Analysis of Sentiment Labeling of Slovene User-Generated Content

Darja Fišer,[*†] **Tomaž Erjavec**[†]
* Department of Translation, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana
† Department of Knowledge Technologies, Jožef Stefan Institutute, Jamova cesta 39, 1000 Ljubljana
E-mail: darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si

Abstract
The paper takes a close look at the results of sentiment annotation of the Janes corpus of Slovene user-generated content on 557 texts sampled from 5 text genres. A comparison of disagreements among three human annotators is examined at the genre as well as text level. Next, we compare the automatically and manually assigned labels according to the text genre. The effect of text genre on correct sentiment assignment is further investigated by investigating the texts with no inter-annotator agreement. We then look into the disagreements for the texts with full human inter-annotator agreement but different automatic classification. Finally, we examine the texts that humans and the automatic model struggled with the most.

Keywords: sentiment analysis, quantitative and qualitative evaluation, user-generated content, non-standard Slovene

1. Introduction

Sentiment analysis or opinion mining detects opinions, sentiments and emotions about different entities expressed in texts (Liu, 2015). It is currently a very popular text-mining task, especially for social networking services, where people regularly express their emotions about various topics (Dodds et al., 2015). A sentiment analysis system for Slovene user generated content (UGC) was developed by Mozetič et al. (2016) and has been, *inter alia*, used to annotate the Janes corpus of Slovene UGC (Erjavec et al., 2015). The first results are encouraging but the results vary both in inter-annotator agreement and accuracy of the system across genres (Fišer et al., 2016), suggesting 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 an error analysis of the incorrectly classified texts, which is the goal of this paper.

The paper is organized as follows. In Section 2 we give a brief presentation of the corpus and its sentiment annotation. In Section 3 we present the results of a quantitative analysis of manual and automatic sentiment annotation on a sample collection of texts. In Section 4 we follow with a qualitative analysis of the texts and their features that make the task difficult for humans as well as those that the algorithm struggles with. The paper ends with concluding remarks and ideas for future work.

2. Sentiment Annotation of Janes

The Janes corpus (Erjavec et al., 2015) is the first large (215 million tokens) corpus of Slovene UGC that comprises blog posts and comments, forum posts, news comments, tweets and Wikipedia talk and user pages. Apart from the standard corpus processing steps, such as tokenization, sentence segmentation, tagging and lemmatization (Ljubešić and Erjavec, 2016) as well as some UGC-specific processing steps, such as rediacritization (Ljubešić et al., 2016), normalization (Ljubešić et al., 2014) and text standardness labeling (Ljubešić et al., 2015), all the texts in the corpus were also annotated for sentiment (negative, positive, or neutral) with a SVM-based algorithm that was trained on a large collection of manually annotated Slovene tweets (Mozetič et al., 2016).

We also produced a manually annotated dataset. This evaluation dataset comprised 600 texts, which were sampled in equal proportions from each subcorpus (apart from blog comments as they have been found to behave very similar to news comments) in order to represent all the text genres included in the corpus in a balanced manner.

The sample was then manually annotated for the three sentiment labels by three human annotators. The annotators marked some texts as out of scope (written a foreign language, automatically generated etc.), so the final evaluation sample consists of 557 texts.

In the following sections the labels assigned by the annotators were compared to each other while the automatically assigned scores were compared to the annotators' majority class, i.e. the sentiment label assigned to each text by the most annotators. In cases of complete disagreement the neutral sentiment is assigned as the majority class.

3. Quantitative Analysis of Sentiment Annotation

In our quantitative analysis we first analyze the difficulty of the task for humans and the algorithm on the evaluation sample. We also compare annotation results with respect to text genres. Finally, we measure the degree of disagreements of the assigned labels in order to measure the severity of the annotation incongruences.

3.1. Comparison Between Manual and Automatic Annotations

First, a comparison of disagreements among the human annotators was computed as well as that of the automatic system with the majority class. Since we are investigating sentiment annotation accuracy from the perspective of the difficulty of the task, measured with the dispersion of annotations by human annotators, we are operating with percentage agreement in this paper. While we have measured inter-annotator agreement with Krippendorff's alpha, which is 0.563 for human annotations and 0.432 for automatic annotations with respect to the human majority vote (cf. Fišer et al., 2016), this measure reports inter-annotator agreement for the entire annotation task and is as such not informative enough for the task at hand in this paper.

