# Why Words Alone Are Not Enough: Error Analysis of Lexicon-based Polarity Classifier for Czech

**Kateřina Veselovská**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
veselovska@ufal.mff.cuni.cz

**Jan Hajič, jr.**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
hajicj@ufal.mff.cuni.cz

## Abstract

Lexicon-based classifier is in the long term one of the main and most effective methods of polarity classification used in sentiment analysis, i.e. computational study of opinions, sentiments and emotions expressed in text (see Liu, 2010). Although it achieves relatively good results also for Czech, the classifier still shows some error rate. This paper provides a detailed analysis of such errors caused both by the system and by human reviewers. The identified errors are representatives of the challenges faced by the entire area of opinion mining. Therefore, the analysis is essential for further research in the field and serves as a basis for meaningful improvements of the system.

## 1 Introduction

After finishing the initial phase of our research in the area of sentiment analysis in Czech during which the collected data resources were manually annotated, we attempted to train two classifiers for automatic polarity detection of a given text: the lexicon-based classifier and the Naive Bayes classifier. Both systems were trained on two different types of the data (see Section 3). As shown in Table 1, the Naive Bayes classifier was consistently outperformed by the primary lexicon-based one (denoted as PC in the table), which on the less complicated data performed comparably to state-of-the-art, see Cui et al. (2006). Acc, R, P and F stand for accuracy, recall, precision and F-measure, respectively.

| Model | Acc | R(-) | P(-) | F(-) | R(+) | P(+) | F(+) | R | P | F |
|---|---|---|---|---|---|---|---|---|---|---|
| baseline | 0.630 | 0 | 0 | 0 | 1 | 0.630 | 0.773 | 0.370 | 0.233 | 0.286 |
| PC, train | 0.960 | 0.964 | 0.935 | 0.949 | 0.958 | 0.977 | 0.967 | 0.960 | 0.961 | 0.960 |
| PC, test | 0.889 | 0.907 | 0.821 | 0.862 | 0.878 | 0.939 | 0.908 | 0.889 | 0.894 | 0.890 |
| Bayes,train | 0.864 | 0.717 | 0.901 | 0.798 | 0.955 | 0.849 | 0.899 | 0.803 | 0.879 | 0.833 |
| Bayes, test | 0.827 | 0.630 | 0.872 | 0.730 | 0.947 | 0.811 | 0.874 | 0.745 | 0.847 | 0.781 |

**Table 1. Baseline, comparing performance on training and test data**

We will briefly describe the system below in Section 4. The results are discussed in detail in Veselovská (2012).

## 2 Related Work

The very first stage of the project has been described in Veselovská et al. (2012). Closely related work using methods that analyze sentiment on a deep level is done by Polanyi and Zaenen (2004), who consider the role of lexical and discourse context of the attitudinal sentences. The importance of discourse, namely interaction between opinions, is also emphasized by Johansson and Moschitti (2013), who demonstrate that relational features, mainly derived from dependency-syntactic and semantic role structures, can significantly improve the performance of automatic systems for a number of fine-grained opinion analysis tasks. There is a number of papers dealing with sentiment analysis from the point of view of compositional semantics. Whereas Choi and Cardie (2008) show that simple heuristics based on compositional semantics can perform better than learning-based methods that do not incorporate compositional semantics, Moilanen and Pulman (2007) explain sentiment classification of grammatical constituents in quasicompositional way. Some work on sentiment analysis in Czech has been also done by Habernal et al. (2013), but so far no authors provided error analysis of Czech polarity classifiers.

## 3 Data

Since our initial motivation was to create a tool for detecting the way news articles might influence public opinion, we firstly worked with the data obtained from the Home section of the Czech news website Aktualne.cz (*http://aktualne.centrum.cz/*) – or more precisely, with the articles primarily concerned with domestic politics, namely the situation before the elections in 2010. Unfortunately, it turned out that the analysis of such texts was a rather difficult task in terms of automatic processing, because Czech journalists mostly avoid strongly evaluative expressions. Moreover, the corpus was not large enough for a full-scale evaluation, as it contained merely 410 segments of texts (6,868 words, 1,935 unique lemmas) which were manually annotated on polarity. Also, the language we were dealing with was not straightforward. Furthermore, the distribution of polarity classes over segments was very nonuniform, with neutral segments occupying 78% of the data and positive segments making up less than 5%. Given the small size of the data, it was practically unachievable to correctly classify positive segments, and those that were classified correctly were usually swamped by positively classified neutral segments. The same problem appeared in case of negative segments, although less severe. Consequently, it was not possible to provide the error analysis based on the results from Aktualne.cz data.

