Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Využití značkování sémantických rolí k analýze sentimentu

Plzeň, 2017 Bc. Marek Šimůnek

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 29. června 2017

Bc. Marek Šimůnek

Abstrakt

Cílem diplomové práce bylo využití významu (sémantiky) věty pro zlepšení úspěšnosti analýzy sentimentu. Jedním z úkolů bylo prozkoumat druhy sémantických rolí a metody jejich automatického získávání. Dále vybrat nejvhodnější systém pro jejich značkování se zaměřením na český jazyk.

Pro získávání sémantických rolí byl vybrán nástroj Treex založený na Pražském závislostním korpusu. Po vytvoření sémantických rolí pro vstupní data proběhla extrakce příznaků. Pro analýzu sentimentu na úrovni dokumentu byly použity metody učení s učitelem.

Nejlepších výsledků dosahoval klasifikátor maximální entropie. Naměřené hodnoty překonaly výsledky předchozích výzkumu metod učení s učitelem.

Klíčová slova: Značkování sémantických rolí, analýza sentimentu, strojové učení

Abstract

The goal of the diploma thesis was to use the meaning (semantics) of the sentence to improve accuracy in sentiment analysis. One of the tasks was to examine the types of semantic roles and the methods of their automatic retrieval. Then choose the most appropriate system for their labeling with a focus on the Czech language.

Treex is based on Prague Dependency Treebank and was used to get semantic roles. Feature extraction was performed after semantic role labeling. For sentiment analysis on document level were used machine learning methods.

The best results achieved the maximum entropy classifier. The obtained values outperformed the results of previous research using supervised methods.

Keywords: Semantic role labeling, sentiment analysis, machine learning

Poděkování

Tímto bych chtěl poděkovat svému vedoucímu diplomové práce panu Ing. Tomáši Herzigovi, za odborné vedení, podnětné návrhy a vstřícný přístup.

Obsah

1 Úvod

Analýza sentimentu, někdy také známá pod termínem „dolování názorů“ (opinion mining), je obor, kterému se v posledních letech věnuje stále více pozornosti. Výrobní nebo obchodní společnosti, politické subjekty, cestovní kanceláře apod. rádi využívají analýzu sentimentu k určení své další marketingové strategie. Současně i jednotliví zákazníci před zakoupením určitého výrobku nebo služby rádi sáhnou k analýze, která jim vyhodnotí nejoblíbenější produkt na trhu. Komerční využití analýzy sentimentu tak poskytuje silnou motivaci pro vytváření stále dokonalejších programů.

Výsledky analýzy sentimentu nejsou pořád dostačující a nabízí prostor pro zlepšení. Hledají se stále nové metody a příznaky, které by pomohly zlepšit přesnost klasifikace.

Cílem diplomové práce bylo využití významu (sémantiky) věty pro pochopení jejího sentimentu. Význam věty vyjadřují sémantické role. Identifikace těchto rolí se nazývá značkování sémantických rolí. Na základě teorií z oboru lingvistiky byly vyvinuty různé nástroje a metody pro automatické získávání sémantických rolí. Proto bylo nejdříve nutné se seznámit s dostupnými nástroji a zvolit nejvhodnější systém.

Analýze sentimentu v českém jazyce není věnován dostatečný vědecký zájem. Z tohoto důvodu se diplomová práce zaměřuje na češtinu. Tomu se musel přizpůsobit výběr systému pro značkování sémantických rolí.

2 Strojové učení

Jak už úvod zmiňuje, tato práce se věnuje problémům analýzy sentimentu a značkování sémantických rolí. Jedno z možných řešení pro tyto úlohy je využití slovníkového přístupu. Ten se spoléhá na specializované lexikony vytvářené odborníky a jejich nevýhodou je špatné zohlednění kontextu. V práci jsou tak použity metody založené na strojovém učení (*Machine learning*).

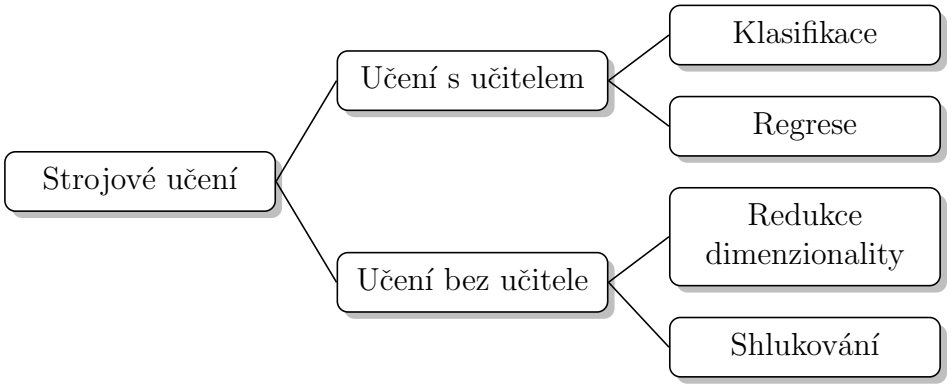
Strojové učení poprvé definoval v roce 1959 Arthur Samuel jako „Obor, který počítačům zajišťuje schopnost učit se bez toho, aby byly explicitně naprogramovány."

Pojem učení lépe vysvětluje z pohledu počítačového programu Tom Mitchell: „O počítačovém programu řekneme, že se učí ze zkušenosti E vykonávat množinu úloh T s úspěšností P, pokud se jeho úspěšnost P na úkolech T zlepšuje se zkušeností E." Zjednodušeně řečeno jde o to, aby výkon (úspěšnost) systému rostl s přibývajícími zkušenostmi (trénovacími daty).

Hlavním cílem ve strojovém učení je schopnost generalizace. Zobecňování ve smyslu natrénování algoritmu tak, aby dostatečně úspěšně dokázal pracovat na dosud nepozorovaných datech, které nebyly součástí trénovací množiny.

2.1 Metody učení

Metody strojového učení lze rozdělit podle způsobu učení. Dále při výběru algoritmu musíme brát v potaz několik dalších faktorů. Jaká přesnost je vyžadována, kolik máme trénovacích dat, jsou označkovaná a jaký je počet příznaků? Všechny tyto faktory nám mohou pomoct pro výběr vhodné metody.



Obrázek 2.1: Nejčastější využití metod strojového učení

2.1.1 Učení s učitelem

Způsob učení s učitelem (supervised learning) má svůj predikční model založený na vstupních datech (trénovací data) a pro ně známém výstupu. Model je tvořen trénovacím procesem, ve kterém se z trénovacích dat odhadne výsledek. Ten se, pokud je predikce nesprávná, upraví na základě přidané informace od učitele na očekávaný výsledek. Trénování probíhá dokud model nedosáhne požadované přesnosti na trénovací množině.

Jedna ze základních úloh učení s učitelem je odhad číselné hodnoty výstupu na základě trénovacích dat (tzv. regrese). Další častým problémem je klasifikace, kde vstup zařadíme do jedné z kategorií.

Naivní Bayes

Naivní Bayesův klasifikátor založený na jednoduchém principu aplikace Bayesovy věty. Tato věta předpokládá, že výskyt všech příznaků, které z textu zvolíme, je statisticky nezávislý na výskytu všech ostatních. Nejzákladnější přístup pro zvolení příznaků je použití modelu bag of words, který obsahuje množinu četností jednotlivých slov a zahazuje pořadí slov nebo gramatické tvary. Dokument D je přiřazen do třídy C za předpokladu

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \tag{2.1}$$

Pokud má dokument stejnou pravděpodobnost zařazení do každé třídy. Znamená to, že marginální pravděpodobnost P(D) je pro všechny třídy stejná a

může být z rovnice vynechána.

$$P(C|D) = P(D|C)P(C) \tag{2.2}$$

Chceme-li maximalizovat pravděpodobnost dokumentu D náležící do třídy C. Zvolíme příznaky reprezentující dokument D a dostaneme:

$$C_{NB} = argmaxP(priznak_1, priznak_2, ..., priznak_n|C)P(C) \tag{2.3}$$

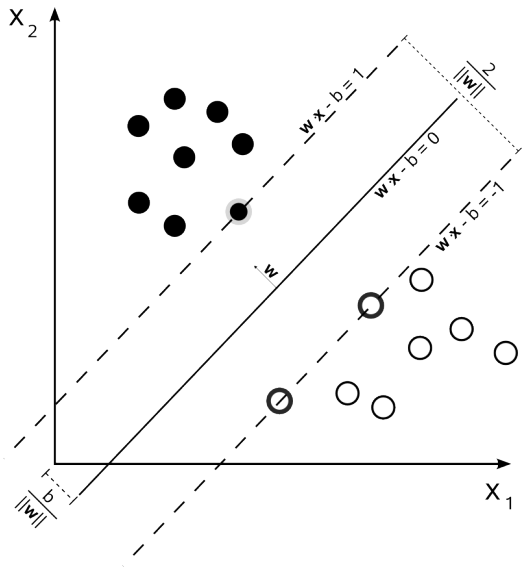
Za naivního předpokladu podmíněné nezávislosti příznaků můžeme zjednodušit rovnici na.[10]

$$C_{NB} = argmaxP(C)\prod_i P(priznak_i|C)P(C) \tag{2.4}$$

Naivní Bayesův klasifikátor nabízí slušné výsledky při snadné implementaci. To je jedním z důvodů, proč je často první volbou pro použití. Stanoví dostačující hranici pro porovnání úspěšnosti s ostatními metodami. Vyžaduje malou množinu trénovacích dat k odhadnutí rozhodovacích parametrů.

Support Vector Machine

Metoda SVM se v klasifikační úloze snaží najít optimální rozdělující nadroviny oddělující trénovací data v prostoru příznaků. Optimální nadrovina odděluje body projekce trénovacích dat, které leží na opačných stranách této nadroviny a zároveň nadrovina má největší vzdálenost k nejbližším bodům jakékoliv třídy. [11] Zjednodušeně řečeno metoda se snaží najít nejširší pás v prostoru mezi třídami tak, aby je od sebe oddělil.



Obrázek 2.2: Nejširší pás bez bodů pro SVM trénovací data ze dvou tříd. Body na okrajích jsou nazývány podpůrné vektory [11]

SVM je využíváno pro řešení mnoha problémů díky své schopnosti generalizace. Tato populární metoda učení s učitelem umí vytvořit rozhodovací hranici pro lineárně i nelineárně separabilní data. [3]

Maximum entropy

Klasifikátor maximální entropie (MaxEnt) je pravděpodobnostní klasifikátor strojového učení. Je založen na principu maximální entropie. Když nejsou žádná vstupní data, klasifikátor předpokládá uniformní rozdělení (maximální entropii). Vytváří si o datech, co nejméně předpokladů. Pozorovaná data entropii sníží, protože nastavují množinu omezení, která by měla být zahrnuta do výsledného rozdělení. Po přidání dat do svého modelu se na zbytku, který ještě nebyl „spatřen“, opět snaží entropii maximalizovat.

Trénovací data v MaxEnt nastavují omezení na podmíněné rozdělení pravděpodobnosti. Najdeme množinu funkcí, které jsou užitečné pro klasifikaci. Pro každý příznak zjistíme očekávanou hodnotu pro trénovací data a zahrneme do pravděpodobnostního rozdělení. To zaručuje nalezení jedinečného pravděpodobnostního rozdělení, které má maximální entropii. Toto rozdělení

je vždy v exponenciální tvaru:

$$P(c|d) = \frac{1}{Z(d)} \prod_{i=1}^{n} e^{\lambda_i f_i(d,c)}, \tag{2.5}$$

kde c je třída, d dokument, $f_i(d, c)$ je funkce příznaků, λ_i odpovídá odhadovanému parametru a $Z(d)$ je normalizační faktor.

Ukázalo se, že metoda maximální entropie může být citlivá na špatný výběr příznaků. Není potřeba mít příznaky pro všechny třídy. [50]

Neuronové sítě

Neuronové sítě jsou založeny na principu biologických struktur, kde základní jednotka je neuron. Neurony jsou navzájem propojené a přenášejí signál na základě aktivačních funkcí.

První pokusy využití neuronových sítí ve zpracování přirozeného jazyka využívaly one-hot vector, který reprezentuje slovo. Každá dimenze vektoru odpovídá jinému slovu. Tento bag of word model nicméně ztrácí sémantickou informaci, což je zvláště důležité v analýze sentimentu.

Nedávný výzkum se soustředil na vytvoření vektoru, který reprezentuje slova a dokumenty s udržením co nejvíce informací. Existuje model word2vec, kde vektorová reprezentace je vypočítána pro každé slovo. Výsledkem je, že slova s podobným významem mají na rozdíl od slov nesouvisejících k sobě v eukleidovském prostoru blíž.

Také další typ neuronových sítí Deep learning zaznamenal obrovský pokrok. Rekurentní neuronové sítě jsou schopny uchovat informaci obsaženou v trénovacích datech a zachytit vztahy mezi slovy. Představitelem je model LSTM (long-short term memory).

LSTM se liší od jiných typů RNN tím, že její buňka obsahuje brány, které více kontrolují zachování nebo potlačení informace v „paměti“. Výstup z buňky je výsledkem několika funkcí. Závisí na tom, jak moc současný vstup ovlivní vytvoření nové paměti, jak se pamět' z předchozích pamětí podílí na nové, a jaká část paměti převládne pro vygenerování výstupu. [16]

2.1.2 Učení bez učitele

Učením bez učitele (unsupervised learning) nazýváme způsob, u kterého pro vstupní data neznáme správný výstup. Systém vytvoří svůj model odvozováním struktur ze vstupních dat. Výsledkem je nalezení obecných pravidel, redukce dimenzionality nebo organizování dat podle jejich podobnosti (shlukování).

Algoritmy jsou založeny na pravděpodobnostních modelech, které učením bez učitele zajistí shrnutí a extrahování klíčových vlastností dat.

Lze zvolit i kombinaci mezi učením s učitelem a bez učitele.

2.2 Zpracování přirozeného jazyka

Zpracování přirozeného jazyka (*Natural language processing* - NLP) zasahuje do více vědních oborů, zejména informatiky, umělé inteligence a lingvistiky. Věnuje se problémům analýzy, porozumění a tvorby jazyka, který lidé používají pro interakci mezi sebou. Vstupem do systému může být psaný text nebo mluvené slovo v přirozeném jazyce člověka a výstupem je odpověď systému ve stejné formě.

Mezi důležité úkoly NLP patří tvorba lidské řeči (speech synthesis), převod mluvené řeči do textu (speech recognition), strojový překlad (machine translation), extrakce a dolování dat z textu nebo automatická sumarizace. Diplomová práce se soustředí na významné úkoly analýzy sentimentu a značkování sémantických rolí (*semantic role labeling*).

