SUPPORT OF INFORMAL CARERS FOR PEOPLE AFTER A STROKE WITH CROWDSOURCING AND NATURAL LANGUAGE PROCESSING

Petr ŠALOUN[*], Barbora CIGÁNKOVÁ[**], David ANDREŠIČ[**], Lenka KRHUTOVÁ[***]
[*] Palacky University Olomouc, Krizkovskeho 511/8, CZ-771 47 Olomouc, Czech Republic, E-mail: petr.saloun@upol.cz
[**] Faculty of Electrical Engineering and Computer Science, VSB - Technical University of Ostrava, Ostrava, Czech Republic, E-mails: {barbora.cigankova.st, david.andresic}@vsb.cz
[***] Faculty of Social Studies, University of Ostrava, Ostrava, Czech Republic, E-mail: lenka.krhutova@osu.cz

ABSTRACT

. For a long time, both professionals and the lay public showed little interest in informal carers. Yet these people deals with multiple and common issues in their everyday lives. As the population is aging we can observe a change of this attitude. And thanks to the advances in computer science, we can offer them some effective assistance and support by providing necessary information and connecting them with both professional and lay public community.

In this work we describe a project called "Research and development of support networks and information systems for informal carers for persons after stroke" producing an information system visible to public as a web portal. It does not provide just simple a set of information but using means of artificial intelligence, text document classification and crowdsourcing further improving its accuracy, it also provides means of effective visualization and navigation over the content made by most by the community itself and personalized on a level of informal carer's phase of the care-taking timeline.

In can be beneficial for informal carers as it allows to find a content specific to their current situation. This work describes our approach to classification of text documents and its improvement through crowdsourcing. Its goal is to test text documents classifier based on documents similarity measured by N-grams method and to design evaluation and crowdsourcing-based classification improvement mechanism. Interface for crowdsourcing was created using CMS WordPress. In addition to data collection, the purpose of interface is to evaluate classification accuracy, which leads to extension of classifier test data set, thus the classification is more successful.

Keywords: *Classification, Text documents, Natural language processing, Documents similarity, N-grams, Crowdsourcing, WordPress, Caretaker, Stroke.*

1. INTRODUCTION

Informal carers deal with many difficult situations when care-taking their relatives. Today we can utilize the technology and computers to support them by providing a specific, personalized content and information. We created a web portal providing all necessary information which should help them to improve their care-taking, but it requires a meaningful and effective navigation over the content made from the most by carers themselves. We face issues like differences between professional and laical nomenclature which we attempt to address using specific language mean ment for navigation over the content called tag cloud. In addition to this, we also provide visual mean attempting to place the carer to a correct position on a care-taking timeline. It is all implemented in a web server running WordPress CMS. The result will be hand-overed to a non-profit organization supporting the target group of informal carers. The project connected beneficients of the care-taking and their organization with IT specialists in order to improve their support and quality of life.

In our effort we utilize so-called natural language processing (NLP), which is usually used in for example information extraction tasks or text classification, where it helps to automate and speed up the classification process. Despite the progress in text classification, humans are usually still more accurate, which opens a space for human-assisted classification, e.g. by means of having human as a reference system. Here, it is also possible to use the collective intelligence of multiple people for such task, which is called crowdsourcing that is experiencing boom in the last years.

In this work, we briefly describe our experience and summarize our previous results of Czech, Slovak and English text documents classification which serve as a base for this work [42]. We also summarize crowdsourcing advantages, methodology and comparison with other approaches.

The aim of the experimental project of agile software development called "Research and development of support networks and information systems for informal carers for persons after stroke" is to create an information systems (IS) for informal carers for a person after stroke (ICs) using modern information technology that allows the users to gain relevant, timely and interconnected information on support networks for prevention of their possible social isolation and exclusion, physical and psychological exhaustion, health disorders and other risk factors associated with their difficult life situation. The developed IS will also help improve awareness of ICs support systems across other long-term care providers – in particular public administrators, general practicioners etc. Intent of the project is to create and validate the pilot IS IC model in Moravian-Silesian Region by 12/2021, which can subsequently be applied in other regions and / or other target IC groups.

