Ph.D. Thesis Review:
"Aspects of Sentiment Analysis"
By Ing. Tomáš Hercig
Review by Alexandra Balahur, Ph.D.

The thesis is devoted to the development and evaluation of different methods and resources for sentiment analysis at different levels and with a special regard towards the treatment of inflected languages, of which Czech was the focal point. Most of the research in automatic sentiment analysis of social media has been performed in English and Chinese, therefore treating low-resourced languages is a noteworthy achievement.

"Aspects of Sentiment Analysis" is an easy to read thesis, all notions related to the topic are clearly defined and explained. The theoretical foundations are thoroughly presented, each task is well defined and relevant examples are given. The same clarity and thoroughness is also reflected with regard to the advantages and disadvantages of each method or algorithm and the situations of where there are suitable to use. The theoretical basis is enriched also in Chapter 4, where the most relevant feature selection. I particularly appreciated the presentation of the theoretical basis of the thesis, as it made the manuscript very fluid to read and self-explanatory.

The state-of-the-art summary is up-to-date, it is complete and very thoroughly researched. The best approaches are presented, in a critical manner, showing the depth of understanding of the subject of the thesis as well as giving appropriate course of action to further research that is taking the insights presented into account.

The thesis contains an extensive comparative evaluation of sentiment analysis methods, using distributional semantics and different machine learning algorithms, settings and features, for aspect based and at document level.

Chapter 7 covers in-depth research on machine learning methods for document-level sentiment analysis of Czech social media and Chapter 9 presents the first attempt at using neural networks for sentiment analysis in Czech. The variety of settings compared is impressive, in both cases. I would have liked though to see these two chapters combined, and a comparison done between the use of "standard" versus neural-network based learning.

Chapter 8 describe the approaches to the aspect-based sentiment analysis task in Czech and English. It is interesting to observe the differences between the two languages and the author has succeeded fully in capturing the differences in performance due to the peculiarities of the two languages. Here, it is important to note the creation of two new Czech corpora within the restaurant domain for the ABSA task: one labelled for supervised training, and the other unlabelled for unsupervised training, which are available to the research community.

Chapter 10 describes an approach to sarcasm detection in Czech and English. Here, although the features used are not new, it was interesting to note the importance of language-specific processing for Czech.

All in all, the minor comments I have are related to a much needed discussion of the large quantity of results obtained throughout the thesis. Apart from the impressive amount of experiments that have been carried out, I would have appreciated a more in-depth analysis of the results and an explanation of why certain algorithms work better and under what circumstances.

Overall, all goals of the thesis have been fulfilled. The results are backed by solid publications in top scientific venues and the high ranks obtained in the international competitions where the systems presented in the thesis participated, as well as linguistic resources that can be further exploited by research in sentiment analysis, especially for Czech.

In summary, the author of the thesis proved to have the ability to perform scientific work and to achieve meaningful results, as well as publish these results in relevant venues. I do recommend the thesis for presentation with the aim of receiving the Ph.D. degree.

Ispra, 16 January, 2018

Alexandra Balahur, Ph.D.
European Commission Joint Research Centre
Text and Data Mining Unit (I.3)

Department of Cybernetics
Faculty of Applied Sciences
University of West Bohemia
Univerzitni 8, 306 14 Pilsen, Czech Republic
Tel.: +420 377 63 2508, +420 377 63 2523

Prof. Ing. Luděk Müller, Ph.D.
Fax: +420 377 63 2502
E-mail: muller@kky.zcu.cz; muller@ntis.zcu.cz
http://ui.kky.zcu.cz; http://www.ntis.zcu.cz/en

Review of the Doctoral Thesis of Ing. Tomáš Hercig "Aspects of Sentiment Analysis"

As follows from the title, the thesis is devoted to application of machine analysis, machine learning and evaluation techniques with the aim to improve the state-of-the art in automatic sentiment analysis.

This Thesis goal is well motivated due to the increasing demand of machine tools dealing with human natural language in an automatic manner.

The main focus of the work lies in the following three research tasks constituting the goals of the Thesis:

1. Deal with specific properties of Czech language in the sentiment analysis environment.
2. Use additional semantic and/or syntactic information to improve sentiment analysis.
3. Explore the influence of figurative language (e.g. sarcasm) on sentiment analysis.