Therefore, we decided to use the auxiliary data: the domestic appliance reviews from the Mall.cz (*http://www.mall.cz/*) retail server obtained from a private company. The Mall.cz corpus is much bigger (158,955 words, 13,473 lemmas). These reviews were divided into positive (6,365) and negative (3,812) by their authors. We found this data much easier to work with, because they are primarily evaluative by their nature and contain no complicated syntactic or semantic structures. Unlike the data from Aktualne.cz, they also contain explicit polar expressions in a prototypical use. Furthermore, they do not need to be tagged for the gold-standard annotation. The Mall.cz data, however, do present a different set of complications: the grammatical mistakes or typing errors cause noise in the form of additional lemmas and some of the reviews are also categorized incorrectly. However, compared to the problems with the news articles, these are only minor difficulties which can be easily solved. For this reason, the Mall.cz data are more suitable for the error analysis task.

## 4 The Lexicon-based Classifier System

There are several steps leading to the effective lexicon-based classifier. During the preprocessing phase, all the data first undergo lemmatization, using a tagger of Hajič (2004). From the tagger output, not only do we retain the lemma but also the part of speech and negation morphological tags. Then, we automatically generate a polarity lexicon from the training data and compute the measurement of how reliable a given lexicon item works as a polarity indicator. From our data, we first need to estimate the probability that, when encountering a given lemma, it is a part of a polar segment. Assuming we have that probability for each lemma we encounter in a given segment, we can by means of some aggregation, for instance a simple sum, easily decide whether to classify the given segment as polar. Then we can analogously determine its orientation. The desired properties of an indicative strength function are satisfied by lemma precision (see Wiebe et al., 2004). Then we need to compute a baseline for our lexicon, i.e. the probability that a randomly chosen word implicates the given polarity.

The classifier uses a standard unigram bag-of-words model, simply summing the indicator strength measurements over all the lemmas in a given segment. Then it selects the polarity class with the highest accumulated value in the desired measure. We have also employed a number of simple filters and other methods in order to improve the automatic annotation: filtering by frequency, weighed filtering by frequency (where the threshold for accepting a lemma as a feature is weighed by the baselines so that smaller polarity classes do not get discriminated), statistical significance filtering (where we accept a lemma if we can exclude the hypothesis that it is evenly distributed across polarity classes at a given level – 0.999, 0.95 and 0.8) or filtering by part of speech. Also, we have attempted to deal with sentence-level negation: first, if a segment contained a negative verb, the values for positive and negative polarity would be reversed for the segment, and a less crude method where we would specify which parts of speech to the right of a negative verb we would like to reverse.

# 5 Error Analysis

## 5.1 System Errors

Unfortunately, the first-aid filtering methods have proven rather useless – even those which appeared promising when we took a closer look into the list of incorrectly detected instances. For example, we found a number of functional words assigned with a wrong polarity. Nevertheless, when we removed them from the classification, the overall results did not improve. Moreover, when we started to eliminate the content words, the results got even worse. In order to reveal the main cause of the mistakes, we had to get back into the data once again.

We discovered various reasons of the system errors which can be divided into following categories. Statistically, the significant source of errors are still the short segments like *"Nothing"*, *"Price"* or *"I don't know"* which appear in both positive and negative reviews. These can by classified by the simple majority vote. If the vote is equal, the lemma classification is based on the baseline.

Also, some of these short segments have pretty high indicative strength for one polarity, but they often appear in the reviews expressing opposite evaluation (so filtering by frequency does not help):

<dg_postnegativetext>Proti:Kvalita.</dg_post negativetext>

*<dg_postnegativetext>Cons:Quality.</dg_post negativetext>*

In these cases the system always assigns the incorrect value. The solution to these problems could be elimination of all one-word answers or assigning the polarity of these items according to the polarity they have in subjectivity lexicon for Czech (see Veselovská, 2013).

One of the most frequented wrongly detected short phrases was *"High price"* tagged by the classifier with a positive instead of negative value. Besides, the classifier sometimes could not detect the domain-dependent evaluation, like *"long washing programs"*. These cases could be solved by using n-grams instead of just unigrams. Using n-grams could also hold for incorrectly detected evaluative idioms ("Je to sázka na jistotu" – *"It is a safe bet"* etc.) which are not listed in the Czech subjectivity lexicon or which are domain-dependent.

