Identifying Fear Related Content in Croatian Texts

Lucia Načinović*, Benedikt Perak† Ana Meštrović*, Sanda Martinčić-Ipšić*

* Department of Informatics, University of Rijeka,
Omladinska 14, 51000 Rijeka, Croatia
{lnacinovic, amestrovic, smarti}@inf.uniri.hr

† Department of Cultural Studies, University of Rijeka,
Slavka Krautzeka bb, 51000 Rijeka, Croatia
bperak@ffri.hr

Abstract

This paper presents the initial work for the task of identifying the emotion of FEAR in texts written in the Croatian language. For the purpose of this analysis, text articles, blogs and online comments were collected from specific Croatian websites and classified into two categories: "*Fear present*" and "*Fear not present*". In the process of classification, supervised machine learning method based on Naïve Bayes algorithm was used. We experimented with different sets of features to compare their impact on the accuracies of the learnt classifiers and to determine which set of features yields better result (accuracy). The first set of features was constructed on a semantic model of embodied metonymic domains of fear derived from the cognitive semantic study of metonymic constructions of fear in Croatian. The second set of features contained only lexical concepts of fear, its synonyms and direct hyponyms. The third set is a combination of both sets. The results are presented in terms of accuracies, indicating that both embodied metonymic and lexically expressed features can be relevant for the emotion identification in texts.

Identifikacija s strahom povezanih vsebin v hrvaških besedilih

V članku je predstavljena začetno delo pri nalogi identifikacije čustva " strah" v besedilih, zapisanih v hrvaškem jeziku. Za namen te analize smo s hrvaških spletnih strani zbrali članke, bloge in spletne komentarje ter jih razvrščali v dve kategoriji: "Prisotnost strahu" in "Odsotnost strahu". Pri razvrščanju smo uporabili nadzorovani postopek strojnega učenja, ki temelji na naivnem Bayesu. Preizkušali smo različne nabore značilk in domen in opazovali njihov vpliv na natančnost naučenih razvrščevalnikov, da bi določili nabor značilk, ki daje najboljše rezultate razvrščanja. Prvi nabor značilk smo razvili na podlagi semantičnega modela vgrajenih metonimičnih domen strahu, ki izvirajo iz kognitivne semantične raziskave strahu v hrvaščini. Drugi nabor značilk je vseboval le leksikalne koncepte za strah, njegove sinonime in hiponime. Tretji nabor značilk pa je združeval oba prejšnja nabora. Dobljeni rezultati razvrščanja izraženi z mero pravilnosti, kažejo da sta obe metonimično in leksično zasnovane značilkerelevantne za identifikacijo čustev v besedilih.

1. Introduction

This paper presents research in the field of sentiment analysis. Sentiment analysis is a computer aided process of identifying different affective states within a particular segment of specific text corpora.

From the epistemological standpoint it can be said that sentiment analysis of a particular emotional category is a lost cause, for how can a machine detect emotions of another when it does not have emotions of its own? However, from the perspective of cognitive science one could argue that the process of emotional analysis in humans is not that dissimilar from those used in computers. After all, we as humans learn to acquire different emotional words for affective states, establishing connections between symbolic labels and our psychological, physiological states and behavioural traits (Davidson et al., 2003; Lewis et al., 2008). Furthermore, we learn how to express them appropriately in communication and elicit those states in others (Fussell, 2002; Barrett et al., 2007).

Categorization and meaning of emotions is complex and dynamic informational process emerging from the interaction of neurobiological, cognitive and symbolic structures (Barrett, 2011; Damasio, 1999). With so much lacking in comparison to human neurobiological structure, it would be irrational to demand from machines (on this level of technology) to *feel* emotions or to *recognize* emotional categories. On the other hand, to *analyze* and *identify* emotional categories on the basis of cognitive networks expressed in the linguistic symbolic structures is a feasible task. Analysis of emotional content in texts can therefore be reduced to the identification of emotional conceptual schemas. The need for conceptual organization of emotions is necessary because the epistemological nature of affective phenomena is highly individual. Therefore, the embodied feeling of a certain emotion is always purely subjective: no one can feel emotion of the other. Insuring incommensurability of affective experience enables cultural relativity in emotional categorization and lexicalization (Wierzbicka, 1999; Boster, 2005).