3 Značkování sémantických rolí

Cílem práce je využití sémantických rolí jako příznaků pro zlepšení analýzy sentimentu. Je nezbytné si na začátku definovat, co je sémantika a co jsou sémantické role.

V oboru lingvistiky je sémantika nauka o významu jazykových jednotek (morfémů, slov, vět...) [5]. Sémantická role označuje kategorii větné fráze z pohledu sémantiky [12]. Vyjadřuje významový vztah pojmenované entity k ději nebo stavu vyjádřenému predikátem (nejčastěji sloveso). Zjednodušeně řečeno určuje jakou „roli“ daná entita „hraje“ jako účastník v dané události. Terminologie, počty a názvy jednotlivých rolí nejsou v teorii lingvistiky jednotné. Mezi nejvíce vyskytují se role patří: [14]

- konatel (agent) - personální původce děje
- patient (patients) - nositel děje, který je dějem nějak zasažen
- bez českého výrazu (theme) - nositel děje, který nemění dějem svůj stav
- adresát (recipient) - personální původce děje
- proživatel, cíl, kauzátor, nástroj, zdroj, místo atd.

Úloha určení a označení sémantických rolí se nazývá *semantic role labeling* (dále jen SRL). Vyřešení SRL poskytne odpovědi na otázky kdo, kdy, co, kde, proč a další v mnoha problémech oboru zpracování přirozeného jazyka. Zejména v oblastech, kde je potřeba nějaký druh sémantické interpretace (např. extrakci informací, odpovídání na otázky, analýze sentimentu). [15]

Následující věta ukazuje označení sémantických rolí. *Pražský strážník* je původce děje. Přísudek *zadržel* je predikát.

Kdo	*udělal co*	*komu*	*kde*
Pražský strážník	zadržel	podezřelého	na místě činu.
Agent	*Predikát*	*Theme*	*Místo*

3.1 Sémantické role v lingvistice

Pro porozumění používaní současných metod a dostupných anotovaných textových struktur dat si stručně vysvětlíme lingvistické pozadí za teorií sémantických rolí.

Zásadní práce od Fillmora s názvem *Case for Cases*[17] dala lingvistům základ pro rozvinutí teorií sémantických rolí. Fillmorova teorie se věnuje pádům v rámci syntaxe a sémantiky, jde tedy o pády v hloubkové struktuře věty (tzv. *deep cases*) [18]. Musíme si dát pozor a „hloubkové“ pády nezaměňovat s gramatickými. Gramatické pády jako nominativ, akuzativ či dativ slouží k realizaci pouze „povrchní“ struktury (*surface structure*). Příkladem rozdílu povrchové a hloubkové struktury jsou věty 1 a 2 . Mají jinou povrchovou, ale stejnou hlubokou strukturu.

(1) „Táta hlídá syna.“
(2) „Syn je hlídán tátou.“

Fillmore ve své práci uvedl myšlenku, že každá věta se skládá ze slovesa a jedné nebo více frází. Fráze spadá do jedné z šesti základních hloubkových pádů: agentive, instrumental, dative, factitive, locative, objective. Jde o univerzální a pravděpodobně vrozené koncepty lidí, které lze zaznamenat ve všech jazycích [14]. Výhodou je zobecnění významu. Oproti tomu nevýhodou je již zmiňovaná nejednotnost v počtu a významu sémantických rolí. Dalším problémem je nejednoznačné přiřazování sémantických rolí, kdy lze přiřadit více rolí jedné frázi.

Dowty zvolil jiný přístup a zmenšil množinu rolí na dvě základní [20]:

- Proto-agent má vlastnosti: volitelné zapojení do události nebo stavu, způsobí událost nebo změnu stavu u jiného účastníka, pohyb (vzhledem k pozici jiného účastníka).
- Proto-Patient má vlastnosti: prochází změnou stavu, kauzálně ovlivněn jiným účastníkem, přírůstkové téma.

Na předchozím příkladu věty 1 je *táta* Proto-Agentem a *syn* Proto-Patientem i v případě věty 2 role zůstanou stejné.

Další pohled, s kterým přišla Levinská [21] se zaměřuje na sémantickou klasifikaci predikátů samotných. Klasifikace je založena na verb alternations, což je případ, kdy stejné sloveso může být použito s jiným počtem a jiným umístěním argumentů (tzv. valencí [13]).

Hypotéza byla, že syntaxe odráží základní sémantiku. Pokud skupina sloves se chová stejně syntakticky, měla by existovat nějaká sémantická spojitost. Levin provedla hypotézu na anglickém jazyce a rozdělila je do 47 nejvyšších kategorií, 193 kategorií druhé a třetí úrovně pro celkem 3100 sloves. Nevýhodou tohoto rozdělení byly sémanticky nehomogenní třídy (slovesa s jiným významem), slovesa, která spadala do více tříd a rozpor v typech alternation.

3.2 Sémantické role v počítačové lingvistice

Teorie sémantických rolí z lingvistiky měla obrovský vliv při tvorbě slovníků zachycující vztahy mezi predikáty a argumenty. [19] Forma reprezentace sémantické struktury není jednotná a každý slovník má svojí vlastní. Většinou však vychází ze syntaktické struktury, která se následně zpracuje do sémantických vztahů.

3.2.1 Závislostní a frázový strom

Syntaktická struktura zachycuje vztahy mezi slovy. Reprezentace jejich struktury se dělí na dva nejčastější případy: závislostní strom a frázový strom.

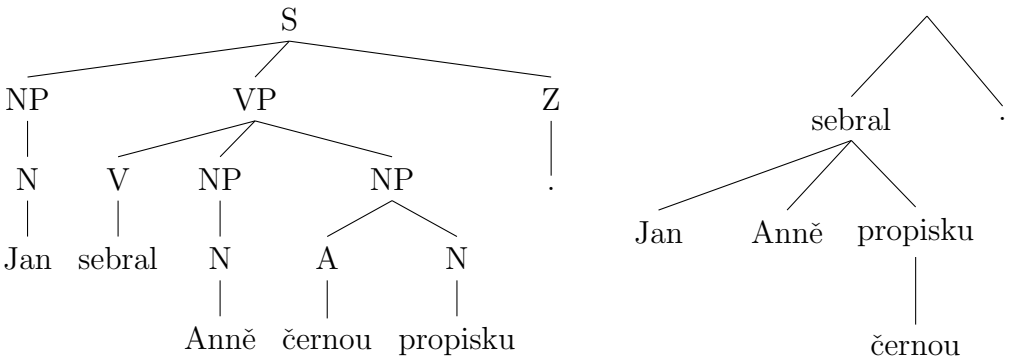
Frázový strom zobrazuje vztahy mezi jednotlivými částmi věty (fráze), které se poté rozkládají na další menší části (konstituenti). Konstituent je jazyková jednotka, ze které se tvoří bloky stejné úrovně do bloků vyšší úrovně [22]. Postup vytvoření frázového stromu se řídí podle frázových přepisovacích pravidel.

Druhý způsob, jak reprezentovat syntaktickou strukturu, je závislostní strom. Uzel v závislostním stromě odpovídá jednomu slovu. Ve stromu jsou závislosti mezi řídícími slovy a jejich podřízenými nebo rozvíjejícími slovy.

Využití frázového stromu je vhodné tam, kde je dané pořadí slov a jasná přepisovací pravidla (angličtina). Známý anglický anotovaný projekt Penn

Treebank používá frázové parsování. Oproti tomu závislostní strom má výhodu ve volném slovosledu a způsobu vyjadřovaní (čeština). Závislostní parsování používá Pražský závislostní korpus (PDT). Musíme brát v potaz, že přesnost určení závislosti v českém jazyce se pohybuje okolo 85% [1].

Na rozdíl od frázového stromu má závislostní strom menší počet uzlů a přehlednější grafické znázornění závislosti na jednotlivých slovech. Na obrázku 3.1 vidíme názorné ukázky obou stromů pro stejnou větu. U obrázku 3.1a) jsou jednotlivé části vět značeny zkratkami S (věta), NP (jmenná fráze), VP (slovesná fráze), V (sloveso), N (podstatné jméno), A (přídavné jméno) a Z (interpunkce).



a) Frázový strom b) Závislostní strom

Obrázek 3.1: Syntaktické stromy pro větu „Jan sebral Anně černou propisku.“

3.2.2 FrameNet

FrameNet je prvním projektem pro automatické SRL. FrameNet je lexikální databáze angličtiny obsahující 1200 semantic frames (sémantických rámců), 13000 lexikálních jednotek a více než 200 tisíc anotovaných vět. Vznik projektu je založen na teorii Frame Semantics (rámcová sémantika) od Fillmora. Hlavní myšlenka spočívá v tom, že význam většiny slov lze pochopit na základě sémantického rámce, popisu typu události nebo entit a participantů v ní. Příkladem může být koncept vaření, kde rámec (frame) se jmenuje *Apply_heat* (aplikovat teplo) a *Cook*(vařit), *Food* (jídlo), *Heating_instrument* (ohřívací nástroj) jsou elementy rámce (frame elements), Slova, která spa-

[1] `http://ufal.mff.cuni.cz/czech-parsing`

dají do tohoto rámce jako jsou *fry* (smažit), *bake* (péct) se nazývají lexikální jednotky (lexical unit) [24].

FrameNet trpí tím, že jeho korpus není reprezentativní vzorek jazyka a skládá se z převážně z ručně vybraných ukázek. Dále sématické role trpí neduhem, že jsou moc situačně konkrétní, než-li obecnější jako u role Agent, Theme a Location, které můžou být použity napříč různými místy a styly. Projekt zapříčinil první pokusy SRL s přístupem statistického strojového učení. Nicméně nyní se jeho dataset tolik nepoužívá. [19]

3.2.3 VerbNet

VerbNet[2] je největší lexikon sloves anglického jazyka. Je organizován podle tříd sloves, které navrhla již zmiňovaná Levin [21]. Postupem času kategorie prošly změnou pro větší syntaktickou a sémantickou soudržnost mezi členy různých tříd.

Lexikon je považován za důležitý zdroj pro počítačovou lingvistiku. Většinou se používá ve spojení s jinými lexikony. Převážně pro svou absenci souvisejících anotovaných rolí v korpusu.

Tabulka 3.1: Zjednodušená ukázka třídy z VerbNetu

Class Hit-18.1	
Roles and Restrictions: Agent[+int_control] Patient[+concrete] Instrument[+concrete]	
Members: bang, bash, hit, kick, ...	
Frames:	
Name	Basic transitive
Example	Paula hit the ball
Syntax	Agent V Patient
Semantics	cause(Agent, E)manner(during(E), directedmotion, Agent) !contact(during(E), Agent, Patient) manner(end(E),forceful, Agent) contact(end(E), Agent, Patient)

3.2.4 PropBank

Proposition Bank (PropBank), inspirující se z projektu VerbNet, používá obecnější sémantické role označované jako prototypy. Jedná se o Dowtyho

[2] `https://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

rozdělení na Proto-Agent a Proto-Patient (sekce 3.1). PropBank anotoval sémantické role pro všechny slovesa z korpusu Penn Treebank (obsahuje hlavně žurnalistické články z Wall Street Journal).

PropBank byl vytvořen se záměrem poskytnutí trénovacích dat ve strojovém učení pro úlohu SRL. Stal se tak důležitým zdrojem pro automatickou tvorbu sémantických rolí.

PropBank má pro každé sloveso množinu rámců (frameset), v kterých argumenty začíná číslovat od nuly. Argumenty doplňují význam slovesa a jsou číslovány od 1-4. Role jednotlivých argumentů se můžou lišit, ale většinou vypadají následovně:[20].

- Arg0 - Proto-Agent
- Arg1 - Proto-Patient
- Arg2 - benefactive / instrument / attribute / end state
- Arg3 - start point / benefactive / instrument / attribute
- Arg4 - end point

Tento postup má zajistit konzistentní značkování argumentů napříč různými syntaktickými tvary.

Dále se PropBank snaží přiřadit funkční tag ke všem modifikátorům a doplňkům slovesa. V následující tabulce 3.2 je zkrácený seznam možných funkčních tagů.

Tabulka 3.2: Zkrácený seznam možných funkčních tagů PropBank

Zkratka tagu	**Funkce**	**Příklad**
TMP	kdy?	včera, 5pm v neděli
LOC	kde?	v koupelně, v časopisu
DIR	kam, odkud?	dolů, z Ameriky
MNR	jak?	rychle, s nadšením
PRP/CAU	proč?	protože..., takže ...
REC	zvratná zájemena a slovesa	se, si, sebou
GOL	konec pohybu, převodu slovesa	Na zem, k Janovi
ADV	smíšené, co nikam nepatří	navzdory, bohužel
PRD	argument se týká něčeho nebo upravuje jiný	snědl maso *syrové*

Ukázka rámce win.01 pro sloveso vyhrát

Role:

Arg0: winner (vítěz)

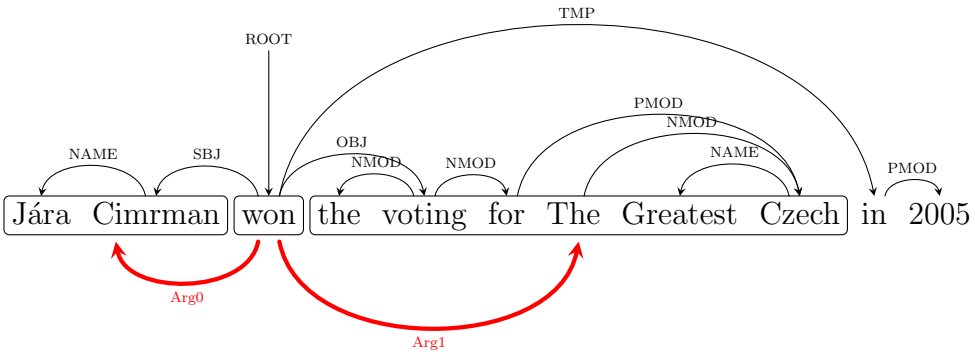
Arg1: thing won (vyhraná věc)

Příklad: Jára Cimrman won the voting for The Greatest Czech in 2005.

Arg0: Jára Cimrman

vztah: win

Arg1: the voting for The Greatest Czech



3.2.5 Pražský závislostní korpus

„Pražský závislostní korpus (Prague Dependendcy Treebank, dále jen PDT) je první počítačový korpus češtiny komplexně syntakticky anotovaný na základě závislostní gramatiky.“[25] Skládá se z velkého množství anotovaných českých textů (převážně z Českého národního korpusu) doplněných o 3 propojené úrovně: rovina morfologická (lemma, tag), rovina analytická (závislosti jednotlivých prvků z morfologické roviny) a tektogramatická rovina (rozbor sémantiky, významu).

Podrobnější vysvětlení jednotlivých vrstev PDT[3] s příkladem si ukážeme na větě.

