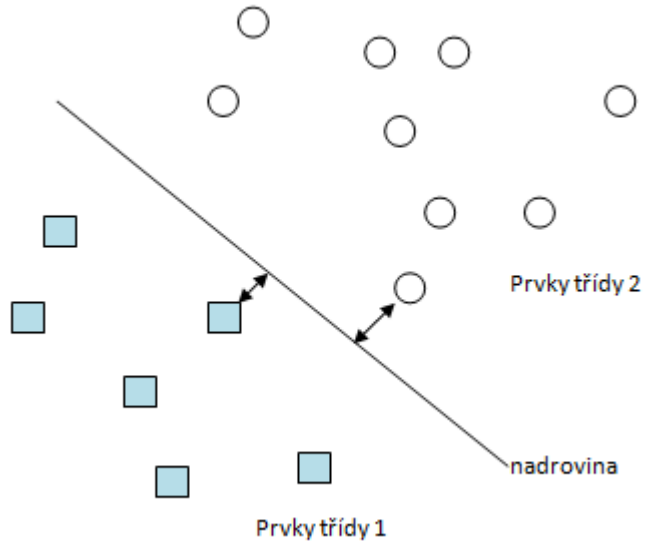
1.5.1 SVM

SVM, neboli Algoritmus podpůrných vektorů (Support Vector Machines) je metoda strojového učení. Algoritmus podpůrných vektorů hledá nadrovinu (rozhodovací hranici), která by v prostoru příznaků optimálně rozdělovala data na požadované třídy. Hlavním požadavkem při hledání nadroviny je maximální vzdálenost mezi nadrovinou a nejbližším prvkem jednotlivých tříd. Jinými slovy, okolo nadroviny by měl být co nejširší pruh bez bodů.

Samotná rozhodovací hranice je určena pouze relativně malým množstvím prvků z trénovací množiny (podpůrnými vektory). Ostatní prvky nemají na utváření modelu klasifikátoru žádný vliv [1]. Tato metoda patří k poměrně novým metodám strojového učení.

Parametr C určuje kompromis mezi složitostí modelu (jeho hladkostí) a mírou odchylek větších než ϵ, které jsou tolerovány optimalizační rovnicí. Například, jeli C příliš velké (nekonečno), potom je snahou optimalizace minimalizovat riziko na základě zkušenosti (počet použitých podpůrných vektorů roste.

Parametr ϵ ovlivňuje šířku oblasti, která se používá k nastavení vzorků trénovacích dat. Hodnota ϵ ovlivňuje počet použitých podpůrných vektorů pro rekonstrukci klasifikátoru. Čím větší ϵ, tím méně se vybere počet podpůrných vektorů. Oba tyto parametry značně ovlivňují složitost i přesnost modelu.



Obr. 2: Ukázka rozdělení prvků tříd nadrovinou nalezenou pomocí SVM

1.5.2 k-NN

k-NN je česky algoritmus k-nejbližších sousedů (k-Nearest Neighbours). Je řazen mezi algoritmy líného učení (lazy learner). V prostoru se nachází prvky trénovací množiny zařazené do dvou tříd. Při určování třídy neznámého prvku je vyhledáno *k* nejbližších sousedů, třída která je nejvíce zastoupena v těchto prvcích je poté přiřazena i neznámému prvku [7].

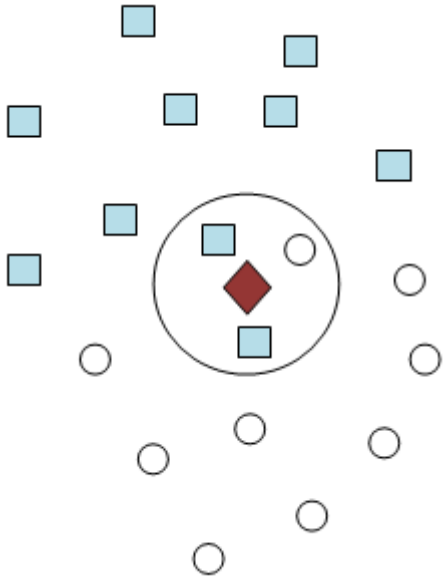
Pro klasifikaci je potřeba následující:

- Množina klasifikovaných (známých) prvků
- Metriku, pomocí které lze vypočítat vzdálenost mezi jednotlivými prvky (např. Euklidovská)

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2} \quad (1)$$

- Číslo *k*

Číslo *k* je počet sousedů, kteří ovlivní výslednou klasifikaci. V praxi je dobré zvolit liché číslo, aby při hlasování nevznikla shoda.