According to the FrameNet Project (Baker et al., 1998), based on Fillmore's frame semantics, core frames of the lexical concept FEAR are: (1) Experiencer - a person or a sentient entity that experiences or feels emotions. (2) Expressor - a body part, gesture, or other expression of the Experiencer that reflects his or her emotional state. (3) State - an abstract noun that describes a more lasting experience by the Experiencer. (4) Stimulus - a person, event, or state of affairs that evokes emotional response in the Experiencer. (5) Topic - a general area in which the emotion occurs.

Using theoretical framework of Cognitive Semantics, we modelled the identification of emotional category FEAR with reference to the embodied Expressor frame and its related metonymic domains, as well as using related lexical concepts of the State frame (Perak, 2011). One of the motivations for this study was to see whether embodied metonymic domains of fear could be relevant features for the identification of fear related content.

Related work on emotion (fear) identification was reported in many recent papers. In (Strapparava & Mihalcea, 2008) the identification of all major emotions from news headlines and blogs was based on WordNet-Affect. Blogs were also used in work by (Aman & Szpakowicz, 2007). Mohammad (Mohammad, 2012) reported on fear/no fear classification of each sentence in the newspaper headlines and blogs comparing the ngram and emotion based lexicon features. Tao (Tao, 2004) detected emotions from text based on lexicon consisting of emotional keywords, modifier words and metaphor words. The extensive work on feature selection for sentiment analysis is reported in (Duric & Song, 2011). The first research regarding sentiment classification in Croatian texts is reported in (Agić et al., 2010).

The process of text classification using Naïve Bayes and the framework for this work is described in section 2. The procedure of document collection is given in section 3. Feature sets extracted from document collection are presented in section 4. The results of the conducted experiment are given in section 5. The paper ends with some concluding remarks.

2. Text classification with Naïve Bayes

In our study, we tried to automatically identify the emotion of FEAR in specific corpora of the Croatian language. The initial task of this research was to classify articles, blogs and comments collected from web-sites into two categories: "*Fear present*" and "*Fear not present*" using supervised machine learning classification method based on Naïve Bayes algorithm. We experimented with different sets of features to compare their impact on the accuracies of the learnt classifiers and to determine which set of features yields better result (accuracy).

There are various approaches to text classification ranging from hand-written rules to unsupervised and supervised automatic machine learning techniques (Manning, 2009). The Naïve Bayes classifier relies on a simple representation of a document as a "bag of words". Another assumption that Naïve Bayes classifier entails is that the feature probabilities are independent of each other given the class. For this initial experiment we used Naïve Bayes classifier.

In text classification, we are given a set of documents d and a fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$. Classes are also called categories or labels. In supervised classification, documents are represented by feature sets which capture the basic information about each input (document) that should be classified. The classes are human defined. The training set consists of m hand-labelled documents with the corresponding class annotations $(d_1,c_1),\ldots,(d_m,c_m)$.

The probability of a document d being in class c is computed as (Manning et al., 2009):

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c. $P(c)$ is the prior probability of a document occurring in class c.

The goal of text classification is to find the best class for the document. In Naïve Bayes classification, the best class is the most likely or maximum a posteriori class c_{map} (Manning et al., 2009):

$$c_{map} = \underset{c \in C}{\operatorname{argmax}}\, \hat{P}(c|d) = \underset{c \in C}{\operatorname{argmax}}\, \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

We do not know the true values of parameters $P(c)$ and $P(t_k|c)$ but we use their estimates $\hat{P}(c)$ and $\hat{P}(t_k|c)$. We estimate the prior probablity by the formula:

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

where N_c is the number of documents in class c and N_{doc} is the total number of documents.

