Recognition of Emotions in Czech Newspaper Headlines

Radim BURGET, Jan KARÁSEK, Zdeněk SMÉKAL

Dept. of Telecommunication, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic

burgetrm@feec.vutbr.cz, karasek.jan@phd.feec.vutbr.cz, smekal@feec.vutbr.cz

Abstract. *With the growth of internet community, many different text-based documents are produced. Emotion detection and classification in text becomes very important in human-machine interaction or in human-to-human internet communication with this growth. This article refers to this issue in Czech texts. Headlines were extracted from Czech newspapers and Fear, Joy, Anger, Disgust, Sadness, and Surprise emotions are detected. In this work, several algorithms for learning were assessed and compared according to their accuracy of emotion detection and classification of news headlines. The best results were achieved using the SVM (Support Vector Machine) method with a linear kernel, where the presence of the dominant emotion or emotions was analyzed. For individual emotions the following results were obtained: Anger was detected in 87.3 %, Disgust 95.01%, Fear 81.32 %, Joy 71.6 %, Sadness 75.4 %, and Surprise 71.09 %.*

Keywords

Emotion corpus, emotion detection, emotion classification, text mining, Czech.

1. Introduction

Emotions are in general defined as a subjective experience associated with mood, temperament, personality, and disposition. Emotions can appear in most parts of human-to-human communication and often provide additional information about a message. For example, if somebody is joking, is scared or needs help and so on. They are present and can be detected in our faces [2], [3], in our voices [4], [25], in our "body language", generally in our behavior, in various biometric signals that can be measured on a human body and even in texts we write.

As some emotion expressions are culturally independent - for example, even in a foreign language where we do not understand the meaning of words it is relatively easy for anyone to recognize anger, scare, surprise, etc. in the message. Similarly in the expression of our face it is not so important if we grew up in Britain, USA or China, most of such expressions are very similar and have a similar meaning. The problems come with text documents. Any document or sentence is strongly dependent on the language it was written in. Even relatively similar languages have often different spelling and often also a bit different meaning and strength of some words. This is the reason why a model trained for one language is useless in another.

There is a variety of models for emotion description and classification. Some researchers prefer to avoid hard categories and define overall emotion as a point in continuous 2D space (see Fig. 1). There are also dozens of other emotion classification models. One of the most basic classifications is that into a positive, negative and neutral content. Also the active and passive attitude is being distinguished. For each category up to several dozen different emotion states are assigned.

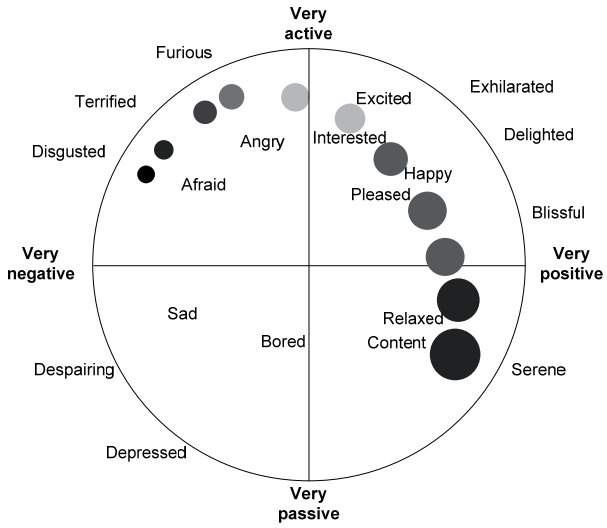

Fig. 1. A two-dimensional representation of emotion, derived from [1].

However, for our purpose we choose a classification model with seven basic emotional states: Anger, Disgust, Fear, Joy, Sadness, Surprise and No emotion (Neutral). The main reason is that this classification model is one of the most commonly used in scientific papers. Therefore, we suppose our work can be easily compared with other works and papers and thus we hope it will have a higher contribution to the research of text emotion classification.

This paper deals with the detection of emotions and identification of a dominant emotion in short text messages (such as headline news) in the Czech language. The Czech language is considered to be one of more complex languages and the main reason why it is difficult to analyze a Czech text is that nouns and verbs can have different forms for the same meaning.

The main contribution of this paper lies in providing a training data corpus of news headlines in the Czech language and providing results of classification of emotions using the SVM method with linear kernel, the SVM method with radial kernel, the SVM method with polynomial kernel with degrees two and three of freedom, *k*-nearest neighbor algorithm, decision trees using the J48 algorithm, Bayes networks, linear regression and linear discriminated analysis. The data are pre-processed to transform texts into vector space models and are evaluated using 10-fold cross-validation.

The lack of current classification models is considered to be a lack of incorporating psychological theories, lack of semantic analysis using dependency trees or linguistic rules, and a lack of database reasoning.

The rest of the paper is structured as follows. Section 2 deals with the related research papers. Section 3 deals with the description of training data and the method of data pre-processing and model training. Section 4 discusses the results achieved from described data and model. Section 5 discusses the results achieved in this paper, and the last section 6 gives conclusions of the paper.

2. Related Works

In this section, some of the research papers related to detection of emotion in the text are briefly presented. The following research papers can be divided into two categories depending on what direction to pursue. Paper [9] can be considered as a common basis of articles which we refer to.

