Analýza sentimentu v slovenských textoch

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

Analýza sentimentu

- Sentiment – „Nálada“ textu
- Pozitívny, negatívny alebo neutrálny
- Viac úrovní

Pozitívny	WOW!!! toto dneska musím vyskúšať!!!!
Neutrálny	je to pekné ale je to len reklama aby si ľudia volali viac a tým si z peňaženky míňali peniaze
Negatívny	Somarina a nie akciové ceny. Tie akciové ceny by malo byť pôvodná cena. Podvod nič viac cele nayko.

Motivácia

- Zistiť náladu pisateľa textu
- Analyzovať popularitu produktov z recenzií a komentárov
- Merať zmeny mienky zákazníkov

Problémy

- Slovenský jazyk
  - Diakritika
  - Ohýbanie slov
  - Synonymá
- Irónia
- Emotikony a emoji
- Štýl písania na sociálnych sieťach

Prístupy k analyzovaniu sentimentu

- Úroveň analyzovania
  - Dokumenty
  - Vety
  - Aspekty
- Klasifikovanie textov (Medhat, 2014)
- Strojové učenie
  - Naivný Bayesov klasifikátor (NB)
  - Klasifikátor maximálnej entropie (ME)
  - Metóda podporných vektorov (SVM)
- Slovníkový prístup (LEX)

Návrh riešenia

Triedy klasifikovania

- Do koľkých kategórií budeme klasifikovať vstupný text?
- 2 (Pozitívny a Negatívny)
- 3 (Pozitívny, Neutrálny a Negatívny)
- 4 (Silne Pozitívny, Pozitívny, Negatívny a Silne Negatívny)
- 5 (Silne Pozitívny, Pozitívny, Neutrálny, Negatívny a Silne Negatívny)

Predspracovanie

- Konverzia textu na črty
- Pomocou čŕt sa klasifikuje text
  1. Diakritikovanie textu
  2. Extrakcia emotikonov a emoji
  3. Lematizovanie
  4. Segmentácia
  5. Odstránenie „stop“ slov

Klasifikovanie textov

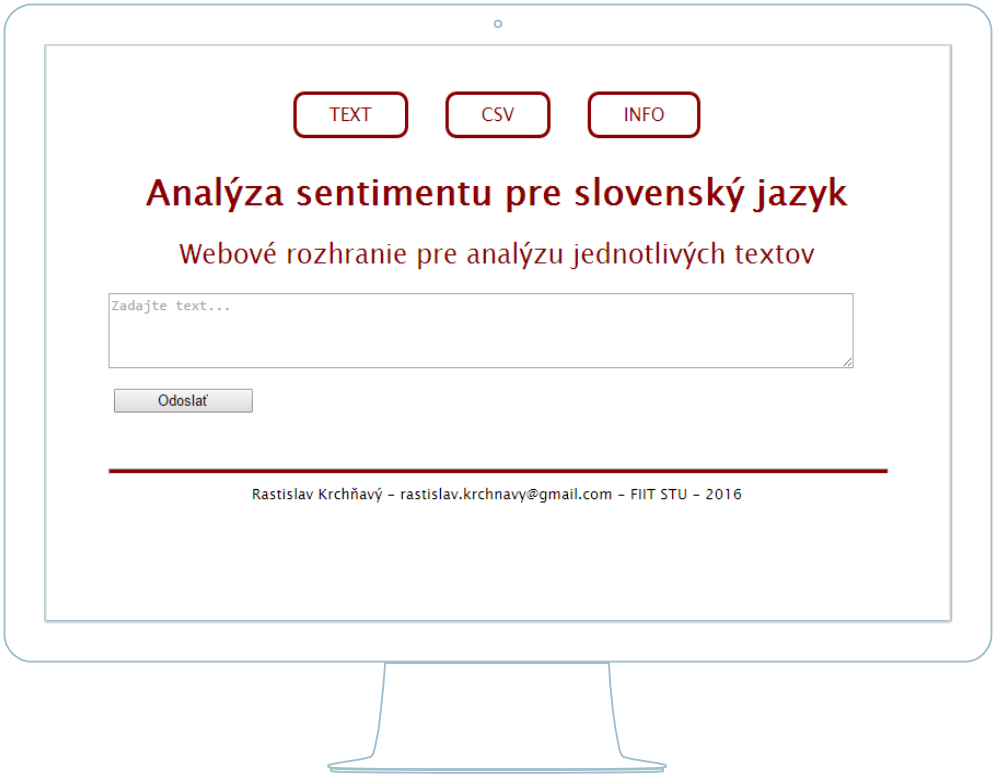
- Natural Language Toolkit (NLTK)
- Dataset od spoločnosti Seesame
  - Cez 1500 komentárov na Facebooku
  - 5 tried
  - Viacero domén (financie, obchod telekomunikácie, ...)
- Krížová validácia

Riešenie

- Nástroj na analyzovanie ľubovoľných textov
- Nasadené na webe
- Funkcie
  - Analyzovanie jednotlivých textov
  - Analyzovanie .csv a .xlsx súborov
  - Webová služba
- Vstupom klasifikátora sú jednotlivé texty
- Výstupom je najpravdepodobnejšia trieda sentimentu