The conditional probability $\hat{P}(t_i|c_j)$ for each term (feature) in class is estimated as the fraction of times term t_i appears in the set of all terms (V) in documents of the class c_j:

$$\hat{P}(t_i|c_j) = \frac{count(t_i, c_j)}{\sum_{t \in V} count(t, c_j)}$$

3. Document collection

For the purpose of our research, documents from various online information sources were collected. We collected a total of 3218 articles (5534367 tokens). The selected web-sources included 4 religious and 5 political portals, 6 blog spaces, 3 web-pages with comments and 4 columns from the daily newspapers (shown in Table 1).

Religion	
	http://www.glas-koncila.hr/
	http://www.hbk.hr/
	http://www.hrvatskipravoslavci.com/
Politics	
	http://www.vlada.hr/
	http://mrak.org/
	http://novapolitika.blog.hr/
	http://pollitika.com/ptracker
	http://www.ivangrubisic.com/
Blogs	
	http://www.blog.hr/
	http://www.bloger.hr/index.aspx
	http://www.monitor.hr/vijesti/kategorija/blogovi/
	http://www.jutarnji.hr/komentari/blogovi/
	http://blog.vecernji.hr/
	http://www.slobodnadalmacija.hr/Blogeri/tabid/54/Default.aspx
Comments	
	http://www.novilist.hr/Komentari
	http://www.index.hr/vijesti/komentatori/
	http://www.jutarnji.hr/komentari/komentari_sub/
Columns	
	http://www.jutarnji.hr/komentari/kolumne/
	http://www.vecernji.hr/kolumne/
	http://www.glas-slavonije.hr/kolumne.asp
	http://www.dubrovacki.hr/pregled/kolumne

Table 1: Web-sources used in document collection

Abovementioned web-pages and respective topics were chosen because their genre and literary style indicated that subjective account of emotions (particularly fear) would be expressed to a greater extent than in the articles from any randomly chosen web-pages. All articles are written in the Croatian language.

Out of 3218 articles, 1507 articles were manually annotated into the categories "*Fear present*" and "*Fear*

not present" by 23 annotators. They were instructed to annotate the articles according to their subjective judgment whether there is fear related content present in text or not. Out of 1507 annotated articles, there were 1263 articles categorized into the category "*Fear not present*" and 244 articles were categorized into the category "*Fear present*" (statistics shown in Table 2).

Category	Number of articles	Tokens
Fear present	244	271966
Fear NOT present	1263	2130275
Total number of articles	1507	2402241

Table 2: Statistics of the document collection used in the research

4. Feature selection

The process of sentiment analysis is designed by selection of salient features. According to (Duric & Song, 2011), the criteria that are useful in selecting salient features for sentiment analysis include: a) features should be expressive enough to add useful information to the classification process, b) all features together should form a broad and comprehensive viewpoint of the entire corpus, c) features should be as domain-dependent as possible, d) features must be frequent enough and e) features should be discriminative enough.

In our classification procedure of fear related content we experimented with three different sets of features. First set of features contained only words that represent embodied metonymical domains of fear, i.e. bodily features of experiencing and expressing fear such as *tresti* (en. *tremble), blijed* (*en. pale*), *hladan* (*en. cold*), *znoj* (*en. sweat*). The embodied metonymical domains were provided by corpus based research of metaphoric and metonymic emotional conceptualization of fear in the Croatian language (Perak, 2011). In this work, Perak identified 2231 metonymical constructions that profile an embodied emotional model of fear. Out of 2231 metonymical constructions 60 words were directly associated with physical manifestation of fear. This result was used as the initial feature set for the automatic fear identification of Croatian texts presented in this paper. The list of 60 words was morphologically expanded using Croatian Morphological Lexicon (Tadić & Fulgosi, 2003) and manually verified. Finally, the first feature set "*Physical manifestation*" contained a list of 505 words related to the physical manifestation of fear.