[3] `http://ufal.mff.cuni.cz/pdt2.0/`

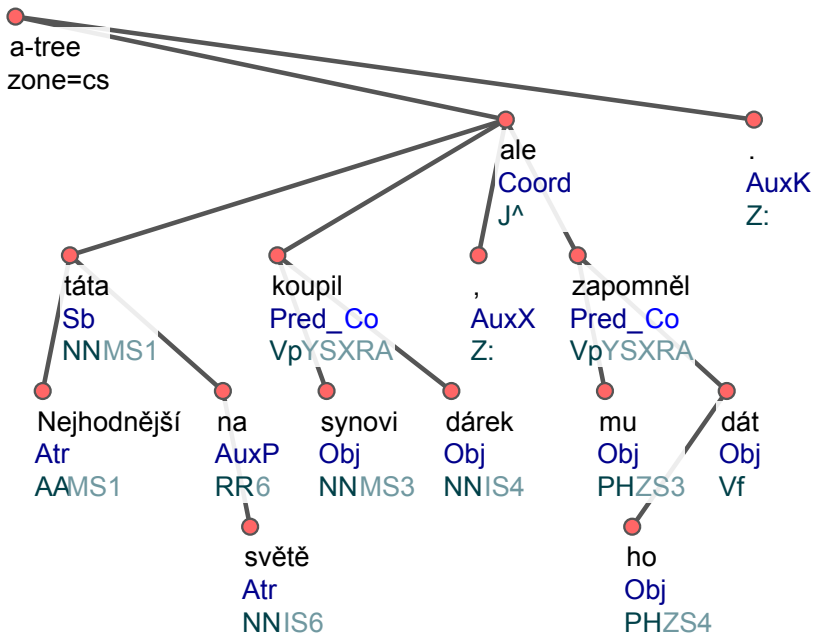
(3) „Nejhodnější táta na světě koupil synovi dárek, ale zapomněl mu ho dát.“

1. Morfologická rovina rozděluje původní text do vět. Na této rovině se slovním jednotkám přiřazuje několik atributů. Nejdůležitější je morfologické lemma (základní tvar slova) a tag, který obsahuje morfologické kategorie jako jsou slovní druh, rod, číslo, čas, zápor, stupeň atd. Další atributy slouží k opravám chyb, identifikování zkratek, cizích slov a frází.

   První slovo věty *nejhodnější* bude mít lemma hodný s přídavnou informací, že se jedná o význam *nezlobivý* a ne od slova být hoden. Dále tag slova *nejhodnější* vypadá následovně: `AAMS1----3A----`, což znamená přídavné jméno (AA), mužský životný(M), jednotné číslo (S), pád (1), stupeň (3), bez záporu (A).

2. Analytická rovina se definuje jako orientovaný strom s kořenem, s ohodnocenými hranami a uzly. To znamená, že každému prvku z morfologické roviny je přiřazen jeden rodič, na kterém je prvek závislý. Ohodnocení hran analytickou funkcí určuje závislostní vztah na rodiči.

   Analytická funkce (afun) na obrázku 3.2 je zobrazena modrým písmem pod slovem v uzlu a může nabývat hodnot jako jsou predikát (Pred), podmět (Sb), předmět (Obj), doplněk (Atv), přívlastek(Atr) atd.

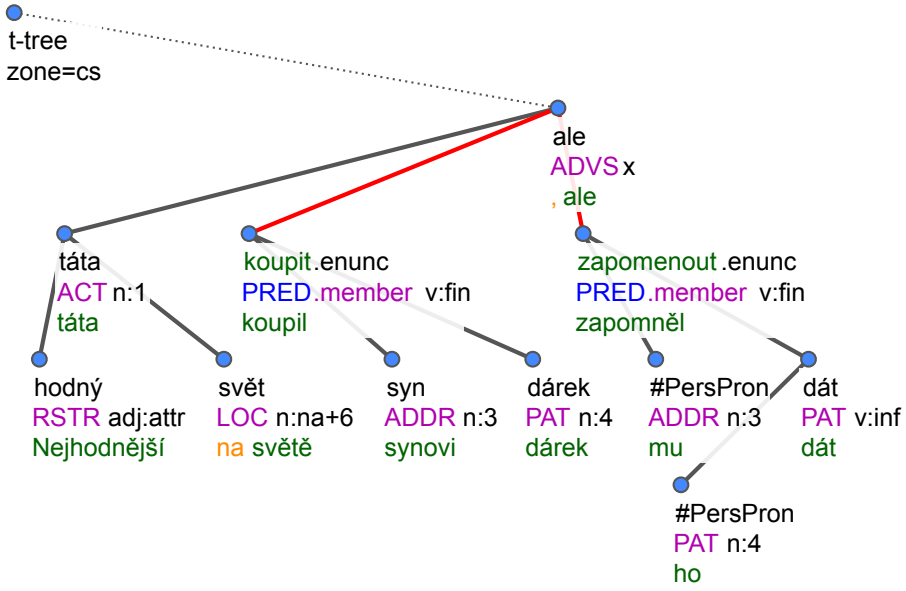


Obrázek 3.2: Ukázka analytické roviny věty 3

3. Tektogramatická rovina zachycuje významovou strukturu věty. Je reprezentována orientovaným stromem s kořenem s ohodnocenými hranami a uzly.

   Uzly zastupují plnovýznamová slova. Nemusí zde být všechny prvky z morfologické roviny (např předložky) a některé uzly můžou být nové (nevyjádřený podmět). Některé uzly mají navíc informace (grammatémy), které nelze odvodit ze struktury. Ohodnocení hran jako v analytické úrovni popisuje typ vztahu. Každé sloveso nebo jistý typ podstatného jména má navíc atribut valenčního rámce (česká obdoba VerbNetu).

   Základními atributy uzlu jsou tektogrammatické lemma (lexikální význam uzlu), grammatémy (význam lexikálních a morfologických kategorií) a funktor (sémantické ohodnocení syntaktického vztahu). Funktor je na obrázku 3.3 v uzlu vyznačen fialově. Z věty vidíme funktory ACT (aktor - nositel děje), ADDR (adresát), ADVS (koordinační struktura vyjadřující odporovací vztah), PAT (patiens - předmět zasažený dějem) atd.



Obrázek 3.3: Ukázka tektogramatické roviny věty 3

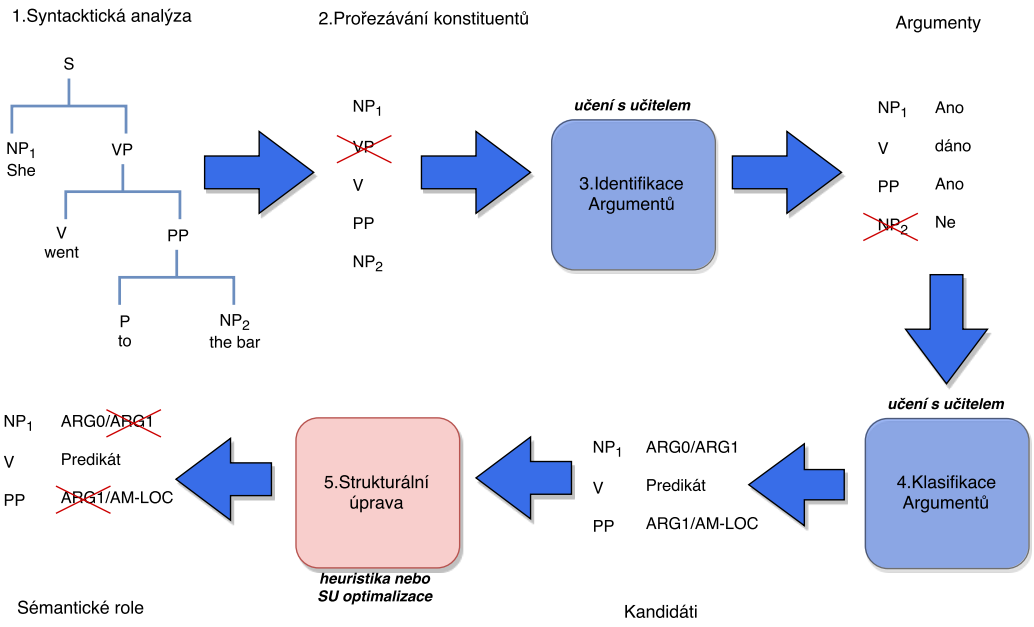
3.3 Automatické metody SRL

Jelikož manuální značkování sémantických rolí je časově velice náročné, hledají se způsoby jak tento proces zautomatizovat. Používají se metody strojového učení a statistický přístup.

3.3.1 SRL učení s učitelem

První přístup, jak řešit automatické SRL je učení s učitelem. Protože máme k dispozici anotovaná data (viz 3.2) můžeme je využít pro učení klasifikátoru.

Obvyklý postup, jak se k úloha řeší, se skládá z následujících kroků. Vstupem je věta. Nejprve se ve větě najde predikát a jeho argumenty (indentifikace argumentů) a pak se označí sémantickými rolemi (klasifikace argumentů)[19].



Obrázek 3.4: Architektura SRL systému [20]

1. Na začátku provedeme syntaktickou analýzu (syntactic parsing), z které dostaneme množinu kandidátů na argumenty pro každý predikát.

2. Kvůli výpočetní náročnosti zpracování kandidátů následuje prořezávání (prunning) argumentů, kde se vyřadí velmi nepravděpodobní kandidáti. Zvláště proto, že téměř většina nejsou argumenty slovesa a data by tak byla nevyrovnaná pro klasifikátor, který by měl velký podíl negativních vzorků (nejsou argumenty predikátu) a malý podíl pozitivních vzorků (jsou argumenty predikátů). Což může působit problémy pro správnou klasifikaci.

   Pro prořezávání se často používá jednoduchá heuristika od Xue, Palmer [32]. Algoritmus začíná tak, že predikát přidá všechny své syntaktické doplňky ze sourozeneckých uzlů (uzle se stejným rodičem) kromě uzlů s koordinační strukturou (spojky).

3. Ve fázi identifikování argumentů se provádí binární klasifikace každého kandidáta, jestli argument je nebo není. Jedná se o další postup jak zmenšit množinu kandidátů. Klasifikace se provádí většinou nezávisle na ostatních uzlech (tzv. lokální model). Pro učení se používají příznaky z věty, derivačního stromu a jiných zdrojů. Konkrétnějším způsobům výběru příznaků se věnuji v sekci 3.3.1.

4. Dalším krokem je klasifikace označených argumentů do 1-N tříd kandidátů. Tato klasifikace může být prováděna společně s identifikací argumentů nebo nezávisle. Většinou se však řeší v návaznosti na sebe, jak znázorňuje obrázek 3.4. Protože identifikace argumentů je úzce spojena ze syntaktickou vrstvou a klasifikace argumentů se sémantickou. Výběr příznaků pro tyto dvě úlohy se tak může značně lišit.

5. Poslední krokem je závěrečná strukturální úprava, kde se používá model společného ohodnocení (joint scoring). Na rozdíl od lokálního modelu se zde využívá hodnocení všech označených sémantických rolí ve vzájemném propojení celého stromu. Opravuje se tak případy chyb lokálního modelu. Například FrameNet a PropBank nedovoluje překrývání argumentů jednotlivých slov. V PropBank se nesmí argumenty opakovat.

   Tyto chyby se dají vyřešit metodou opětovného ohodnocení (re-ranking) [33]. SRL systém vrátí několik možných označení pro všechny slova. Metoda se pak snaží najít takovou množinu označení, která maximalizuje lokální a celkové skóre za podmínky splnění všech omezení. Dalším způsobem je využití pravděpodobnostního modelu pro vytvoření strukturovaného výstupu.

Extrakce příznaků

Vybrání správných příznaků je klíčovým pro dosáhnutí dobrých výsledků úloze SRL. Gildea a Jurafsky [27] předložili výčet použitelných příznaků, které se staly základem pro další práce v SRL.

- *Slovesný rod* - rozdíl mezi činným a trpným rodem má vliv na spojení sémantických rolí a jejich gramatických funkcí. Většinou předmět činných sloves odpovídá sémantickým rolím podmětu sloves v trpném rodě (věta 1 a 2). Slovesný rod se zjišťuje nalezením vzorů, kterými se tvoří. Takže v češtině pro trpný rod by se jednalo o nalezení slovesa *být* s příčestím trpným nebo slovesa se zvratným zájmenem.

- *Frázové typy* - jiná role je realizována jinou syntaktickou skupinou. Taková je myšlenka za využitím tohoto typu příznaku. Ve FrameNetu tvoří téměř polovina jmenné fráze (NP) a čtvrtinu předložkové fráze (PP), dalšími ve výčtu jsou adverbiální fráze (ADVP), částice (PRT) atd.

- *Cesta derivačním stromem* - příznak je určen zachytit syntaktický vztah mezi konstituentem a cílovým slovem (predikát). Příznak je reprezentován řetězcem symbolů se směrem pohybu v uzlech od cílového slova ke konstituentu. Například cesta ve stromu nahoru od frázového slovesa je reprezentována: VP↑.

- *Řídící kategorie* - existuje často spojení mezi sémantickou rolí a její syntaktickou realizací jako podmět nebo předmět. Může nabývat pouze hodnot S (podmět) a VP (předměty slovesa).

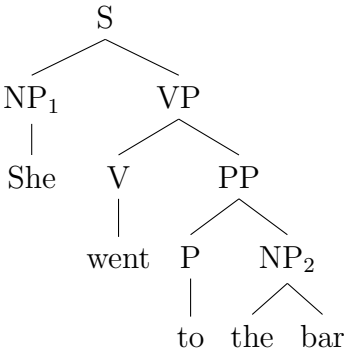
- *Podkategorie* - pro sloveso se využije seznam frázových typů, které jsou jeho dětmi v derivačním stromu. Intuicí za tímto příznakem je využití počtu argumentů slovesa pro omezení počtu možných sémantických rolí.

- *Pozice* - pro opravu chyb při parsování, kdy tento příznak může být využit i bez derivačního stromu. Příznak jednoduše říká, na které straně od predikátu se konstituent nachází. Očekává se, že bude značně korelovat s gramatickou funkcí, jelikož podmět se obvykle objevuje před slovesem a předmět za slovesem.

- *Hlava fráze* - je slovo, které ve frázi nerozvíjí žádný jiný člen fráze. Například v případě: „rychle běžící muž“ je hlavou fráze slovo *muž*. Hlava fráze podstatných jmen se může využít pro zúžení výběru role v sémantickém rámci. Řekněme v komunikačním rámci je pro slova *Petr* nebo *bratr* pravděpodobnější, že budou v roli SPEAKER (mluvčí) než slova *příběh* nebo *otázka*, které jsou nejspíše v roli TOPIC (téma).

Dalšími možnými příznaky mohou být pojmenované entity, shluk sloves, první/poslední slova konstituentu, pořadí/vzdálenost konstituentů, množina argumentů, předchozí role (poslední nalezený typ argumentu), slovní druhy (POS), n-gramy atd.

