Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta

JTDH
30. september 2016

Klemen Kadunc
Marko Robnik-Šikonja

Vsebina

- analiza mnenj: kaj, zakaj, kako
- slovenski sentimentni leksikon
- mnenjski korpus
- ovrednotenje z metodami strojnega učenja
- zaključki

Analiza sentimenta: kaj?

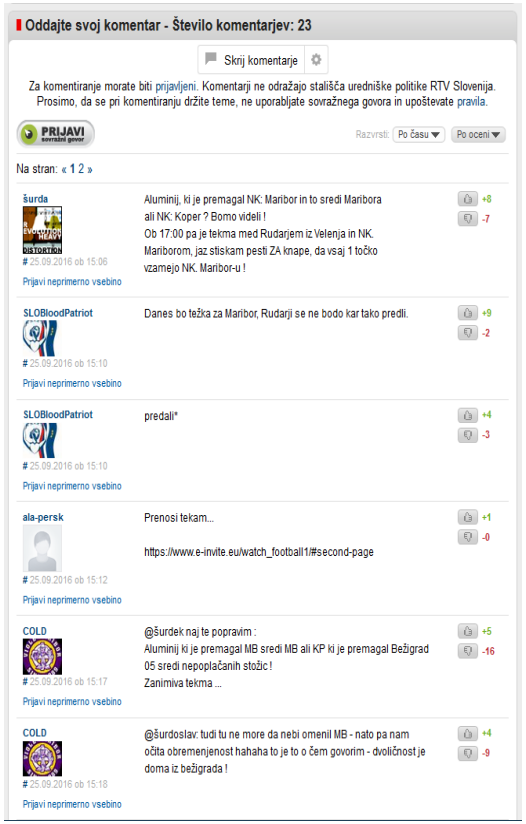
- zaznavanje mnenj v besedilih
  npr. o izdelku, dogodku, stranki
- sinonimi:
  - rudarjenje mnenj,
  - analiza mnenj,
  - rudarjenje sentimenta,
  - zaznavanje mnenj

- polarno in z lestvico
  - pozitivno, negativno, (nevtralno)

- različni nivoji:
  - dokument, odstavek, stavek
  - splošen vtis, o posamezni entiteti , osebi itd.

Analiza sentimenta: zakaj?

- velike količine uporabniško generiranih vsebin:
  - komentarji, mnenja, blogi, forumi
  - novice, tviti, objave na socialnih omrežjih
- vrsta raziskave javnega mnenja
- nujna avtomatizacija
- večina pristopov dela na angleščini, nekaj večjezičnih
- težavnost: nestandarden jezik, okrajšave, emotikoni, večpomenskost besed, različen sentiment v različnih kontekstih, negacija, sarkazem



Analiza sentimenta: kako?

- s pomočjo leksikona sentimenta (potrebno ga je imeti)
- s pomočjo tehnik obdelave naravnega jezika in strojnega učenja
- kombinirano

Sentimentni leksikoni

- seznam besed z označenim sentimentom (polarno, stopenjsko, več kategorij)
- splošni in tematski
- izdelava: ročna, (delno) avto
- več javno dostopnih za angle
  - Hu & Liu (2004), še vedn in 4783 negativnih besed
  - SentiWordNet (Baccianell dopolnitev WordNeta s se
- slovenščina:
  - Rok Martinc (2013) na pc (Nielsen, 2011), vsebuje
  - Mateja Volčanšek (2015) na podlagi seznama General Inquirer (Stone, 1997), 1669 pozitivnih in 1912 negativnih besed.

positive words	negative words
a+	2-faced
abound	2-faces
abounds	abnormal
abundance	abolish
abundant	abominable
accessable	abominably
…	…

Nov slovenski sentimentni leksikon

- na osnovi leksikona Hu & Liu (2004)
- (delno) ročni prevod, pregledan
- Nelematiziran in lematiziran: 1921 pozitivnih, 5143 negativnih besed
- prosto dostopen
- spletno orodje za dopolnjev
- slabosti
  - neupoštevanje kontekst
  - prevod ni zajel nekaterih (neformalna komunikacij
  - brezstopenjski
- dilema: leme ali vse oblike

