Počítačové zpracování emocí

Kateřina Veselovská

Ústav formální a aplikované lingvistiky
MFF UK

6. října 2016, MUNI

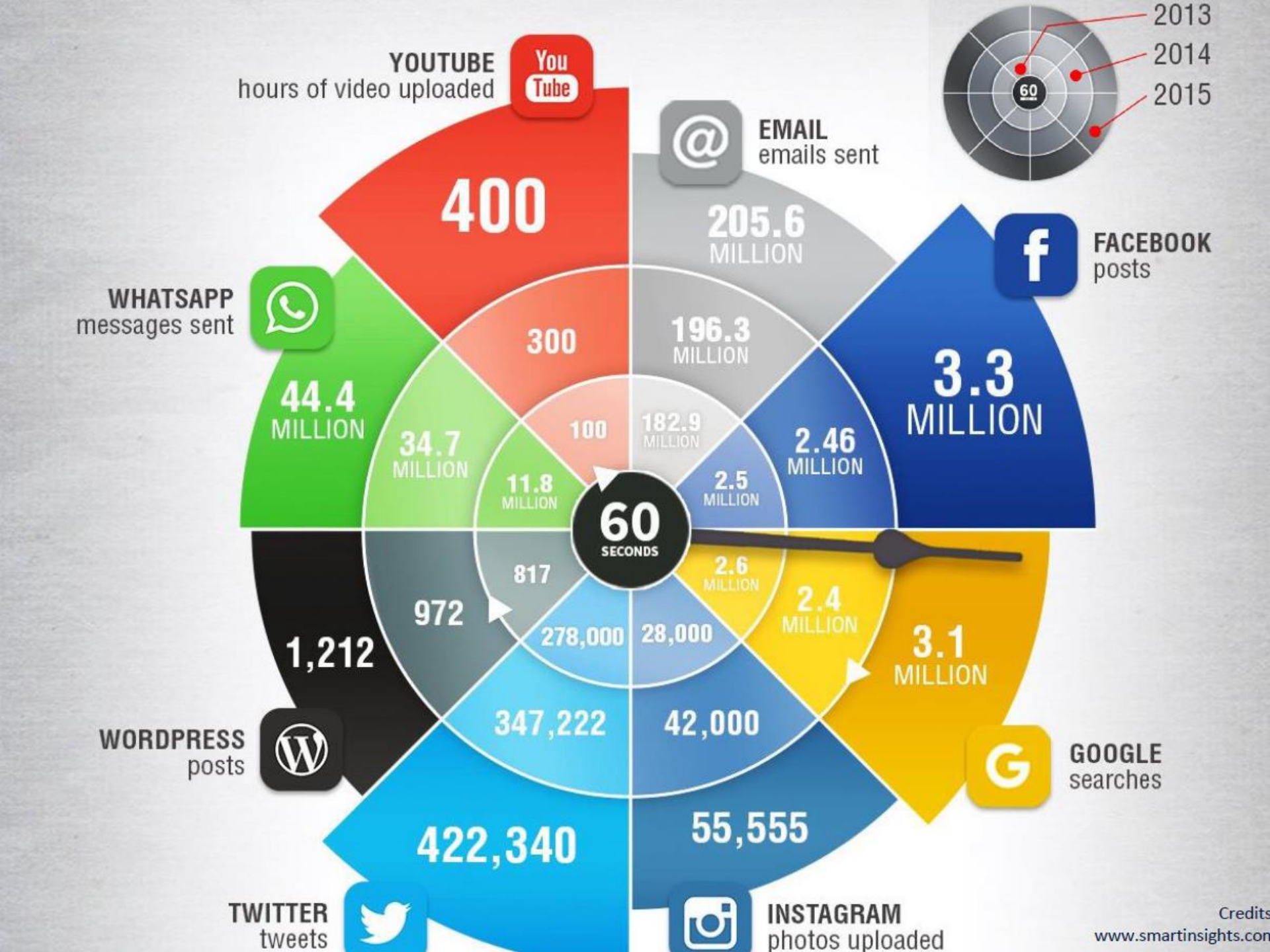
Počítačové zpracování emocí

v lexiku

Kateřina Veselovská

Ústav formální a aplikované lingvistiky
MFF UK

6. října 2016, MUNI


2013
2014
2015
60
YOUTUBE
hours of video uploaded
400
300
100
EMAIL
emails sent
205.6
MILLION
196.3
MILLION
182.9
MILLION
FACEBOOK
posts
3.3
MILLION
2.46
MILLION
2.5
MILLION
WHATSAPP
messages sent
44.4
MILLION
34.7
MILLION
11.8
MILLION
60
SECONDS
817
2.6
MILLION
2.4
MILLION
3.1
MILLION
GOOGLE
searches
WORDPRESS
posts
1,212
972
278,000
28,000
347,222
42,000
422,340
55,555
TWITTER
tweets
INSTAGRAM
photos uploaded
Credits
www.smartinsights.com