Zcela jiný přístup jak vybírat příznaky je použití metod s jádry (kernels) [19]. Tento způsob má objevit skryté struktury v syntaktické reprezentaci kandidátů. Nejvíce zmiňovaný postup se nazývá tree kernel (jádra stromu), který vytvoří podstromy. Podstromy z trénovací množiny pak porovnává se zpracovávaným stromem. Výhoda kernel přístupu spočívá v jednoduchém použití, kdy se získá velké množství příznaku bez manuálního výběru speciálních příznaků. Nevyžaduje přesnou shodu jednotlivých příznaků. Je vhodné pro klasifikátory pracující s jádry (SVM).



Obrázek 3.5: Syntaktický strom

Příznak	Anglicky	Česky
Cíl	went	šla
Hlava fráze	She	Ona
Řídíc kategorie	S	věta
Frázový typ	NP	jmenná fráze
Pozice	left	vlevo
Cesta	V↑ VP↑S↓NP_1	
Podkategorie	VP$\longrightarrow$V PP	
Lemma slovesa	go	jít
Slovesný rod	active	činný

Tabulka 3.3: Příklad příznaků pro slovo *She* (Ona)

Používané klasifikátory

Každý rok se vyhlašují úkoly z oblasti zpracování přirozeného jazyka. Nejlepší řešení těchto úloh se pak prezentují na konferenci učení přirozeného jazyka (CoNLL). V roce 2005 se zadal problém SRL [28]. Cílem bylo identifikovat ve větě argumenty slovesa a označit je sémantický rolemi. Oblíbenými metodami pro řešení byly MaxEnt (viz sekce 2.1.1) a SVM (2.1.1). 15 týmů zvolilo právě zmíněné algoritmy a pouze 6 vybralo jiný přístup.

Nejlepších výsledků dosáhl SNoW (Sparse Network of Winnows), učící se klasifikátor do více tříd. Je to systém, který je speciálně přizpůsoben pro rozsáhlé úlohy pro učení. Učící se architektura využívá síť lineárních funkcí nad předdefinovaným nebo postupně získaným prostorem příznaků. Učení probíhá buď pomocí Winnowova aktualizačního pravidla (Winnow update rule), perceptronu nebo klasifikátoru Naivního Bayese.[30] V předchozím roce vyhrála metoda SVM s polynomiálním jádrem druhého stupně.

V roce 2009 se vyhlásila úloha (CoNLL09), která si dávala za cíl vyřešit spojení syntaktické a sémantické informace pro více jazyků [29]. Základem bylo určit syntaktické a sémantické závislosti a jejich označení. Na předních pozicích v jednotlivých podúlohách se umisťoval čínský tým. Pro svůj systém použili SVM pro klasifikaci predikátů, MaxEnt pro SRL a pro závěrečnou globální optimalizace zvolili Integer Linear Programing (ILP). [31]

3.3.2 SRL kombinace učení s a bez učitele

Pro vysokou úspěšnost SRL u učení s učitelem je třeba rozsáhlý anotovaný korpus. Bohužel ten je velice drahé vytvořit. Navíc označené datasety nepokrývají celý jazyk. Jsou i případy, kdy pro některé jazyky takové zdroje ani nemusí existovat. [20]

Cílem je snížení závislosti na anotovaných datech. To lze vyřešit několika způsoby. Prvním je mezijazykový přenos modelu nebo anotací ze zdrojů bohatých na označená data (obvykle angličtina). Druhým je zvolit kombinaci mezi učení s učitelem a bez učitele (*semi-supervised learning*), kde k označeným datům přidáme neoznačená. Posledním způsobem je učení bez učitele, který by měl automaticky rozpoznat sémantické role z neoznačených dat.

Semi-supervised learning lze zhruba rozdělit do tří skupin. Metody s vytvářením náhradního učitele, kde automaticky anotujeme neoznačená data a využíváme je jako označená. Metody se sdílením parametrů. Neoznačená data se použijí k redukci řídké reprezentace slov. Ve třetí skupině metod se přidávají označená data k řízení modelu bez učitele.

Vytváření náhradního učitele

Pro vytvoření náhradního učitele se musí vybrat neoznačená data, která chceme označit. Vybírání neoznačených dat je založeno na následující hypotéze. Věty s neznámými lexikálními jednotkami budou mít podobnou syntaktickou a sémantickou strukturu jako označená data. Tato podobnost se vyjádří skórem zarovnání mezi označenými a neoznačenými daty. Skóre se skládá ze sémantického a syntaktického ohodnocení. Z neoznačených dat se vybere množina nejpodobnějších vzorků, provede se na nich anotace a přidají se do trénovací množiny. [34]

3.3.3 SRL učení bez učitele

SRL metody učení bez učitele se na rozdíl od učení s učitelem nemusí spoléhat na anotované datasety. Pro fázi automatického identifikování argumentů jsou už vymyšleny výkonné heuristiky (logistický klasifikátor s latentní proměnnou) [35]. Učení bez učitele se tak zaměřuje na problém klasifikace argumentů [36]. Dává si za cíl automaticky odvozovat sémantické role z neoznačených

dat. Obecně se jedná o problém shlukování. Musíme brát v potaz, že pro dané shluky neznáme označení.

Shlukování probíhá tak, že sdružujeme výskyty argumentů s jejich syntaktickým popisem nebo klíči argumentů. Například klíče argumentu z obrázku 3.3 pro slova *ona* jsou `činný:vlevo:podmět`. Shlukuje se podle těchto klíčů.

Lang a Lapata [37] navrhli aglomerativní shlukování argumentů, kde každý klíč argumentů začíná v svém shluku. Poté se postupně spojují a rozdělují ke zlepšení přesnosti reprezentace sémantických rolí. Algortimus nazvali Split-Merge. V každé spojovací fázi (merge) je každý pár ohodnocen, jak pravděpodobná je náležitost ke stejnému shluku. Ti s nejvyšším skorem jsou pak spojeni. Skóre zahrnuje lexikálně podobné argumenty, podobné slovní druhy argumentů a splňuje omezení, že všechny argumenty věty mají jinou sémantickou roli.

Jiný přístup k shlukování je zvolení generativního modelu pro odvozování argumentů. Titov a Klementiev [38] založili svůj algoritmus na Bayesovském modelu. Používají podobné učící signály (klíče argumentů) jako v Split-Merge shlukování. Jsou navrženy dva modely. V prvním se odvozují role pro každý predikát nezávisle na ostatních. V druhém případě se využívá verb alternations (viz 3.1), kde podobná slovesa sdílejí syntaktickou informaci. Vede se matice podobnosti pro páry učících se signálů. Na základě skóre z matice mezi páry jsou klíče argumentů shlukovány dohromady.

3.4 Shrnutí

Výkonnostně učení s učitelem stále převyšují modely učení bez učitele [39]. Nicméně na malých anotovaných datasetech je situace opačná. V reálném využití se musí brát ohled na to, že nejdražší operací je vytvoření závislostního stromu. Dalším náročným úkonem je morfologická analýza. O něco méně prostředků vyžaduje označení slovních druhů a nejlevnější je lemmatizace.

Existují přístupy, kdy se zcela syntaktická analýza vynechá za cenu ztráty přesnosti, ale udržení nízkých nákladů na použité výpočetní prostředky. V tomto směru učení bez učitele přináší slibné výsledky.

4 Analýza sentimentu

Analýza sentimentu, někdy označována jako dolování názoru (*opinion mining*), se zabývá analýzou lidských emocí, pocitů, postojů a hodnocení k určitým entitám. Identifikování a extrakce těchto subjektivních informací se nejčastěji provádí na psaném textu.

Rozvoj analýzy sentimentu po roce 2000 souvisí s růstem vlivu sociálních médií, který se snaží využít jak komerční sféra tak státní organizace. Firmy a organizace chtějí vědět názor zákazníka na jejich produkty nebo služby. Stát zajímá postoj na existující nebo navrhované zákony. Dalšími oblíbenými entitami pro dolování názorů jsou osobnosti, události, problémy nebo témata. [1, s. 29]

Analýza sentimentu je velice komplexní problém z několika důvodů. Lidé často nerozumí, jak je psaný příspěvek myšlen kvůli chybějícímu kontextu. Příspěvek navíc může obsahovat sarkasmus a tím význam posunout opačným směrem. Dalším častým jevem při zpracovávání textu je výskyt gramatických, syntaktických chyb. Fénomén sociálních sítí jsou emotikony a hashtagy, které leckdy nesou celý sentiment. Předzpracování musí být natolik robustní, aby se počítalo nejen s těmito případy.

Tato práce se pokouší o zlepšení výkonnosti analýzy sentimentu zapojením hlubšího porozumění kontextu skrytého uvnitř struktury jazyka. Pomocí využití sémantických rolí pro lepší pochopení významu věty. Nejdříve musíme definovat, co je sentiment.

Definice: **Sentiment** je základní pocit, postoj, hodnocení nebo emoce spojené s názorem. Je vyjádřen trojicí *(y,o,i)*, kde

- y je typ sentimentu
- o je orientace sentimentu
- i je intenzita sentimentu.

(Bing Liu, 2015)

Typ sentimentu se může dělit na několika kategorií. V spotřebitelském světě se dělí na racionální sentiment (z racionálního uvažování , hmatatelných

důkazů, bez emocí) a emocionální sentiment (nehmatatelný, emoční reakce na subjekty).

Někdy problém vyžaduje vyjádřit intenzitu sentimentu (sentiment intensity) na stupnici hodnocení. Tzv. hodnocení sentimentu (*sentiment rating*)[1] bývá nejčastěji na škále pěti bodů (1-5 hvězd), které lze interpretovat:

- velmi pozitivní (5 hvězd)
- pozitivní (4 hvězdy)
- neutrální (3 hvězdy)
- negativní (2 hvězdy)
- velmi negativní (1 hvězda)

Lze vybrat i stupnici s více body pro přesnější rozdělení intenzit sentimentu. Povaha přirozeného jazyka je velmi subjektivní a toto jemnější roztřídění se stává velice složitým problémem.

Posledním z trojice je orientace sentimentu. Může být pozitivní, negativní nebo neutrální. Orientace se taky někdy nazývá polarita, sémantická orientace nebo valence. [1, s. 66] Zmiňované vlastnosti sentimentu si ukážeme na názorném příkladě:

(4) „Tohle auto miluju.“
(5) „Tohle auto má vynikající jízdní vlastnosti.“

Obě věty 4 a 5 mají pozitivní polaritu a velmi pozitivní intenzitu sentimentu. Ukázka 4 má emocionální a věta 5 racionální typ sentimentu.

4.1 Úlohy analýzy sentimentu

Dělení analýzy sentimentu na různé podproblémy je motivováno především reálným využitím. Každá publikace má mírně odlišné dělení, přesto se většina shodne na dvou hlavních. Velký počet prací provádí klasifikaci, regresi nebo hodnocení sentimentu podle polarity (orientace) textu. Druhým častým problémem je klasifikace do dvou kategorii *subjektivní* a *objektivní*.

4.1.1 Polarita sentimentu

Za předpokladu, že text obsahuje souhrnný názor na jednu entitu, je možné klasifikovat tento názor do jedné ze dvou opačných polarit sentimentu. Obvykle se jedná o pozitivní nebo negativní kategorii. Binární klasifikace podle názoru vyjádřeného v textu se nazývá klasifikace polarity (*polarity classification*) [6].

Mnoho vědeckých prací před 15 lety (např. [8]) používalo 2 třídy kvůli zjednodušení problému klasifikace. Záměrně byla vynechána neutrální třída. Do této třetí třídy patří text, který neobsahuje žádný názor (např. objektivní fakt *"Včera jsem si koupil mobil."*) nebo se v něm vyskytuje smíšený či konfliktní sentiment (*"Film nebyl ani dobrý, ani špatný"*). Výzkumníci předpokládali, že vyřešením binárního problému pozitivní vs negativní, automaticky zjistí i třetí kategorii. Neutrální třída se bude nacházet blízko rozhodovací hranice binárního modelu. Nicméně Moshe Koppel a Jonathan Schler [7] ukázali na důležitost zahrnutí neutrální třídy při učení klasifikátoru. Přidání třetí třídy pomáhá lepšímu rozeznání mezi pozitivními a negativními dokumenty. Zvyšuje tak celkovou přesnost klasifikace a lepší identifikování tzv. neutrálních prvků jazyka.

Dosud je obecně zmiňována klasifikace sentimentu na částech textu. Analýza sentimentu může být prováděna na různých logických částí, ze kterých se text skládá. V literatuře a výzkumných pracích se nejčastěji dělí na 3 úrovně. [1, s. 38]

1. **Document level** - Na úrovni dokumentu je úkolem klasifikovat celkový názor dokumentu. Dokument se bere jako celek a nezabývá se analýzou entit nebo aspektů, jež obsahuje. Předpokládá, že dokument (recenze, komentář) vyjadřuje názor o jediné entitě (film, produkt) a obsahuje názor od jediného autora. Proto tento přístup není vhodný, pokud dokument obsahuje více názorů o různých objektech. Klasifikace dokument sentimentu (*document sentiment classification*) je bráno jako jednodušší úkol v analýze sentimentu, protože lze převést na tradiční problém klasifikace textu do kategorií, které odpovídají polaritám sentimentu. Nabízí se pro řešení využít metody učení s učitelem.
2. **Sentence level** - Na úrovni věty chceme u každé určit polaritu (pozitivní, neutrální nebo negativní). Tato úroveň je spjata s pojmem subjektivní klasifikace (*subjectivity classification*), kde rozlišujeme věty obsahující faktické informace (objektivní) od vět subjektivních, kde vyja-

dřujeme názor nebo postoj.

3. **Aspect level** - Předchozí úrovně nám neřeknou, co přesně lidé mají a nemají rádi. Úroveň aspektu se zaměřuje na ještě detailnější určení polarity jednotlivých názorů a hlavně jejich cílů.

   (6) „Já ten mobil zbožňuju, ale baterka vydrží jen jeden den.“

   Například věta 6 vyjadřuje pozitivní reakce na mobil, ale nemůžeme říct, že je úplně kladná. Pro jiného uživatele může být životnost baterky důležitá a mobil by si pravděpodobně nekoupil. V praxi jsou cíle názoru vyjádřeny entitami (např *mobil*) a/nebo jejich aspekty (*baterie mobilu*). Účelem této úrovně je nalezení názoru na entity a/nebo jejich aspektů (*baterie mobilu*).