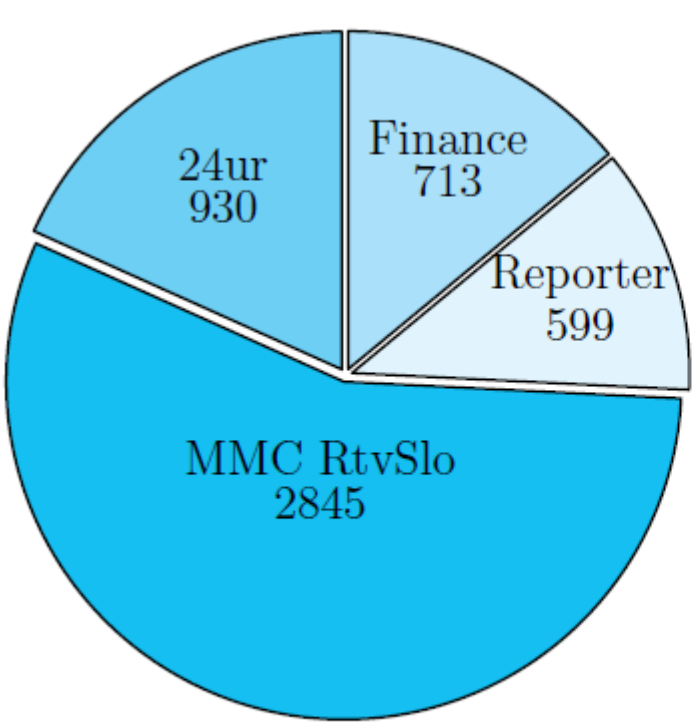
pozitivne besede	negativne besede
adut	abnormalen
aerodinamičen	absurd
agilen	absurden
agilno	absurdnost
aktualen	afektiran
ambiciozen	afnati
…	…

Korpus uporabniških komentarjev

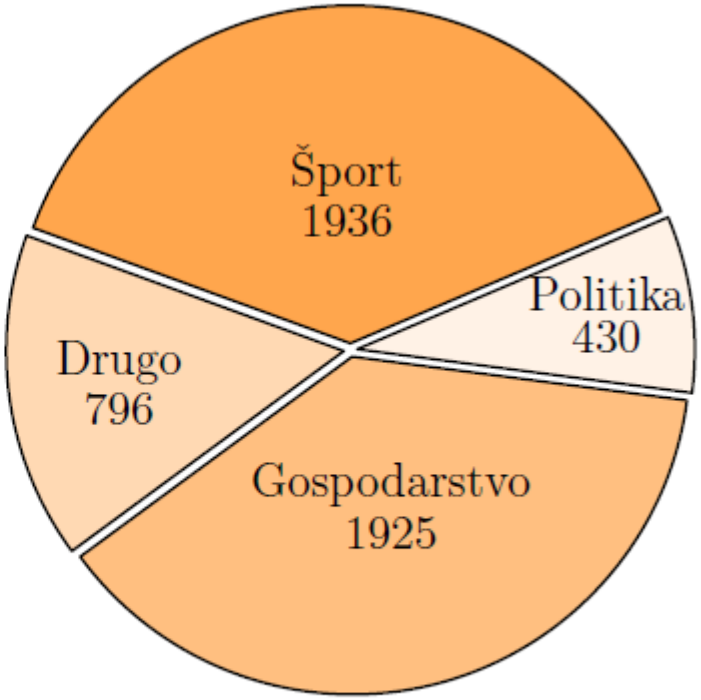
- Iskanje relevantnih komentarjev z Google API.
- Nastavitev poljubne konfiguracije za luščenje komentarjev.
- Oznake: pozitivno, negativno, nevtralno, irelevantno ter dodatno "potrebno poznavanje konteksta".
- Vsak komentar je bil označen s strani treh označevalcev.
- uravnotežen korpus 580 komentarjev vsake vrste

	gosp.	politika	šport	drugo	RtvSlo	24ur	Finance	Reporter	skupaj
pozitivno	129	26	679	64	566	255	54	23	**898**
negativno	262	33	240	53	441	48	75	24	**588**
nevtralno	1420	351	882	638	1614	584	554	539	**3291**
skupaj	**1811**	**410**	**1801**	**755**	**2621**	**887**	**683**	**586**	**4777**