1.1. Informal Carers

For a long time, both professionals and the lay public showed little interest in informal carers. The change in the attitude to informal carers which we are witnessing during the last decades is primarily – although not exclusively – caused by the urgent need for a solution to the demographic trends of population ageing. In parallel with the situation in other countries, informal carers in the Czech Republic have neither been sufficiently identified, nor systematically supported. "While it is possible to define other focus groups by a certain social event (such as maternity by giving birth) or socio-economic characteristics (such as age in senior citizens or lack of employment in the unemployed), caregivers are a group that is largely non-demarcated/ undefined." [29]

This is reflected also in the area of terminological definition of informal care and informal carers, this area is highly divergent and differs in the depth and the breadth of the definition concepts, in the purposes of the definitions, in the definition criteria chosen as well as the in the fields in which (or for which) the terminology is defined. For the purposes of this paper, we will proceed from the similarities found in these definitions. In the absence of terminological consensus (if it can ever be achieved), terminological definitions of informal care – and hence of informal carers – usually have the following in common: ". . . informal care involves lay [. . .] care conducted without any specific professional education, without financial remuneration and outside of one's employment, and is accompanied with a high degree of emotional involvement." [32]. Concurrently, lay care may include the involvement of both family members as well as friends, neighbours, acquaintances or colleagues and the like.

One of the key factors with regard to support provided to informal carers is how well informed these carers are. This information process is often marred by obstacles [30] [31] as carers are exposed to the deficits of an "invisible group", and it is not infrequent that information reaches them in a haphazard manner. These people are not always aware of the fact that as a result of their caregiving, they themselves may belong to a group of people whose physical, psychical, social or economic, etc. health may be at risk. For a number of reasons, unlike care recipients themselves (i.e. people who suffered CVA in our case, or children with a disability and the like), informal carers do not set up formalised associations. With the exception of a type of sickness benefit recently introduced in the Czech Republic under the name "long-term care benefit", informal carers are usually not visible both for the formal systems of support and from the perspective of their potential to form self-help groups. Among other problems, this also significantly hinders research in this area, or as the case may be, identification and searching for potential research respondents.

Informal carers have become the ever-more frequent object of professional investigation (for instance [32] and [33]). Research in this area deals with the informal carers' contribution to the economy [34], or attempts to quantify how well informed carers are [35] [36], it also focuses on the quality of their life [37], satisfaction [38], their situation on the labour market [39], how effective intervention is in reducing their strain [40], or with their emotions and stress [41]. Research which is relevant in relation to the focus group of our project and of this paper furthermore includes research focusing on informal carers in the context of caring for people after Cerebrovascular Accidents [24] [25] [26] [27] [28].

2. TEXT DOCUMENT CLASSIFICATION AND CROWDSOURCING

The main goal of text classification is to assign the given text to some of the pre-defined classes. In the area of text mining, it is also a process of automatic learning of categorization schemas used for direct classification of new, uncategorized documents [1]. Some approaches use different forms of document similarity metric, such as cosine similarity. This metric is then used in learning as well in classification phases. Before the classification itself, it is necessary to perform two steps:

- Transformation of the document to a form that can be parsed. This includes removal of stop words, tags and other pre-processing (see section 6). –
- Extraction of text properties that are then evaluated and their weight is calculated. These properties are then represented as vectors describing a presence of words or syntactic unit [1]. Many classifiers use a bag-of-words (BOW) approach for text representation [1]. It is a simplified text representation used mostly for NLP and information extraction where the document is transformed to a set of individual words with- out grammar structures and words order, but still containing possible words duplicity. During the classification, an occurrence frequency for each word in the bag is calculated so it can be then used as an input for classifier during training.

Today classifiers use either statistical approaches or machine learning and can be divided into two categories: supervised and unsupervised. Further text in this section describes today most common algorithms such as decision trees, N-grams, artificial neural networks and Bayes classifier [1].

2.1. Naive Bayes Classsifier

Naive Bayes Classifier is a probability-based classifier built on top of Bayes theorem (described for example in [2]) saying how conditional probability of some event relates to an opposite conditional probability. Bayes classifier assumes that presence or absence of some attribute of the given class is not dependent on presence or absence of some other attribute [1]. The advantage of Bayes classifier is that it performs well with smaller training data set to determine statistical parameters.

2.2. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is often used for term- weighting (evaluation of individual text attributes). It is a statistical metric that measure an importance of words in the given document [3]. Term Frequency stands for count of the given word in the document divided by the total count of all words in the

document. This normalization is done to eliminate the advantage of long documents in such calculation. The Inverse Document Frequency represents the importance of individual words. It is characterized as a logarithm of count of all documents divided by count of documents containing the given word [3]:

- TF(t) = (count of t in the document) / (total count of words in the document)
- IDF(t) = (total count of documents / count of documents that contain the t)

Matching documents will then have a high frequency of the given word that is also not so much present in other documents. One of the major disadvantages of TF-IDF is its ignorance of key semantic connections between words because it compares documents only based on frequency of individual words. Still, different variations of TF-IDF are often used in search engines for document ranking [1].