The first group deals with the basic classification of emotions in the text. This group analyzes several emotions, similar to this article, or focuses only on one particular emotion. Paper [10] describes experiments concerned with the automatic analysis of emotion in text. It deals with the construction of a large data set and proposes and evaluates several methods for the automatic identification of emotions in text. This group also contains articles dealing with the actual analysis and comparison of different methods, which are used in detection and classification of emotion. On the other hand, this article analyzes short texts, where the detection of emotions is very difficult. Paper [11] explores the task of automatic classification of text by hierarchical and flat classification methods. This article presents a novel method arranging neutrality, polarity and emotions hierarchically. A novel, bootstrapping technique for identifying paraphrases is presented in [12]. Based on our experiments, an approach based on decision trees achieves a lower accuracy than the SVM method, which is used in this article. The bootstrapping approach learns extraction patterns for six classes of emotions (similar to those used in this paper), based on the highest scoring extraction patterns. An experiment showing how an annotation task can be set up so that untrained participants can perform emotion analysis with high agreement even when not restricted to a predetermined annotation unit and using a rich set of emotion categories is presented in [13] (emotional perception of fairy tales). Machine learning techniques instead of human experts are used in [14] to extract emotions in Music. The classification is based on a psychological model of emotion that is extended to 23 specific emotion categories. The result can be used in online music database which offers browsing by musical emotion (mood). Paper [15] is focused similarly as this paper is, and deals witch reader emotion classification of simplified Chinese news headlines. The SVM method was used as the classification method. Paper [16] also deals with a headline emotion classification. This approach is based on frequency and co-occurrence information collected from the World Wide Web. Analysis of emotions in Thai text is described in [17] and [18]. The first of these papers deals with emotion word analysis and the second paper applies latent semantic analysis to classify emotions in Thai text. Unfortunately, in paper [17] the size of the training set is not defined, and a deeper analysis of training data is missing. Paper [18] proposes a novel approach that takes advantage of bi-words occurrence to classify emotion hidden in a short sentence.

The second group deals with the text published on web pages, and tries to capture the feelings of an individual blogger. Paper [19] presents a system that learns to recognize emotions based on textual sources and tests it on a large number of blog entries tagged with moods by their authors. The paper shows how a machine-learning approach can be used to gain insight into the way writers convey and interpret their own emotions. An annotation scheme and the result of this scheme on a corpus of blogs are presented in [20].

The second group also deals with the detection of emotions in real-time communication, specifically in instant messaging software, in email communication, etc. Paper [21] deals with how people express emotions and detect emotions during text-based communication. Nonverbal communication is a significant factor of common communication. In the text-based communication this factor is omitted and it is harder to determine emotions of a person we communicate with. In contrast to this paper, authors relied on four strategies to express Happiness versus Sadness, including Disagreement, negative affect terms, punctuation, and verbosity. Paper [22] addresses the task of affect recognition from text messaging. The purpose of this work is to improve social interactivity and affective expressiveness of computer communication. To visualize textual affective information, an avatar displaying emoticons, social behavior, and natural idle movements is used in this paper. Text-driven rule-based system for emotion cause detection is proposed in paper [23]. By analyzing the corpus data, seven groups of linguistic cues are identified and two sets of linguistic rules for detection of emotion causes are generalized. Using the linguistic rules, a rule-based system for emotion cause detection is developed. This study should lay the ground for a future research into the inferences of implicit information and the discovery of new information based on cause-event relation. Study [24]

focuses on automatic emotion detection in descriptive sentences and how this can be used to tune a facial expression parameter for 3D character generation. Classification accuracy achieved with SVM is used in this study. The proposed automatic feature selection algorithm implemented in this study helped to detect new words from the training corpus which were relevant to the classification tasks but were not considered by researchers.

3. Training Process Description

3.1 Training Data

As mentioned before, emotions are a matter of subjective feelings and as such they can be hardly described and objectively measured. And this is a problem. When we are to deal with training an emotion classification learning model, an absolute necessity is to have quality training data. A classification model can have only such good accuracy as the training data had. The classification model used in this paper is depicted in Fig. 2.

For the purpose of emotion classification and identification in text data there are already several English training data sets available (such training data are often called a corpus). They are often not directly linked to emotions but in general to semantic meaning. SemEval 2007[1], SensEval-2[2], SensEval-2[2], Senseval-3[2], Senseval-4[2], WordNet-Affect[3] and others can be given as examples.

Annotation	Newspaper Headlines
Fear, Anger	Czechs and Poles get missile warning (Češi a Poláci dostali protiraketové varování)
Fear, Anger	Iraqi insurgents attack U. S. base (Iráčtí povstalci napadli základnu Spojených států)
Surprise, Joy	Breakthrough in breast cancer research will save 1,000 women a year (Průlom ve výzkumu rakoviny prsu zachrání 1000 žen ročně)
Fear, Sadness	Indonesian bird flu death (Indonéské oběti ptačí chřipky)
Fear, Anger	Apple's renewed attacks on Windows Vista (Apple obnovil útoky na Windows Vista)
Joy, Anger	Miss Brazil wins lawsuit (Miss Brazílie vyhrála soudní spor)

Tab. 1. Sample of examples from the annotated text.