The other set of features contained only lexical concepts of fear, its synonyms and direct hyponyms such as *strah* (en. fear), *prestrašiti* (en. scare), *horor* (en. horror), etc. which were extracted from the English WordNet and translated to Croatian. The resulting set contained 270 words which formed the second feature set "*Lexically expressed fear*".

The third set of features labelled "*Combination*" with 775 words is the union of all words from the first and the second feature sets.

5. Experiment

In our initial experiment, the Naïve Bayes classifier was trained on the document collection (described in section 3) using NLTK - Natural Language Toolkit (Bird et al., 2012). For each feature set described in section 4, 10 runs of Naïve Bayes training/testing was performed. For each run the document collection was randomly divided into training and test set in the ratio 9:1. The accuracy was computed as the average accuracy of ten different runs for each feature set. The results are shown in Table 3.

Feature set used in classification	Accuracy
Physical manifestation	0.8
Lexically expressed fear	0.83
Combination	0.81

Table 3: Accuracies for three different features sets

We obtained the best accuracy with the feature set that contains words that lexically express fear (0.83). The accuracies of the other two classifiers were slightly lower.

During the first experiment, we also identified the most informative features, i.e. features that have the biggest ratios of conditional probabilities. For each feature set, 30 most informative features were selected in order to have feature sets of the same size. Then, the procedure of classifier training and testing was repeated with the 30 most informative features from each feature set in order to compare the performance of reduced feature sets of the same size.

The results with the 30 most informative features are shown in Table 4.

Feature sets reduced to 30 most informative features	Accuracy
Physical manifestation	0.8
Lexically expressed fear	0.82
Combination	0.83

Table 4: Accuracies for reduced feature sets

Ten most informative features in *"Physical manifestation"* feature set are: cold *hladnu*, limbs *udovi*, face *licem*, pale *blijedi*, small *malima*, green *zeleni*, bitter *gorko*, green *zelena*, limb *ud*, fat *debelim.*

Ten most informative in *"Lexically expressed fear"* feature set are: alert *uzbunu*, horror *grozote*, alert *uzbuna*, chill *jeza*, panic *panika*, fear *straha*, terror *terora*, fears *strahove*, panic *panici*, afraid *boji.*

Ten most informative features in *"Combination"* feature set are: alert *uzbunu*, bitter *gorka*, horrors *grozote*, limbs *udovi*, face *licem*, pale *blijedi*, seen *vidljivi*, green *zeleni*, panic *panika*, alert *uzbuna*.

The best results of identification were achieved with the 30 most informative features from the combined feature set (embodied metonymic profiles/features and explicitly lexically expressed fear). So far, the results support the usage of physically expressed features in fear identification.

The performance comparison of "*Physical manifestation"* and "*Lexically expressed fear*" feature sets shows that the embodied profiles of the physiological processes of perception, representation and reaction can be relevant features for the identification and elicitation of the concept fear even without explicit lexicalization of the

category. The same principle should be considered for identification of other emotions in Croatian texts as well.

6. Conclusion and future work

In this work the articles, blogs and comments collected from the Croatian web-sites were classified into two categories: "*Fear present*" and "*Fear not present*". Supervised machine learning classification method based on Naïve Bayes algorithm was used. This classification process was designed to automatically determine whether there is fear related content in the text or not. We experimented with different sets of features in order to compare their impact on the accuracies of the learnt classifiers and to determine which set of features yields better result in terms of achieved accuracy. The experiment that we conducted was the initial attempt in fear identification in Croatian texts.