Furthermore, it could be advantageous to apply a coefficient for the initial and terminal position of words in a given segment. According to the reviews, it seems that the words occurring at the beginning or in the final parts of the text are more predictive towards the overall polarity:

<dg_postpositivetext>Pro: Je to výkonný a kvalitní vysavač, vím to, protože jsem ho měla víc jak deset let, ale bohužel se častým používáním porouchal a nechtěla jsem ho nechat opravovat, tak jsem si koupila nový. Ten starý vysavač funguje pořád jako vysavač, nejdou s ním čistit koberce. Půjčovala a půjčuje si ho celá rodina i příbuzný, je fakt dobrý, mohu ho doporučit.</dg_postpositivetext>

*<dg_postpositivetext>Pros: It is a high-performance and quality vacuum cleaner, I am sure, because I had it for more than ten years, but unfortunately it got destroyed by the frequent use and I did not want to have it fixed, so I bought a new one. I still use the old one, but it is not possible to clean the carpets with it. The whole family borrows it constantly, it is really good and I can only recommend it. </dg_postpositivetext>*

Moreover, the system is at the moment not able to treat emoticons: it considers every part of the smiley to be a separate word. To find positive and negative emoticons could help to detect given sentiment much better, as outlined in Read (2005).

There are also errors that can be improved using some simple linguistic features. We have already worked with sentential negation, using the rule roughly saying that all the negated verbs switch the overall polarity of the given sentence. But there are still plenty of rules which could be further implemented. Mostly, this concerns syntactic features. We found many incorrectly detected adversative constructions like:

<dg_postpositivetext>Pro: Není to žádný luxusní model, ale na chalupu stačí.
</dg_postpositivetext>

*<dg_postpositivetext>Pro: It is not a luxurious model, but for the cottage it will do.
</dg_postpositivetext>*

The “but” sentences can be as well solved by the rule, as indicated already in Hatzivassiloglou and McKeown (1997).

Also, there were many incorrectly evaluated concessive or conditional sentences in the data:

<dg_postpositivetext> Přestože neplní hlavní funkci kvůli které jsem ho kupoval (uklidit jednu místnost po druhé během naší nepřítomnosti), tak se jedná o jednoho z nejlepších robotů v nabídce na našem trhu. <dg_postpositivetext>

*<dg_postpositivetext> Although it is not suitable for the function I bought it for (to clean the rooms one by one when we are not at home), it is still one of the best available robots. <dg_postpositivetext>*

These problems might be eliminated by creating a stop-words list of items signalling non-evaluative part of the sentence.

### 5.2 Errors Caused by Human Annotators

Quite often, the reviewers were not evaluating given product, but they were rather commenting on something completely else:

<dg_postpositivetext>Pro: nemohu hodnotit, zboží jsem pro poškození vrátil </dg_postpositivetext>

*<dg_postpositivetext>Pro:I cannot review this, I sent the goods back since it was damaged. </dg_postpositivetext>*

or:

<dg_postpositivetext>Pro: Meteostanici mám jako dárek pro manžela, vyzkoušela jsem ji jen krátce při převzetí, tak se ještě nemůžu spolehlivě vyjádřit</dg_postpositivetext>

*<dg_postpositivetext>Pro: I bought the meteostation as a present for my husband and I tried it out just quickly after I received it, so I cannot review it yet.</dg_postpositivetext>*

On the other hand, we also noticed cases when the system classified the review correctly anyway:

<dg_postpositivetext>Pro: Přednosti tato pračka nemá.</dg_postpositivetext>

*<dg_postpositivetext>Pro: This washing machine has no pluses. </dg_postpositivetext>*

This kind of problems is tightly connected to pragmatics, but it might be partly solved by the reliable target detection.

The very common instances on which the classifier failed were the reviews in which people quoted other reviewers:

<dg_postpositivetext>Pro: Někdo píše SNAD dobrá značka???? Tato značka je mezi mraznicemi a ledniceni jednoznačná 1 </dg_postpositivetext>

*<dg_postpositivetext>Pro: Anyone said QUITE good brand???? This brand is number one among freezers and fridges </dg_postpositivetext>*

This is the matter of reliable finding of different sources of evaluation.