As training data (so called corpus in the context of text classification) approximately 1000 randomly chosen news headlines in the Czech language were used. Each headline was labeled by eight independent persons and each emotion class was marked from 0 % to 100 % according to how a person subjectively felt the strength of a particular emotion. Individual headlines were displayed separately in random order. Thus we hope that the impact of previous texts on the evaluation of a current result was reduced. An average value over answers of all eight persons was used as the resulting value. This was accomplished for each emotion: Anger, Disgust, Fear, Joy, Sadness, and Surprise. In case the level of resulting emotion strength did not exceed at least 20 %, it was considered not to contain any emotion and "No emotion" class was assigned. This approach is similar to the way SemEval 2007[1] were evaluated. The samples of Czech news headlines labeled by the independent persons are showed in Tab. 1.

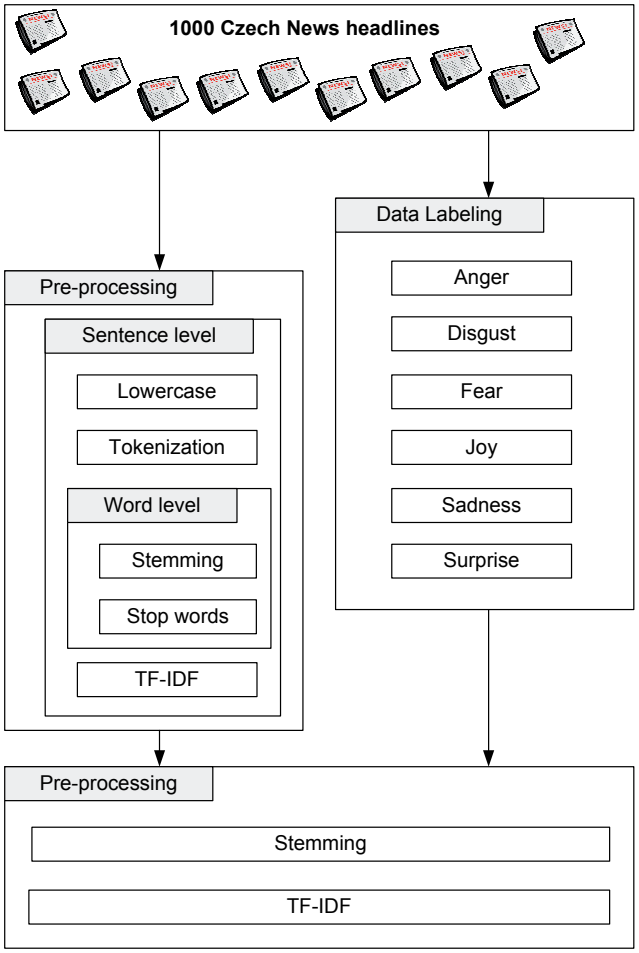

Fig. 2. The classification model for emotion recognition in texts.

In the corpus the presence of emotions (i.e. the resulting emotion strength was at least 20 %) and the dominant emotion (i.e. the emotion with the highest strength) were identified. If some of the emotion exceeded 20 % it was considered to have no emotions.

A comparison of the frequency distribution of all emotions throughout the training set is shown in Fig. 3. An evaluation of the frequency of each emotion is shown in Fig. 5 (Appendix B). The histograms in Fig. 5 were obtained on the basis of subjective evaluation of the strength of each emotion by individual listeners. Emotion of Surprise shows a significantly different behavior compared with other emotions. Listeners perceived this emotion more intensely than other emotions. This is probably caused by the effort of journalists to arouse readers' interest and encourage them to read the whole article. In this case it seems most appropriate to use emotion of Surprise.

[1] Semantic Evaluations, http://nlp.cs.swarthmore.edu/semeval/
[2] Semantic Analysis of Text, http://www.senseval.org/
[3] WordNet Domains, http://wndomains.itc.it/index.html

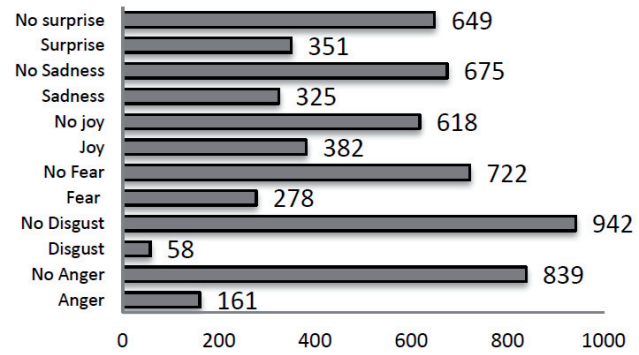

Fig. 3. Histogram of analyzed emotion frequency in the training data set.

3.2 Data Pre-processing

Before the data were provided for a learning algorithm, they needed to be pre-processed first. The pre-processing was accomplished in several steps. First, all the texts were converted to lower-case. Thus, for example, the word "pain" is considered to have the same meaning, no matter whether it is written as "Pain" or "pain". In the next step, the sentence is tokenized. The so-called tokenization means a process when the sentence is divided into words. To cope with various morphologies of the same word, the so-called stemmer is used, which converts the word to one common form of the word. For that purpose Dolamic's stemmer was used [5].