Obr. 3: Ukázka rozdělení prvků pomocí k-nejbližších sousedů

1.5.3 Random Forest

Algoritmus náhodných lesů generuje soubor náhodných klasifikačních stromů. Pro přiřazení prvku do příslušné třídy, dostane každý strom v lese za úkol zařadit prvek do třídy. Les poté provede hlasování a zařadí prvek do třídy, která byla při hlasování zvolena nejvíce. Největší předností této metody je snadná interpretace člověkem. Nevýhodou je relativně malá míra přesnosti a nepříliš uspokojivé výsledky. Při vytváření rozhodovacího stromu se vychází z trénovacích dat. Při posuzování stromů se berou v potaz vztahy:

- GINI index *(p(j | t)* je relativní frekvence třídy *j* v uzlu *t*.)

$$GINI(t) = 1 - \sum_j [p(j|t)] \quad (2)$$

Případně (n je počet všech záznamů v rodičovském uzlu *p*, *ni* je počet záznamů spadající pro část stávajícího uzlu *i*).

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \quad (3)$$

- Entropie (míra neuspořádanosti), (*p(j | t)* je relativní frekvence třídy *j* v uzlu *t*.)

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t) \quad (4)$$

Nevýhodu je, že mají tendenci vytvářet komplexní stromy s velkým množstvím členění (prořezávání stromu (tzv. pruning)).

1.6 Metodiky hodnocení úspěšnosti

1.6.1 Přesnost (Accuracy)

Přesnost při klasifikaci je poměr počtu správně zařazených dokumentů do dané kategorie ku počtu všech zařazených dokumentů do dané kategorie [7]:

$$A = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100 \quad [\%] \quad (5)$$

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

1.6.2 Celková vytíženost (Total recall)

Výtěžnost při klasifikaci je poměr počtu správně zařazených dokumentů do dané kategorie ku počtu všech relevantních dokumentů v rámci dané kategorie [7]:

$$R = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100 \quad [\%] \quad (6)$$

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

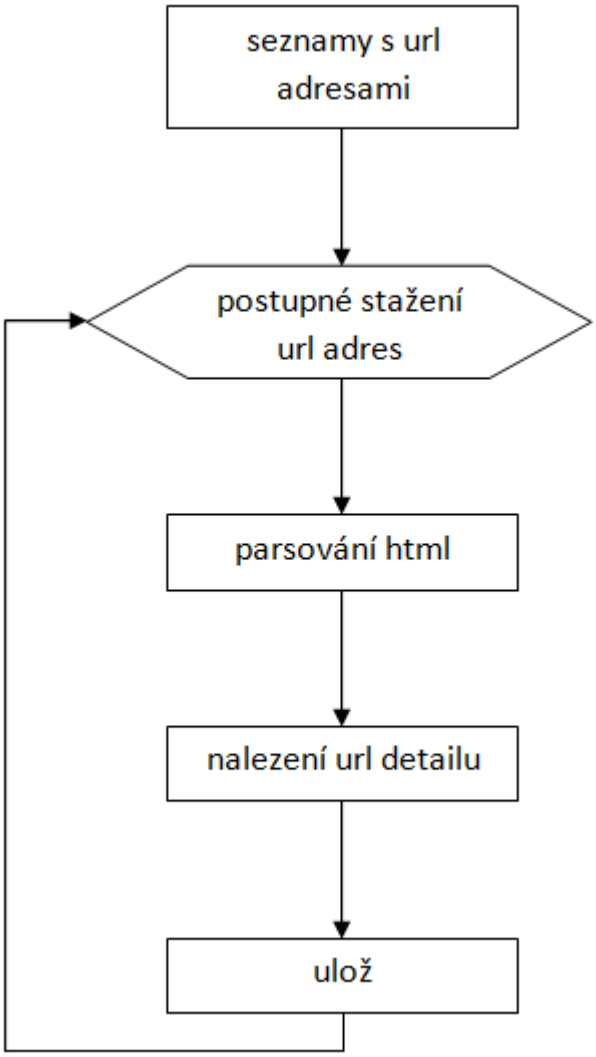
2 Praktická část

V této části byly vybrány čtyři klasifikátory: k-NN, SVM, FLM a Random Forest k simulaci. Výsledky simulací pro tyto klasifikátory zpracovat a vyhodnotit výslednou úspěšnost. Dále potom vybrat nejvhodnější klasifikátor a natrénovat model pro rozpoznávání emočního náboje ve třech různých jazycích. Pro tyto účely byla jako první jazyk vybrána čeština, další potom anglický a německý jazyk z důvodu hojného využívání ve světě.

2.1 Vytvoření databáze

Ze všeho nejdříve bylo nutné vytvořit databázi negativních a pozitivních komentářů. Z vybraných webových serverů na hodnocení filmů bylo potřeba vybrat a uložit všechny komentáře. U jednotlivých filmů měli uživatelé možnost ohodnotit vybraný snímek nulou až deseti hvězdičkami, při čemž hodnocení deset znamenalo maximální spokojenost u daného snímku (u českého serveru bylo možné hodnocení nula až pět hvězd) a následně vložit i komentář, jak se jim film líbil nebo naopak. Na základě toho byl skript automaticky schopen určit, do jaké skupiny text patří – zda jde o pozitivní text (8 a více hvězdiček), nebo zda se jedná o negativní text (2 a méně hvězdiček).