The best accuracy was achieved with the feature set that contains words that lexically express fear: In the first experiment, the performances of the feature sets could not be compared since there were far more features in the third set ("*Combination"*) which comprises of all the features from the first ("*Physical manifestation"*) and the second feature set ("*Lexically expressed fear"*). The second experiment was conducted with the purpose of comparing accuracies depending on feature sets of the same size. In the reduced feature set experiment the best accuracy is obtained with combined feature set (embodied metonymic profiles/features and explicitly lexically expressed fear).

The results encourage further research in combining cognitive interpretation of various sensory-motor, visceral, causative or culturally related domains that lead to a particular kind of affective experience with the lexicon of semantically related words of the emotional category for emotion recognition in Croatian texts. The first step will be verifying the corpus annotation by repeating the annotation of the same articles by more annotators in order to obtain the agreement of the annotators. We also plan to lemmatize corpus. Afterwards, the experiments including other algorithms and feature sets are planned. Additional features such as the most frequent phrases and derived metonymical and metaphorical constructs will be considered as well. We would also like to extend the research by experimenting with the identification of other emotions in order to identify different types of emotional content in Croatian texts.

7. References

Agić. Ž., Ljubešić. N., Tadić. M., 2010. *Towards Sentiment Analysis of Financial Texts in Croatian*. Proc. 7th International Conference on Language Resources and Evaluation. Valletta: 1164-1167

Aman. S., Szpakowicz. S., 2007. *Identifying Expressions of Emotion in Text*, In Proc. 10th International Conference on Text, Speech and Dialogue. Plzen, Czech Republic. LNCS 4629. Springer. 196-205.

Baker. C.F., Fillmore. J., Lowe. J.B., 1998. *The Berkeley FrameNet Project*. In Proc. 36th Annual Meeting of the Association for Computational Linguistics (ACL '98),. Association for Computational Linguistics. 86-90.

Barrett Feldman. L., 2011. Constructing emotion. *Psychological Topics*, 20/2: 359-380.

Barrett Feldman. L., Lindquist, K., & Gendron, M. 2007. Language as a context for emotion perception. *Trends in Cognitive Sciences*, 11, 327-332.

Bird. S., Klein. E., Loper. E., 2012. *Natural Language Processing with Python*. O'Reilly.

Boster. J., 2005. Emotion categories across languages. In: Cohen, H. and Lefebvre, C. (ed.) *Handbook of categorization in Cognitive science*: 187-222. Amsterdam, NL: Elsevier.

Damasio, A., 1999. *The feeling of what happens body and emotion in the making of consciousness*. New York. Harcourt.

Davidson. R.J., Scherer, K., Goldsmith, H., 2003. *Handbook of Affective Sciences*. New York. Oxford University Press.

Duric. A., Song. F., 2011. *Feature Selection for Sentiment Analysis Based on Content and Syntax Models*. Proc. 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 96-103.

Fussell, S., 2002. *The Verbal Communication of Emotions. Interdisciplinary Perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Manning. D.C., Raghvan. P., Schütze. H., 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

Mohammad. S., 2012. *Portable Features for Classifying Emotional Text*, In Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: HLT.Montreal,Canada.

Lewis, M., Haviland-Jones, J., Barrett Feldman, L., 2008. *Handbook of Emotions.* New York: The Guilford Press.

Perak. B., 2011. *The role of Embodied Cognition in Conceptualization of the Emotional Categories.* Context, Review for Comparative Literature and Cultural Research. 9; 193-1-212-20

Strapparava. C., Mihalcea. R., 2008. *Learning to identify emotions in text*. In Proc. ACM symposium on Applied computing. New York. 1556-1560.

Tadić. M., Fulgosi. S., 2003 *Building the Croatian Morphological Lexicon*. Proc. EACL2003 Workshop on Morphological Processing of Slavic Languages. Budapest. 41-46.

Tao. J., 2004. *Context Based Emotion Detection from Text Input*. 8th International Conference on Spoken Language Processing, ICSLP2004. Jeju. 1337-1340.

Wierzbicka, A. (1999) Emotions across languages and cultures. Cambridge, UK: Cambridge University Press.