Emotion	Words
Anger	greedy (nenasytný), hatred (nenávist), hostility (nepřátelství), hateful (odporný), disdain (pohrdat), angry (rozzlobený), mad (šílený), devil (trápit), harass (sužovat), envy (závist), jealousy (žárlivost)
Disgust	horror (děs), disgust (hnus), sinful (hříšný), noisome (nechutný), dishonest (nekalý), amoral (nemorální, nemravný), obscene (obscénní), detestable (odporný), repulsive (ohavný), wicked (zlomyslný), sicken (znechutit)
Fear	fearful (bázlivý), awful (děsný), horrible (hrozný), hysterical (hysterický), cruel (krutý), diffidence (nesmělost), panic (panika), horrific (příšerný), fear (strach), scare (úlek), anxious (úzkostlivý), terror (zděšení), upset (zneklidněný)
Joy	good (dobrý), amorous (milostný), great (ohromný), fascinating (okouzlující), celebrate (oslavovat), comforting (povzbuzující), glad (rád), romantic (romantický), pleased (spokojený), happy (šťastný), satisfaction (uspokojení), comforting (utěšující), merry (veselý), adoring (zamilovaný)
Sadness	misery (bída), dismay (hrůza), poor (chudobný), regret (litovat), glum (mrzutý), bad (nepříjemný), grim (neradostný), dismal (ponurý), guilty (provinilý), mourning, sadness (smutek), rueful (smutný), glooming (šero), dark (temno), tearful (uplakaný), oppression (útisk), sorry (zarmoucený)
Surprise	terrific (báječný), fantastic (fantastický), admiration (obdiv), amaze (ohromit), surprising (překvapivý), wonderful (skvělý), astonish (udivit), wonderment (úžas), marvel, miracle (zázrak)

Tab. 2. An example of affected words for each category of emotion.

Some words, which are insignificant to classification and potentially can have negative impact for classification accuracy (so-called stop-words), were removed from the set of tokens. These stop-words are listed at the end of this document in Appendix A. Finally, the resulting set of tokenized sentences was passed through Term Frequency–Inverse Document Frequency (TF-IDF) to determine word relevance according to each particular emotion class [6]. Tab. 2 represents an example of affected words for each category of emotions.

3.3 Learning Model Evaluation

To assess the accuracy of identification and classification of emotions a 10-fold cross-validation [7] technique is used. The cross-validation is based on the principle that the training set is split into 10 disjunctive subsets. Nine subsets are used for training and the remaining subset is used for testing. This is repeated 10 times; each subset is left out of training and subsequently used for testing. Thus we get an evaluation of the model on the whole training set and the evaluation is independent of the training data. This technique is especially profitable in the case of a limited number of training data. A total of 9 different learning algorithms were used and evaluated as the learning algorithm. The list of these algorithms is provided in Tab. 3. Detailed information on the algorithms used can be found in [7].

Symbol	Learning Algorithm
A	Support Vector Machine with linear kernel [7]
B	Support Vector Machine with radial kernel [7]
C	Support Vector Machine with polynomial kernel (degree = 3) [7]
D	Support Vector Machine with polynomial kernel (degree = 2) [7]
E	k-Nearest Neighbour (k = 5) [7]
F	Decision trees - J48 algorithm [7]
G	Bayes Net [7]
H	Linear regression [7]
I	Linear Discriminant Analysis [7]

Tab. 3. Training algorithms used.

4. Results

In this paper an identification of the presence of an emotion (Anger, Disgust, Fear, Joy, Sadness, Surprise and No emotion) and classification according to the most dominant emotion present is analyzed.

The accuracy of detection of emotional states: Anger, Disgust, Fear, Joy, Sadness and Surprise are depicted in Fig. 4. The detection was evaluated using nine different learning algorithms as described in Tab. 3. Each column marked with letters A, B, C, D, E, F, G, H, and I in the figure stands for its respective algorithm listed in Tab. 3. Resulting averaged accuracy and standard error deviation among ten cross-validation evaluation is depicted for each column. This standard error is computed as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)} \qquad (1)$$

where μ stands for mean value, X_i for accuracy of evaluation on a partial result of cross-validation, N stands for a number of folds of cross-validation, i.e. 10 in our case.

As can be read from the figure, the most accurate results for Anger detection gave the SVM method with a linear kernel [7]. The results were 87.3% ± 5.65%. Thus the model can identify an angry emotion in the content of news headline with this accuracy.

In case of detection of Disgust emotion in the data, the algorithm with the highest accuracy was the SVM method with radial kernel [7]. The resulting accuracy was 95.01% ± 2.73%. Compared with other emotion groups the emotion Disgust is of relatively low frequency, but a relatively high degree of accuracy with a relatively low dispersion of mean deviation is achieved. At a closer look at randomly selected training data it is clear that there are words that probably are not too numerous in the other emotional categories and, also, they can be expected to evoke exactly the Disgust emotion. They are words like porn, cancer, campaign, be over, resign, etc. It is also interesting to note that keywords associated with politics are also relatively important in terms of Disgust emotion, regardless of whether the result is positive or negative.

Detection of Fear achieved the highest accuracy with the SVM method with linear kernel [7]. The resulting accuracy was 81.32 % ± 2.13 %.

Detection of Joy achieved the highest accuracy with the SVM method with linear kernel [7]. The resulting accuracy was 71.60 % ± 2.22 %.

For Sadness the highest accuracy was achieved with the SVM method with linear kernel [7]. The resulting accuracy was 75.41 % ± 4.60 %.

The last emotion Surprise also achieved the best accuracy with the SVM method with linear kernel and results are 71.09 % ± 2.21 %.

Averaged values for these values are 80.28.