2.3. Latent Semantic Analysis (LSA)

LSA (also known as Latent Semantic Indexing - LSI) is a technique used for NLP. It is based on analysis of relationship between set of documents and words contained in them. In contrast to classic natural language processing or artificial intelligence approaches, LSA is not using any human-created dictionary, knowl- edge base, grammar or syntactic parser. The input of LSA is just a text divided into meaningful parts such as sentences or paragraphs [4]. LSA uses mathemati- cal approach called Singular Value Decomposition (SVD). It is a method of linear algebra in which a regular matrix is decomposed to 3 smaller matrices such that matrix multiplication of these matrices must return the original matrix. The whole process is described for example in [5].

2.4. Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning method that is usually used in binary classification and regression analysis. In is based on a con- cept of decision planes that defines decision borders [6]. SVM uses a mechanism called hyperplanes in multidimensional space which divides objects of individual classes. The main idea of SVM is to allow linear division of objects of different classes using object transformation that is being done by mathematical functions called kernel function [6]. It is then crucial to find the most fitting hyperplane (plane with maximal margin), that is, find the place in which the the distance between closest points to the plane is as large as possible. In order to describe the hyperplane, we need just points that lies at the edge of maximal margin. These points are called support vectors [7]. Other points are not relevant to the hyper- plane. SVM method is therefore capable to find those training samples which are most relevant to finding the hyperplane. The size of the training set required for classifier learning is therefore much smaller [7]. We recognize several types of SVM that differ by used iterative algorithm for error function minimization. They are described for example in [6].

2.5. N-grams

N-gram is defined as a tuple of N items that belongs to some sequence of e.g. words or characters. Sequence of two items is called bigram, sequence of three items then trigram. From four, it is called generally as N-gram. N-grams are usu- ally used for text representation where words are used in the sequence. Another possible usage is document classification based on document similarity. During the classification, sequence of e.g. characters is used. The beginning and the end of the word is then marked by some special character such as underscore [8].

In general, a set of N-grams for a string of length k will contain k+1 N-grams. Great advantage of classification using N-grams is its independence on document language, because there is no need for text pre-processing such as stemming or lemmatization. It is also quite tolerant to grammar errors and typos.

On the other hand, a large number of generated N-grams can be considered as a drawback. On the other hand, this can be reduced by e.g. removing stop words or by using stemming or lemmatization (or some other text length reduction), but by doing this, we lose the advantage of language independence.

In [20] authors for example used character N-grams and unigram indices for Twitter tweets classification. They confirmed language independence but also conclude that although character n-grams of 4-6 characters length leads to classification models with decent performance, the manually indicated tokens (a.k.a. crowdtagging) combined with a Decision Tree classifier outperform any other feature set-classification algorithm combination [20].

3. CROWDSOURCING

The concept of crowdsourcing can be defined as a business practice where the given activity is outsourced to a crowd [14]. Another definition can be found in [15] where author says that crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. The most important part of this definition is the undefined network of people. Everyone can then get task assigned to him or her. The only selection that is done in such process is selection of achieved results. Results are also often just aggregated.

Crowdsourcing theoretical roots were defined in [22]. It is based on an idea of collective intelligence. This concept can be understood as ”all together we are smarter than just one of us” [16]. It is a concept also known as wisdom of the crowd. In [14] authors attempts to answer the 8 basic questions about crowdsourcing. As for advantages of crowdsourcing, we can name for example releasing core company employees for other work and lower expenses. A nice description of crowdsourcing pros can be found in [17]. One of the most difficult tasks in crowdsourcing usage is finding the right crowd motivation [18].

3.1. Examples of Use

Several large companies such as Waze, Lego, Samsung, Lays or Greenpeace suc- cessfully used the crowdsourcing in real world applications [23]. In [20] authors used crowdsourcing to obtain tokens for sentiment analysis of tweets and used them as a feature set which turned out to perform best in compare to other feature sets established by other means (e.g. N-grams). Similarly in [21] authors compared various kinds of low-level features, including those extracted through deep learning and conclude that keywords suggested by the crowd (called crowd lexiconherein that are based on crowdtagging), established through a crowd- sourcing platform can be effectively used for training sentiment classification models for short texts (tweets and Facebook comments) and that those models are at least as effective as the ones that are developed through deep learning or even better [21].

4. PROOF OF CONCEPT RESULTS

Our classifier was tested on two data sets and then briefly with use of crowd- sourcing.

4.1. Data Sets

Both data sets contained X text documents in Y classes. Each was then split into training and test set. After processing of each set, the classification accuracy was evaluated.

4.1.1. Language Data Sets

As mentioned in [42], our first data set contained technical texts from DATAKON conferences in different languages. The aim of this data set was to confirm language independence of the classifier. We used Czech, Slovak and English texts here. Each category contained 40 texts with 60 up to 20 words. Both training and test data set contained 20 texts for each class. Our classifier successfully classified all 60 documents in test data set. Perfect accuracy was expected for English, but we expected worse numbers for Czech and Slovak that are very similar to each other. The language independence of N-grams method was therefore confirmed.