Vytvoření databáze bylo rozděleno do dvou kroků. V prvním kroku bylo třeba získat url adresy, ze kterých se v další fázi čerpalo. Seznam url adres není nikde dostupný, je tedy potřeba shromáždit co nejvíce url adres, které směřují na detail produktu (filmu). Byly tedy staženy pomocí vytvořeného skriptu, který prohledává, zda se někde v html kódu neobjevují url adresy ke stažení. Html kód byl parsován knihovnou Jsoup, která umožňuje jednat s jednotlivými prvky html kódu (textové data) jako se samostatnými objekty a rychle a snadno prohledávat strukturu dokumentu. Následně nalezena adresa url detailu, která byla uložena do příslušného textového souboru. Na Obr. 4 je vyobrazen vývojový diagram pro získání databáze s url adresami.



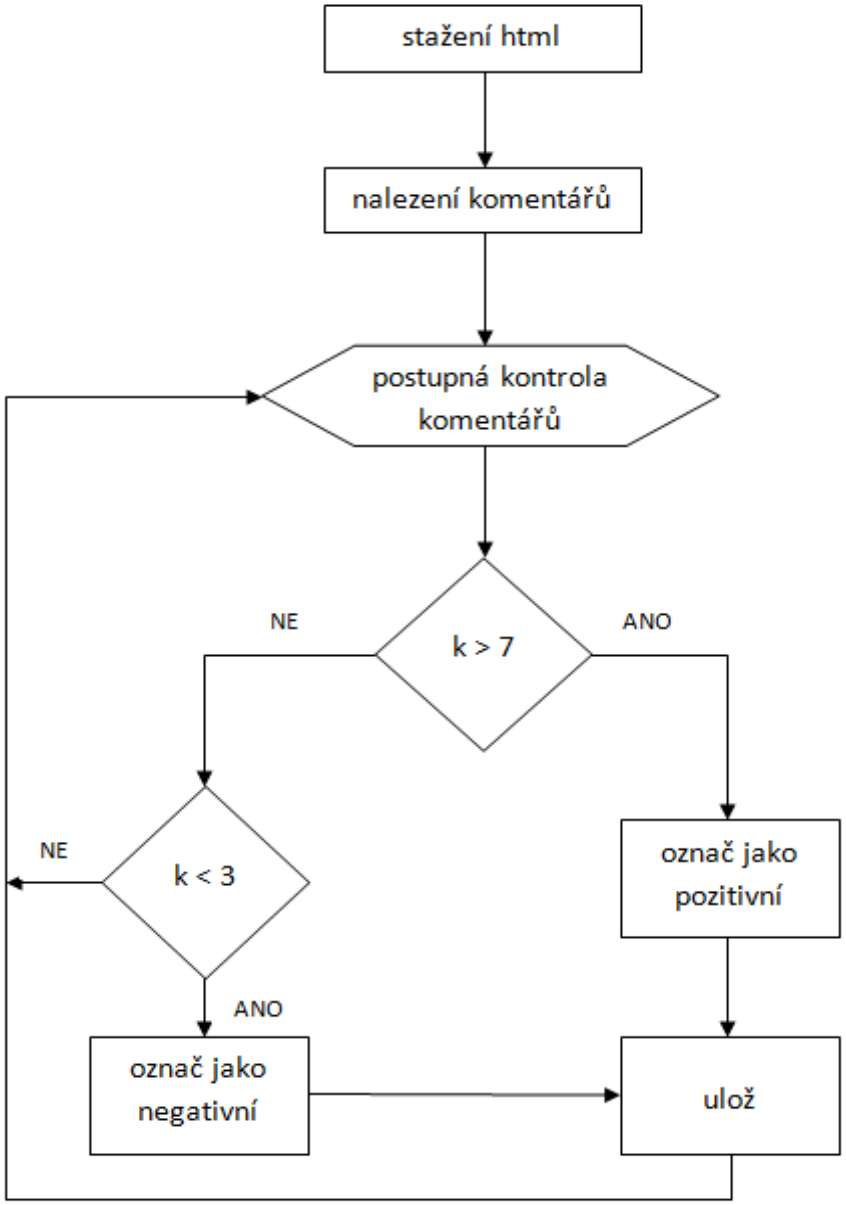
Obr. 4: Vývojový diagram pro získání url adres

Jednotlivé databáze textů byly shromažďovány pomocí skriptu, psaném v programovacím jazyku Java. Po úspěšném zpracováním url adres, se mohlo přikročit k dalšímu kroku a to sestavení trénovacích a testovacích množin pro simulaci.

Po načtení adresy detailu produktu (filmu) se získaný řetězec znaků opět parsuje podobným způsobem jako bylo výše uvedeno, následně je vyhledán blok s komentářem, který je potřeba získat. Poté probíhá kontrola, zda je hodnocení komentáře vyšší než sedm, pokud ano, komentář se označí jako pozitivní a následně uloží do příslušného textového souboru, dále je vybrán další komentář a celý proces se opakuje. Pokud je hodnocení komentáře nižší nebo rovno sedmi, přistupuje se k další podmínce, kde se testuje, zda je komentář nižší než tři. Když tuto podmínku splňuje, testovaný komentář

se opět označí, tentokrát ale jako negativní. Uložen je opět do textového souboru. Takto se pokračuje, dokud jsou k dispozici nalezené komentáře. Vývojový diagram na obr. 5 vyobrazuje hlavní část programu pro vytvoření databáze textu.