The results in Table 1 shows that the task was easier for some texts in the sample both for humans and for the system as annotators' labels range from perfect agreement to an empty intersection. While at least two human annotators provided the same answer on nearly 97% of the sample, all three annotators agreed on less than half of the texts, which is a clear indication that the task is not straightforward and intuitive for humans, suggesting that better guidelines and/or training are needed to obtain consistent and reliable results in the annotation campaign. As could be expected, texts that were difficult to annotate for humans also proved hard for the system. Namely, the system chose the same label as the annotators in the majority of the cases (65 %) only for those that humans were in complete agreement. Where the annotators disagreed partially or completely, there is substantially less overlap with them and the system (46% - 33%).

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

Table 1: Comparison between automatic (to majority class) and manual annotations.

3.2. Comparison Between Text Genres

In order to better understand which text types are easy and which difficult for sentiment annotation, we compared the labels assigned by the annotators and the system according to the genre of the texts in the sample. In texts for which annotators are in perfect agreement, the biggest overlap between the system and the majority vote of the annotators is achieved on news comments. These are followed by blog posts which, together with the news comments, represent over half of all the texts receiving the same sentiment label by both humans and the model.

The effect of text genre on the difficulty of correct sentiment assignment was further investigated by looking at the genre of those texts for which there was no agreement among the human annotators, i.e. texts which were annotated as negative by one annotator, positive by another and neutral by the third. The results of this analysis are presented in Table 3 and are consistent with the previous findings in that sentiment in forum posts and tweets is the most elusive while being the least problematic on Wikipedia talk pages and in news comments.

	All annotators agree				*2/3 annotators agree*				*All annotators disagree*			
	Different		*Identical*		*Different*		*Identical*		*Different*		*Identical*	
Type	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
blog	14	16	34	21	38	24	27	20	3	25	1	17
forum	23	**26**	29	18	42	**26**	20	15	4	**33**	1	17
news	12	14	48	**30**	21	13	32	**24**	2	17	0	0
tweet	23	**26**	24	15	25	16	21	16	2	17	4	**67**
wiki	15	17	25	15	33	21	33	24	1	8	0	0
total	87		160		159		133		12		6	18

Table 2: Comparison between automatic and majority vote per text genre.

Since the system was trained on tweets, one would expect them to receive the highest agreement, which is not the case. A possible reason for this is that sentiment is more explicitly expressed in news comments than in tweets, whereas blogs might be easier because they are longer which again makes them easier for sentiment identification. Forum posts, on the other hand, seem to be the hardest overall, which is addressed in more detail in Section 4.

Text type	*Disagreement*	
blog	4	21%
forum	6	**32%**
news	2	11%
tweet	6	**32%**
wiki	1	5%
total	19	100%

Table 3: Disagreement among the annotators per text genre.

3.3. Comparison of the Degree of Disagreements

Since not all incongruences between the system and the true answer are equally bad from the application point of view, we looked into the degrees of disagreements for the texts receiving the same label by all three annotators and a different one by the system. As can be seen from Table 4, the automatic system has a clear bias towards neutral labels, i.e. more than half of the mislabeled opinionated texts were marked as neutral by the algorithm. Mislabeling neutral texts as opinionated is seen in about a third of the cases. The worst-case scenario, in which negative texts are labeled as positive or vice versa and therefore hurts the usability of the application the most, is quite rare (12%). The behavior of the system on texts with partial human agreement is consistent with the findings above in assigning sentiment of opposite polarities which again represents the smallest part of the sample (8%). Neutralizing negative and positive texts occurs on 40% of the sample, which is slightly lower than for the texts on which all the annotators agree. The most prevalent category are neutral texts mislabeled as negative which is seen in 34% of the cases, substantially more than above.

Differences	*Annotators agree, system disagrees*		*2/3 annotators agree, system disagrees*	
neg → neut	29	**33%**	36	23%
neg → pos	7	*8%*	7	*4%*
neut → neg	14	16%	54	**34%**
neut → pos	13	15%	29	18%
pos → neut	20	**23%**	6	17%
pos → neg	4	*5%*	27	*4%*
Total	87	100%	159	100%

Table 4: Discrepancies between automatic and majority human vote.

4. Qualitative Analysis of Sentiment Annotation

In this section we present the results of qualitative analysis of the biggest problems in sentiment annotation observed in the evaluation sample. We first examine all the texts for which there was no agreement among the human annotators and then focus on the texts that humans found easy to annotate consistently but the system failed to annotate correctly.

4.1. Toughest Sentiment Annotation Problems for Humans

By examining the texts which received a different label by each annotator we wished to investigate the difficulty of the task itself, regardless of the implementation of an automatic approach. In the evaluation sample of 557 texts there were 18 such cases: 6 tweets, 5 forum posts, 4 blog posts, 2 news comments, and 1 Wikipedia talk page.

