# Advantages and Limits of Text Mining Software for Analysis of Students' Satisfaction in Online Education
**(case study)**

**Zdena Lustigova[1], Veronika Novotna[2]**

## Abstract
*The article deals with the analysis of students' feedback while studying at different online learning environments. Information about satisfaction is often available in the form of textual (frequently multilingual) information, hidden in users' reviews, chat rooms, tea rooms and other unspecified and unstructured ways of feedback. To read such materials is often time-consuming and according to the quantity almost impossible. Text mining helps us 1/ to classify and to categorize the type of responses (complaints positive, negative, irrelevant, disease, etc.), usually on the base of sentiment analysis, 2/ to reveal the most frequent problems, 3/ to discover similarities and patterns and/or 4/ to identify similar text records (clusters).*
*Authors do not present the "big data" approach, based on powerful (and expensive) software. They focus just on part of the whole large scale of users' reflection to present the basic problems educational researchers might meet while working with available software tools (Statistica, Semantria) The unstructured text they processed (9 870 students' reviews/chats/remarks within 32 online courses) was created and published in 12 languages. Authors describe the problems they met, especially in the area of multilingual information processing. Despite all effort, nearly half of languages failed to process.*

## 1. Introduction
Text mining is often used to analyse open-ended responses from Web survey questionnaires, from general questionnaires with open answers, complaints, reviews, WoMs in social networks, etc. Etc. etc.., simply wherever there is a space for the client's own opinion.
Text mining problems and tasks generally belong to the data mining tasks. Text mining tools tries to process unstructured text and extract valuable information from it. Within the textual variables we usually at first look for keywords (most frequent words) and in the following step do their frequency analysis. Cases (specific clients, logs, etc.) where these keywords occurred, are indexed, and then returned to a file (database) as a new numeric variable, which we use in the context of classification methods. Another type of task is then comparison of documents according to the frequency of particular words or so called „terms „in matrix of terms or frequency matrix. (1, 2, 3)
Depending on the number and structure of words we can identify the theme and meaning of the written material (document). It might be multi-page book, thousands of reviews or thesis, but also, for example, a web page or pages to exactly the desired level, social networks, SMSs and many, many others.

### 1.1 Opinion mining and sentiment analysis
Sentiment Analysis also known as subjectivity analysis, opinion mining, and appraisal extraction [3] is an application of natural language processing, computational linguistics and text analytics to identify and retrieve certain information from the text.
The term sentiment analysis was for the first time used by Nasukawa in [2] and the term opinion mining by Dave and co-authors in 2003. Many other authors provide us within their earlier works on interpretation of sentiment adjectives, viewpoints, subjectivity, metaphors, and affects. Another researches about sentiment analysis and opinion mining were conducted after the year 2000 by Das and Chen, 2001; Morinaga 2002; Pang, 2002; and many others.
In education both methods can be used not only for mining and classifying opinion but also for the elimination of threatening or otherwise coloured reviews, written by individuals, who present (for

[1] Charles University**,** Czech Republic
[2] Charles University, Czech Republic

example in the context of blogs or other publicly accessible chat rooms) their insanity rather than opinion.
Text mining applications like sentiment analysis helps us to classify and categorize the type of response (complaints positive, negative, irrelevant, disease, etc...) and other similar characteristics. Teacher or tutor then can only deal with certain types of responses, which either provide valuable information for further development of online course and teaching strategies, or help to eliminate potential threads of any kind. They do not waste time with irrelevant remarks or answers. (e.g. [4],[5],[6],[7],[8],[9])

### 1.2 Automatic classification of texts

Even more interesting application of textmining tools is the identification of similarities within the specific text records based on cluster (cluster) analysis. Textual records are classified and sorted into clusters according to their level of similarity.

### 1.3 Text mining in educational area

Papers and articles, dealing with text mining, are either too general (see Survey of text mining application from 2013), either written by computational specialist or statisticians, and thus highly professional, but not too helpful for teachers, tutors or educational researchers. (see Journal of Educational data mining).
In our explorative research we do not focus on the "big data" approach, based on powerful (and expensive) software like IBM Whatson, even if we consulted its possibilities with IBM specialists, too. We decided to work with easily available SWs or freeware and to reflect just a part of the whole large scale of basic problems educational researchers might meet while working with available software packages like Statistica [1].