Porazdelitev komentarjev



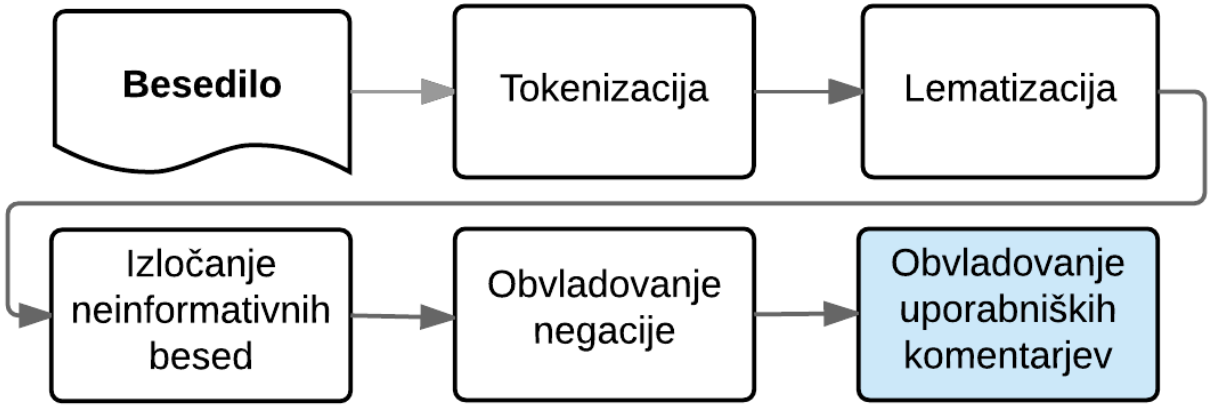
(a) Po spletnih virih



(b) Po kategorijah

Strojno učenje za analizo sentimenta

- predprocesiranje besedil



- vektorizacija besedil kot vreče besed
- klasifikatorji: logistična regresija (LR), metoda podpornih vektorjev (SVM), večvrednostni naivni Bayesov klasifikator (MNB) in binarni naivni Bayesov klasifikator (BNB)

Testni scenariji

- 10-kratno prečno preverjanje na uravnoteženem korpusu
- osnova: večinski razred 580 komentarjev po posameznih kategorijah
- testiranje predobdelave besedil: različni tokenizatorji, lematizacija, specifike spletnih komentarjev
- priprava značilk: model z n-grami, vključevanje sentimentnih leksikonov, uteževanje značilk s tf in tf-idf
- izbira značilk hi^2
- različni klasifikatorji

Uporabnost sentimentnih leksikonov

Model	LR	SVM	MNB	BNB
unigrami	61,8	59,7	63,2	48,9
unigrami + KSS	**62,9**	**60,6**	64,5	49,8
unigrami + GIS	61,5	59,5	64,4	49,4
unigrami + SWN	61,0	59,8	63,4	49,2
vsi skupaj	62,2	60,5	**65,2**	**50,3**

- razlike statistično niso značilne

Izbira najboljše metode

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova	54,5	57,6	51,5	54,3	54,5
LR	63,6	68,1	61,3	61,6	63,7
SVM	63,2	69,0	62,1	58,6	63,2
MNB	**65,5**	68,6	66,8	60,6	**65,3**
BNB	60,1	65,0	56,7	58,4	60,0

- neuravnotežen korpus: 588 nevtralnih, 898 pozitivnih, 3291 negativnih komentarjev
- večinski razred, CA = 68,9%
- MNB
  CA = 76,2%, F1 pos = 60,0%, F1 neg = 85,4%

Primerjava z angleščino

- primerjava v enakih pogojih (trirazredni sentiment, spletni komentarji)
- približno enake vrednosti CA, med 60% in 70%
- leksikoni sentimenta rahlo izboljšajo klasifikacijo

Zaključki

- izdelan je javno dostopen sentimentni leksikon primerljiv z angleškimi in orodje za njegovo vzdrževanje
- izdelana je analiza predprocesiranja, predstavitve značilk in klasifikacijskih metod za slovenščino
- možne izboljšave:
  - razdvoumljanje besed (SloWNet)
  - uporaba kakovostnih večjezičnih slovarjev
  - avtomatsko določanje sentimentnih besed v različnih kontekstih
  - obravnava negacije in oblikoslovna analiza
  - testiranje na večjem korpusu