Parametre programu

- Použitý klasifikátor
- Výber slovníka
- Kombinácia klasifikátorov
- Použitie diakritkovača a lematizátora
- Odstraňovanie slov na základe slovného druhu
- Označovanie záporu vo vete



Nástroj dostupný na:
https://analyza-sentimentu.herokuapp.com/

Výsledky

81,67 %

Najvyššia dosiahnutá správnosť v 2 triedovej klasifikácií – klasifikátor ME

Celkové výsledky experimentu

Triedy	2	3	4	5
Správnosť (Accuracy)	81,67 %	66,25 %	50,23 %	40,93 %
Klasifikátor	ME	ME	ME + LEX	ME

ME – klasifikátor maximálnej entropie,
ME + LEX – klasifikátor maximálnej entropie v kombinácií so slovníkovým prístupom
Baseline (náhodný klasifikátor) – 2 triedy 50 %, 3 triedy 33 %, 4 triedy 25 %, 5 tried 20 %

Zhrnutie

- Slabá podpora pre SK jazyk
- Navrhli sme riešenie pre analyzovanie sentimentu
- Experimentovanie na viac ako 1500 príspevkov
- Podobná úspešnosť ako pri riešeniach v iných jazykoch
- Dobrý výsledok nakoľko sami ľudia súhlasia so sentimentom textu na 79% (Onegva, 2010)

Podrobné výsledky

Trieda	Silne negatívny (-2)	Negatívny (-1)	Neutrálny (0)	Pozitívny (+1)	Silne pozitívny (+2)
Presnosť (P)	78,97 %		-	86,42 %	
Pokrytie (R)	91,11 %		-	69,98 %	
F1 miera	84,61 %		-	77,34 %	

Tabuľka presnosti, pokrytia a F1 miery pre 2 triedy klasifikácie

Trieda	Silne negatívny (-2)	Negatívny (-1)	Neutrálny (0)	Pozitívny (+1)	Silne pozitívny (+2)
Presnosť (P)	64,84 %		43,48 %	72,53 %	
Pokrytie (R)	86,88 %		13,07 %	69,11 %	
F1 miera	74,26 %		20,10 %	70,78 %	

Tabuľka presnosti, pokrytia a F1 miery pre 3 triedy klasifikácie

Podrobné výsledky

Trieda	Silne negatívny (-2)	Negatívny (-1)	Neutrálny (0)	Pozitívny (+1)	Silne pozitívny (+2)
Presnosť (P)	44,44 %	52,14 %	-	44,02 %	53,71 %
Pokrytie (R)	26,89 %	77,49 %	-	30,87 %	44,73 %
F1 miera	33,51 %	62,34 %	-	36,29 %	48,81 %

Tabuľka presnosti, pokrytia a F1 miery pre 4 triedy klasifikácie

Trieda	Silne negatívny (-2)	Negatívny (-1)	Neutrálny (0)	Pozitívny (+1)	Silne pozitívny (+2)
Presnosť (P)	39,60 %	41,69 %	34,38 %	37,44 %	41,90 %
Pokrytie (R)	24,79 %	69,21 %	21,57 %	28,52 %	41,45 %
F1 miera	30,49 %	52,04 %	26,51 %	32,38 %	41,67 %

Tabuľka presnosti, pokrytia a F1 miery pre 5 tried klasifikácie

Otázky

1. Pre spracovanie textu strojom je problémom nejednoznačnosť slov (pojmov) v texte. Ako by sa dal tento problém vyriešiť?

> Problém nejednoznačnosti slov by sa da vyriešiť určením významu slova použitého vo vete. Význam slova možno zistiť kontextu jeho použitia vo vete alebo v celom príspevku. Klasifikátor by následne dokázal rozlíšiť sentiment rovnako napísaných slov podľa ich použitia vo vete.

Otázky

2. Pri určovaní sentimentu sú užitočné nielen informácie o polarite slova, ale významné sú aj intenzifikátory. Ako by ste ich vedeli využiť v navrhnutom prototype?

> Intenzifikátory (slová ako „veľmi“, „strašne“, „celkom“) a slová viažuce sa k nim je možné nájsť v texte a následne označiť ako slová, na ktoré kládol pisateľ dôraz. Následne by klasifikátor pri určovaní triedy sentimentu vyhodnocoval tieto slová s vyššou váhou ako obyčajne.

Otázky

3. V analýze sa spomína zopár možností reprezentácie dokumentov. Myslíte si, že by pomohlo pri určovaní sentimentu zohľadniť aj dôležitosť slova (napr. pomocou TF-IDF alebo odstránením stop-slov)?

> TF-IDF priraďuje rozdielnu váhu slovám vo vete (najvyššiu váhu majú najmenej sa vyskytujúce). Nakoľko odstránenie stop slov je v práci implementované a v niektorých prípadoch pomáha zlepšiť výsledok, je pravdepodobné, že aj metóda TF-IDF by vylepšila úspešnosť klasifikácie.