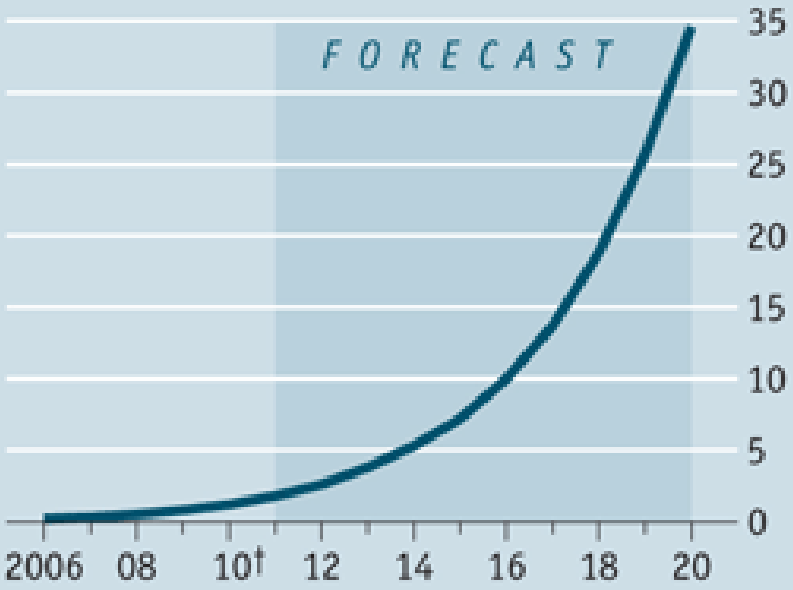
"80% of all data is dark and unstructured. We can't read it or use in our computing systems. By 2020, that number will be 93%."

John Kelly, IBM Senior VP

Too much information

Worldwide digital data created and replicated

Zettabytes*



Source: IDC *1 zettabyte = 1 trillion gigabytes †Estimate

IDC: Companies that can figure out a way to properly collect, synthesize and use **unstructured data** can improve their **bottom line**, **reduce costs** and help organizations **respond** more quickly to **changing market** and **customer sentiment**.

Odkud přicházím

- **Ú**stav **F**ormální a **A**plikované **L**ingvistiky MFF UK

- tým vědeckých pracovníků, programátorů, pedagogů a studentů, kteří společně pracují na široké škále témat spojených s oborem počítačové lingvistiky

- projekt :(SEANCe :) = sentiment analysis in Czech

Odkud přicházím

- strojový překlad, dialogové systémy, závislostní korpusy (PDT, PCEDT, PDTSC…)

https://ufal.mff.cuni.cz/teaching/prospective-students/

Emoce v jazyce

Rozpor:

- emocionální prožitek je společností vnímán jako pozitivní hodnota (“žít” znamená “cítit”)
- život s nedostatkem emocí je vnímán jako “plochý”, neuspokojivý
- neschopnost interpretovat komplexní emoce jako mentální handicap (Aspergerův syndrom aj.)

X

- v mezilidské komunikaci jsou emoce (jak pozitivní, tak negativní) obvykle prostředkem manipulace

Emoce v jazyce

Sociolingvistika

- interpersonální funkce – sociální vztahy mezi lidmi jsou vyjadřovány jazykově
- textové strategie

Emoce v jazyce

Pragmalingvistika

- faces – positive, negative - to, co vystavujeme k veřejnému (pozitivnímu) hodnocení, to, čím se vůči společnosti vymezujeme negativně
- strategie “ohrožování tváří” a strategie “zachování tváře”

Emoce v jazyce

Forenzní lingvistika

- míra emocionality/expresivity textu jako distinktivní rys
- predikce lži
- ztotožňování pachatelů trestných činů

Emoce v jazyce

- teorie komunikace
- Speech Act theory – expresivní ilokuční akt – vyjádření postoje a emocí
- Grice's Maxim of Manner
- RaBR a další praktické komunikační techniky – úspěšná a účelná komunikace má být prosta emocí. Explicitní hodnocení (i pozitivní) vnímáno negativně.

Emoce v jazyce

Manipulativní techniky v médiích

- teorie argumentace, logické (argumentační) fauly: zesměšňování, zastrašování, lichocení, vyvolání zášti, kvalifikující jazyk...

“Každá slušná rodina chodí do kostela.”

Emoce v jazyce

Míra subjektivity textu jako měřítko “dobré/špatné” žurnalistiky

J.X.D.: *“Příběh je to tak strašně podobný těm drogovým, že se zdá být termín „nesubstanční závislost“ používaný některými lékaři jako oprávněný. Podle mého je to však blbost.”*

Emoce v jazyce

- objektivní žurnalistika x komentáře, názory, sloupky vyjadřující stanovisko redakce x subjektivní blog
- subjektivní žurnalistika postupně začíná převládat – důsledky?
- kategorie “pravda” ustupuje kategorii “ztotožnění s převládajícím názorem”?