Následně byly tyto databáze rozděleny na trénovací data (training data, 66 %) a testovací data (testing data, 33 %). Ve výsledku obsahovala trénovací databáze 2000 komentářů s pozitivním nábojem a 2000 negativních. V testovací databázi bylo 1000 souborů pro obě emoční skupiny.



Obr. 5: Vývojový diagram hlavního programu

Tab. 1: Použité hodnocení pro pozitivní a negativní text

	negativní (počet hvězdiček)	**pozitivní (počet hvězdiček)**
český text	0, 1	4, 5
anglický text	0, 1, 2	8, 9, 10
německý text	0, 1, 2	8, 9, 10

2.2 Výsledky jednotlivých klasifikátorů

2.2.1 Fast Large Margin (FLM)

Tento klasifikátor byl simulován pro dvě velikosti trénovací a testovací databáze z důvodu posouzení zda má velikost databáze vliv na celkovou úspěšnost a výslednou přesnost.

Tab. 2: Přesnost a vytíženost u menší databáze komentářů pro FLM

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	784	176	81,67
pozitivní, předpovídaný	216	824	79,23
celková vytíženost [%]	78,40	82,40	

Jak lze vidět v Tab. 1 celková vytíženost u negativních komentářů byla 78,40% a u pozitivních 82,40%, což znamená, že tento klasifikátor lépe poznal a zařadil komentáře s pozitivním nábojem. Výsledná přesnost (accuracy) pak byla 80,40%.

U databáze s 5 000 trénovacích dat a 2 500 testovacích dat pro obě emoční skupiny. Byly výsledky podobné, jen s nepatrně vyšší úspěšností. Přesnost byla 81,86%. A můžeme tedy říct, že objem trénovacích dat a testovacích dat má pozitivní vliv na hodnocení úspěšnosti. Výsledky simulace je možno vidět v Tab. 2.

Tab. 3: Přesnost a vytíženost u větší databáze komentářů pro FLM

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	1989	396	83,40
pozitivní, předpovídaný	511	2104	80,46
celková vytíženost [%]	79,56	84,16	

2.2.2 Random Forest

Z Tab. 3 je na první pohled vidět, že úspěšnost simulace nebyla tak vysoká jako u FLM. Důvodu může být mnoho, například špatné nastavení výstupních parametrů, špatný výběr vstupních dat. Celková přesnost byla 51,55%. Lze tedy říct, že klasifikátor jen odhadoval na základě typu, do jaké třídy data patří. Random Forest není vhodný pro klasifikaci textu.

Tab. 4: Přesnost a vytíženost u databáze komentářů pro klasifikaci pomocí Random Forest

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	690	659	51,15
pozitivní, předpovídaný	310	341	52,38
celková vytíženost [%]	69,00	34,10	

2.2.3 k-NN

U klasifikátoru k-NN byla přesnost pouze 50,55%. Tato metoda tedy není vhodná pro klasikami textu do dvou emociálních tříd.

Tab. 5: Přesnost a vytíženost u databáze komentářů pro klasifikaci pomocí k-NN

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	995	984	50,28
pozitivní, předpovídaný	5	16	76,19
celková vytíženost [%]	99,50	1,60	

2.2.4 SVM

U algoritmu podpůrných vektorů bylo simulováno měření pro více parametrů. Měněny byly parametry C, ϵ a zkoumáno, jestli tyto parametry ovlivňují tvorbu modelu klasifikátoru. Pro parametr C byly vybrány hodnoty C = 0,5, 5 a 10. Následně vybrána nejvhodnější hodnota parametru a poté tato hodnota odsimulována pro různé ϵ.