As can be seen from Table 5, there are significant discrepancies in annotator behavior. While Annotators 1 and 2 chose positive and negative labels equally frequently (A1: 9 negative, 8 positive, 1 neutral; A2: 8 negative, 8 positive, 2 neutral). Annotator 3 was heavily biased towards the neutral class (A3: 1 negative, 2 positive, 15 neutral). The automatic system lies in between these two behaviors (S: 5 negative, 7 positive, 6 neutral), sharing the most equal votes on individual texts with Annotator 1 (44%) and the fewest with Annotator 2 (22%). This suggests that annotators did not pick different labels for individual texts due to random/particular mistakes but probably adopted different strategies in selecting the labels systematically throughout the assignment. While Annotators 1 and 2 favored the expressive labels even for the less straightforward examples, Annotator 3 opted for a neutral one in case of doubt. These discrepancies could be overcome by more precise annotation guidelines for such cases.

Source	*Ann1*	*Ann2*	*Ann3*	*System*	*Note*
blog	-	+	0	-	mixed
blog	-	+	0	-	mixed
blog	-	+	0	-	mixed
blog	+	-	0	0	mixed
forum	-	+	0	+	mixed
forum	+	-	0	-	context
forum	+	-	0	0	context
forum	+	-	0	+	context
forum	+	0	-	+	context
news	-	+	0	+	context
news	+	-	0	-	mixed
tweet	-	+	0	0	sarcasm
tweet	-	+	0	0	mixed
tweet	-	+	0	0	mixed
tweet	0	-	+	0	short
tweet	+	-	0	+	mixed
tweet	+	-	0	+	sarcasm
wikip.	-	0	+	+	mixed

Table 5: Analysis of the difficult cases for the human annotators.

A detailed investigation of the 18 problematic texts showed that 3 out of 6 tweets contain mixed sentiment in the form of message and vocabulary distinctive for one sentiment, which is then followed by an emoticon of a distinctively opposite sentiment. 2 tweets were sarcastic and 1 simply too short and informal to understand what the obviously opinionated message was about ("*prrrr za bič :P / prrrr for the whip :P*").

4 out of 5 forum posts are lacking a wider context (the entire conversation thread) which is needed in order to find out whether the post was meant as a joke or was sarcastic. Some annotators annotated it as is, others assumed sarcasm or opted for a neutral label. 1 forum post contained mixed sentiment.

All 4 blog posts were relatively long and contained mixed sentiment. For example, a post that contains a description of a blogger's entire life starts off with very positive sentiment that then turns into a distinctly negative one after some difficult life situations. While some annotators treated this text as neutral as it contained all types of sentiment, others treated it as negative since negative sentiment is the dominant one in terms of amount of text it appears in with respect to other parts, in terms of strength with which it is expressed, and/or in terms of the final position in the text, suggesting it to be the prevailing sentiment the author wished to express.

1 news comment was lacking context and 1 contained mixed sentiment, which is also true with the Wikipedia talk page that is complaining about a plagiarized article but in a clearly constructive, instructive tone that is trying not to complain about the bad practice but teach a new user about the standards and good practices respected by the community.

4.2. Toughest Sentiment Annotation Problems for Computers

In the second part of the qualitative analysis we focus on the 87 texts from the sample which were labeled the same by all three annotators but differently by the system. With this we hope to see the limitations of the system when trying to deal with the cases most straightforward for humans. The sample consisted of 23 forum posts and 23 tweets, 15 Wikipedia talk pages, 14 blog posts and 12 news comments. As said in Section 3, almost all of the discrepancies (87%) were neutral texts that were mislabeled as opinionated by the system or vice versa. Serious errors, i.e. cross-spectrum discrepancies were rare (4.6% true negatives mislabeled as positive and 8% true positives mislabeled as negative).

Problematic feature	*No.*	*%*
no feature identified	22	25.29
neg. vocabulary	18	20.69
+ vocaulary	10	11.49
cynical	10	11.49
emoticons	7	8.05
too short	5	5.75
quote	5	5.75
foreign/specialized vocabulary	5	5.75
non-standard text	2	2.3
names	2	2.3
mixed sentiment	1	1.15
Total	87	100.00

Table 6: Analysis of the problematic text features for the sentiment annotation algorithm.

We performed a manual inspection of the erroneously annotated texts and classified them into one of 10 the categories representing possible causes for the error. As Table 6 shows, in over a quarter of the analyzed texts, no special feature was identified and it really is not clear why the system made an error there as the sentiment in them is obvious. The most common characteristics of the mislabeled texts, which occurred in 43% of the analyzed sample, were lexical features, i.e. the vocabulary typical of negative/positive messages, foreign and specialized vocabulary, proper names and non-standard words that are most likely out-of-vocabulary for the model and therefore

cannot contribute to successful sentiment assignment. E.g. a perfectly neutral discussion on Wikipedia was labeled as negative due to the topic of the conversation (*quisling, invader, traitor*). Similarly, many posts with objective advice to patients on the medical forum which contain a lot of medical jargon were mislabeled as negative.