4.2 Výběr příznaků

Jako příznaky pro tyto metody lze vybrat z několika způsobů [1, s. 201]:

- *N-gramy.* N-gram je posloupnost jednotlivých n slov. Příznakem je jedno slovo (unigram), dvě po sobě jdoucí slova (bigram) a tak dále. U nich se zpravidla počítá frekvence výskytu v textu. Tím zjistíme, jakou důležitost sehrávají pro jednotlivé třídy, do kterých je chceme klasifikovat. Často se vyřazují n-gramy s malým počtem výskytů, protože se může jednat o šum, které by výsledek klasifikace zhoršovaly.

  Jedná se o nejpoužívanější způsob vybrání příznaků v tradiční textové klasifikaci a v analýze sentimentu se ukázaly jako vysoce efektivní.

- *Slovní druhy* (Part of speech – POS). Slovní druh slova lze vzít jako příznak. Zvláště přídavná jména určují polaritu názoru nebo sentimentu a tak jsou brány jako speciální příznak. Nebo lze použít všechny tagy slovních druhů a jejich n-gramy jako příznaky. Všechny možné tagy slovních druhů jsou získávány z treebanky, která popisuje syntaktickou a sémantickou strukturu jazyka. Tyto tagy můžeme získat z textu některou z metod zpracování přirozeného jazyka (např. Viterbiho algoritmus).

- *Slovníkové metody.* Zejména citově zabarvená, pozitivní nebo negativní slova můžou být přirozeným příznakem pro určení sentimentu. Kromě

jednotlivých slov ze slovníku můžeme využít i fráze ovlivňující sentiment. Dalšími příznaky může být použití pravidel, která využívají jazykové konstrukce k vyjádření nebo naznačení sentimentu. K těmto všem příznakům se často ještě přidávají slova (*sentiment shifters*), která nenesou žádnou informaci o sentimentu, ale ve spojení s pozitivním nebo negativním slovem mohou význam velmi ovlivnit. Z této skupiny jsou nejdůležitější zejména slova, která obracejí smysl některého z předchozích slov.

- *Syntaktická závislost.* Využívá závislostní syntaxe slov jako příznaky, které jsou získávány z parsování nebo závislostního stromu.

4.3 Analýza sentimentu v češtině

Český jazyk patří do skupiny flektivních jazyků. To znamená, že vyjadřuje gramatické funkce ohýbáním (tj. skloňování, časování, předpony a přípony). Čeština navíc obsahuje celkem 42 písmen. Všechno toto přispívá k větší časové náročnosti oproti anglickému jazyku při zpracování textu.

Jelikož tato diplomová práce se zaměřuje na češtinu, uvedu hlavní rozdíly jazykových vlastností oproti angličtině. Z hlediska větné syntaxe má čeština volný slovosled. Anglický jazyk má tendenci k nominálnímu (jmennému) vyjadřování. Podstatná a přídavná jména zastupují velkou část informací ve větě. V českém jazyce je velká část informací vyjadřována slovesy [41].

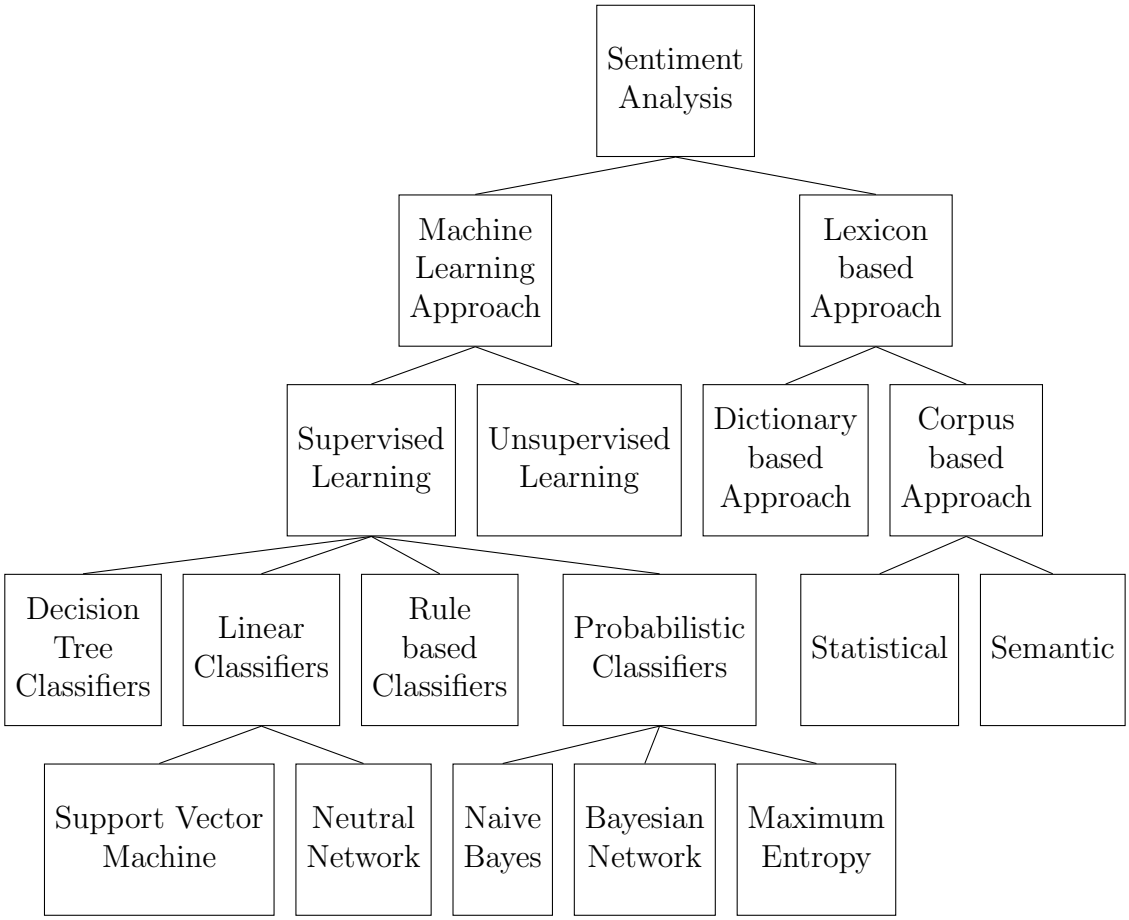
Dalším problémem českého jazyka je nedostatek označkovaných dat, které jsou potřeba pro metody strojového učení. Čím více vstupních dat, tím lepší přesnost a jednodušší ověření správnosti používaných technik.

4.4 Metody analýzy sentimentu

Klasifikace sentimentu rozdělujeme na dvě hlavní kategorie: slovníkový (lexicon-based) přístup a metody strojového učení (machine learning). Někdy se používá kombinace obou možností. Na obrázku 4.1 vidíme rozdělení nejvýznamnějších technik, které se v analýze sentimentu používají.

S rozvojem oboru strojového učení se věnuje stále větší pozornost neuro-

novým sítím a metodám učení bez učitele [23].



Obrázek 4.1: Přehled metod analýzy sentimentu [23]

4.5 Evaluace sentimentu

Jelikož je nezbytné ověřit výkonnost vybraných příznaků a metod a upravovat je náležitě podle jejich úspěšnosti, musíme tuto „úspěšnost“ nějak kvantifikovat.

Pokud probíhá klasifikace do více tříd hodnotíme systém nejdříve po jednotlivých třídách.. Při klasifikaci pro třídu A mohou nastat možnosti:

- *Skutečně pozitivní* (TP) - nález správné třídy A, systém označil A a ve

skutečnosti patří do A

- *Skutečně negativní* (TN) - správné odmítnutí zařazení do třídy A, systém neoznačil náležitost do třídy A a ve skutečnosti nepatří do A
- *Falešně pozitivní* (FP) - falešný poplach (chyba prvního typu), systém označil náležitost do třídy A, i když ve skutečnosti do třídy A nepatří
- *Falešně negativní* (FN) - chyba druhého typu, systém neoznačil náležitost do třídy A, i když měl

Zmíněné možnosti se pak použijí pro výpočet přesnosti, úplnosti a F-míry.

		Skutečnost	
		Pozitivní	Negativní
Odpověď systému	Označený	TP	FP
	Neoznačený	FN	TN

Přesnost

Úplnost

- *Precision* (přesnost) - přesnost určení třídy.
- *Recall* (úplnost) - schopnost nalezení třídy.
- F-measure (F-míra) - harmonický průměr Precision a Recall

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

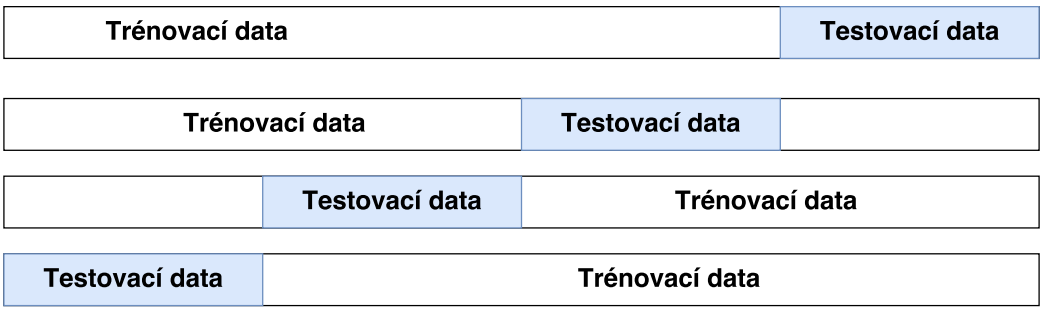
$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{4.3}$$

4.5.1 Křížová validace

Měření chyby nebo výkonnosti na jedné množině testovací dat není statisticky přesné. Chceme eliminovat případy, kdy algoritmus dosahuje zkreslených výsledků na trénovacích datech, a které nejsou reprezentativní.

Křížová validace má zjistit, jak moc použité metody ovlivňují nezávislé vzorky dat. Data jsou rozdělena na k disjunktních podmnožin stejné velikosti. Jedna podmnožina slouží jako trénovací a zbylé jako testovací. Výkonnost modelu vzniklého z trénovacích dat je měřena na testovací množině. Každá podmnožina je použita právě jednou jako testovací. Postup testování modelu se provádí k-krát, pokaždé na jiných vzorcích (viz obrázek 4.2).



Obrázek 4.2: Ukázka 4-násobné křížové validace

Odhad chyby vybraného modelu se zjistí jako průměr chyb pro jednotlivé testovací množiny. Výhodou této metody je využití všech označkovaných dat. Naproti tomu výpočetně náročné, protože požadovaný model je třeba vypočítat k-krát.

Z empirických zkušeností se nejčastěji používá 10-násobná křížová validace (10-fold cross validation) jako kompromis mezi velikostí trénovacích a testovacích dat (90% ku 10%). [26]

5 Vstupní data

Pro ověření užitečnosti nově vybraných příznaků bylo potřeba vybrané metody otestovat na datech. Pro zvolené učící se algoritmy s učitelem byla potřeba označkovaná data. Zadavatelem mi byl dodán dataset recenzí z filmového portálu CSFD, příspěvky ze sociální sítě Facebook a produktové recenze z internetového obchodu Mall.

Doménově se tak jednalo o rozmanitou skupinu dat, kde se dá předpokládat užívání jiných výrazových prostředků. Nejvíce odlišné jazykové prostředky měly vstupní data z Facebooku. Na sociálních sítí se často používá tzv. „nepřirozený“ jazyk. Je to jazyk využívající hovorové výrazy, gramatické a pravopisné chyby, smajlíky, novotvary nebo zkratky. [43]

Nepřirozený jazyk působí značné problémy pro parsery. Výsledkem je značná chybovost v označených slovech a nepřesné závislostní stromy.

Tabulka 5.1: Přehled velikostí datasetů pro analýzu sentimentu na úrovni dokumentu

Název datasetu	Počet dokumentů	Průměrná délka	Pozitivní	Negativní	Neutrální
CSFD filmové recenze	91.4k	51	30.9k	29.7k	30.8k
MALL produktové recenze	145.3k	19	103k	10.4k	31.9k
Facebook příspěvky	10k	11	2.6	2k	5.2k

5.1 CSFD.cz

Dataset CSFD pro klasifikaci sentimentu dokumentu obsahuje přes 90 tisíc filmových recenzí. Třídy pozitivní, neutrální a negativní jsou zastoupeny poměrně vyrovnaně. Ze všech zkoušených datasetů obsahují v průměru největší počet slov na dokument.

Problém určení sentimentu vzniká u příliš krátkých recenzí, kde chyběl dostatečný kontext. Je těžké určit sentiment u recenze *„Co dodat.“* bez dalších znalostí. V tomto případě se jednalo o pozitivní recenzi. Opačným extrémem jsou příliš dlouhé recenze, kde autor popisuje průběh filmu a vlastní názor na film je až na konci recenze.

(7) „Začínáme, jak jinak, než písní Imagine, kterou z videoklipu zpívá John Lennon. Po první sloce ho však střídá černá zpěvačka Yolanda Adamsová, kterou doprovází Billy Preston ... *(pokračuje v dlouhém popisu dokumentu a recenze končí)* ... Děkuji, ale nikdy víc. Jsem rád, že i to utrpení při psaní tohoto článku je konečně za mnou.“

Rozlišení vět, které souvisí s celkovým názorem je velmi obtížné a vyžaduje zapojení sofistikovanějších postupů.

Metody předzpracování pro tento dataset musí být navíc robustní, protože součástí byla i recenze s arabským textem prolínající české zhodnocení.

5.2 Mall.cz

Největší počet dokumentů obsahuje dataset internetového obchodu Mall.cz. Produktové recenze jsou děleny opět do tří kategorií: pozitivní (100k), neutrální (30k), negativní (10k). Recenze na rozdíl od CSFD dat obsahují více gramatických chyb, překlepů a vět bez sloves, což může způsobovat problém při sémantické analýze a automatickém hledaní lemmat.

Většina příspěvků je zastoupena pozitivní kategorii. Obsahují stručný popis produktu nejčastěji vyjádřeny slovy spokojenost, doporučuji, kvalitní, perfektní atd. Problémem je rozlišit mezi kategoriemi neutrální (věta 8) a pozitivní (věta 9). Kde sémantická informace nepomůže a jedná se o čistě subjektivní názor.

(8) „Spokojenost, jednoduchá montáž.“

(9) „S výrobkem jsem spokojen.“

5.3 Facebook.cz

Poslední nejmenší testovaný dataset má 10 tisíc příspěvků. Příspěvky mají nejmenší průměrný počet slov na dokument. Sentiment je často obsažen ve smajlících nebo znásobených samohláskách (věta 10).

(10) „ááá... uplně si říká o zulíbání ♥ :O :****“

Zároveň příspěvky většinou nedrží větné stavby, jejímž příkladem je věta: „*dobrou....mazlíčci:-))*“. Dále obsahují mnoho pravopisných chyb, překlepů a neobvyklých slovních tvarů (např ňunínek). Tento nepřirozený jazyk činí velké problémy při dělení věty na jednotlivá slova a samotném hledání sémantických rolí, kdy je obtížné najít predikát a jeho argumenty. To způsobuje velkou chybovost SRL. Takto chybně označené role zhoršují výsledek při analýze sentimentu.

