IDENTIFYING EXPRESSION OF EMOTIONS IN CZECH TEXT

Radek Červenec
Master Degree Programme (2), FEEC BUT
E-mail: xcerve07@stud.feec.vutbr.cz

Supervised by: Radim Burget
E-mail: burgetrm@feec.vutbr.cz

Abstract: Since the amount of information stored in the form of electronic text documents has been rapidly growing over the past few years, there is an increasing demand for tools enabling to automatically analyze these documents and benefit from their emotional content. The purpose of this paper is to introduce the implemented system for automatic analysis of emotions in Czech text based on a machine learning method.

Keywords: emotion detection, emotion recognition, feature selection, SVM, text mining

1 INTRODUCTION

A substantial portion of the available information is stored electronically, in the form of text databases such as news article archives, blogs, research paper databases, digital libraries, Web pages etc. As the sizes of these kind of databases are rapidly growing and human abilities to effectively analyze them are limited, there has been increased interest in text mining methods to derive high-quality information from text. The methods were eventually applied to automatic emotion recognition from text which provide an insight into author's intent and enable to obtain additional information conveyed by text. Many beneficial applications can be found in human-computer interaction as computers acquire some of the emotional skills that people need to appear more intelligent to user interacting with them. In general, automatic emotion recognition can be fast and cheap way to analyze large text document collections with important emotional context such as opinion mining, market analysis, text processing in Safer Internet Centers etc.

2 CORPUS ANNOTATION

Given the lack of of performed experiments in the field of emotion detection from czech texts and lack of available corpus in general, a training data set had to be manually created and annotated. The emotion corpus consisted of texts drawn from various blogs, online technical support archives and message boards related to major national and world wide events. These type of sources were selected because they are potentially rich in emotion content. They usually contain suitable examples of used expressions of emotions. Additionally they provide a wide variety in writing styles, choice and combination of words, as well as topics expressed in text. The texts were annotated at paragraph level and at sentence level. Annotators listed keywords and terms that were crucial for their decisions. Training set items had also source and domain specification as other mandatory parameters. The main objective of such a system was to create an universal corpus that would be independent of the method of emotion detection. In order to simplify and speed up the process of annotation, a simple computer program was developed to enable annotation using graphical user interface.

3 EMOTION RECOGNITION SYSTEM

The proposed system is depicted in Fig.1. The following section presents individual steps that were taken during preprocessing and model learning.

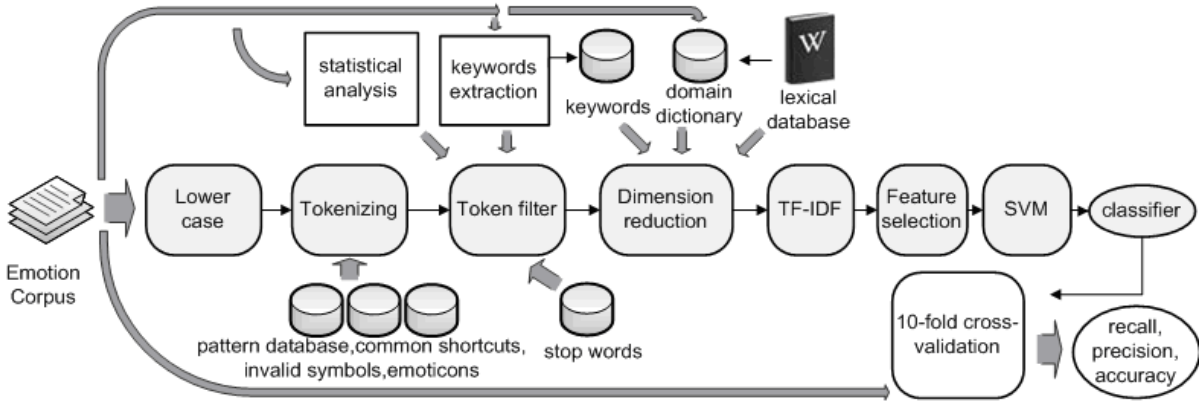

Figure 1: The proposed system for emotion recognition

3.1 TEXT PREPROCESSING

In terms of preprocessing methodologies, structured text representation was extracted from original unstructured sources. First of all, sentences were converted to lower-case. Then the stream of text symbols were segmented into indivisible meaningful units called tokens. The tokenizer was implemented with respect to special character sequences that had to be preserved. Those tokens that were irrelevant to or even could degraded text categorization task were dropped. A list of stop words was created[1] for that purpose. Additionally, statistical analysis helped to remove outliers, these were tokens that occurred only once or twice in the whole corpus. A lexical database *Czech WordNet* was the key component for token transformation. The database groups words into synsets and define semantic relations between them. Hypernymy (generalized forms) was found to reduce dimension. However such a transformation was not always desirable, the keywords marked during corpus annotation formed domain dictionary and those words were left out from the transformation process. In order to build a model of classifier, the text representation was changed. An TF-IDF algorithm presented in [2] was applied. It enabled to determine word relevance in terms of a particular document as well as within the corpus.