4.1.2. Psychological Data Set

Contrary to the first data set, this data set contains not so balanced count of texts for each class. Its aim was to investigate how the classifier will perform with not so well structured data. Also, these texts contains psychological topic. They are sorted to classes which borders are not so clear as in case of previous data set. These texts are often difficult to classify even by human. The expected accuracy of classification therefore was not high. The data set contained 87 documents in 3 classes. As discussed in [42], there was one misclassified document in each class. In general, these errors were related to documents difficult to be classified even by human as they contain several topics at once.

4.1.3. Crowdsourcing

The classifier accuracy was also tested by implemented crowd- sourcing interface. Our crowd contained people from OSU[1] and VŠB-TUO[2] universities. Topics of contributions inserted into the interface were suggested as life of non-formal care takers and its influencing as a consequence of care taking. This resulted into 3 classes [42]. Training set was provided by OSU. It consisted 180 one- or two-sentence texts classified into 4 categories. Crowd that creates texts for classification (and also performing classification accuracy testing) con- sisted from students of Faculty of Medicine of OSU. During a test phase, correct class was assigned to the text in case of error. We suggested an approach to extend the training data set by incorrectly classified contributions and observed an increasing trend of classification accuracy, yet on a very small example.

5. REAL-WORLD APPLICATION FOR INFORMAL CARERS

We utilized our approach in a real-world application for informal carers. The idea is to classify each new user post to one of the 4 pre-defined classes corresponding to identified phases of informal carers timeline:

1. The initial shock - first encounter with a stroke.
2. The time of the acute care.
3. The time of the post-acute care (rehabilitation).
4. The time of home care-taking.

These new posts will be automatically classified using the N-grams algorithm and then the user will be kindly asked to correct the class if necessary. Misclassified posts will then be used to extend the training set and re-training the classification model. Key words of each class will then be displayed in a tag cloud (see Fig. 1) where each tag (key word) will lead to a list of articles and posts related to its class (see Fig. 2) so the user can access the content and find useful information related to his/her particular phase of the informal carer's timeline.

Currently we achieve 50% classification accuracy using very small training data set made of the timeline description and initial moderated discussion. We expect that further learning using crowdsourcing as well as extension of stop words dictionary will increase the classification accuracy.

[1] https://www.osu.cz/

[2] https://www.vsb.cz/

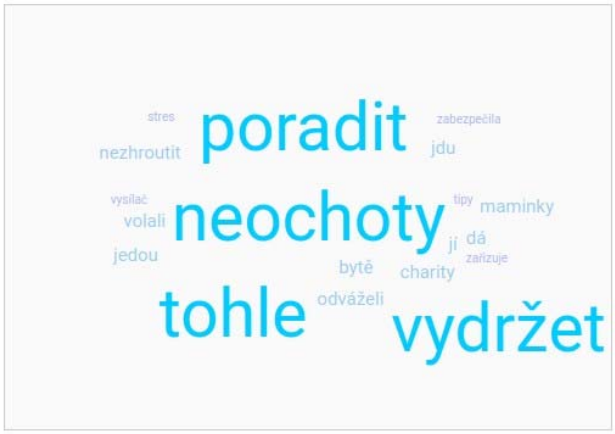

Fig. 1 Tag cloud on the web for informal carers

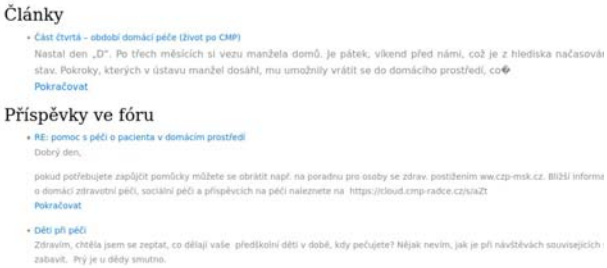

Fig. 2 Tagged content on the web for informal carers

6. PROCESSING OF TEXTS WRITTEN IN NATURAL LANGUAGE

NLP software requires consistent knowledge base such as large dictionary, gram- mar rules, ontology and synonyms etc. [10]. NLP process consists of several phases using different methods to ”decrypt” multiple language unclarities, e.g. tagging of part of the speech or understanding and recognition of the natural language [10]. These phases can be [9]: morphological analysis, syntax analysis and semantic analysis. Morphological analysis processes a single word as the smallest atomic unit. Using dictionary, it assigns a basic form to a word, word class and other morphological categories. Syntax analysis, on the other hand, processes whole sentences and formal description of their structures. Semantic analysis determines the meaning of word or a broader sentence. From these methods, morphological analysis is the one most explored and most algorithmizable [9]. On the other hand, semantic analysis is generally most difficult due to words homonymy.