The results obtained by different algorithms show that some algorithms, such as various variants of the SVM method, consistently give relatively good results. On the other hand, algorithms such as the *k*-NN or Bayesian-Network methods often fail. This behavior can probably be explained by the nature of the SVM methods. The SVM methods inherently tend to look for solutions with ideal hyperspace distribution. As a result, even in the case of a multi-dimensional limited training set, the SVM method gives, in comparison with other methods, interesting results (about a thousand headlines used in the training set is still a small amount compared with the number of words occurring in Czech). It is obvious that the most successful variant was obtained by SVM methods with linear or radial kernel. It is probably due to the nature of training data. The linear kernel can easily separate classes on the basis of individual words in a linear way. On the other hand, the radial kernel should be able to easily find clusters, which lead to classification into one of the appropriate emotion classes.

True / Prediction	Joy	Fear	Surprise	No emotion	Anger	Sadness	Disgust	Precision (%)
Joy	203	67	86	148	29	94	15	32.0
Fear	10	24	5	9	2	6	0	42.8
Surprise	12	5	21	10	1	5	3	36.8
No emotion	29	18	21	42	5	28	0	29.4
Anger	3	2	4	6	7	3	1	26.9
Sadness	10	7	8	13	4	29	0	40.9
Disgust	0	0	1	0	1	0	3	60.0
Recall (%)	76.0	19.5	14.4	18.4	14.3	17.6	13.6	

Tab. 4. Classification according to dominant emotion.

When detecting a dominant emotion in the sentence, Support Vector Machine with linear kernel was used. This learning model was selected in respect to very good results obtained in the previous classifications. The results are described in Tab. 4, where the true and predicted values are described. The number in tables stands for the total number of instance, which matches the predicted class in a leftmost column and the true class in the topmost row. The precision is calculated as:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \cdot 100\% \tag{2}$$

where N_{TP} stands for a number of true positives, N_{FP} stands for a number of false positives.

The recall is a fraction of the sentences that are relevant to the prediction. It is computed according to (3).

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \cdot 100\% \tag{3}$$

where N_{FP} stands for a number of false positives, N_{FN} stands for a number of false negatives and rest of the symbols have the same meaning as in (2).

5. Discussion

To discuss the results achieved, several other papers were chosen. Unfortunately, currently there is not any similar work for the Czech language. Therefore, the results will be discussed only with respect to works involved in the English language [8]. In paper [8] hybrid approach combining keyword based approach and Knowledge-Based Artificial Neural Network (KBANN) emotion classifier are proposed and evaluated. In this work 75 % of the corpus were used for training and 25 % for testing. The emotions used are Anger, Fear, Hope, Sadness, Happiness, Love, Thank and Neutral. The accuracy is tested only on data without emotional keyword (as they declare it is approximately 10% of the whole corpus) and the results are Anger: 63 %, Fear: 60 %, Hope: 47 %, Sadness: 65 %, Happiness: 61 %, Love: 56 %, Thank: 45 % and Neutral: 65 %.

