Analysis of Sentiment Labelling of Slovene User-Generated Content

Darja Fišer[†*] and Tomaž Erjavec[*]

* Department of Translation, University of Ljubljana

† Department of Knowledge Technologies, Jožef Stefan Institute

E-mail: darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si

CMC 2016 Ljubljana

Overview

1. Introduction
2. Sentiment annotation of the Janes corpus
3. Quantitative analyses
4. Qualitative analyses
5. Conclusions

Introduction

- Sentiment analysis: a popular text-mining task, especially for social networking services
- A sentiment analysis system for Slovene user-generated content (UGC) was developed by Mozetič et al. (2016)
- It has been also used to annotate the Janes corpus of Slovene UGC
- Results vary both in inter-annotator agreement and accuracy of the system across genres → further improvements of the system are needed
- One of the steps towards this goal is a *qualitative analysis of (dis)agreement among the annotators and error analysis of the incorrectly classified texts*

The Janes corpus

- The Janes corpus is the first large (215 million tokens) corpus of Slovene UGC
- It comprises blog posts and comments, forum posts, news comments, tweets and Wikipedia talk and user pages
- Linguistic annotation: tokenisation, rediacritization, normalisation, sentence segmentation, tagging and lemmatisation
- Also: text standardness labelling

Janes sentiment labelling

- The texts in the corpus were annotated for sentiment: negative, positive, or neutral
- The annotation was performed with a SVM-based algorithm
- SVM was trained on a large collection of manually annotated Slovene tweets (Mozetič et al., 2016)
- But these tweets are not available (company financing)

- We produced a manually annotated dataset of 600 texts
- Sampled in equal proportions from each subcorpus
- The sample was manually annotated by 3 annotators
- Some texts marked as out of scope:
  the final evaluation sample consists of 557 texts

Manual vs. SVM annotations

- Krippendorff's α = 0.563 for humans, 0.432 for SVM
- 3 agreed on ~50%, 2+ annotators agreed on ~97%
- SVM agreement given manual agreement:

Manual / SVM	3 annotators agree		2/3 annotators agree		0 annotators agree	
identical	160	65%	133	46%	6	33%
different	87	35%	159	54%	12	67%
total	247	44%	292	52%	18	3%

Difficulty of text genres

- Taking texts with perfect human agreement, the best SVM results are on news comments, followed by blog posts
- Observing texts with no human agreement, the least problematic are Wikipedia talk pages and news comments, worst are forum posts and tweets (sic!)
- Possible reasons:
  - sentiment more explicitly expressed in news comments than in tweets
  - blogs might be easier because they are longer

Types of SVM errors

Investigating texts that received the same label by 3 annotators but a different one by SVM:

- Automatic system has a bias to the „neutral“ label: ~50% of mislabelled texts were marked as neutral by the algorithm
- Mislabelling neutral texts as opinionated: ~33%
- Worst case: negative texts labelled as positive or vice versa: only 12%

Qualitative analysis:

Toughest problems for humans

Observe texts which received a different label by each annotator:

- Annotators 1 and 2 chose positive and negative labels equally frequently but Annotator 3 was heavily biased towards the neutral class
- This suggests that despite receiving the same guidelines annotators adopted different strategies in selecting the labels systematically throughout the assignment
- Need for more precise annotation guidelines
- SVM shares the most equal votes with Annotator 1 (44%) and the fewest with Annotator 2 (22%)

Qualitative analysis: Toughest problems for SVM

Observed texts which received the same label by all annotators but a different one by the system:

- ~25%: no special feature was identified, it is not clear why the system made an error, as the sentiment is obvious
- ~43%: lexical features, most likely OOV for the model
- The rest: quotes, parts of discussion threads, fragmentary, truncated messages, URL links, emoticon and emojis, cynical texts and texts with mixed sentiment

Conclusions

- Quantitative and qualitative analysis of sentiment annotation of the Janes corpus, to enable better understanding of the task of sentiment annotation in general and facilitate improvements of the system in the future
- Main observations:
  - blogs are easiest
  - tweets and forum posts are much harder for the system than for humans
- Planned improvements:
  - provide annotators with more comprehensive guidelines
  - provide the automatic system with training data from the worst performing text types
  - less clear how to improve automatic labelling of sarcastic, ironic and cynical tweets