Tab. 6: Přesnost a vytíženost pro různá C a konstantní ϵ = 0,002 u SVM

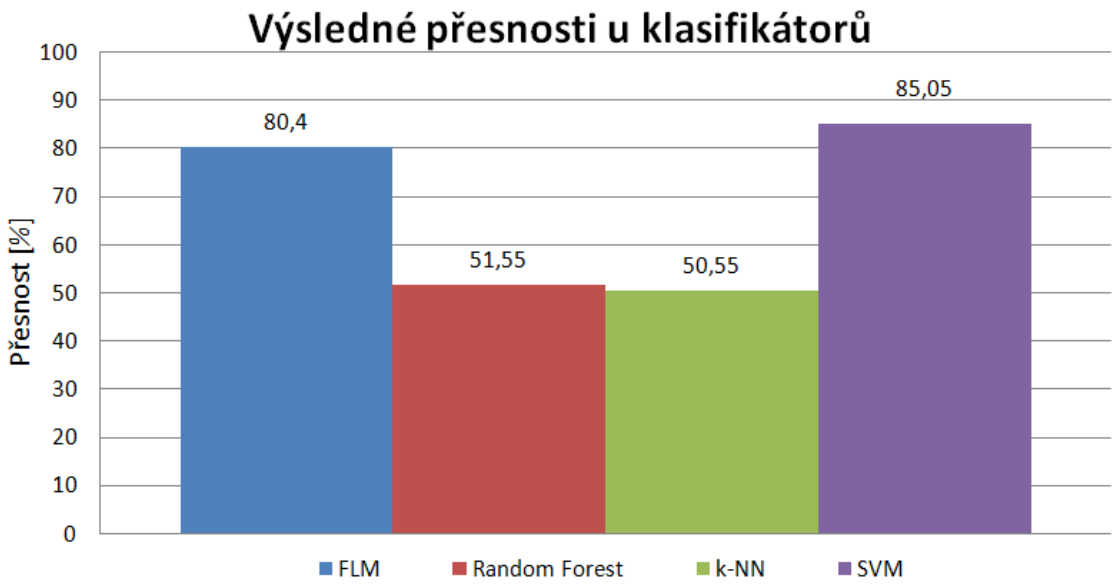
	vytíženost u neg. [%]	**vytíženost u poz.** [%]	**přesnost u neg.** [%]	**přesnost u poz.** [%]
C = 0,5	85,70	84,40	84,60	85,51
C = 5	76,30	83,90	82,58	77,97
C = 10	72,20	84,80	82,61	75,31

V Tab. 5 jsou vytíženosti a přesnosti jednotlivých tříd. Pro C = 0,5 byla výsledná přesnost 85,05 %, u C = 5 přesnost 80,10 % a u C = 10 byla přesnost 78,50 %. Obecně lze tedy říci, že čím větší C, tím menší přesnost. Z naměřených výsledků je vidět, že ze zvolených parametrů je nejvhodnější C = 0,5, jelikož byl při simulaci nejúspěšnější a měl nejvyšší přesnost. Dále byl zvolen parametr ϵ pro C = 0,5, ϵ = 0,125, 0,5 a 4

Tab. 7: Přesnost a vytíženost pro různá ϵ a konstantní C = 0,5 u SVM

	vytíženost u neg. [%]	**vytíženost u poz.** [%]	**přesnost u neg.** [%]	**přesnost u poz.** [%]
ϵ = 0,125	85,70	84,40	84,60	85,51
ϵ = 0,5	86,20	83,80	84,18	85,86
ϵ = 4	0,00	100,00	0,00	50,00

Testování přesnosti a vytíženosti pro ϵ = 0,125, 0,5 a 4 je v Tab. 1.6. Jak je vidět nejlépe ze všech měřených hodnot dopadla simulace pro ϵ = 0,125, kde výsledná přesnost byla 85,05 %. S podobným výsledkem skončilo i měření pro ϵ = 0,5 s přesností 85,0%. Mimo zůstala simulace pro ϵ = 4, kde umělá inteligence vyhodnotila všechny vstupní data jako pozitivní emoci a logicky měla v polovině případů pravdu. Tato hodnota ϵ je zcela nevhodná.



Obr. 6: Závislost přesnosti na použité metodě klasifikace

Na Obr. 4 je vyobrazen graf přesností pro obě emoční třídy v průměrných hodnotách. Nejvhodnější se ukázala být metoda podpůrných vektorů, jako nejhorší byly metody náhodných lesů a metoda nejbližším sousedů.

2.3 Měření klasifikátoru SVM pro tři jazyky

V předchozím bodě bylo zjištěno, že nejúspěšnější metody pro klasifikaci textu je metoda podpůrných vektorů. Další kroku práce bude tedy použita tato metoda.

Jak již bylo zmíněno, simulované jazyky jsou čeština, angličtina a němčina. Z databází, které byly uloženy, se použila opět jen část vzorků pro testování. Důvodem je časová náročnost simulací. Pro tyto účely se simulace prováděly se souborem, který obsahoval 5 000 trénovacích dat a 2 500 testovacích dat.

Tab. 8: Počet všech získaných poznámek

	Pozitivní hodnocení	**Negativní hodnocení**
Český text	1 183 254	183 098
Anglický text	940 982	235 334
Německý text	194 454	50 419

2.3.1 Simulace textů v českém jazyce

Pro odsimulování textů psaných v českém jazyce byla stažena a uložena databáze s pozitivními i negativními hodnoceními. Z nejmenovaného českého serveru na hodnocení filmů bylo posbíráno celkem 1 366 352 poznámek, jak je možno vidět v Tab. 7, což jsou obrovská kvanta textu. Trénování pomocí takového objemu dat by trvalo měsíce, proto bylo rozhodnuto o omezení rozsahu databáze na přiměřenou velikost souboru pro testování. Díky předchozímu rozboru se jako nejvhodnější klasifikátor zvolila metoda podpůrných vektorů (SVM).