Emoce v jazyce

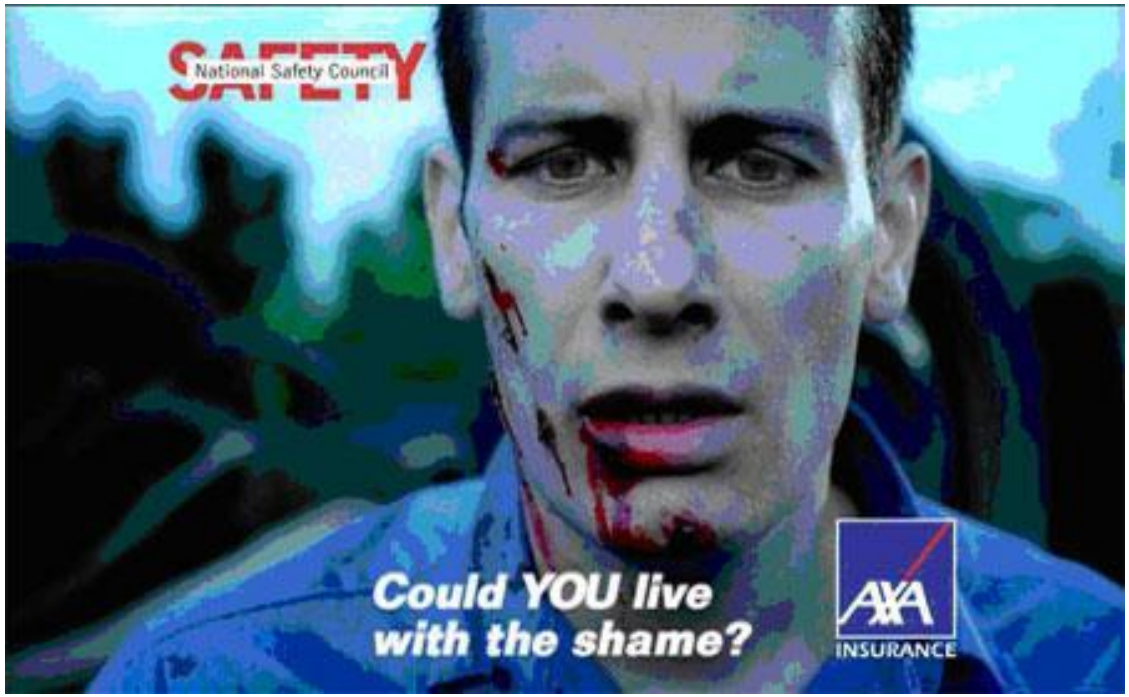
Marketingové strategie:

emocionalizace produktů

Emoce v jazyce

Marketingové strategie:

emocionalizace produktů



Emoce v jazyce

PR: mluvčí společnosti

- prezentuje cizí postoje bez odkazu na vlastní postoje
- musí vyjádřit hodnocení informací, ale bez emocionality
- vytváří maximálně pozitivní obraz společnosti minimálně emocionálními prostředky

Emoce v jazyce

- Analyzovat a interpretovat emocionalitu (subjektivitu, expresivitu, atd.) v textu je v rámci analýzy a interpretace komunikace důležité.

- Proto se zabýváme tím, jaké prostředky jsou pro vyjadřování emocionality (subjektivity, postojů, atd.) jsou v češtině dostupné, obvyklé, a jak se liší od prostředků jiných jazyků.

Sentiment Analysis

Sentiment Analysis

Proč to děláme

- hodnocení produktů
- průzkum veřejného mínění
- monitoring sociálních sítí
- intenční analýza
- forenzní lingvistika (kyberšikana, hanobení rasy a národa…)
- predikce trendů v marketingu
- vývoj akciových trhů
- predikce výsledků voleb

Jak to děláme

- kvalitativně:

- případové studie
- kritická analýza diskurzu

- rozhovor
- zúčastněné pozorování

Jak to děláme

- kvantitativně:

- větší soubory dat/respondentů
- obsahová analýza

- dotazníky
- strojové učení

Jak to děláme

- kvantitativně:

- větší soubory dat/respondentů
- obsahová analýza

- dotazníky
- strojové učení

Jak to děláme

- kvantitativně:

- větší soubory dat/respondentů
- obsahová analýza

- dotazníky
- strojové učení

↑ TADY VYUŽÍVÁME LINGVISTIKU

Jak to funguje

Miloš Zeman je český prezident.

vs.

Miloš Zeman je nejlepší prezident všech dob.

Detekce polarity

Miloš Zeman je nejlepší prezident všech dob.

vs.

Miloš Zeman je nejhorší prezident všech dob.

Miloš Zeman stál po boku Martina Konvičky.

Konečná bilance bruselských bombových útoků je 31 mrtvých a 250 zraněných.

Anotátorská shoda

Kappa ≈ 0.66

Hlasování prostou většinou

[Hadi v letadle jsou nejlepší hadí film od první Anakondy, je to parádní podívaná se spoustou skvělejch efektů – hlavně když dohánějí pasažéry ke strašné, bolestivé smrti.] +

Hlasování prostou většinou

- slovníky hodnotících výrazů
  - bootstrapping
  - syntaktická blízkost
  - překlad
- pro češtinu SubLex 1.0
  - 4 626 hodnotících výrazů
  - morfologie, POS/NEG orientace
  - dostupný online

