ANALÝZA SENTIMENTU V SLOVENSKÝCH TEXTOCH

Rastislav Krchňavý
Vedúci projektu: Ing. Marián Šimko PhD.
Konzultant: Mgr. Matej Hruška

Téma

- Analýza sentimentu
- Pre slovenský jazyk
- Z príspevkov z diskusií, sociálnych sietí, recenzií
- Z textu chceme vyjadriť „číselnú hodnotu“
- Štatistika
- Prečo? – ľudia chcú vedieť názory na ich produkt, firmu

Sentiment

- Subjektívny postoj
- Klasifikácia pozitívny/negatívny
- Kategórie (radosť, satisfakcia, hnev, smútok,...)
- Chceme ho určiť z textu
- Chceme ho zmerať jeho veľkosť

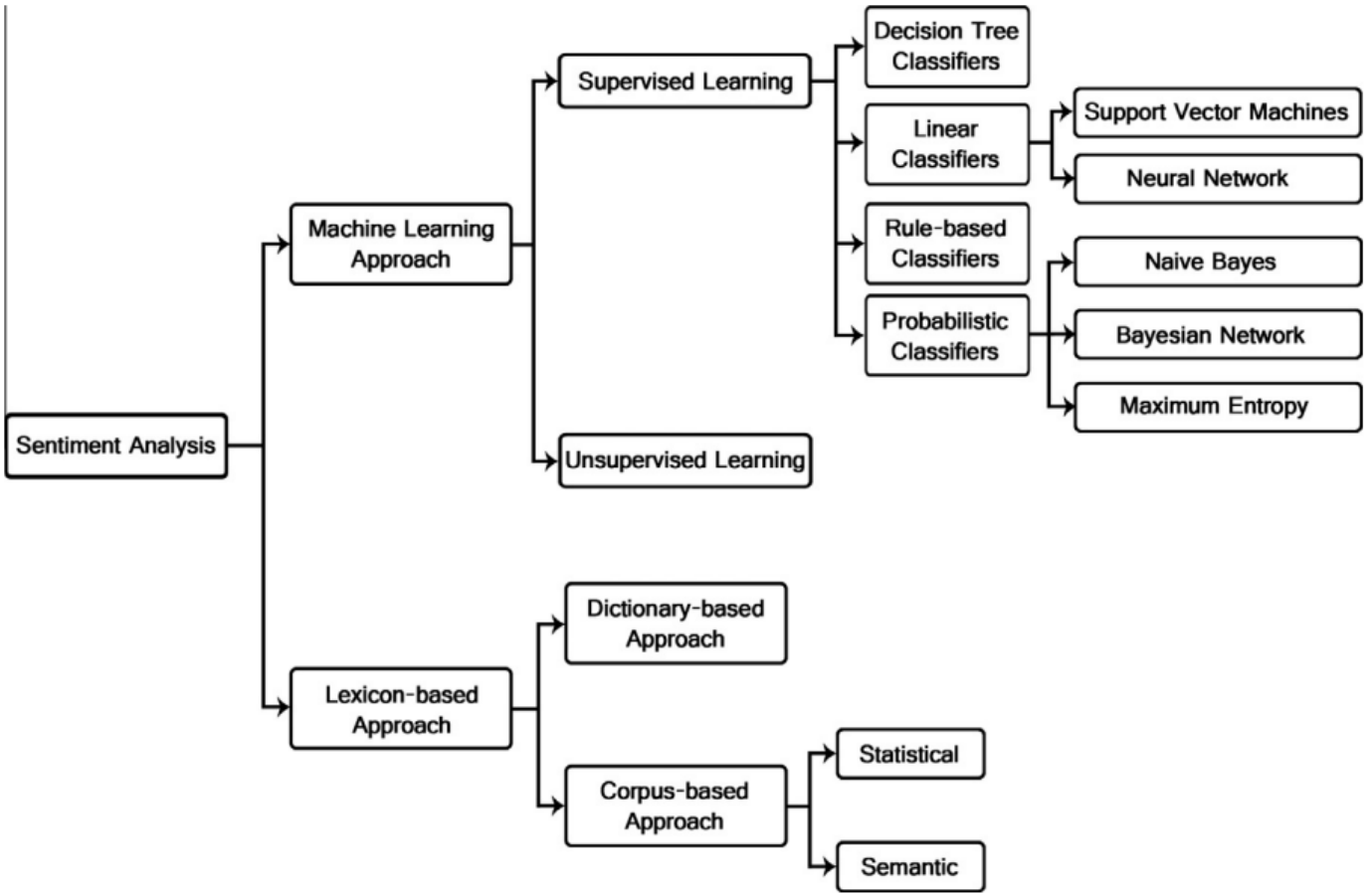
Príklady z datasetu

- Negatívny:
- Fico nie je normálny ani ostatni s nim.Včera som šlapala domov z prace 10 km , MHD bolo paralyzované, kto je zato zodpovední...Dajme im my pokuty, keby cestári fungovali,tak by sasa to nestalo.Napadne trocha snehu a už sú hotoví,21. storočí.Vážení občaniam myslite na tieto prešlapy aj pri volbách, pripomenme im to...lrebo hlásajú, že oni sú pripravení na sneženie, ale kedy, v lete?
- Pozitívny:
- Janka, citam vsetky tvoje clanky jednym dychom. Palawan je moj sen, ale v sezone(v zime) su rezorty neskutocne drahe, a v lete dost prsi a slnka je menej(aj ked horuco je stale).Z Europy je to daleko a treba zostat aspon 10-14 dni, aby to vobec oplatilo. Pre vas, co ste na Filipinach, to je ale na skok, dufam, ze navstivis viacej filipinskych ostrovov a zdokumentujes nam to. Velmi sa tesim na tvoje dalsie clanky Máš pravdu, menej ako 14 dní nemá zmysel a letenky sú dosť drahé.

Predspracovanie

- Slovenčina
  - tvary slov
  - dvojitý zápor
  - diakritika
  - zamlčaný podmet
- Emotikony
- Odstrániť interpunkciu
- Previesť na základnú tvar – lemu

Metódy analýzy sentimentu



Zdroj: http://www.sciencedirect.com/science/article/pii/S2090447914000550

Metódy analýzy sentimentu