Příklad pozitivní testované věty v českém jazyce: *„ Nejlepší film co jsem kdy viděl. Dokonalá hudba, herci, střih, kamera, efekty, režie, pro mě plných 100%. Neuvěřitelné co se Zemeckisovi povedlo natočit.*

Příklad negativní testované věty v českém jazyce: *„Ubohé, nevkusné, v žádném případě vtipné ani ničím zajímavé. Lidi, proč může někdo takového vůbec natočit.“*

Tab. 9: Přesnost a vytíženost u textů psaných v českém jazyce

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	4377	769	85,06
pozitivní, předpovídaný	623	4231	87,17
celková vytíženost [%]	87,54	84,62	

U Tab. 8 je možno vidět, že umělá inteligence vyhodnocovala testovaná data poměrně přesně. Průměrná přesnost pozitivní i negativní text dohromady byla 86.06 %. Po srovnání výsledků s předchozí simulací, že počet vstupní dat má pozitivní dopad na výsledky simulace.

2.3.2 Simulace textů v anglickém jazyce

Stejně jako u textů v českém jazyce, bylo nashromážděno velké množství negativních a pozitivních dat viz Tab. 7. U této simulace se použila taktéž metoda podpůrných vektorů z důvodu zjištění, zda má na dosažené výsledky vliv různý jazyk

testovaných dat. Při pohledu na Tab. 9 je hned jasné, že simulace pro anglický jazyk dopadla mnohem příznivěji a s přesností 93,77 % se umístila na prvním místě s ohledem na výslednou úspěšnost.

Tab. 10: Přesnost a vytíženost u textů psaných v anglickém jazyce

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	4610	232	95,20
pozitivní, předpovídaný	390	4768	92,44
celková vytíženost [%]	92,20	95,35	

Příklad pozitivní testované věty v anglickém jazyce: *„This movie was extremely enjoyable and I recommend it. The screenplay is intelligent and very funny and aside from a slow stretch or two, I really liked the movie. "*

Překlad: *„Tento film byl velmi příjemný a doporučuji ho. Scénář je inteligentní a velmi zábavný, kromě několika pomalých úseků, opravdu jsem si film užil. "*

Příklad negativní testované věty v anglickém jazyce: *„This movie is a terrible waste of time. I have never seen a movie move so slowly and so without a purpose. "*

Překlad: *„Tento film byl neskutečná ztráta času. Nikdy jsem neviděl film, který by se pohyboval tak pomalu a bezúčelně. "*

2.3.3 Simulace textů v německém jazyce

Při shromažďování textů v německém jazyce se ani zdaleka nepodařilo dosáhnout takového počtu dat jako v prvních dvou případech. Důvod bude pravděpodobně ten, že většina korespondentů dává přednost populárnějšímu serveru na hodnocení snímků v anglickém jazyce než málo známému serveru v německém jazyce. I tak se ale podařilo získat dostatečné množství vzorků pro simulaci Tab. 7. Z naměřených výsledků je vidět Tab. 10, že ani simulace německých dat nedopadla špatně. S přesností 82,17 % zaostává za simulací s textem v anglickém jazyce o 11 %.

Tab. 11: Přesnost a vytíženost u textů psaných v německém jazyce

	negativní	**pozitivní**	**třída přesnosti** [%]
negativní, předpovídaný	4260	1043	80,34
pozitivní, předpovídaný	740	3957	84,25
celková vytíženost [%]	85,20	79,15	

Příklad pozitivní testované věty v německém jazyce: *„Der Film mit dem so genialen Hauptdarsteller Eddie Marsan...! Von mir voll die 10 Punkte...! Ich fand Ihn faszinierend vom Anfang bis zum so genialen wundervollen Schluß- Traurig wunder schön. "*

Překlad: *„Film s brilantním hercem jako je .... . Ode mě plných 10 hvězdiček. Zjistil jsem, že film je od začátku do konce fascinující a má geniální a nádherný závěr. "*

Příklad negativní testované věty v německém jazyce: *„NEIN! Wie dieser Film eine so spannende Prämisse innerhalb von 10 Minuten kaputt macht, unfassbar. Sehr merkwürdige Szenen, teilweise völlig sinnlos, merkwürdige Darsteller, von der Hauptdarstellerin mal abgesehen und dazu bricht der Film mit seinen eigenen Regeln. "*