Czech SubLex 1.0

A	báječný	polarity=positive
V	balamutit_:T	polarity=negative
N	banalita	polarity=negative
A	banální	polarity=negative
N	bankrot	polarity=negative
N	barbar	polarity=negative
A	barbarský	polarity=negative
D	barbarsky_^(*1ý)	polarity=negative
N	barbarství	polarity=negative
N	bastard	polarity=negative
V	bát	polarity=negative
V	bavit_:T	polarity=positive
N	bázeň	polarity=negative
D	bázlivě_^(*1ý)	polarity=negative
N	bažina	polarity=negative
N	bdělost_^(*3ý)	polarity=positive
I	běda	polarity=negative
V	bědovat_:T	polarity=negative
V	belhat_:T	polarity=negative
V	běsnit_:T	polarity=negative
N	bestie	polarity=negative
N	bezduchost_^(*3ý)	polarity=negative
A	bezduchý	polarity=negative
D	bezdůvodně_^(*1ý)	polarity=negative
N	bezmoc	polarity=negative
D	bezmocně_^(*1ý)	polarity=negative
N	bezmocnost_^(*3ý)	polarity=negative
A	bezmyšlenkovitý	polarity=negative
N	beznaděj	polarity=negative
D	beznadějně_^(*1ý)	polarity=negative
A	beznadějný	polarity=negative
D	bezohledně_^(*1ý)	polarity=negative
N	bezohlednost_^(*3ý)	polarity=negative
D	bezostyšně_^(*1ý)	polarity=negative
A	bezpečnostní	polarity=positive
N	bezpečnost-1_^(*5ý-1)	polarity=positive

Hlasování prostou většinou

Hadi v letadle jsou nejlepší hadí film od první Anakondy, i když jsou občas nechutně digitální a chovají se nemožně.

Strojové učení

- supervised learning
  - Naive Bayes
  - Support vector machines
  - MaxEnt
  - slovníkové klasifikátory
- unsupervised learning
  - Turneyův třístupňový algoritmus

Strojové učení

- nedostatek českých dat – příprava anotovaných datasetů pro trénování klasifikátorů

Data

- Aktuálně: sekce Domácí (428 segmentů)
- ČSFD.cz: filmové recenze (531 vět)
- Mall.cz: recenze bílého zboží (10 177 recenzí)
- Facebook: statusy (10 000 postů)

Aktualne.cz

Podepsala se na něm lehká viróza, se kterou bojuje několik dnů. To byl jediný důvod k té lehké indispozici.

Csfd.cz

Tohle je náhodou áčkový béčko a já jim to žeru i s navijákem.

Mall.cz

Je to výkonný a kvalitní vysavač. Moje tchýně ho měla deset let, bohužel s ním ale zacházela nešetrně, ona tak ostatně zachází s celou rodinou, je to neuvěřitelně protivná herdekbaba, ale co člověk nadělá, špatné příbuzné si nevybíráme, že. Ta ženská nám pije krev každým dnem víc. Každopádně, když se stroj porouchal, nechtěla ho nechat opravovat, a tak nám ho věnovala. I porouchaný vysavač ale pořád funguje jako vysavač, nejdou s ním jen čistit koberce. Půjčovala a půjčuje si ho celá rodina i příbuzný, je fakt dobrý, mohu ho doporučit.

Facebook

No to si děláte #@?!!

Anotace

- na úrovni slov – *Je to [skvělý]+ film.*
- na úrovni vět – *[Je to skvělý film.]+*
- na úrovni dokumentů

[Je to skvělý film. Ale rozhodně bych o něm neřekla, že je nejlepší, co jsem kdy viděla. Obsazení herců je dobré, tedy až na Froda, v některých částech mi už docela lezl na nervy.]?

Anotace

- na úrovni slov
- na úrovni vět

  ↑ tady klasifikátory fungují
- na úrovni dokumentů

Anotace

- manuální a automatická
  (automatická deterministicky odvozená z manuální)
- dva anotátoři na úrovni slov, větná rovina odvozená
- případně manuálně úroveň věty

Anotace – zdroj

Kyborský generál Grievous nesnáší Jedie.

Anotace – hodnocení

Kyborský generál Grievous nesnáší Jedie.

Anotace – cíl

Kyborský generál Grievous nesnáší Jedie.

Měření shody

- na větě: kappa, f-score

  (unlabeled, labeled, polarita)

- na slovech: f-score na překryvu

  kyborský generál Grievous x Grievous

- kappa kolem 0.66, f-score 0.60 – 0.94

Nástroje

- preprocessing: lemmatizace
- klasifikátory: Naive Bayes
  slovníkový klasifikátor
- bag-of-words model
- word-based features
- filtrování

Nástroje

- lemmatizace
  - lemmatizace, desambiguace
- POS tag a negace
  - tagger Morče

Nástroje

Klasifikátor 1

- features: přítomnost lemmatu v daném segmentu

Klasifikátor 2

- features: přítomnost lemmatu v daném segmentu

  + pravidla

Nástroje

- oba klasifikátory odhadují prediktivní sílu jednotlivých lemmat vzhledem po POS/NEG polaritě

- trénování ~ vytváření slovníku lemmat + jejich prediktivní síly

Nástroje

Filtrování

- na základě frekvence
- na základě POS
- zohlednění negace

+ lze přidat externí lexikon

Výsledky

Model	Acc	Recall	Precision	F-score
Baseline	0.63	0.31	0.23	0.29
NB + rules train	0.96	0.96	0.96	0.96
NB + rules test	0.89	0.89	0.89	0.89
NB train	0.86	0.80	0.88	0.83
NB test	0.83	0.75	0.85	0.78

⇒ i s jednoduchým klasifikátorem lze dosáhnout dobré úspěšnosti na poměrně malém vzorku dat

Error Analysis

- chyby lidí
- chyby systému

Chyby lidí

Porušování konverzačních maxim

- kvantity
- kvality
- relevance
- způsobu

Chyby lidí

Kategorie pozitiva:

Meteostanici mám jako dárek pro manžela, zatím jsem ji nevyzkoušela, ale myslím, že je super.