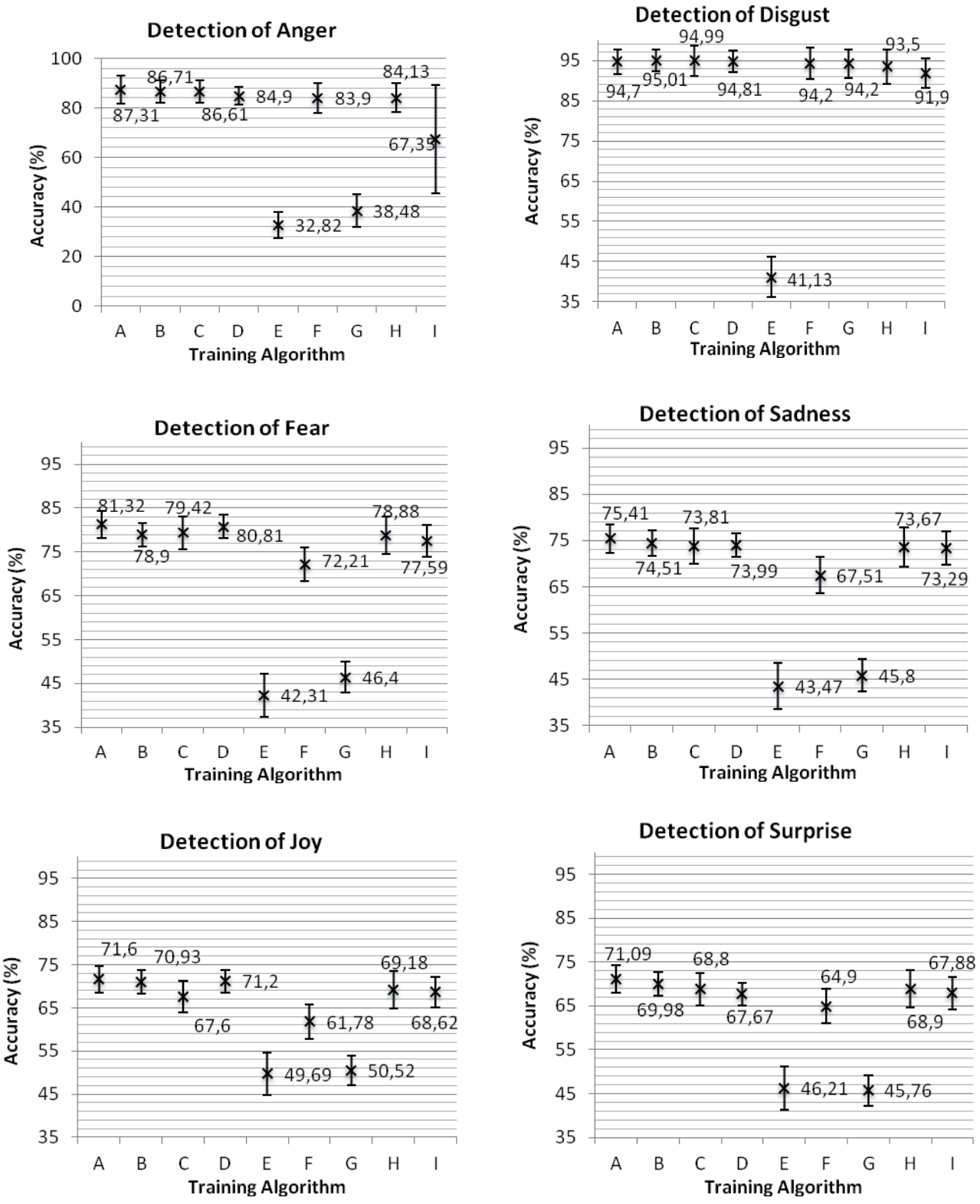

Fig. 4. Accuracy of emotion detection from short sentences. The training algorithms used are described in Tab. 3.

Languages where there are a large number of users (especially English or Chinese) have a huge advantage in that high quality and relatively large training sets are already available, from which you can start and carry out extensive analyses. For the Czech language, such sources exist on a limited scale only, in particular in terms of classification of emotions. Another problem associated with the Czech language is the necessity of having a good Stemmer,

which converts all the words to the root form. Compared with the previously mentioned Chinese or English text the Czech text is more complicated. In the case of processing Czech texts, higher demands are also placed on pre-processing. In particular, Stemmer and Lemmatizer guarantee the same identical form of words of the same form. The accuracy of classification can also be influenced by cultural differences and the level of abstraction. In countries where there is censorship, it can be expected that authors might be quite restrained and without motivation to embellish their newspaper headlines with abstract terms. On the basis of examples where the classification fails, some representatives were identified, on which an analysis of the failure causes was realized. Moreover, the examples were chosen subjectively to identify the causes of problems and suggest possible directions to be followed in future work. One example of failed classification is the phrase "Media zobrazují dívky jako sexuální objekty" (Media Show Girls as sex objects). This sentence was incorrectly classified to contain Disgust emotion. During the deeper analysis it was found that the training set was too small. Similar phrases appear in all categories with the same frequency but, unfortunately, in a very limited extent. Evaluated as having the greatest significance was the word "sex". Another example is the phrase "Kanaďané obchodovali s dětským pornem" (Canadians have traded child porn). In this case, the classifier identified this sentence as containing less than 20 % intensity of Disgust emotion. This was due to the fact that other emotions prevailed over the Disgust emotion, In spite of this we think that the classifier was correct and properly classified the Disgust emotion. Another typical example where the classification often fails is when the words themselves require a certain degree of the knowledge of semantics, based on cultural practices. As a representative of this group was chosen the phrase "Začni pozdě, do důchodu jdi bohatý" (Start late, retire rich). In this case, the classifier did not classify the emotion of Surprise.

6. Conclusion

This paper provides a description and evaluation of learning models of identification and classification of emotions in Czech news headlines. As mentioned in section 3, it is relatively difficult to accurately identify each emotion. One of the reasons can be that it is closely related to semantic understanding of a text. For example, the headline "Jestřábi se ujímají vedení" (the Hawks are taking on leadership) can have a different meaning in a sports' magazine and in a hunters' journal. However, even when someone fully understands the semantic content of the message, two different people can still have invoked totally different emotions. For example the message that "Euro kleslo na nejnižší úroveň" (the Euro has dropped to the lowest level) can be bad news for people living in Euro zone but for people having most of their deposits in the US dollars it can be good news. Based on an analysis of examples where the greatest problems in classification appear, it is clear, that future work should consist in using a significantly larger training set or combining the approach described in this paper with the ontological base of the Czech language such as the Czech WordNet database. Any method that would be able to take into account the semantic meaning of sentences and identify whether the event being described is usual or extraordinary, would also bring significant improvements.

Acknowledgements

The access to the MetaCentrum supercomputing facilities provided under the research plan MSM6383917201 is appreciated.

This work was supported within the framework of Research Project SIX (CZ.1.05/2.1.00/03.0072).

References

[1] COWIE, R., DOUGLAS-COWIE, E., SAVVIDOU, S., MCMAHON, E., SAWEY, M., SCHRÖDER, M. Feel trace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion.* Newcastle (Northern Ireland, UK), 2000.

[2] PŘINOSIL, J., SMÉKAL, Z., ESPOSITO, A. Combining features for recognizing emotional facial expressions in static images. In *Proceedings of Conference Information: International Conference on Verbal and Nonverbal Features of Human and Human-Machine Interaction, Lecture Notes in Artificial Intelligence*, 2007, vol. 5042, p. 56 - 69.

[3] PŘINOSIL, J., SMÉKAL, Z. Robust real time face tracking system. In *Proceedings of the 32nd International Conference on Telecommunications and Signal Processing - TSP2009*. Dunakiliti (Hungary), August 26-27, 2009, p. 101 - 104. ISBN 978-963-06-77169-5h.

[4] ATASSI, H., SMÉKAL, Z. Real-time model for automatic vocal emotion recognition. In *Proceedings of the 31st International Conference on Telecommunications and Signal Processing - TSP2008*. Parádfürdö (Hungary), September 3 - 4, 2008, p. 21- 25. ISBN 978-963-06-5487-6.