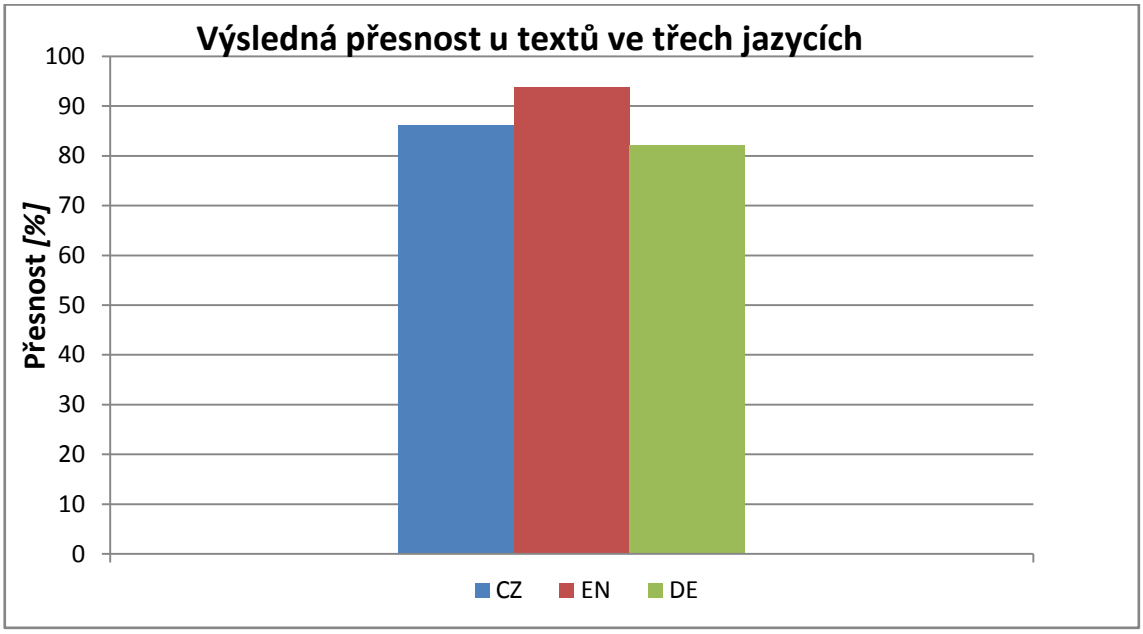
Překlad: *„Ne! Tento film je nezajímavý, nesouvislý a nepochopitelný. Velmi podivné scény, některé úplně nepochopitelné a podivné. Snímek s vlastními pravidly. "*

2.3.4 Porovnání výsledků

Úspěšnost všech testovaných jazyků byla značně vysoká. Nejlepších výsledků dosahuje anglický jazyk, kde byla průměrná úspěšnost pro pozitivní i negativní text necelých 94 %. U českého jazyka se přesnost odhadu zastavila na 86 procentech a pro německý jazyk přesnost 82% Obr. 4.

Důvodů, proč nejlépe dopadla simulace pro anglický jazyk, může být několik. Jedna z příčin bude kvalitou testovaných subjektů. U anglického jazyka mělo hodnocení často i několik desítek vět, či souvětí. Naopak hodnocení v českém a německém jazyce

bylo krátké, většinou obsahovalo maximálně jednu nebo dvě jednoduché věty, v některých případech jen pár slov. Dalším důvodem těchto výsledků je pravděpodobně omezená slovní zásoba anglického jazyka oproti českému, kde se používá mnohem více nesprávně gramaticky napsaných slov, hovorových výrazů a různých zdrobnělin. V českém jazyce se také používá více přídavných jmen a přívlastků.



Obr. 7: Závislost přesnosti na jazyku textu

2.4 Příčiny pro chybnou klasifikaci

Důvodů pro chybnou klasifikaci textu je mnoho. Jedna z příčin může být, že text je psán ironicky. Charakteristické jsou tím, že vyjadřují něco jiného, než je v nich ve skutečnosti napsáno. Pro strojové zpracování je tento text velmi těžké odhalit. Další důvodem je text, který psán s chybami, hovorově nebo v cizím jazyce. Může i docházet k záměně a pochopení významu textu. V anglických a německých textech byl pouze minimální preprocesing, pouze odstranění speciálních znaků a čísel. Také nebyla použita kontrola pravopisu (spellchecker) jako u jazyka českého.

3 Závěr

V rámci této práce byl proveden teoretický rozbor metod na zpracování textu a popis jednotlivých klasifikátorů použitých práci. Byly zde probrány základy text miningu a jeho uplatnění v praxi.

Hlavním přínosem práce je změření zadaných klasifikátorů při použití na česky, anglicky a německy psaných textech, zpracování výsledků, zhodnocení jejich úspěšnosti a porovnání jednotlivých klasifikátorů.

Dosažené hodnoty přesnosti (accuracy), jenž je jedním z měřítek úspěšnosti pro klasifikaci do definovaných emočních tříd, byly následující: u klasifikátoru Fast large margin byla výsledná přesnost 80,40 % u menší databáze se vstupními daty a 81,86 % u databáze s větším rozsahem. Tudíž bylo ověřeno, že úspěšnost klasifikace dat do tříd je závislé na počtu vstupních dat. Dále bylo zjištěno, že metoda náhodných lesů a k-nejbližších sousedů se nehodí na klasifikaci textu do emočních tříd. Ukázalo se že, nejlepší metoda na získávání znalostí z textu je algoritmus podpůrných vektorů, kde výsledná přesnost pro parametr C = 0,5 byla 85,05 %.

Při srovnání textů v českém, anglickém a německém jazyce bylo zjištěno, že nejvhodnější jazyk na identifikaci emocí pomocí umělé inteligence je jazyk anglický, u kterého byl odhad, zda jde o negativní nebo pozitivní emoci, přesný skoro v 94 procentech případů. U českého jazyka byla úspěšnost 86 % a nakonec přesnost 82 % u jazyka německého.

Vytvořené databáze textů budou použity v dalším výzkumu.

Seznam použitých zdrojů

[1] 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.