Before almost every text processing, several pre-processing steps must be done, such as transformation to a lower-case form, removing of special characters, stopwords and tokenization. Another usual steps are stemming and lemmatization.

6.1. Stemming and Lemmatization

Stemmers and Lemmatizers are attempting to find the common base or root of each word in the text. These tools are useful for e.g. counting the frequency of words in text because they allow to unite different forms on a words with the same meaning. Stemmers are working with individual words without context and thus cannot distinguish between different meaning of words. They are simply cutting prefixes and suffixes (and leaving just stems). For more details, please see e.g. [11]. On the other hand, lemmatization is working with morphological analysis of words. Lemmatization tools are working with grammar rules for the document language. More details can be found in e.g. [12].

6.1.1. Stemmers and Lemmers for Czech Language

Czech language is in general one of those more difficult for stemming and lemmatization. Czech language uses a lot of prefixes and has more complex inflection. Due to this there are not many usable frameworks or software libraries.

One solution offers Apache Lucene[3]. This search engine offers Czech language analyzer that contains set of Czech stop words, stemmer and tokenizer that can be enhanced by filters for e.g. lower-case transformation. The only disadvantage is the absence of Czech lemmatizer. This can be compensated by Czech morphological analyzer developed by Masaryk university[4] called Majka[5]. In its base settings, it assigns to each word [13]: basic form and grammar mark, all words related to the same lemma and all possible words with diacritic.

7. CLASSIFIER DESCRIPTION

We used a classifier built on top of our prototype [42]. The implementation is based on N-gram-based text categorization described in [8] and consists from two phases: training and classification. In the training phase, training documents are tokenized, stop words are excluded, n-grams with length from 2 to 5 are generated and ordered based on their frequency of appearance and finally class profile made of first 600 n-grams is generated and stored. In the classification phase, the profile of unknown document is calculated (similar to the training phase, just for a single document). Then the distance between unknown document profiles and profiles in database are calculated using out-of-place method. And at last, unknown document is assigned with a class with shortest distance.

In opposite to [8] our classifier utilizes a reduction of count of words in document by removing the stop words. Using this reduction, it is not necessary to start in the profile

[3] https://lucene.apache.org/
[4] https://www.muni.cz/
[5] https://www.muni.cz/vyzkum/publikace/935762

class at the position 300 (as suggested in [8]) but it is possible to start from the beginning of the list. Also, our classifier works with longer profiles, mostly because of planned classification of psychological text. Their classes have usually a very thin border so we can expect the need of more N-grams [42].

Beside the classification, our application also determines key words of each class. These key words will then be displayed to selected users with a kind request to use them in their contribution. By showing key words only to some users, we create two user groups that will serve as referential groups to confirm the following hypothesis: classification will perform better if contributions for classification contain pre-defined key words.

The calculation of key words is realized by TF-IDF algorithm modified (in respect to [19]) to class purpose. The calculation will look like following:

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (1)$$

where n_{ij} is frequency of term i in documents of class j.

$$IDF_i = \log(\frac{|D|}{|d:t_i \in d|}) \quad (2)$$

where t_i is term and D is set of all classes. TF_{ij} is quotient of term frequency n_{ij} to count of all word in documents of the given class. IDF_i is then logarithm of quotient of classes count to count of classes containing term t_i.

Five key words with greatest weight per category from those obtained by this method are selected and stored to database. We distinguish several user roles for crowdsourcing user interface. Logged users can insert their contributions (a.k.a. posts) where some of them will be kindly asked to use pre-selected keywords based on selected contribution category. These contributions will be then classified by our classifier and in case of discrepancy, user can correct the category. Such contribution will then be added to the training set (by Editor via manual data synchronization) in order to improve classifier's accuracy.

8. CONCLUSIONS

The aim of this work was to create a text document classifier based on text document similarity with further usage of crowdsourcing in order to increase classification accuracy as a mean of support of informal carers for people after a stroke. After an analysis of classification algorithms, N-grams algorithm was chosen, mainly for its language independence but also for is easy implementa- tion. The classifier was then connected with the crowdsourcing interface of the web portal built on top of WordPress. The accuracy was tested on two data sets and then by crowdsourcing interface proving its language independence and mis- classifying only border-line cases difficult even for a human. Eventually, classifier accuracy was left to the users themselves using our crowdsourcing interface. In order to improve the accuracy, extending the training data set (especially with incorrectly classified texts) was implemented.

The web portal was launched in 2021 and the project itself connected beneficients of the care-taking and their organization with IT specialists in order to improve their support and quality of life. It provides a clear visual timeline of informal carer's phases with necessary information and effective navigation to content made by other informal carers from the same phase of the timeline.