Chyby lidí

Kategorie pozitiva:

Nevím, jak jsem mohla bez sušičky být. Haní ji jen ten, kdo ji nemá, nebo zhrzená manželka, když jí nechce manžel sušičku koupit. Úspora času, sice něco se musí žehlit, ale minimálně. Za sobotu jsem stihla usušit ložní prádlo, včetně obalů z matrací a lůžkovin (polštáře, deky) a ještě jsem měla spoustu času.

Chyby systému

Krátké segmenty v obou kategoriích:

Nic. Cena. Nevím. Prostě myčka.

Chyby systému

Výrazy příznačné pro jednu kategorii, které se objeví ve druhé:

Kvalita.

Chyby systému

Doménová závislost:

Dlouhé prací programy.

Negace

[Není to žádný luxusní model.] –

Adverzativní koordinace

[Není to žádný luxusní model, ale na chalupu stačí.] +

Intenzifikátory

Je to pěkně blbý.

Idiomy

Za málo peněz hodně muziky.

Idiomy

Janchor ★★★★

The Force is with you – but you are not a Jedi yet.

Pojmenované entity

Grafická podoba

**SUPEEEER*!!!:-DD*

Vulgarismy

Barman je @#%$!!

Vulgarismy

Barman je @#%$!!

> barman|být|??????

Vulgarismy

- *ho.no, nas*at*
- *KUA, OMG, WTF*
- *Kaďousek, Peacha*

Vulgarismy

Je tu tma jak v prdeli. (neutrální)

S Jardou je vždycky prdel. (pozitivní)

A je to v prdeli. (negativní)

Vulgarismy

Je to kurva dobrý.

Detekce cílů hodnocení

Kurva dobrá baterie, ale displej stojí za hovno.

Data

Alza.cz: IT produkty

a) krátké segmenty
b) delší recenze

Data

Krátké segmenty:

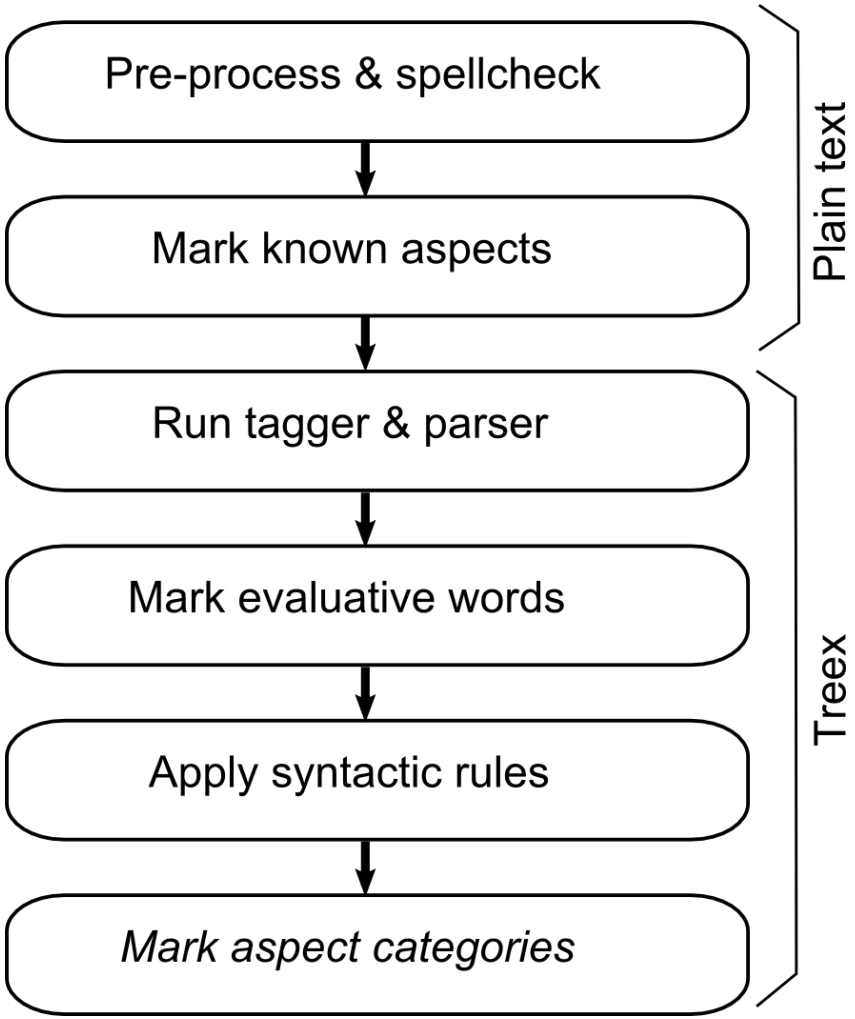
- 1000 pozitivních + 1000 negativních, ručně značené cíle

