Text mining in reports from student stays in Czech enterprises

Abstract— **This research aims at finding differences between the organizations during innovation processes We analyze reports from student stay in the Czech enterprises with respect of innovation in the enterprise, or of potential of innovations. In this paper we analyze the student reports by means of natural language processing tools. We show which words are frequently used and how representative the students reports are. We show how text classification can be used for characterizing a company and what are the keywords - with respect to the company, with respect to the innovation. We compare the results with an observation of an expert.**

Company; Innovation process; Text mining

I. INTRODUCTION

The constant and rapid changes occurring in the markets force the companies look for the new ways how to survive and to be competitive. The customers make higher demands on the products they buy and on the other services. Products have to be introduced on the market faster and have to meet individual demands. Therefore it is necessary to bring new ideas and approaches to business and innovation is necessary in today conditions.

A waste of text documents, like newspaper and magazine articles, reports etc., written by managers, newspapermen but also customer blogs or web pages – are available for most of companies. In this paper we explore some of them – reports written by students of Technical University of Liberec after their stay at a company. Organization of long term industrial trainee is realized during 17 year with cooperating companies. These companies include the automotive, ICT, services. At present, this cooperation is supported and financed by the European Social Fond Operational Program Human Resources 3.2. THEORY AND PRACTISE - support for university students to obtain internships for employers, transferring practical experience in teaching (CZ.1.07/2.2.00/07.0321). We use text mining methods for discovery of patterns that may be typical for company, or may characterize the company as innovative.

The current business environment is characterized as highly turbulent, influenced by modern information and communication technologies, globalization, short innovation and production cycles and employees' mobility. It is not easy to compete in such an environment; organizations have to utilize their corporate resources to the greatest possible extent. Such resources include finance, employees, tangible assets, technologies and also knowledge.

As stated by P. Drucker (2001): "The most important, and indeed the truly unique, contribution of management in the 20th century was the fifty-fold increase in the productivity of the 'manual worker' in manufacturing. The most important contribution management needs to make in the 21st century is similarly to increase the productivity of 'knowledge work' and the 'knowledge worker." In a similar manner, Nonaka (1995) states: "In an economy whose only certainty is uncertainty, knowledge is the only source to gain permanent competitive advantage."

Why do some entrepreneurs find successful business opportunities while others do not? One of the reasons is that they try to find new opportunities and try to innovate their business. Innovation research has progressed over more recent years. We can find a number of factors at three levels of analysis—the individual, work group, and the organization. (West, 2001, 2002; King & Anderson, 2002).

Most innovations will be a mixture of emergent processes, adopted and adapted procedures which are in common usage elsewhere, and ideas which become sharpened over time by realistic limitations imposed by the organization (e.g., profitability, practicality of use, way of knowledge sharing, …), and so innovation researchers have almost exclusively focused upon cases and processes of relative novelty in organizations (West, 2002).

In this paper we analyze the student reports by means of natural language processing tools. We show which words are frequently used and how representative the students reports are. We show how text classification can be used for characterizing a company and what are the keywords - with respect to the company, with respect to the innovation. We compare the results with an observation of an expert.

In the following text we first describe the input data in Section 2. Section 3 concerns the most frequent words. In Section 4 we show how relevant the student reports are. In the following section we show which kind of reports are the most relevant and which the less relevant. Section 6 concerns prediction of innovative potencies of a company. We conclude with summary of the results and plans for future work.

II. DATA COLLECTION

The reports document the realization during one year student's education industrial trainee. This industrial trainee is part of studying bachelor course Computing and Business on Faculty of Economics Technical University of Liberec.

The reports concern four companies. As a part of information may be private we will assigne letters A, B, C and D to them. A (iteg) is SME that…, B (autocont) is one of the biggest hardware sellers and software houses in Czech republic. C is a car producer, D (flores) is SME that …

Reports include information about internal and external situations of company. It means how employees communicate between themselves, with customers and suppliers, how the management motivates employees, how the innovation processes are realized, how the company develop their internal and external processes, how the information and communication systems are used and some other detail information.

We analyzed 138 texts that describe companies A, B, C and D from different point of view. The document collection contained 29074 tokens (words and diacritics) and 7323 different words.

III. DESCRIPTIVE CHARACTERISTICS

For each company we first built an ordered list of words. Then we compared those list and looked for words that are frequent for a company and infrequent for the others. As each report describe a company from one of focuses (e.g. overview, technology, management, motivation, communication, collaboration, software use, education, brief evaluation of the stay etc.) we also analyzed the frequent words with respect to the focus.