Analýzu sentimentu na sociálních sítí navíc komplikuje jazykový prostředek sarkasmus. Ten se na rozdíl od produktových nebo filmových recenzí na sociálních síti vyskytuje daleko častěji. Ve větě, kdy říká někdo něco pozitivního, je ve skutečnosti negativní. (např. věta 11).

(11) „ jsou fakt výborný“

6 Implementace aplikace

Cílem diplomové práce je zjistit, jaký vliv má využití značkování sémantických rolí v analýze sentimentu na celkovou výkonnost klasifikace. Hypotéza byla taková, že využitím hluboké struktury věty (3.1) se výkonnost systému zlepší. Jde o to, vybrat takovou sadu příznaků, která bude natolik robustní, aby fungovala na všech typech dat.

Nejprve bylo potřeba ze vstupních dat získat sémantické role. Poté vybrat sadu příznaků, které nejvíce ovlivňují sentiment. Vyzkoušet na nejvhodnějších metodách strojového učení a zhodnotit výsledky.

6.1 Popis použitých knihoven

Jelikož je časově i implementačně náročné psát si vlastní optimalizované algoritmy strojového učení, hledala se vhodná knihovna obsahující metody strojového učení. Pro získání sémantických rolí byl zvolen framework *Treex*. Navíc použití volně dostupných knihoven značně urychlí vývojovou fázi. Pro samotnou analýzu sentimentu jsem využil knihovnu *Brainy*.

6.1.1 UDPipe

Zaměřením práce byla analýza sentimentu v češtině. Pro získání sémantických rolí bylo nutné vybrat systém, který umí vystihnout závislostní vztahy a sémantický význam věty pro český jazyk. Ze zvažovaných možností se nabízel parser *UDPipe*[1] , poskytující tokenizaci, lemmazitaci, značkování a závislostní analýzu. Závislosti jsou získany pouze na syntaktické úrovni v UDPipe, která by se dala srovnat s analytickou rovinou v PDT (sekce 3.2.5). Nezachycuje význam věty, který chceme získat.

[1]UDPipe, `http://ufal.mff.cuni.cz/udpipe`

6.1.2 Treex

Kritériím pro hledaný systém nejvíce vyhovoval vysoce modulární framework Treex, jehož implementace je provedena v jazyce Perl. Vznikl pro účely strojového překladu s využitím projektu PDT. Postupem času se stal základem pro vývoj ostatních problémů z oboru NLP. Framework obsahuje modul, jež provede analýzu na tektogramatické rovině. Získáme tím sémantickou strukturu ze vstupních dat.

Architektura Treexu se skládá z jednotlivých modulů nazývaných bloky. Blok je základní procesní jednotka. Odpovídá jednotlivým podúlohám NLP jako jsou tokenizace, lemmatizace, tagování, analýza na všech rovinách PDT a tak dále. Sekvence bloků tvoří scénář. Vstup je zpracován podle scénáře, kde jsou postupně použity jednotlivé bloky za sebou. Výchozí výstup je XML soubor ve formátu frameworku. Lze však zvolit jiné zapisovací bloky a zahrnout je do scénáře.

Framework má velký počet závislostí a poměrně komplikovaný postup instalace. Proto jsem využil technologii Docker, která řeší zmíněné problémy poskytnutím abstrakce rozhraní operačního systému. Aplikace s jejími závislostmi je zabalená do kontejneru, který pak může běžet v Dockeru. Dostupný balíček pro Treex se nachází na `https://hub.docker.com/r/ufal/treex/`.

Pro získání hloubkové struktury jsem zvolil scénář využívající bloky specializované na český jazyk. Scénář vypadá následovně:

1. Nastavení globální proměnné na češtinu pro ostatní bloky. Načtení vstupu a segmentace na věty podle regulárních pravidel, kdy konec věty je rozpoznám pokud končí tečku, otazník nebo vykřičník a je následován velkým písmenem.

2. Vytvoření morfologické vrstvy - rozdělení každé věty na list tokenů opět podle sady regulárních pravidel. Obecná pravidla pro všechny jako je odstranění přebytečných mezer, ponechání url adres, rozdělení znásobených znaků (např. vykřičníků) a tak dále. Pro češtinu se specificky dělí přípona *li* a rozsah čísel s pomlčkou.

   Poté jsou jednotlivé tokeny označeny tagem blokem využívající nástroj MorphoDiTa[2] (morfologický slovník a značkovač). Do MorphoDiTa je načten model pro český jazyk. Tagy jsou doplněny do uzlů na analytické

[2] `http://ufal.mff.cuni.cz/morphodita`

vrstvě. Následuje oprava nejčastějších chyb u morfologických tagů. Pro češtinu se opravují různé zkratky, rozdělující čárka pro číslice, nerozpoznaná lemmata a tak dále.

3. Vytvoření analytické vrstvy - probíhá pomocí závislostního parseru. Parser je založen na hledání maximální kostry grafu. Poté se na analytickém stromě opraví chyby, které vznikli při parsování (oprava pádů podle předložek, uzlů s více než dvěma dětmi atd.)

4. Vytvoření tektogramatické vrstvy - označí se hrany z analytické vrstvy pro spojení (převážně pomocná slovesa, předložky a podřadící spojky). Vytvoří základ tektogramatického stromu, který se pak postupně doplňuje o jednotlivé atributy.

   Nastaví se řídící člen stromu. Nejdůležitější atribut funktor se stanoví ve dvou fázích. V první fázi se určuje několikanásobný větný člen a shodné přívlastky. Poté se provádějí opravy struktury, proběhne druhá fáze určování funktorů. Dále se doplní zbytek atributů (typ uzlu, všechny typy gramatémů) a nakonec se strom zapíše do souboru.

Výsledkem je XML soubor s analytickým a tektogramatickým stromem pro věty ze vstupních dat.

6.1.3 Brainy

Pro analýzu sentimentu byla potřeba sada klasifikačních algoritmů. Zvolil jsem knihovnu strojového učení Brainy. Knihovna je napsaná v jazyce Java a byla vydaná pod GPL licencí.

Důvodem pro zvolení Brainy byl její důraz na propracovaný systém správy příznaků. Lze v ní snadno definovat a měnit sadu příznaků používaných pro metody strojového učení. Navíc jsem se s knihovnou seznámil během studia a to mi umožnilo lépe využít možnosti tohoto nástroje.

Brainy je rozdělen do tří hlavních komponent: [42]

- Správa příznaků - definuje rozhraní pro samotné příznaky i jejich sadu. Rozhraní `Feature` reprezentuje šablonu příznaku a převádí uživatelem definované objekty do vektoru. Třída `FeatureSet` sloučí všechny vektory příznaků do jednoho a vytvoří matici představující data z vektoru příznaků.

- Matematický aparát - skládá se ze dvou částí. První část poskytuje implementace algoritmů lineární algebry. Druhá část obsahuje optimalizační algoritmy.
- Algoritmy strojového učení - nejdůležitější část zahrnuje rozhraní pro řešení všech důležitých problémů, jako je klasifikace, regrese, shlukování a sekvenční značkování.

6.2 Struktura aplikace

Aplikace je rozdělena podle balíčků na jednotlivé logické celky.

Všechny balíčky jsou součástí adresáře `cz.zcu.fav.nlp.simunek.semal`

- `features` – balík obsahuje typy příznaků, které se získávají z dokumentů
- `model` – balík obsahuje metody pro klasifikaci sentimentu. Každá metoda implementuje rozhraní `SentimentAnalyzer`, která obsahuje metody *train* pro fázi trénování a metodu *analyze* pro fázi klasifikace na testovacích datech.
- `data` – balík reprezentuje strukturu dat používaných v aplikaci jako jsou dokumenty, sémantické role atd.
- `preprocessing` – balík pro předzpracování vstupních dat
- `util` – pomocný balík pro načítání dat a udržování slovníku

6.2.1 Načtení Treex dat

Načítání dat získaných z Treex provádí třída `TLayerFileHandler`. Kvůli velikosti XML souboru získaný z Treex byl vybrán SAX parser. Ukládány jsou pouze atributy tektogramatického a analytického stromu. Nejsou ukládány všechny uzly pro ušetření paměti. Po přečtení celého souboru jsou jednotlivé uzly stromů ještě filtrovány. Zejména jsou zahozeny uzly, které nemají závislost s predikátem nebo jiným uzlem potencionálně nesoucí sentiment (podstatné jméno, přídavné jméno, příslovce).

6.3 Předzpracování

Předzpracování dokumentů má výrazný vliv na analýzu sentimentu. Je proto nutné zvolit takovou variantu, aby distribuční prostor nebyl příliš veliký a slova v jiném morfologickém tvaru byla stejná a zároveň jsme neztratili příliš mnoho informace o sentimentu. Tím se sníží výpočetní náročnost. Také se redukuje šum a odstraní se nevýznamný obsah pro výsledek klasifikace.

Používané metody pro předzpracování jsou:

- *Tokenizace*. Proces rozdělení vět na menší části tzv. tokeny. Token je slovníkový tvar slova.
- *Odstranění stopslov*. Odstraníme slova, která nenesou samy o sobě žádný význam. V českém jazyku jsou to většinou předložky a spojky.
- *Case folding*. Slova je potřeba normalizovat, aby stejné slovo obsahující rozdílná velká a malá písmena byla prezentována jako totožná. Řešením je převedení všechny písmena do lower-case nebo upper-case podoby.
- *Stemming*. Účelem stemmingu je převedení slov na jejich kmen. Odstraňují se předpodny a přípony. Slova v různých morfologických tvarech tak budou reprezentovány stejně.

Po takovém předzpracování musíme brát v potaz, že jsme ztratili určitou část informací o sentimentu za cenu snížení velikosti distribučního prostoru. Tyto informace by měly být zachovány v příznacích sémantických rolí.

Pro předzpracování n-gramů byly zvoleny dva způsoby. Lišily se ve způsobu tokenizování a stemmování. Pro první způsob byl zvolen český analyzátor (`CzechAnalyzer`)[3] z knihovny Apache Lucene . Jeho postup pro zpracování věty je následující. V první fázi proběhne standardní tokenizace (podle Unicode Text Segmentation Algorithm). Poté převedení na malá písmena, odstranění českých stopslov. Na závěr se provede stemmatizace český stemmerem.

Druhý způsob používá tokenizátor Ark-tweet-nlp [40], který je určený pro tokenizaci krátkých zpráv ze sociálních sítí. Výrazně tak zvýšil úspěš-

[3] `https://mvnrepository.com/artifact/org.apache.lucene/lucene-analyzers-common`

nost u pro dataset z Facebooku. Pro stemmatizaci High Stemmer Precision (HPS)[49]. Oba nástroje výrazně zvyšují úspěšnost u datasetu Facebook.

6.4 Extrakce příznaků

Výběr příznaků je klíčový pro úspěšné klasifikování sentimentu. Tato důležitá část práce představí množinu možných příznaků, které lze získat. Základní přehled už byl nastíněn v sekci 4.2. Práce se však zaměřuje na využití sémantické informace obsažené v textu pro zlepšení analýzy sentimentu.

Nejprve je třeba si stanovit základní sadu příznaků, se kterou pak půjdou srovnávat vybrané „sémantické“ příznaky.

6.4.1 N-gramy

Jak už bylo zmíněno n-gramy mají velký význam v analýze sentimentu. Jedná se o nejjednodušší a poměrně účinný příznak ve zpracování přirozeného jazyka. Proto poskytuje výchozí metriku pro srovnání.

6.4.2 Slovní druhy

Slovní druh (Part of speech - POS) pomáhá rozlišit mnohoznačný význam slova. Například pro slovo září ve větách 12 a 13 se zcela liší polarita věty.

(12) „Herec ve filmu září.“
(13) „V září vychází film.“

Slovní druh tak přidá další informaci k unigram příznaku . Dále se slovním druhem se používá tag (sekce 3.2.5). Ten však obsahuje až moc informací, které neovlivňují sentiment. Výjimkou jsou kategorie zápor a stupeň, které se podílejí na sentimentu a ztratí se při stemmatizaci nebo lemmatizaci. Přitom můžou polaritu sentimentu úplně změnit nebo zesílit intenzitu.

6.4.3 Character n-gramy

Character (znakové) n-gram je posloupnost o n znacích za sebou. Tento typ příznaku by měl zachytit sentiment ze smajlíků, slov se zvýšeným počtem samohlásek, opakující se interpunkce či jiná forma posloupnosti znaků vyjadřující emoce. Velmi vhodný pro „nepřirozený“ jazyk používaný na sociální síti. Krásnou ukázkou je věta 14, která činí parserům potíže. Pro unigram slov by zachytili pouze slova super a jůůůů, přičemž jů s jiným počtem samohlásek by nevyhodnotilo jako stejné slovo.

(14) „ jůůůůů....to je super...!!! :)“

Délka character n-gramů byla zvolena mezi 3 až 6, tak aby obsahovala smajlíky a interpunkce a zároveň aby nebyl prostor příznaků příliš velký.

6.4.4 Sémantické příznaky

Zpracovaná vstupní data knihovnou Treex byla převedena do analytického a tektogramatického stromu. Stromy reprezentují syntaktickou a sémantickou strukturu dat. Úkolem bylo vybrat takové atributy z obou stromů, které by pomohly při analýze sentimentu.

Bylo by časově náročné vyzkoušet všechny kombinace příznaků. Proto jsem čerpal z dokumentace PDT [45], kde jsem vybíral vzory, které by mohly nést informaci o sentimentu. Dalším zdrojem byla disertační práce Veselovské [44], která rozebírá lingvistickou strukturu sentimentu v českém jazyce.

Tektogramatické lemma

Nejprve se zaměříme na expresivitu slova na lexikální úrovni. Slovo dokáže vyjádřit sentiment samo o sobě. Zíma [46] rozdělil typy tohoto vyjádření do tří kategorii expresivity: inheretní expresivní (rozlišitelný bez větné souvislosti) , adherentní (vznikající významovou změnou), kontextový (stylisticky odlišné slovo v porovnání s kontextem). Veselovská stanovuje podobné kategorie pro emocionalitu, která je podmnožinou expresivity zaměřené na vyjádření sentimentu.

Inherentně emocionální slova jsou slova rozlišitelná bez větné souvislosti. Vyjadřují větší intenzitu sentimentu než slovo, od kterého byly odvozeny. Nemusí být jenom odvozeny. Často vynikají svou hláskovou podobou (šťabajzna) nebo typem slovního tvaru. Příkladem jsou přípony -ák (pašák), -idlo (zlobidlo), -isko (psisko), -izna (barabizna) atd.