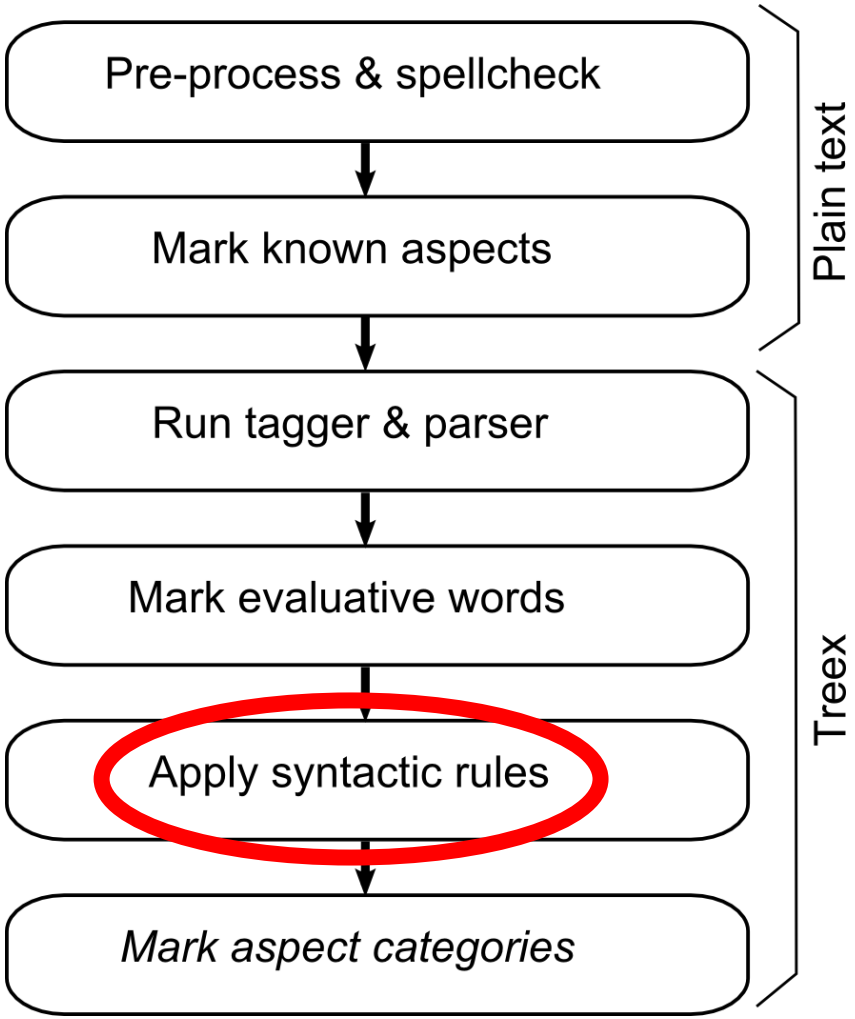
Data

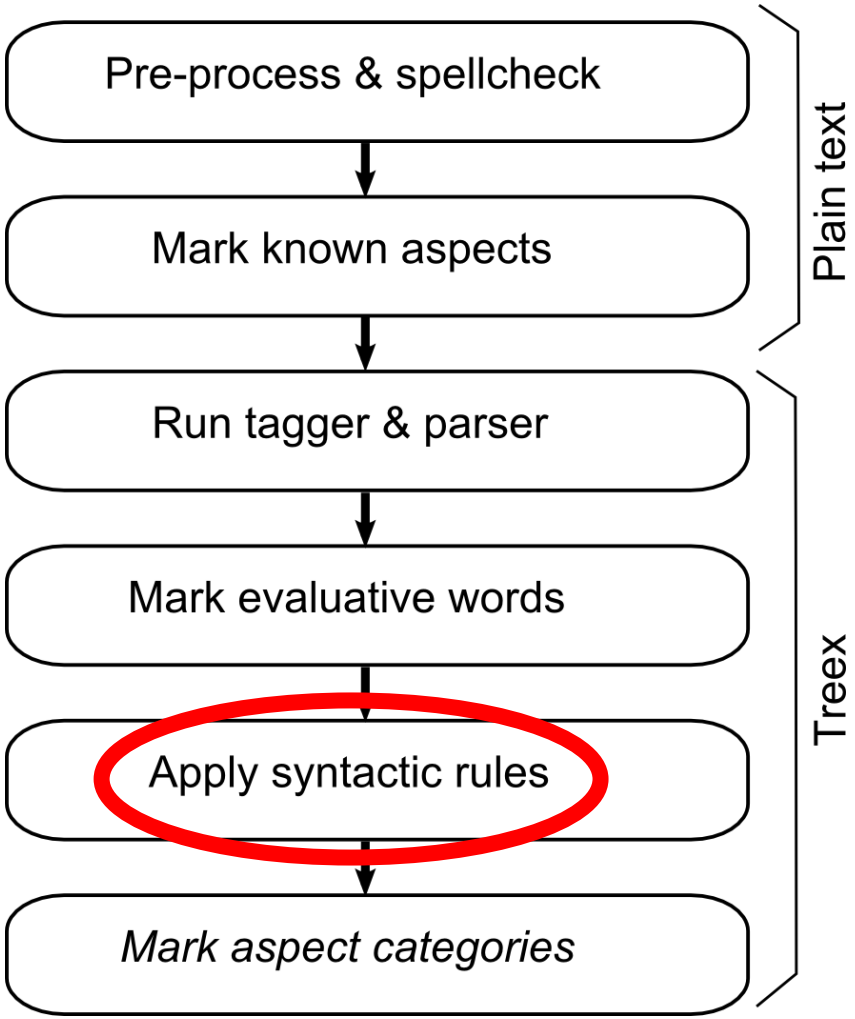
Dlouhé segmenty:

Velké zklamání. Tahle sluchátka hrají opravdu špatně – zvuk je příliš hluboký, zatemněný a vibrace ještě zvětšují pocit "hutnosti" a těžkosti. Na hry dobré, na hudbu nepoužitelné. Již po hodině používání mě nesmírně bolely uši. Možná to bylo způsobeno tím, že mám velkou hlavu... Čímž se dostávám k dalšímu negativu. Sluchátka jsou opravdu malá a lidé s většími hlavami prostě nemají šanci je pohodlně nosit. Ani mikrofon nepotěší…

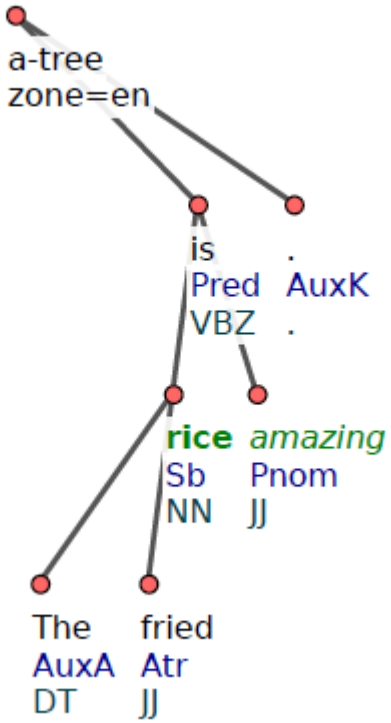
- 100 pozitivních, 100 negativních, nejdelší 7057 znaků


Pre-process & spellcheck
Mark known aspects
Run tagger & parser
Mark evaluative words
Apply syntactic rules
Mark aspect categories
Plain text
Treex


Pre-process & spellcheck
Mark known aspects
Run tagger & parser
Mark evaluative words
Apply syntactic rules
Mark aspect categories
Plain text
Treex


Pre-process & spellcheck
Mark known aspects
Run tagger & parser
Mark evaluative words
Apply syntactic rules
Mark aspect categories
Plain text
Treex

Syntaktická pravidla



The fried rice is amazing.