3.2 FEATURE SELECTION

The process of feature selection was divided into two steps in which an optimal set of features was selected from highly dimensional data set. In the first step, features were filtered based on the value of Pearson's correlation coefficient. In the second step, a genetic algorithm was applied as the heuristic to find the optimal set of features. Each individual in population of solutions was represented as a set of features and was assessed by 10-fold cross validation [2] whose outcome in form of F-score served as the fitness function.

3.3 TRAINING MODEL

Since the SVM had proved to be promising learning techniques [2], it was selected to build the model of classifier. LIBLINEAR [1], a library for large-scale linear classification served as the basis for the SVM implementation with Radial Basis Function (RBF) kernel.

[1] Stop words source http://www.fit.vutbr.cz/research/prod/index.php.cs?id=133&notitle=1

4 CLASSIFICATION RESULTS

Precision, recall and F-score were used to evaluate the performance of the proposed system [2]. The main goal of the classification task was to identify a presence of emotions in sentences. The sentences were classified into three defined classes: negative+vulgar, positive and neutral. Various classification scenarios were created to justify proposed steps during preprocessing and model training. The bar charts (Fig. 2) gives unite F-score of three emotion classes over various classification scenarios. Overall, the proposed system (scenario *complete*) achieved higher values of the F-score than the other scenarios over all emotion classes. This pointed out that the proposed steps for preprocessing and model training did not negatively influence classification results. Apart from *neutral* emotion class, there was only slight decrease in F-score values while the WordNet was removed from the system (scenario *no WordNet*). That is due to low search effectiveness in the lexical database as it was approximately 30 %. On the other hand, the feature selection impacted classification results for *negative+vulgar* class and *positive* class as F-score droped dramatically while it was removed from system (scenarion *no feature selection*).

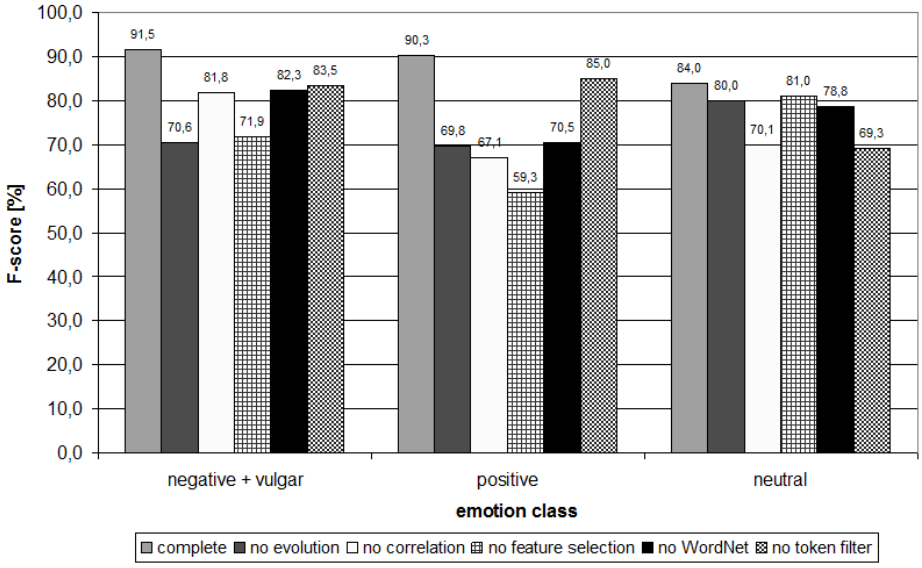

Figure 2: F-score of three emotion classes for various classification scenarios.

5 CONCLUSION

Working with various classification scenarios it was shown that proposed system achieved the highest values of F-score. One possible reason for misclassification of some sentences could be also a subjectivity, which have to be present as the samples are evaluated by people. Another problem was ironic texts that are generally very difficult to analyze as they convey a meaning exactly opposite from their literal meaning. There are still possibilities for improvements, the system could be extended by integration of a lemmatizer that determines the lemma for a given word and improves search effectiveness in lexical database which might result in better accuracy.

REFERENCES

[1] FAN, Rong-En, et al. LIBLINEAR. *Journal of Machine Learning*. 2008, 9, pp. 1871-1874.

[2] FELDMAN, R.; SANGER, J. *The Text Mining Handbook*. Cambridge : Cambridge University Press, 2007. 410 p. ISBN 978-0-521-83657-9.