Patří sem citově zabarvená slova, zdrobněliny, vulgarismy. Nelze říci, že převládá jedna z kategorie pozitivní nebo negativní. A jejich automatická detekce je velmi obtížná. Morfologická vrstva PDT neposkytuje zachycení informace o expresivitě. A lze jen těžko rozlišit sentiment mezi *stará barabizna* a*starý dům*. Spoléhat se tak musíme na n-gram model.

Pokud v předzpracování zvolíme stemmatizaci slov, připravíme se o sentiment, které slovo nese. Nicméně se musí zohlednit velikost dat pro učení. A při nedostatečném počtu unikátních slov jsou data velmi řídká a nenacházíme při klasifikování ukázky z trénování. Inherentně emocionální slova navíc nejsou příliš častá. Pro n-gram příznaky jsem použil stemmatizaci. Pro sémantické příznaky jsou použita lemmata.

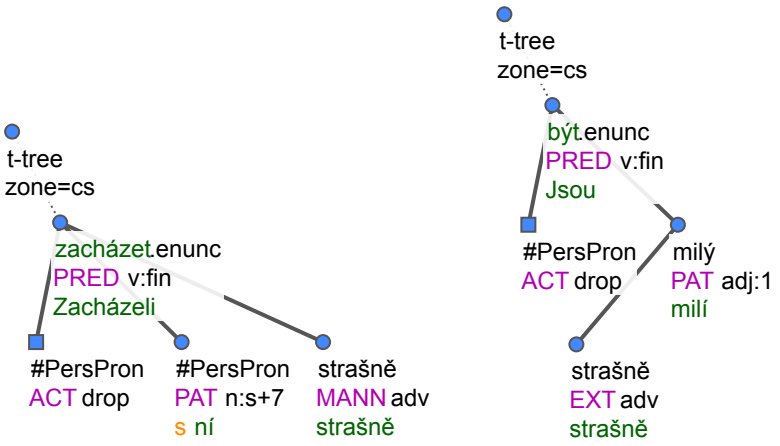
Funktor

Sémantické ohodnocení vztahu slov neboli funktor je jeden ze základních atributů, které by měly přispět ke zlepšení analýzy sentimentu. Jeden z případů, kde dokáže rozlišit význam slov je v adherentní emocionalitě. Adherentní emocionalita dává známým slovům jiný význam. Například užití slova *robota* ve smyslu těžké práce. Patří sem i slova s dvojznačnou polaritou (nejčastěji příslovce). Buď doplňují děj nebo zesilují intenzitu slova, na které jsou navázané.

Uvedeme si příklad, kde značkování sémantických rolí pomůže rozpoznat sémantických význam slova *strašně*. Ve větách 6.1 můžeme vidět správné identifikování sémantické role na tektogramatické vrstvě. Ve větě 6.1a) je přiřazen funktor MANN [4] a závislost na slovese. Funktor MANN vyjadřuje doplnění děje. Ve druhé větě 6.1b) má slovo strašně určen funktor EXX [5] (uvedení míry či intenzity). Příznak se bude skládat z funktoru, samotného slova a jeho přímé závislosti.

[4] „MANN (manner) je funktor, který vyjadřuje způsob uvedení kvalifikační či hodnotící charakteristiky děje, případně vlastnosti vyjádřené řídícím slovem.“ [45]

[5] „EXT (extent) je funktor, který vyjadřuje způsob uvedení míry či intenzity děje, vlastnosti či nějaké vnější okolnosti vyjádřené řídícím slovem.“ [45]



a) „Zacházeli s ní strašně.“ b) „Jsou strašně milí.“

Obrázek 6.1: Tektogramatické stromy pro ukázky

Příznak nemusí úplně zachycovat všechny případy dvojznačné emocionality. Objevil jsem případy, kdy Treex špatně vyhodnotil funktor i slovo, na kterém mělo být závislé.

Důležité je zachytit v jakém vztahu jsou spolu věty v souvětí. Pro souvětí odporovací, ve kterém obsah věty první je popřen nebo v nesouladu s větou druhou. Tento druh souvětí může otočit polaritu sentimentu. Pro odporovací souvětí je definován funktor ADVS. Nejčastějšími zástupci jsou spojky *ale*, *jenž*, *avšak* atd. V tektogramatickém vrstvě je funktor kořen stromu. V práci přidávám každý kořen tektogramatického stromu, kterým je nejčastěji predikát, ale v případě souvětí i spojka.

Grammatémy

Gramatémy z velké části odpovídají významu lexikálních a morfologických kategorií (podobné jako POS tag) uzpůsobeny na tektogramatickou rovinu. Jako slibný příznak ze všech gramatémů se jevil *deontmod*. Určuje deodentickou modalitu slovesa. Ta popisuje, zda je děj nutný, povinný, záměrný, možný apod. Jenže empirické výsledky potvrdili jako u POS tagu, že jejich využití nemá na výsledky klasifikace téměř žádný vliv.

Podobný atribut gramatémům *sentmod* zachycuje způsob slovesa (oznamovací, rozkazovací, přací atd.) také nepřinesl zlepšení.

Predikát

Většinu významu věty nese predikát s jeho argumenty. Je to tak nejdůležitjší příznak ze SRL. Predikát lze rozdělit do dvou kategorií. V první kategorii nese sentiment už v sobě a není třeba zjišťovat argumenty (žvanit, amputovat). Druhá kategorie sloves přenáší sentiment do svých argumentů [44].

Pro první případ nám bude stačit pro identifikování sentimentu n-gram model. Pro druhou skupinu je potřeba určit cíl sentimentu pro pochopení významu.

Na příkladu se zcela mění sentiment pro slovo *film*. Ve větě je cílem pozitivního názoru.

Zdroj			*Cíl*
Petr	miluje	akční	film.
ACT	PRED	RSTR	PAT

Pokud se slovo *film* a *milovat* vyskytuje v jiném vztahu, už o něm nemůžeme říci, že je názor pozitivní. To dosvědčuje ukázka, kdy změna cíle ovlivní sentiment. Pro pochopení sentimentu je důležité si ukládat, jaký predikát se objevil. A z toho odvozovat, jakou roli mají argumenty.

Zdroj			*Cíl*
Film	miluje	laciné	efekty.
ACT	PRED	RSTR	PAT

Sloveso *milovat* má vzor, kde ACT (agent) je zdroj sentimentu a PAT (patient) je cíl sentimentu. Při dostatečném počtu dat se algoritmus naučí tyto vzory identifikovat. Identifikace těchto vzorů je klíčová při aspektové analýze sentimentu. Pomáhá i na úrovni dokumentu.

Při extrakci příznaků je vybírán vždy predikát s jeho argumenty. Přidány jsou i jejich vzájemné závislosti.

Vybraná množina příznaků SRL

Ze zmíněného výčtu sémantických příznaků jsem vybral takovou množinu, která měla nejlepší úspěšnost na vstupních datasetech.

Byl vybrán predikát a jeho přímé děti z tektogramatického stromu (argumenty). Z analytického stromu se k tektogramatickým uzlům zjistil slovní druh a byly přidány všechny podstatná jména, přídavná jména, slovesa a příslovce. Z těchto vybraných uzlů je používán jejich funktor a lemma. Do množiny byly ještě zařazeny rodiče těchto uzlů.

6.4.5 Rozhraní

Každá třída, která reprezentovala typ příznaků implementovala rozhraní `Feature`.

```
public interface Feature<T> extends Serializable {
    int getNumberOfFeatures();
    void extractFeature(ListIterator<T> var1,
           FeatureVectorGenerator var2);
    void train(InstanceList<T> var1);
}
```

Rozhraní obsahuje tři metody. První vrací počet příznaků. Druhá metoda získává příznaky z testovacích dat a generuje vektor příznaků. Prvky vektoru nabývaly binárních hodnot 1 pro přítomnost příznaku a 0 pro absenci příznaků. Třetí metoda `train` z trénovacích dat prováděla extrakci příznaků a ukládala do hashmapy.

6.4.6 Klasifikátor

Z knihovny Brainy byly vyzkoušeny 3 metody učení. Třída `MaxEntTrainer` reprezentuje klasifikátor maximální entropie. Pro odhad parametrů pro modelu používá metodu s omezenou pamětí L-BFGS [**?**]

SVM implementuje třída `SVMTrainer`, která hledá parametry pro optimální rozdělující nadrovinu pomocí adaptivního gradientního sestupu. Bylo použito SVM s lineárním jádrem. Regularizační faktor C je nastaven na 1.

Naivní Bayesův klasifikátor představuje třída `NaiveBayesGaussianTrainer`.

Příznaky pro všechny klasifikátory jsou binární. Příznaková funkce $f_i(d, c)$

$$f_i(d, c) = \begin{cases} 1 & \text{pro výskyt slova} \\ 0 & \text{jinak} \end{cases}$$

6.4.7 Serializace objektů

Aby se nemusely příznaky z tektogramatického a analytického stromu při každém spuštění znova parsovat, využil jsem serializaci objektů pro zrychlení vývoje. Po zpracování Treex souboru `TLayerFileHandler` se vrátí seznam dokumentů, u kterých jsou uloženy vybrané sémantické příznaky (viz sekce 6.4.4). Dokumenty pak lze serializovat a deserializovat prostřednictvím pomocné třídy `FileLoader`.

K samotné serializaci je využita knihovna *fast-serialization* (FST)[6], která je několikanásobně rychlejší než nativní Java rozhraní. Výsledný serializovaný objekt má menší velikost.

[6] `http://ruedigermoeller.github.io/fast-serialization`

7 Výsledky

Účinnost sémantických rolí pro analýzu sentiment byla testována na dokumentové úrovni. Všechny pokusy byly provedeny po 10-vzorkové křížové validaci. Výsledek je měřen F-mírou (vzorec 4.3) pro lepší porovnání s ostatními datasety. Pro vstupní data byly naměřeny výsledky klasifikace do různého počtu tříd - klasifikace do 3 kategorií (pozitivní, negativní a neutrální) a klasifikace do 2 kategorií (pozitivní a negativní).

Z vyzkoušených algoritmů dosahoval nejlepší výkonnosti klasifikátor maximální entropie. Metoda SVM byla u CSFD a Mallu horší ve všech zkoušených sadách příznaků. Navíc vzhledem k větší velikosti dat oproti Facebooku trval trénovací proces téměř 10x déle než u MaxEnt.

Doba běhu jedné validace se u MaxEnt pohybovala od jedné minuty (Facebook) až do 5 minut (CSFD). Čas značně ovlivňovalo aktuální zatížení RAM. Pro CSFD je vyžadováno spustit aplikaci minimálně s 6GB z důvodu zpracování SRL. Program byl testován na počítači MacBook Pro 2.7 GHz Intel Core i5, 8GB RAM.

Výsledky byly pro CSFD a Mall prováděny na n-gramech z balíčku *luceneNgrams* a třídě *SemanticFeature* s minimálním počtem výskytů pět. Character n-gramy byly s minimálním počtem výskytu 150. Pro Facebook byly použity n-gramy v balíčku *features* a Character n-gramy s práhem pěti výskytů.

7.1 Klasifikace do tří kategorií

Výsledky jsem naměřil pro různé množiny příznaků pro zjištění vlivu na konečný výsledek. Pro porovnání v tabulce 7.1 uvádím i nejlepší výsledky z práce [47].

Množina příznaků složená pouze z unigramů stanovuje základní F-míru pro porovnání. Už u modelu unigramů můžeme vidět velké rozdíly v použitém způsobu předzpracování (sekce 6.3). Pro předzpracování s CzechAnalyzatorem dosahoval pro unigramy u Facebooku o 4% a u CSFD o 3% horší F-míru než u HPS.

Tabulka 7.1: F-míra pro analýzu sentimentu na dokumentové úrovně pro 3 kategorie v %

Množina příznaků	Facebook		CSFD	Mall
	MaxEnt	SVM	MaxEnt	MaxEnt
Unigram	66	67.1	75.1	67.6
Unigram + bigram	65.5	65.5	76.5	73.1
Uni + bi + trigram	65.8	66.1	77.1	74.2
Uni + character-gram	70.5	**69**	77.1	73.3
Uni + SRL	66.1	66.3	**78.6**	75.1
Uni + bi + SRL	66.8	65.2	78.6	75.4
Uni + bi + tri + SRL	67.0	65.4	78.2	**76.6**
Uni + char + SRL	**71.3**	69	78.1	75.3
nejlepší supervised [47]	69.4	68	78.5	75.4
konfidenční interval	±1.0		±0.3	±0.2

Přestože byl zvolen horší způsob předzpracování, byly pro všechny vstupní datasety překonány nejlepší výsledky z [47]. Nicméně jsou blízko hranici konfidenčního intervalu.

Pro Facebook množina příznaků SRL pomohla ke zlepšení jen mírně. Je to dáno velikostí dat, kde se nenalezly potřebné sémantické vzory. Převážně chyba spočívá v nepřesném zpracování a označení rolí framework Treex, který není uzpůsoben pro jazyk sociálních sítí. Navíc Treex obsahuje anotovaná data převážně z novinových článků, což přispělo k chybovosti. Zlepšení úspěšnosti klasifikace dat z Facebooku výrazně pomohly character n-gramy. Je to dáno tím, že sentiment je v příspěvku vyjádřen smajlíky, interpunkcí nebo tyto znakové n-gramy zachytí překlepy a časté zkratky.

Musíme brát v potaz, že horní hranice F-míry není 100%. V porovnání s lidmi, kteří dosáhli F-míry 86% a 92% v[47], nedosahuje strojová analýza sentimentu špatných výsledků.

Množina příznaků SRL přispěla k lepšímu výsledku u CSFD recenzí. Z testovaných datasetů měl slovník největší počet slov a tak se nalezené „sémantické vzory“ objevily v trénovacích i testovacích datech. Zlepšení není tak znatelné.

Největší zlepšení oproti unigramům dosáhly produktové recenze z e-shopu Mall. Zlepšení je výrazné i v porovnání s ostatními příznaky. Pro hlubší

Tabulka 7.2: Počty zařazených recenzí do tříd systémem vůči skutečné třídě pro dataset Mall. Na diagonále jsou správně zařazené recenze

		Skutečnost		
		negativní	neutrální	pozitivní
Odpověď systému	negativní	6972	1571	804
	neutrální	1968	19569	6804
	pozitivní	1451	10814	87392

analýzu výsledku si zobrazíme tzv. confusion matrix, která obsahuje odpovědi systému korespondující se skutečností. Zjistíme, že je velký počet špatně klasifikovaných i za použití příznaků sémantických rolí mezi třídami pozitivní a neutrální. Jak už jsem zmiňoval v sekci 5.2, recenze obsahovaly velmi podobné slovní spojení a takřka stejný slovník, na kterém selhává n-gram model. Velmi jemné rozdíly zachytily sémantické role, a proto přinesly u unigramů téměř 10% zlepšení a u kombinace unigram, bigram, trigram 2% zdokonalení.