[5] DOLAMIC, L., SAVOY, J. Indexing and stemming approaches for the Czech language. *International Journal on Information Processing and Management,* 2009, vol. 45, no. 6, p. 714 - 720.

[6] RAMOS, J. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the 1st Instructional Conference on Machine Learning*. Piscataway, New York, 2003.

[7] HAN, J., KAMBER, M., PEI, J. Data mining: concepts and techniques. Second edition. *The Morgan Kaufmann Series in Data Management Systems*.

[8] SEOL, Y. S., KIM, D. J, KIM, H. W. Emotion recognition from text using knowledge-based ANN. In *The 32nd International Technical Conference on Circuits/Systems, Computers and Communications*, 2008, p. 1569 - 1572.

[9] COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., VOTSIS, G., KOLLIAS, S., FELLENZ, W., TAYLOR, J. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 2001, vol. 18, no. 1, p. 32 - 80.

[10] STRAPPARAVA, C. MIHALCEA, R. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on*

Applied Computing. Fortaleza, Ceara (Brazil), March 16 - 20, 2008. SAC '08. ACM, New York, NY, p. 1556 - 1560.

[11] GHAZI, D., INKPEN, D., SZPAKOWICZ, S. Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles (California), June 05 - 05, 2010. ACL Workshops. Association for Computational Linguistics, Morristown, NJ, p. 140 - 146.

[12] KESHTKAR, F. INKPEN, D. A corpus-based method for extracting paraphrases of emotion terms. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles (California), June 05 - 05, 2010. ACL Workshops. Association for Computational Linguistics, Morristown, NJ, p. 35 - 44.

[13] VOLKOVA, E. P., MOHLER, B. J., MEURERS, D., GERDEMANN, D., BÜLTHOFF, H. H. Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles (California), June 05 - 05, 2010. ACL Workshops. Association for Computational Linguistics, Morristown, NJ, p. 98 - 106.

[14] YANG, D., LEE, W. S. Music emotion identification from lyrics.. In *the 11th IEEE International Symposium on Multimedia ISM '09.* December 14 - 16, 2009, p. 624 - 629.

[15] JIA, Y., CHEN, Z., YU, S. Reader emotion classification of news headlines. In *Natural Language Processing and Knowledge Engineering*, September 24 – 27, 2009, p. 1 – 6.

[16] KOZAREVA, Z., NAVARRO, B., VAZQUEZ, S., MONTOYO, A. A headline emotion classification through web information. In *Proceedings of SemEval-2007*. Prague (Czech Republic), June 2007.

[17] YIMNGAM, S., PREMCHAISAWADI, W., KREESURADEJ, W. Thai emotion words analysis. In *Eighth International Symposium on Natural Language Processing*. October 20 - 22, 2009, p. 211 - 215.

[18] INRAK, P., SINTHUPINYO, S. Applying latent semantic analysis to classify emotions in Thai text. In *2nd International Conference on Computer Engineering and Technology (ICCET).* April 16 - 18, 2010, vol. 6, p. V6-450 - V6-454.

[19] LESHED, G., KAYE, J. Understanding how bloggers feel: recognizing affect in blog posts. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. Montréal (Québec, Canada), April 22 - 27, 2006, p. 1019 - 1024.

[20] AMAN, S., SZPAKOWICZ, S. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*. Pilsen (Czech Republic), September 03 - 07, 2007. *Eds.* Matoušek, V., Mautner, P. *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg, p. 196 - 205.

[21] HANCOCK, J. T., LANDRIGAN, C., SILVER, C. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. San Jose (California), April 28 - May 03, 2007, p. 929 - 932.

[22] NEVIAROUSKAYA, A., PRENDINGER, H., ISHIZUKA, M. Recognition of affect conveyed by text messaging in online communication. In *Proceedings of the 2nd International Conference on Online Communities and Social Computing*. Beijing (China), July 22 - 27, 2007. SCHULER, D. Ed. *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg, p. 141 - 150.

[23] LEE, S. Y., CHEN, Y., HUANG, C. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles (California), June 05 - 05, 2010, p. 45 - 53.

[24] CALIX, R. A., MALLEPUDI, S. A., BIN, C., KNAPP, G. M. Emotion recognition in text for 3-D facial expression rendering. *IEEE Transactions on Multimedia*, 2010, vol. 12, no. 6, p. 544 - 551.

[25] PRIBIL, J., PRIBILOVA, A. Statistical analysis of spectral properties and prosodic parameters of emotional speech. *Measurement Science Review*, 2009, vol. 9, no. 4, p. 95 - 104.

About Authors ...

Radim BURGET (MSc 2006, Ph.D. 2010) is an assistant at the Department of Telecommunications of the Brno University of Technology. His research interests include application of evolutionary computation methods, data mining and text processing.

Jan KARÁSEK received his BSc degree in Teleinformatics (2007), MSc degree in Telecommunications and Informatics (2009) and MSc degree in Business and Management (2010) from the Brno University of Technology. He is currently post graduate student at the Department of Telecommunications of the same university. His current research interests include application of evolutionary computation methods, data mining and text processing.

Zdeněk SMÉKAL (MSc 1973, Ph.D 1978) is a professor at the Department of Telecommunications, Brno University of Technology. He has for a long time been engaged with the principles of discrete and digital processing of one- and multi-dimensional signals with application in digital signal processors. He specializes in particular in problems of processing speech signals, their extraction from noisy background, vocal tract modeling, and TTS synthesis.