The second common source of errors were the inter- and hyper-textual features that are typical of user-generated content, such as quotes from other sources, parts of discussion threads, fragmentary, truncated messages, URL links and emoticon and emoji symbols. The remaining issues include cynical texts and texts with mixed sentiment that have already been discussed in Section 4.1.

5. Conclusions

In this paper we presented the results of a quantitative and qualitative analysis of sentiment annotation of the Janes corpus. These insights should enable better understanding of the task of sentiment annotation in general as well as facilitate improvements of the system in the future. The results of the first analysis show that overall, blogs have proven to be the easiest to assign a sentiment to as both humans and the automatic assignment achieve the highest score here. The sentiment of the blog posts we examined was straightforward to pin down by the annotators due to text length and informativeness, through which it becomes clear which sentiment is expressed by the author.

For humans, the second easiest are tweets, whereas the automatic system preforms worse on them than on news comments and Wikipedia talk pages. This is especially interesting as the automatic system was trained on tweets and would therefore be expected to perform best on the same type of texts. A detailed examination of the problematic tweets shows they are extremely short, written in highly telegraphic style or even truncated and therefore do not provide enough context to reliably determine the sentiment. Furthermore, messages on Twitter are notoriously covertly opinionated, often sarcastic, ironic or cynical, making it difficult to pin down the intended sentiment.

The results of the second analysis are consistent with the first in that texts which contain vocabulary that is typically associated with a particular sentiment but used in a different context or communicative purpose makes the sentiment difficult to determine. As for the forum posts which are much harder for the system to deal with than for humans, highly specialized vocabulary on the medical, science and automotive forums (which in addition to terminology is full of very non-standard orthography and vocabulary) would most likely be beneficial in the training data for the model to learn on. Based on the analysis reported on in this paper, we plan to improve inter-annotator agreement by providing the annotators with more comprehensive guidelines that will inform the annotators about how to treat the typical problematic cases. We will try to improve the automatic system by providing it with training material from the worst performing text types. It is less clear how to improve the quality of the automatic labeling of sarcastic, ironic and cynical tweets that are a very common phenomenon.

6. Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842. 2014-2017).

7. Bibliography

Sheridan Dodds, P., Clark, E. C., Desu, S., Frank, M. R., Reagan, A. J., Ryland Williams, J., Mitchell, L., Decker Harris, K., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F. and Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proc. of the National Academy of Sciences*. 112(8): 2389–2394.

Erjavec, T., Fišer, D., and Ljubešić, Nikola (2015). Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. *Zbornik konference Slovenščina na spletu in v novih medijih*. 20–26. Ljubljana, Znanstvena založba Filozofske fakultete.

Fišer, D., Smailović, J., Erjavec, T., Mozetič, I., and Grčar, M. (2016). Sentiment Annotation of Slovene User-Generated Content. Proc. of the *Conference Language Technologies and Digital Humanities*. Ljubljana, Faculty of Arts.

Kilgarriff, A. (2012). Getting to Know Your Corpus. *Proc. of 15th International Conference on Text. Speech and Dialogue (TSD’12)*. Brno, Czech Republic. September 3-7 2012, 3–15, Springer Berlin Heidelberg.

Liu, B. (2015). *Sentiment Analysis: Mining Opinion,. Sentiments, and Emotions.* Cambridge University Press.

Ljubešić, N., Erjavec, T., and Fišer. D. (2014). Standardizing tweets with character-level machine translation. *Computational Linguistics and Intelligent Text Processing*. LNCS 8404, 164–175, Springer.

Ljubešić, N., Erjavec, T., (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proc. of 10th International Conference on Language Resources and Evaluation (LREC’16).* European Language Resources Association (ELRA).

Ljubešić, N., Erjavec, T., and Fišer, D. (2016). Corpus-Based Diacritic Restoration for South Slavic Languages. *Proc. of 10th International Conference on Language Resources and Evaluation (LREC’16)*. European Language Resources Association (ELRA).

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. *Proc. of 10th International Conference on Recent Advances in Natural Language Processing Conference (RANLP'15). 7–9 September 2015*, 371–378. Hissar. Bulgaria.

Martineau, J., and Finin, T., (2009). Delta TFIDF: An improved feature space for sentiment analysis. *Proc. of 3rd AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, 258–261.

Mozetič, I., Grčar, M., and Smailović, J., (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE*. 11(5):e0155036.

Vapnik, V. N., (1995). *The Nature of Statistical Learning Theory*. Springer.