7.2 Klasifikace do dvou kategorií

Při klasifikaci do tříd pozitivní a negativní se pro klasifikátor stírá vliv použití sémantických rolí kromě recenzí datasetu Mall. Je to způsobeno, tím že v recenzích nebo příspěvcích jsou znatelné rozdíly v použitých výrazových prostředcích a nepotřebujeme další informaci od sémantických příznaků pro určení správné třídy.

Tabulka 7.3: F-míra pro analýzu sentimentu na dokumentové úrovně pro 2 kategorie v %

Množina příznaků	Facebook	CSFD	Mall
unigram	85.5	89.3	86.3
uni + SRL	86.6	89.6	89.3
konfidenční interval	±0.9	±0.2	±0.2

7.3 Zhodnocení výsledků

Sémantické role zlepšily úspěšnost analýzy sentimentu na všech typech datasetů. Nicméně nešlo o výrazné zlepšení kromě produktových recenzí.

U dat z Facebooku se projevila velká chybovost systému Treex při vytváření sémantické vrstvy. Treex navíc nebyl stavěn na nepřirozený jazyk (sekce 5) využívaný na sociálních sítích. Bylo by nutné si vytvořit vlastní bloky na tokenizaci a parsování. Dále se jednalo o malý dataset, kde se systém nemohl naučit sémantické struktury pro neznámé vzorky.

U datasetu z CSFD vidím problém v dlouhých recenzí, kde sémantické role nebraly v potaz globální kontext dokumentu a největší celek, co dokázaly pojmout, bylo souvětí. Recenze většinou popisovala nějakou část filmu (její entity) a celkový názor o filmu byl vyjádřen hned na začátku nebo na konci recenze.

Dataset Mall měl u svých recenzí průměrnou délku jedné věty. Sémantické role tak jsou pro nejvhodnější právě pro tento typ dat. Předpokládám, že sémantické role by přinesly výrazné zlepšení i u sentimentu aspektu, když ve své podstatě drží informaci o zdroji a cíli věty.

Další výzkum by se mohl věnovat právě oblasti analýzy sentimentu aspektu.

8 Závěr

Cílem práce bylo prozkoumat druhy sémantických rolí a metody jejich automatického získávání. Dále vybrat nejvhodnější systém pro jejich značkování se zaměřením na český jazyk. Pro zlepšení úspěšnosti analýzy sentimentu jsem měl využít sémantické role. Důležitou částí práce bylo najít takové sémantické role, které zachycují sentiment. Poté jsem zvolil metody strojového učení a implementoval vlastní rozšíření využívající sémantické role pro analýzu sentimentu. Posledním krokem bylo zhodnotit vliv sémantických rolí při analýze a porovnat s ostatními používanými příznaky.

Teoretická část popisuje základní principy strojového učení. Podrobně rozebírá sémantické role v počítačové lingvistice a způsoby jejich značkování. Po vysvětlení způsobů automatického značkování sémantických rolí následuje seznámení s problematikou analýzy sentimentu.

V realizační části byl pro získání sémantických rolí vybrán nástroj Treex založený na Pražském závislostním korpusu. Po vytvoření sémantických rolí pro vstupní data proběhla extrakce příznaků. Pro analýzu sentimentu na úrovni dokumentu byly použity metody učení s učitelem.

Nejlepších výsledků dosahoval klasifikátor maximální entropie. Naměřené hodnoty překonaly dosavadní výsledky výzkumu metod učení s učitelem [47]. Nejvýraznějšího zlepšení úspěšnosti sentimentu jsem zaznamenal na produktových recenzí internetového obchodu Mall.

Časově nejnáročnější částí práce bylo získání sémantických rolí a následné vybrání vhodných příznaků pro analýzu sentimentu. Předmětem dalšího pokračování výzkumu se nabízí využití metod učení bez učitele, který by našly v sémantických rolí skryté struktury.

Literatura

[1] BING, Liu. *Sentiment Analysis and Opinion Mining.* 1st Edition. Cambridge University Press. 2015. ISBN-10: 1107017890

[2] NIGAM, Kamal; LAFFERTY, John; MCCALLUM, Andrew. *Using maximum entropy for text classification.* In: IJCAI-99 workshop on machine learning for information filtering. 1999. p. 61-67.

[3] ZHOU X., ZHANG X., WANG B. (2016) Online Support Vector Machine: A Survey. In: Kim J., Geem Z. (eds) Harmony Search Algorithm. Advances in Intelligent Systems and Computing, vol 382. Springer, Berlin, Heidelberg

[4] BRYCHCIN, Tomáš; HABERNAL, Ivan. Unsupervised improving of sentiment analysis using global target context. In: Proceedings of the international conference recent advances in natural language processing RANLP. 2013. p. 122-128.

[5] HAVRÁNEK, B. *Slovník spisovného jazyka českého* [online], ÚJČ AV ČR. [cit. 2017-06-11]. Dostupné z: [`http://ssjc.ujc.cas.cz`]

[6] PANG, BO; LEE Lillian. *Opinion Mining and Sentiment Analysis* [online]. [cit. 2017-05-20]. Dostupné z: [`http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf`], 4.1.2 Subjectivity Detection and Opinion Identification

[7] SCHLER, Jonathan. *The importance of neutral examples for learning sentiment* [online]. [cit. 2017-06-01]. Dostupné z: [`http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.9735`]

[8] Shanahan, James G., Yan Qu, Janyce Wiebe. *Computing Attitude and Affect in Text, Springer*, 2005

[9] Knittlová, Dagmar.*Překlad a překládání*, Olomouc: UP, 2010.

[10] Jurafsky, Dan. *Text Classification and Naïve Bayes* [online]. [cit. 2017-02-25]. Dostupné z: [https://web.stanford.edu/class/cs124/lec/naivebayes.pdf]

[11] Ekštein, Kamil. *Support Vector Machines* [přednáška]. Plzeň. Strojové učení. 2012

[12] PALA, Karel*Stručný terminologický slovník*, [online]. [cit. 2017-06-03]. Dostupné z: [https://nlp.fi.muni.cz/cs/Terminologie]

[13] Panevová, Jarmila; Karlík, Petr. *Nový encyklopedický slovník češtiny.* [online]. [cit. 2017-06-13]. Dostupné z: [https://www.czechency.org/slovnik/VALENCE]

[14] Dvořák, Věra. *Nový encyklopedický slovník češtiny*, [online]. [cit. 2017-06-03]. Dostupné z: [https://www.czechency.org/slovnik/S%C3%89MANTICK%C3%81%20ROLE]

[15] Carreras,Xavier ;Màrquez Lluís. *CoNLL-2005 Shared Tasks Semantic Role Labeling*, [online]. [cit. 2017-06-03]. Dostupné z: [http://www.lsi.upc.es/~srlconll/]

[16] HONG, James; FANG, Michael. Sentiment analysis with deeply learned distributed representations of variable length texts. Technical report, Stanford University, 2015.

[17] Charles Fillmore. The case for case. In E.Bach and R.T.Harms, editors, *Universalsin Linguistic Theory*, 1968. str. 1–88

[18] KONEČNÁ, Kristýna.*Fillmorova teorie pádů*, [online]. [cit. 2017-06-03]. Dostupné z: [http://oltk.upol.cz/encyklopedie/index.php5/Fillmorova_teorie_p%C3%A1d%C5%AF:_Obecn%C3%A1_charakteristika]

[19] Marquez, L.; Carreras, X.; Litkowski, K. C.; Stevenson S. *Semantic Role Labeling: An Introduction to the Special Issue*, [online]. [cit. 2017-06-03]. Dostupné z: [http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.34.2.145]

[20] PALMER, Martha; TITOV, Ivan; WU, Shumin.*Semantic Role Labeling Tutorial at NAACL 2013*, [online]. [cit. 2017-06-03]. Dostupné z: [http://ivan-titov.org/teaching/srl-tutorial-naacl13/]

[21] Levin, B.1993. *English Verb Classes and Alternations: A Preliminary Investigation.* The University of Chicago Press, Chicago, IL.

[22] MEĽČUK, Igor Aleksandrovič. *Dependency syntax: theory and practice.* SUNY press, 1988.

[23] Medhat, Walaa.*Sentiment analysis algorithms and applications: A survey*, [online]. [cit. 2017-06-03]. Dostupné z: [http://www.sciencedirect.com/science/article/pii/S2090447914000550?np=y&npKey=78b015f6b8b69ce88b4aea6ee91eef4d968d9a3e573d252112d5817e5518bbda]

[24] Berkley .*What is FrameNet?*, [online]. [cit. 2017-06-03]. Dostupné z: [https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet]

[25] Panevová, Jarmila.*PRAŽSKÝ ZÁVISLOSTNÍ KORPUS*, [online]. [cit. 2017-06-03]. Dostupné z: [https://www.czechency.org/slovnik/PRA%C5%BDSK%C3%9D%20Z%C3%81VISLOSTN%C3%8D%20KORPUS]

[26] Kohavi, Ron.*A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, [online]. [cit. 2017-06-07]. Dostupné z: [http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf]

[27] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

[28] Carrera Xavier.*CoNLL-2005 Shared Task Semantic Role Labeling*, [online]. [cit. 2017-05-15]. Dostupné z: [http://www.lsi.upc.es/~srlconll/st05/st05.html]

[29] Jan Hajič.*CoNLL-2009 Shared Task:Syntactic and Semantic Dependencies in Multiple Languages*, [online]. [cit. 2017-05-15]. Dostupné z: [https://ufal.mff.cuni.cz/conll2009-st/task-description.html]

[30] Punyakanok, V.; Koomen, P.; Roth D. ;Yih W. , *Generalized Inference with Multiple Semantic Role Labeling Systems* [online]. [cit. 2017-06-10]. Dostupné z: [http://www.lsi.upc.es/~srlconll/st05/papers/punyakanok.pdf]

[31] CHE, Wanxiang; LI, Zhenghua; LI, Yongqiang; GUO Yuhang; QIN, Bing, LIU, Ting.*Multilingual Dependency-based Syntactic and Semantic Parsing*, [online]. [cit. 2017-06-03]. Dostupné z: [https://ufal.mff.cuni.cz/conll2009-st/results/papers/15_conll09st.pdf]

[32] XUE, Nianwen; PALMER, Martha. *Calibrating Features for Semantic Role Labeling*. In: EMNLP. 2004. p. 88-94.

[33] TOUTANOVA, Kristina; HAGHIGHI, Aria; MANNING, Christopher D. *Joint learning improves semantic role labeling*. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005. p. 589-596.

[34] Hagen Fürstenau; Lapata, Mirella . Graph alignment for semi-supervised semantic role labeling. *In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009. p. 11-20.

[35] LANG, Joel; LAPATA, Mirella. *Unsupervised induction of semantic roles. In: Human Language Technologies*: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. p. 939-947.

[36] TITOV, Ivan.*Semi-supervised, unsupervised and cross-lingual approaches*, [online]. [cit. 2017-05-15]. Dostupné z: [`http://techtalks.tv/talks/semantic-role-labeling-part-3/58422`]

[37] Lang J.; Lapata, M. *Unsupervised semantic role induction via split-merge clustering*. ACL. 2011

[38] TITOV, I. ;KLEMENTIEV A.. *A Bayesian Approach to Unsupervised Semantic Role Induction*. EACL. 2012

[39] GORMLEY, Matthew R., et al. *Low-Resource Semantic Role Labeling*. In: ACL (1). 2014. p. 1177-1187.

[40] GIMPEL, Kevin, et al. *Part-of-speech tagging for twitter: Annotation, features, and experiments*. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011. p. 42-47.

[41] Nováková, Eva.*Nominální tendence v angličtině a jejich české ekvivalenty ve vybraných funkčních stylech*, [online]. [cit. 2017-06-03]. Dostupné z: [`http://theses.cz/id/5oj6ry/DP_text_Novkov_2.pdf`]

[42] Konkol, M. *Brainy: A machine learning library*. ICAISC 2014. LNCS, vol. 8468, Part II, pp. 490-499. Springer, Heidelberg, 2014.

[43] BLAMEY, Ben; CRICK, Tom; OATLEY, Giles. RU:-) or:-(? Character-vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora. In: SGAI Conf. 2012. p. 207-212.

[44] Veselovská, K. *On the Linguistic Structure of Emotional Meaning in Czech*, [online]. [cit. 2017-06-03]. Dostupné z: [`https://is.cuni.cz/webapps/zzp/detail/85092/`]

[45] UFAL. *The Prague Dependency Treebank 2.0*, [online]. [cit. 2017-06-03]. Dostupné z: [`http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html`]

[46] Zima, J. (1961). *Expresivita slova v současné češtině: studie lexikologická a stylistická*. Nakl. Československé akademie věd.

[47] HABERNAL, Ivan; PTÁCEK, Tomás; STEINBERGER, Josef. *Sentiment Analysis in Czech Social Media Using Supervised Machine Learning*. In: WASSA@ NAACL-HLT. 2013. p. 65-74.

[48] NOCEDAL, Jorge. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 1980, 35.151: 773-782.

[49] BRYCHCÍN, Tomáš; KONOPÍK, Miloslav. *HPS: High precision stemmer. Information Processing & Management*, 2015, 51.1: 68-91.

[50] NIGAM, Kamal; LAFFERTY, John; MCCALLUM, Andrew. *Using maximum entropy for text classification*. In: IJCAI-99 workshop on machine learning for information filtering. 1999. p. 61-67.

A Uživatelská příručka

Program předpokládá následující jména souborů pro vstup

- Facebooku - *gold-posts.txt* pro příspěvky, *gold-labels.txt* pro značky a *gold-posts.txt.ArkTokenized* pro tokenizované příspěvky mezerou
- Mall - 3 soubory v nichž jsou oddělené kategorie: *positive.txt*, *negative.txt* a *neutral.txt*, tokenizované soubory mají všechny k názvu příponu .ArkTokenized (*positive.txt.ArkTokenized*)
- CSFD - stejné názvy jako pro Mall

Pro spuštění aplikace je zapotřebí nainstalovaná Java minimální verze 1.8.

A.1 Sestavení aplikace

Aplikace obsahuje závislosti a jsou uvedeny v konfiguračním souboru `pom.xml`

Aplikace lze sestavit včetně potřebných závislostí příkazem:

```
mvn package assembly:single
```

A.2 Spuštění aplikace

Aplikace se spouští příkazem

```
java -jar <JAR path> <dataset> <featureSetFile> <
    method > <datasetFolder>
```

- dataset má možnosti *mallcz*, *csfd* nebo *fb*
- featureSetFile - seznam možných množin příznaků je ve složce featuresSet

- method má možnosti *maxent*, *svm* nebo *bayes*
- datasetFolder je volitelný parametr, pokud se soubory datasetů nenechází v stejné složce jako spustitelný soubor