Sentiment Annotation of Slovene User-Generated Content

Darja Fišer,[†*] Jasmina Smailović,[*] Tomaž Erjavec,[*] Igor Mozetič,[*] Miha Grčar[*]

[†]Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si
[*]Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
jasmina.smailovic@ijs.si, tomaz.erjavec@ijs.si,
igor.mozetic@ijs.si, miha.grcar@ijs.si

JTDH 2016

Overview of the talk

- Introduction
- The Janes corpus
- Automatic sentiment annotation
- The sentiment of Janes
- Evaluation
- Conclusions

Introduction

- Sentiment analysis (or opinion mining) is a type of text analysis which detects opinions, sentiments and emotions about different entities
- Useful for analysing public opinion about companies, products, political parties, stocks, etc.
- Much research, esp. due to the massive use of social on-line platforms, where people regularly express their emotions about various topics
- We report on applying a pre-trained sentiment labelling system to the Janes corpus of Slovene user-generated content (UCG)

Automatic sentiment labelling

- We use the Support Vector Machine (SVM) algorithm to train a sentiment classification model
- More precisely, we use TwoPlaneSVMbin, a 3-class extension of the basic 2-class SVM, needed to categorize texts into three sentiment classes: negative, neutral and positive
- TwoPlaneSVMbin is a combination of two binary SVM models, where one model separates the negative examples from the neutrals-or-positives while the other separates the positives from the neutrals-or-negatives

Training set

- In 2014 we acquired 100.000 Slovenian tweets during a joint project with the company Gama System
- The tweets were manually labeled as negative, neutral or positive
- Majority class used to train the sentiment classification model (marked class in case of ties)

The Janes corpus

The Janes corpus is the first large (200 mil. tokens) UGC corpus of Slovene.
The corpus is composed of the following text types and sources:

- Blog posts from 2 platforms (rtvslo, publishwall.si)
- Comments on the blog posts
- Posts from 3 forums (avtomobilizem, med.over.net, kvarkadabra)
- Comments on news articles from 3 news sites (rtvslo.si, mladina.si, reporter.si)
- Tweets of 8,749 Slovene users (2013-07 – 2015-12);
- Pagetalk and usertalk pages from the Slovene Wikipedia

Janes sentiment by text type

All texts were annotated for sentiment with TwoPlaneSVMbin:

Subcorpus	negative	neutral	positive
News comments	74%	16%	11%
Blog posts	72%	18%	11%
Blog comments	69%	16%	15%
Forum messages	55%	27%	17%
Tweets	34%	38%	28%
Wikipedia talk pages	17%	23%	60%

- Bloggers and commentators mostly express their disagreement and frustration with daily politics and other events
- Forum members and Twitter users focus more on sharing information, news and knowledge
- Wikipedia editors prioritise community building efforts with supportive, encouraging and inclusive communication

Sentiment by Key Words

We performed an analysis of 100 top-ranking key lemmas for each text type:

- Key lemmas were manually classified as positive, negative or neutral
- How well do the sentiment labels of the key lemmas predict the SVM assigned text sentiment label for each text type?
- Answer:
  - Twitter is the best (sentiment annotation was trained on tweets)
  - News and blog comments are second best (not very different from tweets)
  - Forum posts and wiki comments are next (longer, deal with specialized topics, have a different communicative purpose and target audience)
  - Blogs are worst (the most different as a text genre)

Sentiment by Part of Speech

We performed an analysis of the PoS of the 100 top-ranking key lemmas for each text type and sentiment:

- Different PoS are indicative for different sentiments:
  - verbs ← negative sentiment subcorpora
  - adjectives ← positive sentiment tweet and forum corpora
  - adverbs ← positive sentiment news and blog comments
  - adverbs ← blog comments and forum posts
  - proper nouns ← neutral sentiment subcorpora, esp. news comments
  - abbreviations ← neutral news comments
- Different linguistic means are used for communicating different sentiment:
  - negative messages are expressed directly (nouns, verbs)
  - positive messages expressed descriptively (adjectives, adverbs)
  - neutral, factual an informative content is characterized by frequent mentions of persons and their titles

Sentiment Lexica

100 top-ranking key lemma lists for each text type were used to build sentiment lexica:

- The negative sentiment lexicon contains 263 words:
  - 44% are nouns, 25% adjectives, 25% verbs, and 6% adverbs
  - the only two words that appear in all five negative sentiment subcorpora are the verb *sovražiti* (hate) and the adverb *brezveze* (nonsense)
- The positive sentiment lexicon contains 146 words:
  - 40% are nouns, 29% adjectives, 14% adverbs, 9% interjections, 7% verbs
  - the only word that appears in all five positive sentiment subcorpora is the interjection *bravo* (well done)

Evaluation of the Sentiment Scores

- Manual evaluation of the automatically assigned sentiment scores on a sample of the corpus
- Random 600 texts, 120 from each text type (40 per source)
- Each text manually assigned a sentiment score by 3 annotators
- The annotators could mark a text as out of scope: final evaluation sample = 555 texts
- Krippendorff's Alpha (OK $>$ 0.4):

	All	Wiki	News	Blog	Forum	Tweet
Humans	0.563	0.464	0.513	0.594	0.464	0.547
Auto-major	0.432	0.402	0.394	0.446	0.245	0.372
n	555	107	115	115	119	99

Conclusions

- We presented a sentiment classification system trained on Slovene tweets and its application on the Janes corpus of Slovene user-generated content
- The analysis of sentiment-specific keywords gives interesting insight into the vocabulary that is typically used to express different sentiment
- Evaluation results show that automatic sentiment classification is consistent with human judgements and that there are considerable differences among the performance of the system across the text types
- The sentiment annotation accuracy could still be significantly improved but the current annotation is already useful for e.g., selecting only those texts that have predominantly negative, neutral or positive sentiment and performing on them targeted linguistic analyses.