ACKNOWLEDGMENTS

The project "Research and development of support networks and information systems for informal carers for persons after stroke" implemented by the University of Ostrava and VSB - Technical University of Ostrava (project ID: TACR/TL02000050, project duration: 2019 – 2021) was supported by the Technology Agency of the Czech Republic. We also want to express gratitude to the application guarantors of the project – The Association for Rehabilitation of People after Cerebrovascular Accidents and the Moravian-Silesian Region.

REFERENCES

[1] KARMAN, S. Senthamarai – RAMARAJ, N.: Similarity-Based Techniques for TextDocument Classification. Int. J. SoftComput, 2008, 3.1: 58-62.

[2] OPITKA, P. – ŠMAJSTRLA, V.: "PRAVDĚPODOBNOST A STATISTIKA," [In Czech] (Probability and statistics) 2013. [Online]. Available: https://homen.vsb.cz/oti73/cdpast1/KAP02/PRAV2.HTM. [Accessed on 4. 3.2018].

[3] "Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining," [Online]. Available: http://www.tfidf.com/. [Accessed on 25. 12. 2017].

[4] LANDAUER, Thomas K. – FOLTZ, Peter W. – LAHAM, D.: An introductionto latent semantic analysis. Discourse processes, 1998, 25.2-3: 259-284.

[5] HAJEK, P. et al.: Možnosti využití přístupu indexování latentní sémantiky při předpovídání finančních krizí. POLITICKÁ EKONOMIE, [In Czech] (Possible use of indexed latent semantic approach for financial crisis prediction) 2009, 6: 755.

[6] "Support Vector Machines (SVM)," TIBCO Software Inc, [Online]. Available: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Accessed on 28. 12. 2017].

[7] ŽIŽKA, J.: "Studijní materiály předmětu FI:PA034," [In Czech] (Study materials to FI:PA034) [Online]. Available: https://is.muni.cz/el/1433/podzim2006/PA034/09 SVM.pdf. [Accessed on 29. 12. 2017].

[8] CAVNAR, William B. et al.: N-gram-based text categorization. Ann arbor mi,1994, 48113.2: 161-175.

[9] HABROVSKA, P.: "Vybrané kapitoly z počítačového zpracování přirozeného jazyka," 2010. [In Czech] (Selected chapters from natural language processing) [Online]. Available: http://www.inflow.cz/kratce-o-zpracovani-prirozeneho-jazyka.

[10] SCAGLIARINI, L. – VARONE, M.: "Natural language processing and text mining," 11 April 2016. [Online]. Available: http://www.expertsystem.com/naturallanguage-

processing-and-text-mining/. [Accessed on 15. 12. 2017].

[11] KODIMALA, S.: Study of stemming algorithms. 2010.

[12] RISUENO, T.: "The difference between lemmatization and stemming," 28. 1.2018. [Online]. Available: https://blog.bitext.com/what-is-the-difference-betweenstemming-and-lemmatization/. [Accessed on 4. 3. 2018].

[13] ŠMERK, P. – RYCHLÝ, P.: "Majka – rychlý morfologický analyzátor," [In Czech] (Majka - quick morphological analyzer) 2009. [Online]. Available: https://www.muni.cz/vyzkum/publikace/935762. [Accessed on 15. 12. 2017].

[14] ESTELLÉS-AROLAS, E. – GONZÁLEZ-LADRÓN-DE-GUEVARA, F.: Towards an integrated crowdsourcing definition. Journal of Information science, 2012, 38.2: 189-200.

[15] SCHENK, E. – GUITTARD, C.: Crowdsourcing: What can be Outsourcedto the Crowd, and Why. In: Workshop on Open Source Innovation, Strasbourg, France. 2009.

[16] AITAMURTO, T. – LEIPONEN, A. – TEE, R.: The promise of ideacrowdsourcing–benefits, contexts, limitations. Nokia Ideas project White Paper, 2011, 1: 1-30.

[17] KALSI, M.: "Crowdsourcing through Knowledge Marketplace," 3. 3. 2009. [Online]. Available: http://blog.spinact.com/knowledge as a service/2009/03/crowdsourcing-throughknowledge-marketplace-.html. [Accessed on 2018 3. 4.].

[18] KAUFMANN, N. – SCHULZE, T. – VEIT, D.: More than fun andmoney. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In: AMCIS. 2011. p. 1-11.

[19] VRL, NICTA. An unsupervised approach to domain-specific term extraction. In: Australasian Language Technology Association Workshop 2009. 2009. p. 94.

[20] TSAPATSOULIS, N. – DJOUVAS, C.: Feature extraction for tweet classification: Do the humans perform better? In: Proceedings of the 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2017), pp. 53-58, Bratislava, Slovakia, July 2017.