E.g. for B (a software house) it was the word *solution* that was the most frequent between substantives and adjectives. This word almost not appeared in text about other companies. Similarly for D, it was the word *product* when focused on Motivation. For C, the car producer, the word *company* was much more frequent then for A, B and D.

It correspond with an expert observations. B is now more focusing into selling solutions instead of selling hardware and software. D (flores) is a new company that is developing a new product. In the C car factory, all processes are standardized and it takes usually longer time to change it.

IV. RESULTS

The observation described in the previous section brought only indirect evidence for relevancy of the reports. That is why we looked for direct evidence. We would try to learn classifier that would recognize which report concerned which company.

The learning set contained 138 document. Each document was represented as a bag of words, i.e. each example was represented by a vector of length of 7323 (the number of unique words in the collection). Each item of the vector was equal to a numeric characteristics (importance) of the particular word. We tested three possibilities, from the simplest ot he most complex – Boolean (word appeared/not appeared in the document), a frequency of the word, and term-frequency/inverse-document-frequency (TF/IDF). We chose a word frequency, for its simplicity, because the overall accuracy was almost the same as for TF/IDF.

We used decision trees, Bayes learning, SVM and instance-based learning and 10-fold cross validation. Overall accuracy was between in range of 67-88% with the highest accuracy for multinomial naïve Bayes classifier. When all words that contain the company ame and names of its proprietary products, accuracy decreased to 84% but that decrease is quite small.

We also checked whether there is difference between men and women, however the difference was not strong enough.

V. WHAT KIND OF A REPORT IS RELEVANT

It was observed by the expert that some of document might be more important for company recognition then the other, and that it depends on the focus that the author used. In this experiment we always removed one kind of a focus. The biggest difference in accuracy was observed for two cases – for overview and for brief evaluation of a stay. After

removing the text of brief evaluation the accuracy little increased, on about 3%. It may be explained by the fact that this text usually does not bring any information that concern the company itself. After removing an overview, what is actually an introduction of a company and brief description of the goal of the stay, the accuracy decreased on about 7%.

VI.

VII. CAN WE PREDICT HOW INOVATIVE THE COMPANY IS?

Innovations are principal for long-term growth of a company. Two companies, actually SMEs - flores and iteg - are very active. On the other side – two big companies rather concentrate to conservative solutions. In the last experiment we checked whether this fact – innovation - can be discovered automatically from the reports.

We built two classes, the first containing two SMEs, the other containing the rest. We again used the same pre-processing methods as in the previous experiments and the same learning algorithms. For multinomial naïve bayes the overall accuracy reached 88% and we can conclude that the potential of innovations can be induced form the text.

We also analyzed which words appeared to be most important for this discrimination: positive keywords, i.e. words that are frequent for SMEs concern *projects*, *presentation*. On the opposite side, words *company*, *helpful* have been typical for conservative companies.

VIII. CONCLUSION

In this paper we analyzed reports from student stays in four Czech enterprises. We showed what words are the most typical, how accyrate is the prediction of a company from the text and how accurately the innovative potencies may be predicted.

As future work we plan to extend the document collecrtion with the information from the web (web presentations, news that concerns a company). We also intend to employ natural language processing tools – morphological disambiguation and shallow syntax analysis.

ACKNOWLEDGMENT

We thank This work has been partially supported by … Masaryk University .

REFERENCES

Drucker, P.F. (2001), Management in 21st century, Management Press, Praha, ISBN 80-7261-021-X, p. 129

King, N., & Anderson, N. (2002). Managing innovation and change: A critical guide for organizations. London:

Thompson

Nonaka, I., Takeuchi, H. (1995), The Knowledge Creating Company, New York, Oxford Press, 1995, ISBN 0-19-509269-4

West, M. A. (2001). The human team: basic motivations and innovations. In N. Anderson, D. S. Ones, H. K.

West, M. A. (2002). Sparkling fountains or stagnant ponds: an integrative model of creativity and innovation

implementation within groups. Applied Psychology: An International Review, 51, 355–386

[1] Han J., Kamber M.: *Data Mining: Concepts and Techniques*. Elsevier 2006.

[2] Mitchell T.: *Machine Learning*. McGraw Hill, New York, 1997.

[3] … Proceedings of Znalosti 2010 Czech-Slovak AI conference, Jindřichův Hradec, 2010.

[4] Witten I.H., Frank E.: *Data Mining. Practical Machine Learning Tools and Techniques*. Elsevier 2005