Appendix A – List of Stop Words[4]

dnes,cz,timto,budes,budem,byli,jses,muj,svym,ta,tomto,tohle,tuto,tyto,jej,zda,proc,mate,tato,kam,tohoto,kdo,kteri,mi,nam,tom,tomuto,mit,nic,proto,kterou,byla,toho,protoze,asi,ho,nasi,napiste,re,coz,tim,takze,svych,jeji,svymi,jste,aj,tu,tedy,teto,bylo,kde,ke,prave,ji,nad,nejsou,ci,pod,tema,mezi,pres,ty,pak,vam,ani,kdyz,vsak,ne,jsem,tento,clanku,clanky,aby,jsme,pred,pta,jejich,byl,jeste,az,bez,take,pouze,prvni,vase,ktera,nas,novy,tipy,pokud,muze,design,strana,jeho,sve,jine,zpravy,nove,neni,vas,jen,podle,zde,clanek,uz,email,byt,vice,bude,jiz,nez,ktery,by,ktere,co,nebo,ten,tak,ma,pri,od,po,jsou,jak,dalsi,ale,si,ve,to,jako,za,zpet,ze,do,pro,je,na

[4] Source: http://www.ranks.nl/stopwords/czech.html

Appendix B – Analysis of the Training Set

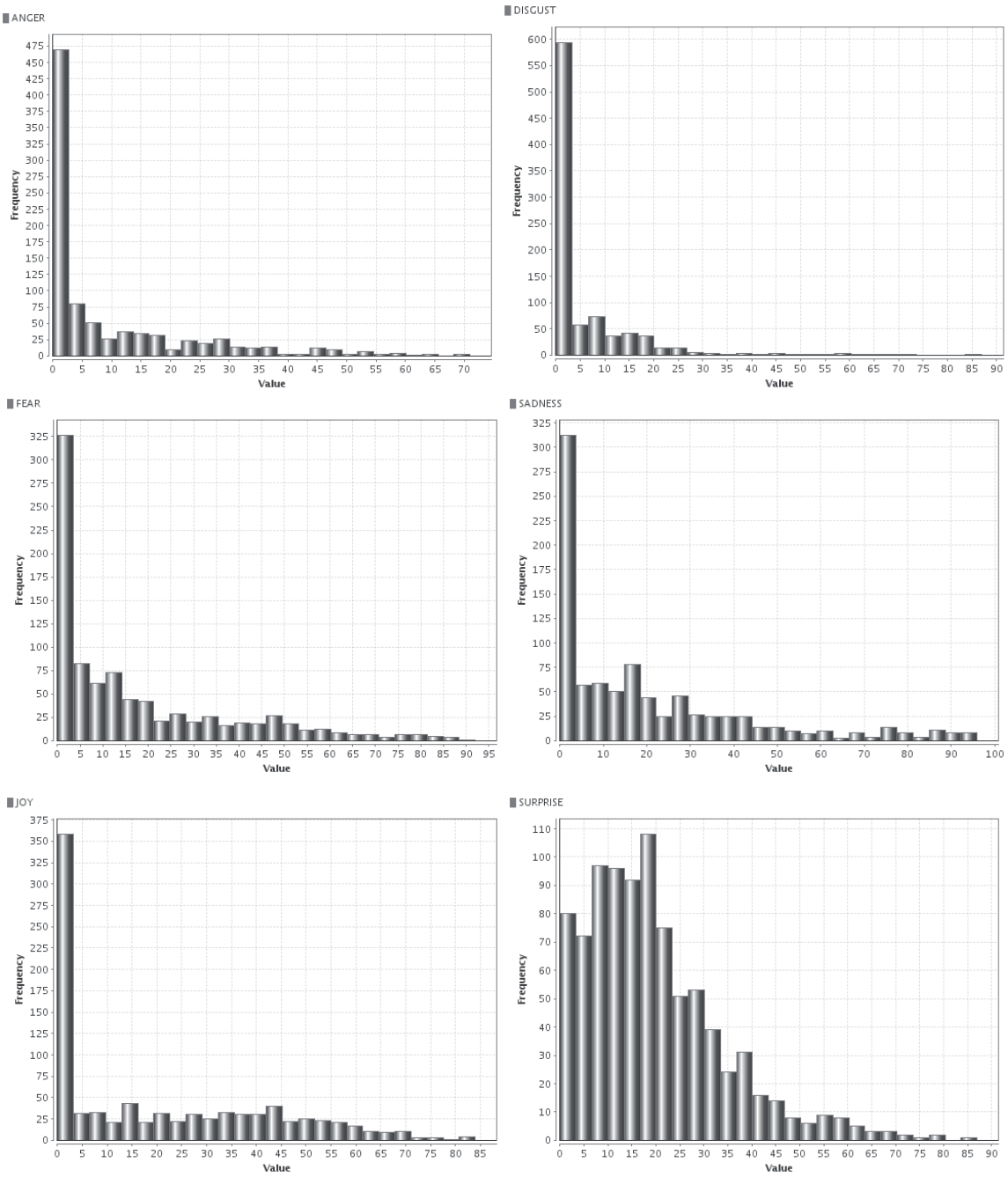

Fig. 5. Histograms of subjective strength of emotion frequency in the training set.