[21] TSAPATSOULIS, N. – DJOUVAS, C.: Opinion mining from socialmedia short texts: Does collective intelligence beat deep learning? Frontiers in Robotics and AI, vol. 5, article 138, January 2019, DOI: 10.3389/frobt.2018.00138

[22] SUROWIECKI, J.: The Wisdom of Crowds. 2005.

[23] KEARNS, K.: "9 Great Examples of Crowdsourcing in the Age of Empowered Consumers," 10. 7. 2015. [Online]. Available: http://tweakyourbiz.com/marketing/2015/07/10/9-great-examplescrowdsourcing-age-empowered-consumers/. [Accessed on 10.3. 2018].

[24] ALBRECHT, D. – WOLLENSAK, T. – ERNST, C. – BECKER, C. – HAUTZINGER, M. – PFEIFFER, K.: (2015). Costs of informal care in a sample of German geriatric stroke survivors. Eur JAgeing 13(1): 49–61. DOI: 10.1007/s10433-015-0356-x.

[25] CHEN, P. – FYFFE, D. C. – HREHA, K.: (2017). Informal caregivers' burden and stress in caringfor stroke survivors with spatial neglect: an exploratory mixed-method study. Top Stroke Rehabil 24(1): 24–33. DOI: 10.1080/10749357.2016.1186373.

[26] IMARHIAGBE, F. A. – ASEMOTA, A. U. – ORIPELAYE, B. A. – AKPEKPE, J. E. – OWOLABI, A. A. – ABIDAKUN, A. O. et al.: (2017). Burden of Informal Caregivers of Stroke Survivors: Validation of the Zarit Burden Interview in an African Population. Ann Afr Med 16(2): 46–51. DOI: 10.4103/aam.aam 213 16.

[27] JOO, H. – DUNET, D. O. – FANG, J. – WANG, G.: (2014). Cost of informal caregiving associatedwith stroke among the elderly in the United States. Neurology 83(20): 1831–1837. DOI: 10.1212/WNL.0000000000000986.

[28] PENDERGRASS, A. – BEISCHE, D. – BECKER, C. – HAUTZINGER, M. – PFEIFFER, K.: (2015). An abbreviated German version of the Sense of Competence Questionnaire among informal caregivers of relatives who had a stroke: development and validation. Eur J Ageing 12(3): 203–213. DOI: 10.1007/s10433-015-0342-3.

[29] GEISSLER, H,: (2021). Neformální péče v datech [Informal care in data]. In: Fryč V,Chmelová M, Adámková P (Eds) (2021). Neformální péče v teorii a praxi. Sborník odborných statí. Praha: Pasparta publishing, pp. 56–64.

[30] GEISSLER, H. – HOLEŇOVÁ, A. – HOROVÁ, T. – JIRÁT, D. – SCHLANGER, J. – SOLNÁŘOVÁ, D. et al.: (2015a). Výstupní analytická zpráva o současné situaci a potřebách pečujících osob a bariérách pro poskytování neformální péče v ČR [Analytical report on current situation and needs of carers and barriers for informal caretaking in Czech republic]. Praha: Fond dalšího vzdělávání.

[31] GEISSLER, H. – HOLEŇOVÁ, A. – HOROVÁ, T. – JIRÁT, D. – SOLNÁŘOVÁ, D. – SVOBODOVÁ, K. et al.: (2015b). Neformální péče ve vybraných státech Evropské unie. Komparativní rešerše a identifikace příkladů dobré praxe [Informal care in selected states of European Union. Comparative research and identification of best practices]. Praha: Fond dalšího vzdělávání.

[32] HOROVÁ, J. – BÁRTLOVA, S. – HAJDUCHOVÁ, H. – MOTLOVÁ, L. – TREŠLOVÁ, M. – ZÁŠKODNÁ, H. – BRABCOVÁ, I.: (2021). Mezinárodní přehled podpory neformálního (rodinného) pečovatelství [International overview of informal (family) care-taking support]. Sociální práce/Sociálna práca 21(2): 20–43.

[33] HUBÍKOVÁ, O.: (2021). Rozvoj sociální práce zaměřené na neformální pečující. In: FRYČ, V. –

CHMELOVÁ, M. – ADÁMKOVA, P.: (Eds) (2021). Neformální péče v teorii a praxi [Informal care in theory and practice]. Sborník odborných statí. Praha: Pasparta publishing, pp. 84–12.

[34] MOSQUERA, I. – VERGARA, I. – LARRAÑAGA, I. – MACHÓN, M. del Río M, CALDERÓN, C.: (2015). Measuring the impact of informal elderly caregiving: a systematic review of tools. Qual Life Res 25(5): 1059–1092. DOI: 10.1007/s11136-015-1159-4.