The Thesis is well organized and clearly presented: After a short introduction, a good theoretical background survey is given in Chapter 2. The original contributions of the Thesis are presented in the subsequent Chapter 3 which also summarizes the work outcomes, evaluates the thesis goals, and gives the conclusion and recommendations for future work.

The author demonstrated extensive knowledge of the state of the art in the research field the Thesis is dealing with and he has the scientific ability to further contribute to the research in this research area. The experimental results validate the suitability of methods proposed by the author.

All goals of the Thesis were fulfilled.

Moreover, the contribution of the Thesis exceeds the original Thesis goals:

- The author's team created several large-scale labeled corpora: (Facebook posts, movie reviews, product reviews, two Czech corpora within the restaurant domain for the aspect-based sentiment analysis task: one labeled for supervised training, and the other unlabeled for unsupervised training, Czech sarcasm corpus of manually-labeled tweets, a large English corpus automatically labeled tweets).

- Several feature selection algorithms where evaluated and compared.

- The influence of named entity recognition on sentiment analysis was investigated.

- The first application of artificial neural networks for sentiment analysis in Czech was introduced.

- The first automatic detection of sarcasm in Czech was evaluated with results outperforming the state-of-the-art methods.

- The experiments with how the figurative language affects the sentiment analysis outcomes were design and conducted.

Furthermore, I highly appreciate the state-of-the-art results achieved by the author in:

- SemEval tasks (namely the Task 11 SemEval-2015, sentiment analysis task of SemEval 2016 in nine experiments among the constrained systems and in two experiments among the unconstrained systems),

- Czech aspect-based sentiment analysis with word clusters created using semantic models.

To have no doubt that the Thesis fulfills all partial requirements for granting Ing. Tomáš Hercig the Doctoral degree I would like also to pose several questions and comments that (or most important ones of them) I hope could be answered by the author during the thesis defense:
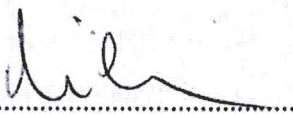
1. The kernel used in SVM is a linear one, so the comparison with MaxEnt may not play in its favor. What was the reason of using this kind of kernel?

2. On page 92 the author states he applied: "Static Word2Vec [Mikolov et al., 2013a] vectors". However, there two approaches: Skip-gram and CBOW. Which of them was used?

3. Page 82: What are BoG features where the groups correspond to various sentiment lexicons? (The BoG acronym is missing in the list of semantics features).

4. I again would highly appreciate if ALL abbreviations/acronyms used in the text were defined in a list of abbreviations.

5. In addition to the missing list of abbreviations, some parts of the work are not thoroughly explained and sufficiently described. For example, in Section 7.33. "Feature Selection Experiments" I could not explicitly found the result of the feature selection process. The reader can only guess on the basis of description under the Table 7.7 that FS1 = Unigrams; FS2 = Uni + bigrams; FS3 = Uni + big + POS features; FS4 = Uni + big + POS + emot. Analogously for Figures 7.5 and 7.6.

6. Similarly, the author states that "Overall, a significant improvement from 73.38% (baseline) to 73.85% was achieved", but what is the baseline is not stated in the text (unigram feature without preprocessing?). Page 61.

7. The English corpus collected automatically for Sarcasmus detection contains "#sarkasmus" hashtags. However, probably not all tweets although they are sarcastic are featured with this tag. How did the author deal with this challenge?

8. The description of the structure of CNN 2 is confusing (no definition of l; l_k; k is probably not k: the length of embedding vectors is $k = 3$?). Moreover, Figure 9.2 is probably inconsistent with the CNN's structure description on the previous page (e.g. number of kernels 32 in the text vs. the depicted part of the convolution layer in the picture, the variable $n,...$). Pages 93-94.

In conclusion, in my opinion, the Thesis meets the requirements of a scientific work; the subject of the thesis is relevant to the current needs of the scientific community; the methods proposed by the author are adequate, lead to valuable results and can significantly contribute to the research field tackling with the problem of automatic sentiment analysis.

The significance of the contribution of the Thesis is supported by the fact that some of the results were published in several peer-reviewed journals.

In summary, I declare that the author has proven the ability to perform scientific work and to achieve meaningful results. So, I do recommend the thesis for presentation with the aim of receiving the Ph.D. degree.

Plzeň, February 16, 2018

Prof. Ing. Luděk Müller, Ph.D.
University of West Bohemia

2