## 2. Research method

The aim of this particular research was to test the possibilities of text mining tool, which is part of the software package Statistica, in the area of unstructured text processing. The text was in the form of students' remarks, reviews, messages and notes in chat rooms and other informal areas (spaces) of 3 online learning systems we had to our disposal.
We used 9 870 students' reviews/chats/remarks, written in 12 languages within 32 online courses (including Czech, Slovak, Russian, Hungarian, Polish, German and other East and Middle European languages but also English, both Britain and American version, used as a general communication language).
As an output we required

- To test the capacity of Statistica text miner for Multilanguage information processing
- To identify the quality of information hidden in particular languages
- To classify positive, negative, ambiguous and irrelevant reviews related to particular courses and teachers, as well as to certain categories of courses
- To identify the positives and negatives as a keywords, calculate their frequency in relation to particular topic/course/group of students/teacher

In the next phase, we tried to apply cluster analysis, and find clusters of students who complain about the same problem or who have similarly positive feeling about some services, or about other factors (clarity, atmosphere, level of "noise", etc…). Cluster analysis should help reveal groups of people who, though perhaps seemingly inconsistent, reflect similar problems. This part of the research is still not finished and evaluated.

## 3. Data processing and selected results

Before the analysis was necessary to standardize the data, encode and eliminate errors. The most time consuming part was the unification of sources (identification of online course and particular source of review, marking countries and cities, tutors, or fixing errors caused by the data transfer from a web server to an Excel file.
The newly created data structure, including unstructured text, was imported into STATISTICA software and processed. Statistica Sw basic controls were found fairly intuitive and user-friendly. The tool focused on text mining is fairly detailed processed, however, in several respects, not too tight. Offered

STOPLIST (lists of words, the tool ignores) therefore mainly connectors, adverbs and other words, are almost dysfunctional because they lack even basic coupling, not only contextually adjusted. It is therefore necessary to rewrite the stoplists for each language and context.
The inbuilt algorithm that distinguishes all minor differences and verb forms, results in a relatively large amount of the non-words. The advantage is the possibility to set the minimum word length, and of course, the language of which the analysis will be based (the STATISTICA 13), the length of root words, the minimum number of repetitions of words in the dataset and other factors.
For the classification of positive, negative, ambiguous and irrelevant reviews related to particular courses and teachers, and for the calculation of their frequency in relation to particular topic/course/group of students/teacher we used Semantria.

## 4. Conclusions

In our dataset of students' reviews, chats and remarks there were 12 languages. Reliable results were obtained just from analysis of those, based on Latin alphabet. The analysis of unstructured text written in Cyrillic, in spite of our and invited experts effort basically did not make sense. Since approximately 25 % of unstructured text was written in Russian, the above mentioned problem was considered as serious.
The second major problem we found was the unavailability of high quality context based stoplists even in common languages, such as the English or German. To create them is necessary but time consuming activity.
The actual results of the analysis (mainly derived from assessment written in English and several other European languages) were not surprising, but helped to reveal certain behavioural patterns (e.g. habit of replying too late at certain teachers, or groups and subgroups of students tending to cheat, etc.).

## References

[1] StatSof data miner recipes. (2013) Available from: http://www.statsoft.cz/file1/PDF/newsletter/2013_03_05_StatSoft_Data_miner_recipes.pdf 02.17 2014
[2] Nasukawa, T., Nagano, T. (2003). *Text Analysis and Knowledge Mining System.,* IBM Systems Journal 40, no. 4, 967-984.
[3] Pang, Bo, and Lillian Lee. Opinion mining and sentiment analysis.Now Pub, 2008.
[4] Novak, Jeremy, and Michael Cowling. "The implementation of social networking as a tool for improving student participation in the classroom." (2011).
[5] Litman, D., J., and Forbes-Riley,K. (2004). Predicting student emotions in computer-human tutoring dialogues. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.Association for Computational Linguistics.
[6] Cummins, S., Burd,L. and Hatch, A.(2010).Using Feedback Tags and Sentiment Analysis to Generate Sharable Learning Resources Investigating Automated Sentiment Analysis of Feedback Tags in a Programming Course. *Advanced Learning Technologies* (ICALT), IEEE 10th International Conference.
[7] Agrawal, R. et al. (2012). Data mining for improving textbooks. ACM SIGKDD Explorations Newsletter 13.2 (2012): 7-19.
[8] Poulos, A., and Mahony,M.J. (2008). Effectiveness of feedback: the students perspective. *Assessment and Evaluation in Higher Education.* 33.2. p.143-154.
[9] Denker, K.J. (2013). Student Response Systems and Facilitating the Large Lecture Basic Communication Course: Assessing Engagement and Learning. *Communication Teacher* 27.1 (2013).