[35] ARAÚJO, O. – LAGE, I. – CABRITA, J. – TEIXEIRA, L.: (2015). Intervention in informal caregiverswho take care of older people after a stroke (InCARE): study protocol for a randomised trial. J Adv Nurs 71(10): 2435–2443. DOI: 10.1111/jan.12697.

[36] MORRIS, S. M. – THOMAS, C.: (2002). The need to know: informal carers and information. Eur J Cancer Care 11(3): 183–187. DOI: 10.1046/j.1365-2354.2002.00337.x.

[37] WHITE, C. L. – LAUZON, S. – YAFFE, M. J. – WOOD-DAUPHINEE, S.: (2004). Toward a model ofquality of life for family caregivers of stroke survivors. Qual Life Res 13(3): 625–638. DOI: 10.1023/B:QURE.0000021312.37592.4f.

[38] SIMON, C. – LITTLE, P. – BIRTWISTLE, J. – KENDRICK, T.: (2003). A questionnaire to measuresatisfaction with community services for informal carers of stroke patients: construction and initial piloting. Health Soc Care Community 11(2): 129–137. DOI: 10.1046/j.1365-2524.2003.00408.x.

[39] CASADO-MARÍN, D. – GARCÍA-GÓMEZ, P. – LÓPEZ-NICOLÁS, A.: (2011). Informal care and labour force participation among middle-aged women in Spain. SERIEs 2: 1–29. DOI: 10.1007/s13209-009-0008-5.

[40] NEMČÍKOVÁ, M. – KATRENIAKOVÁ, Z. – DOBRÍKOVÁ, P. – NAGYOVÁ, I.: (2020). Efektívne intervencie pre znižovanie záťaže neformálnych opatrovatělov osôb s demenciou pri Alzheimerovej chorobe: systematický prehľad [Effective interventions for lowering the load of informal carers for people with demencia and Alzheimer's disease: systematical overview]. Sociální práce/Sociálna práca 20(6): 120–140.

[41] PESANTES, M. A. – BRANDT, L. R. – IPINCE, A. – MIRANDA, J. J. – DIEZ-CANSECO, F.: (2017). An exploration into caring for a stroke-survivor in Lima, Peru: Emotional impact, stress factors, coping mechanisms and unmet needs of informal caregivers. eNeurologicalSci 6: 33–50. DOI: 10.1016/j.ensci.2016.11.004.

[42] SALOUN, P. – ANDREŠIČ, D. – CIGÁNKOVÁ, B. – ANAGNOSTOPOULOS, I.: "Crowd Sourcing as an Improvement of N-Grams Text Document Classification Algorithm," 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, 2020, pp. 1-6, DOI: 10.1109/SMAP49528.2020.9248454.

Received November 24, 2021, accepted December 2, 2021

BIOGRAPHIES

Petr Šaloun was born in Olomouc, Czech Republic. He graduated and passed the rigorous exam at the Faculty of Science Palacky University in Olomouc, Czech Republic in 1986. Ph.D. gained in the field of Computer Science and Engineering at the Technical University in Prague in 2002. He is Associated Professor in Computer Science at the both: Faculty of Education, Palacký University Olomouc, Czech Republic, and Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, Czech Republic. Professional orientation of Petr Saloun is the processing of large data peta-byte (astromical observation), information systems, personalized, adaptive systems, electronic and web publishing technology for e-learning, sequential and parallel parsing of formal languages and compilers and object oriented programming in C++ and Python, including their teaching in bacelor level of study. During his career, Petr Saloun published tens of articles in the proceedings of both domestic and foreign professional and research conferences, he is the author of three books and four university textbooks. He also collaborated on a number of research and educational projects in the field. Petr is ACM professional member

Barbora Cigánková was graduated in 2018 (Ing.) at the department of Computer science of the Faculty of Electrical Engineering and Computer Science at VŠB - Technical University of Ostrava. As she left the university, she has been working in the software engineering area.

David Andrešič was born on 6.4.1988. In 2016 he graduated (Ing.) at the department of Computer science of the Faculty of Electrical Engineering and Computer Science at VŠB - Technical University of Ostrava. Since 2016 he is applying for a PhD at Department of Computer Science. His scientific research is focusing on big data analysis by means of unconventional algorithms.

Lenka Krhutová was born in Ostrava, Czech Republic. She graduated Special Pedagogy at the Faculty of Pedagogy Palacky University in Olomouc, Czech Republic in 1987. Ph.D. gained in the field of Social Work at the University of Ostrava in 2008. Since 2015 she is Associated Professor Social Work in Faculty of Social Studies University of Ostrava and head of Department of Health and Social Studies. Her professional orientation is focus on disability, social work, coordinated rehabilitation, long-term care and integrated health social work