Some of the reviews contained besides other things the implicit evaluation:

<dg_postpositivetext>Pro: Nevím, jak jsem mohla bez sušičky být. Haní ji jen ten kdo ji nemá, nebo zhrzená manželka, když jí nechce manžel sušičku koupit. Úspora času, sice něco se musí žehlit, ale minimálně. Za sobotu jsem stihla usušit ložní prádlo, včetně obalů z matrací a lůžkovin (polštáře, deky) a ještě jsem měla spoustu času.</dg_postpositivetext>

*<dg_postpositivetext>Pro:I don’t know how I could have lived without the dryer. Only those who don’t have it defame it, or the turned down wives whose husbands don’t want to buy it for them. It saves time, some things still need to be ironed, but very little. I dried the bed linen during Saturday, including the mattress and bed linen cases (pillows, blankets) and I still had plenty of time.</dg_postpositivetext>*

Unfortunately, the implicit evaluation is again connected to pragmatics and so far it seems to be one of the most difficult subtasks in sentiment analysis in general. However, the reviewers (at least on the Mall.cz retail server) did not tend to use it more often than prototypical explicit evaluation.

## 6 Conclusion and Future Work

We have analyzed different types of classifier errors on the real evaluative data and suggested various improvements. In the next step of the research, we would like to use n-grams to find the domain-dependent evaluative constructions and evaluative idioms. Also, we would like to detect the unmarked neutral segments by employing the simple heuristic model – e.g. when the system detects expressions like *"I don't know"*. If the segment has less than five words, it will be classified as neutral.

In addition, we realized that it is necessary to implement the detection of emoticons and treat particular parts of adversative constructions separately. Moreover, it seems unavoidable to apply the model for the reliable detection of targets and sources of evaluation, e.g. by employing methods for detecting thematic concentration of the text (see Čech et al., 2013).

## References

Choi, Yejin & Claire Cardie (2008). *Learning with compositional semantics as structural inference for subsentential sentiment analysis.* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp 793-801).

Cui, Hang, Mittal, Vibhu, & Datar, Mayur (2006). *Comparative experiments on sentiment classification for online product reviews*. In AAAI (Vol. 6, pp. 1265-1270).

Čech, Radek, Popescu, Ioan-Iovitz & Gabriel Altmann (2013). *Methods of analysis of a thematic concentration of the text.* Czech and Slovak Linguistic Review. (in press).

Habernal, Ivan, Ptáček, Tomáš & Josef Steinberger (2013). *Sentiment Analysis in Czech Social Media Using Supervised Machine Learning.* In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp 65-74).

Hajič, Jan (2004). *Disambiguation of rich inflection: computational morphology of Czech*. Karolinum.

Hatzivassiloglou, Vasileios & Kathleen McKeown, (1997). *Predicting the semantic orientation of adjectives.* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp. 174-181). Association for Computational Linguistics.

Johansson, Richard & Alessandro Moschitti (2013). *Relational features in fine-grained opinion analysis.* Computational Linguistics 39 (3).

Liu, Bing (2010). *Sentiment Analysis and Subjectivity.* Invited Chapter for the Handbook of Natural Language Processing, Second Edition. Marcel Dekker, Inc: New York.

Moilanen, Karo & Stephen Pulman (2007). *Sentiment composition*. In Proceedings of the Recent Advances in Natural Language Processing International Conference (pp 378-382).

Polanyi, Livia & Annie Zaenen (2004). *Contextual valence shifters.* In Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.

Read, Jonathon (2005). *Using emoticons to reduce dependency in machine learning techniques for sentiment classification*. In Proceedings of the ACL Student Research Workshop (pp. 43-48). Association for Computational Linguistics.

Veselovská, Kateřina (2012). *Sentence-level sentiment analysis in Czech*. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics.

Veselovská, Kateřina, Hajič, Jan Jr. & Jana Šindlerová (2012). *Creating annotated resources for polarity classification in Czech.* In Proceedings of KONVENS (pp. 296-304).

Veselovská, Kateřina (2013). *Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification*. In Proceedings of SLOVKO 2013.

Wiebe, Janyce, Wilson, Theresa, Bruce, Rebecca, Bell, Matthew, & Martin, Melanie (2004). *Learning subjective language.* Computational linguistics, 30 (3) (pp. 277-308).