Model

- linear-chain conditional random fields
- rekurentní neuronová síť

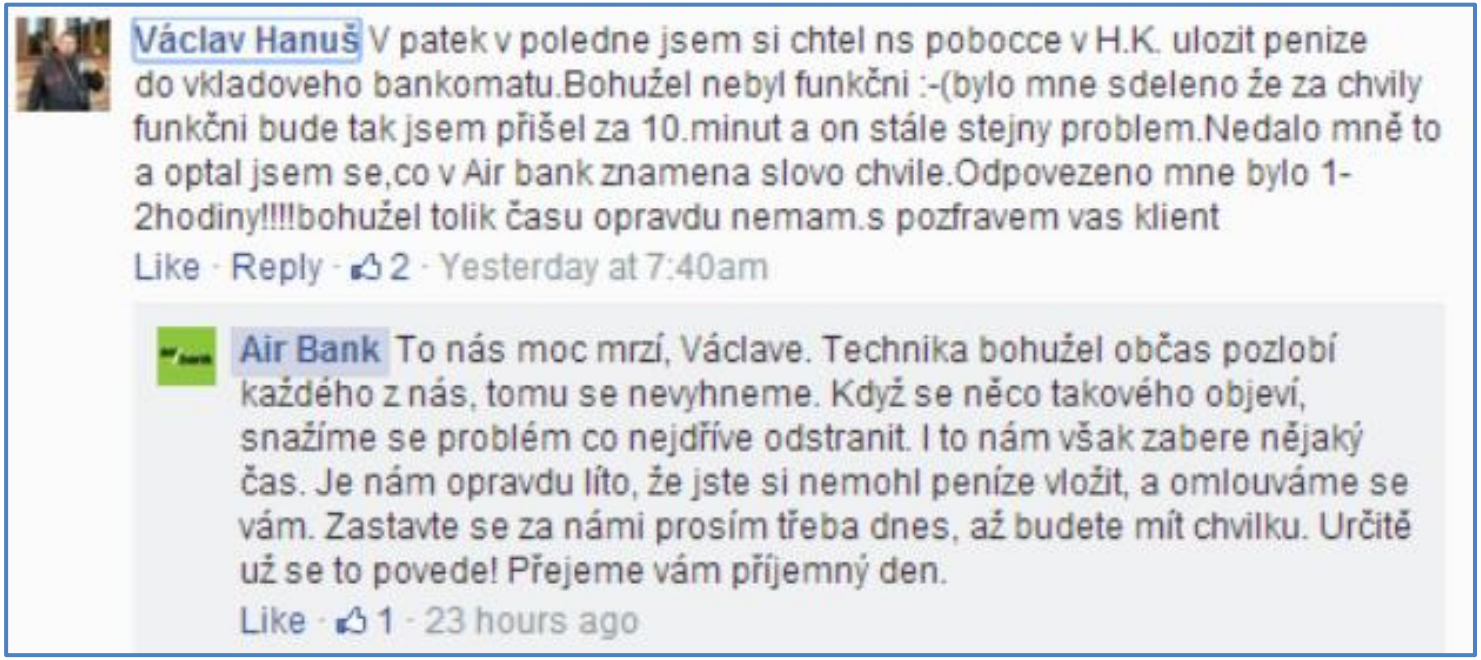
Výsledky

Feature set	Short segments			Long segments		
	P	R	F	P	R	F
surface	85.22	36.85	51.45	47.18	8.05	13.76
+ morpho-syntactic	75.88	54.17	63.21	40.17	23.08	29.31
+ sublex	78.19	55.09	64.64	58.74	18.99	28.70
+ rules	76.54	57.69	65.79	51.74	21.39	30.27

Výsledky

Domain	Restaurants						Hotels	Laptops
Language	English	Spanish	French	Dutch	Russian	Turkish	Arabic	English
Network	59.3	58.8	49.9	53.9	64.8	61.0	47.3	27.0
Baseline	58.0	62.2	54.8	54.7	60.8	34.4	49.4	35.0

Personalizovaný marketing



Václav Hanuš V patek v poledne jsem si chtel ns pobocce v H.K. ulozit penize do vkladoveho bankomatu.Bohužel nebyl funkčni :-(bylo mne sdeleno že za chvily funkčni bude tak jsem přišel za 10.minut a on stále stejny problem.Nedalo mně to a optal jsem se,co v Air bank znamena slovo chvile.Odpovezeno mne bylo 1-2hodiny!!!!bohužel tolik času opravdu nemam.s pozfravem vas klient

Like · Reply · 2 · Yesterday at 7:40am

Air Bank To nás moc mrzí, Václave. Technika bohužel občas pozlobí každého z nás, tomu se nevyhneme. Když se něco takového objeví, snažíme se problém co nejdříve odstranit. I to nám však zabere nějaký čas. Je nám opravdu líto, že jste si nemohl peníze vložit, a omlouváme se vám. Zastavte se za námi prosím třeba dnes, až budete mít chvilku. Určitě už se to povede! Přejeme vám příjemný den.

Like · 1 · 23 hours ago

Zákaznický feedback

Koupit si tento fotoaparát a spoléhat na jméno firmy se mi vůbec nevyplatilo. Např. proti mému staříčkému fotoaparátu je toto naprostý propadák. Průměrný telefon s fotoaparátem udělá stejné, ne-li lepší fotografie. Ani komunikace s centrem podpory nestojí za nic. Výrobek mě zklamal a víckrát už si žádný produkt této firmy bezpochyby nekoupím.

Competitive intelligence

Ze spořícího účtu studenta jsem přešla na běžný účet když jsem začala vydělávat. Jakožto mladý člověk který začne vydělávat jsem potřebovala mít přehled o pěnězích abych věděla kdy smím utratit. A tady to začalo. Platit za to, že se smím laskavě podívat na svůj zůstatek účtu? Platit za to, že smím vůbec vybrat z vlastního účtu? Platit za to, že ho po připsání výplaty používám? Zvolila jsem na doporučení známého tuhle banku, která u nás byla relativně nová a nelituji. Když jsem si zbývajících 250,-kč chtěla převést na nový účet u konkurence a stávající zrušit, paní na mě přísně zamrkala přes brýle co si to jako takhle mladá dovoluji "A smím se zeptat co vás táhlo k nové bance?" "Mají lepší internetové bankovnictví, dotaz na zůstatek, výběry z kteréhokoliv bankomatu..."

Churn analysis

Blahopřeju vám k anti-péči o zákazníka. Jsem u tohoto operátora od roku 2002 a jediné co mi umíte nabídnout jsou běžné – a stále dost nevýhodné – tarify. Člověk si aspoň uvědomí, jak moc potřebujeme Evropskou unii (jejíž instituce jako jediná z relevantních subjektů tlačí ceny dolů). Po špatných zkušenostech hodlám přejít ke konkurenci.

Future work

- extrakce cílů v angličtině pomocí neuronových sítí
- psycholingvistické experimenty
- analýza suprasegmentálních rysů emocionálního vyjadřování
- analýza multimodálních dat

AI-complete problems

tichý vysavač x tichý budík

Go read the book!

USA: *Awesome!* 

SWE: *Inget speciellt* 


I
Haters!


SAVE BEACH
Hotel

Datové zdroje

www.lindat.cz

www.cs.uic.edu/~liub/

http://nlp.stanford.edu/sentiment/

www.ufal.cz/~veselovska/

veselovska@ufal.mff.cuni.cz