- Naive Bayes – rozhoduje o sentimente na základe pravdepodobnostného klasifikátora
- Maximum Entropy – rozhoduje o sentimente na základe najvyššej entropie
- Support Vector Machines – pomocou strojového učenia sa snaží nájsť rovinu rozdeľujúcu dokumenty do kategórii
- Neurónové siete – natrénované pre rozoznávanie sentimentu
- Slovníkové metódy - priemer alebo iný štatistický ukazovateľ podľa počtu slov alebo vzorov z konkrétnej kategórie sentimentu

Porovnanie metód

Paper	Dataset	Technique (precision, %)
Pang et al. [1]	IMDB	NB (81.5), ME (81.0), SVM (82.9)
Turney [2]	Epinions	PMI(66)
Dave et al.[6]	Amazon, CNET	SVM (85.8-87.2), NB (81.9-87.0)
Hu and Liu [4]	Amazon, CNET	Lexicon (84.0)
Abbasi et al. [12]	U.S. & Middle Eastern web forum postings	SVM(95.55)
A. Khan et al. [5]	IMDB,Skytrax, Tripadvisor	Lexicon(86.6)
Zhang et al. [7]	Luce, Yoka	Lexicon(82.62)
Fang et al. [16]	Multi-Domain Sentiment Dataset	ML + Lexicon (66.8)
Zhang et al. [14]	Twitter	ML + Lexicon (85.4)
Mudinas et al. [15]	CNET, IMDB	ML + Lexicon (82.30)

Zdroj: http://www.researchgate.net/publication/267764884_A_COMPARATIVE_STUDY_OF_SENTIMENT_ANALYSIS_TECHNIQUES

Hypotéza

- Vybraná metóda analýzy sentimentu je schopná určiť sentiment textu napísaného v slovenskom jazyku a poskytnúť výsledky, ktorých presnosť je podobná presnosti pri použití iného (svetového) jazyka.
- Naive Bayes ~ 80-85%
- Support Vector Machines ~ 80-85%
- Slovníkove ~ 75-80%
- *záleží od datasetu

Moje riešenie

- Naive Bayes
- 2 triedy
- Pozitívny
- Negatívny
- Trénovacia množina
- Natural Language Toolkit

- For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Zdroj: https://web.stanford.edu/class/cs124/lec/naivebayes.pdf

Diskusia a návrhy

- Existujúce riešenia
- Knižnice
- Možné problémy
- Datasety
- ...