[2] SCHLEGEL, Katja; GRANDJEAN, Didier; SCHERER, Klaus R. Emotion recognition: Unidimensional ability or a set of modality-and emotion-specific skills?. Personality and Individual Differences, 2012, 53.1: 16-21.

[3] FRANCISCO, Virginia; GERVÁS, Pablo. EmoTag: An Approach to Automated Markup of Emotions in Texts. Computational Intelligence, 2013, 29.4: 680-721.

[4] 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.

[5] ULDRICH, Miloš. Text mining aneb Kladivo na nestrukturovaná data. IT Systems 12/2011: Text mining [online]. Časopis IT Systems, 2011, 12/2011 [cit. 2014-12-09]. ISSN 1802-615X. Dostupné z: http://www.systemonline.cz/clanky/text-mining-kladivona-nestrukturovana-data.htm

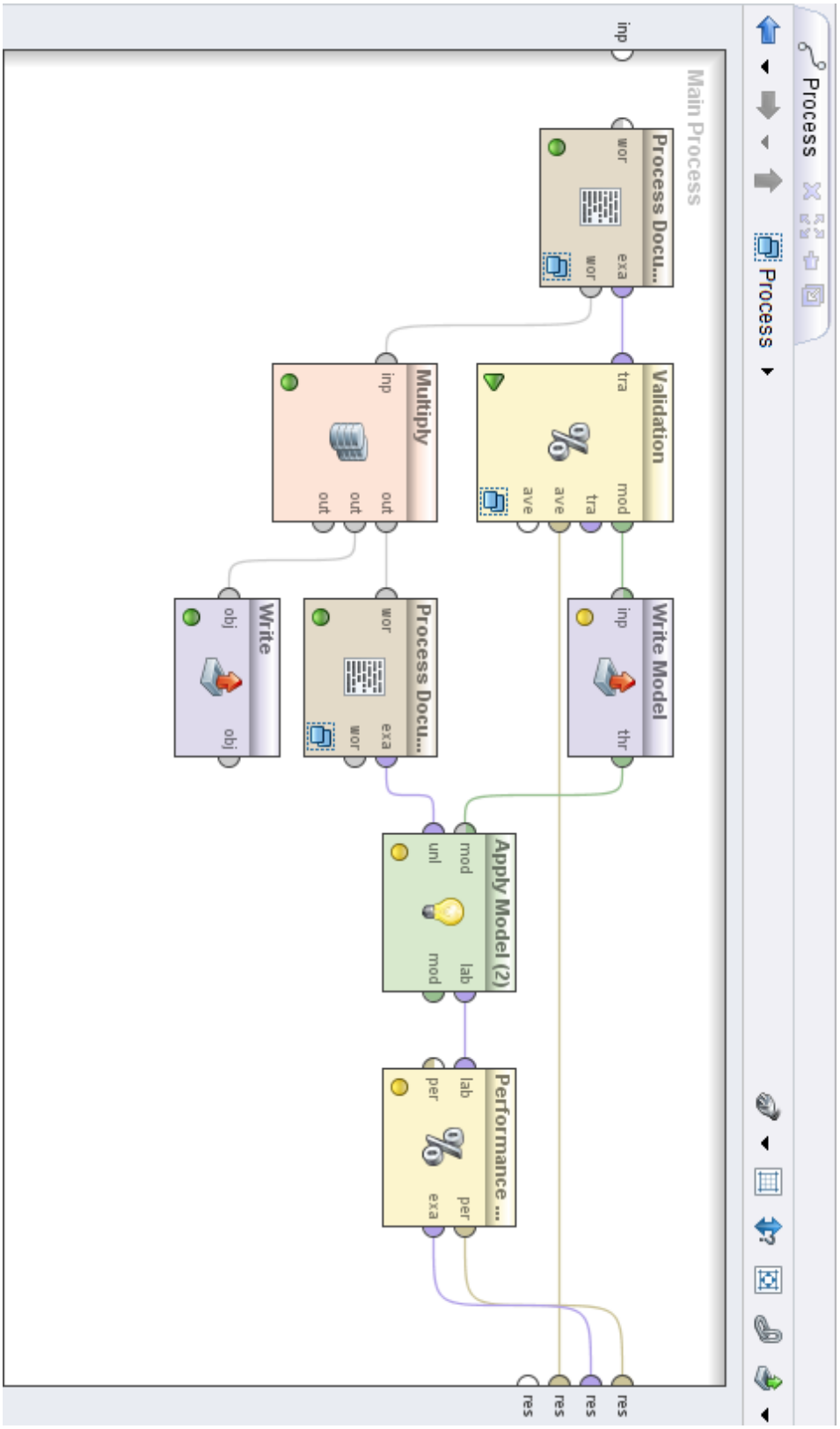
[6] SEDLÁČEK, Petr. Faculty of Informatics MU [online]. 2003 [cit. 2014-12-12]. Text mining a jeho možnosti (aplikace). Dostupné z WWW: http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

[7] 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.

Seznam tabulek

Seznam obrázků

Příloha A – Ukázka prostředí RapidMiner



Příloha B – Výběr klasifikátoru v prostředí RapidMiner