Tomasz Stanisławek, Jarosław Protasiewicz, Marek Kozłowski, Agata Kopacz

A classification of the questionnaire of reviewers and applicants

Ekonomiczne Problemy Usług nr 106, 321-343

2013

ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 781 EKONOMICZNE PROBLEMY USŁUG NR 106 2013

TOMASZ STANISŁAWEK
JAROSŁAW PROTASIEWICZ
MAREK KOZŁOWSKI
AGATA KOPACZ
Information Processing Institute

A CLASSIFICATION OF THE QUESTIONNAIRE OF REVIEWERS AND APPLICANTS

Introduction

A questionnaire is a research instrument consisting of a series of questions in order to gather information from respondents. Usually, a questionnaire consists of a number of questions that the respondent has to answer in a set format. A questionnaire[1] can be defined as a series of processes, that extract useful information in order to solve problems, by asking people involved in the problem the same question, collecting data as answers to the questions, and analyzing them. Questionnaires are mainly conducted for statistical analysis of the responses.

A form of the questionnaire consists of open-ended and closed-ended questions. A closed-ended question limits respondents with a given number of options from which they must choose to answer the question. The response options for a closed-ended question should be exhaustive and mutually exclusive. An open-ended question asks the respondent to formulate his own answer.

[1] H. Inui, M. Murata, K. Uchimoto, H. Isahara, *Classiffication of open-ended questionnaires based on surface information in sentence structure*, In Proceedings of the 6th NLPRS2001, pp. 315-322, 2001.

This kind of question gives the answering person a scope of information that seems appropriate to them. A respondent's answer to an open-ended question is afterwards coded into a response scale or multi-label categorized.

The open form of the questionnaire consists of one style of responding to the questions. This open form is also called a free descriptive questionnaire, since, in that style, the respondents freely describe answers to the prepared questions. This format has been distinguished from the fixed-alternative, in which answers are of a closed form[2].

Questionnaire data, that consist only of closed answers is relatively easy to handle, because they are structured. Researchers have proposed many methods for analyzing these kinds of answers, using such multivariate analysis techniques as cluster analysis and correspondence analysis. Questionnaire data that includes open answers is much more difficult to analyze automatically. At first, they are segmented (split into sequences of sentences) and tokenized (sentences are divided into lists of words). Next, texts represented as vectors of tokens are processed by text mining methods such as text-clustering techniques or the self-organizing map technique. The idea here is to view each answer as a vector of words and to use similarity measures to cluster the vectors. Those kind of methods are effective for summarizing answers, but they are inefficient in extracting target characteristics. Other researchers have proposed methods for analyzing open answers on the basis of associations between the words. The approach is based on calculating associations between word pairs based on their co-occurrences in open answers and then visually present the words and associations on a two-dimensional map[3]. In the paper[4] authors are focused on the open questions in the questionnaire and discuss the problems encountered during the analysis of the responses to such questions, from the viewpoint of statistical NLP. Combining statistical analyses and information retrieval techniques in which the context of questionnaires is discussed[5].

---

[2] *Ibidem.*

[3] K. Yamanishi, H. Li., *Mining open answers in questionnaire data*, IEEE Intelligent Systems 2002.

[4] L. Lebart, A. Salem, L. Berry, *Exploring Textual Data*, Kluwer Academic Publishers 1998.

[5] S. Hirasawa, F. Shih, W. Yang, *Student questionnaire analyses for class management by text mining both in Japanese and in Chinese*, In Proc. 2007 IEEE International Conference on System, Man and Cybernetics 2007.

Authors introduce the methods of data mining and text mining (e.g. LSI, EM algorithms) in order to cope with questions answered by a fixed format and those by a free format. Apart from using traditional classifiers, there are also works focused on applying the association rules techniques to analyze questionnaire data[6]. Based on fuzzy techniques they discover fuzzy association rules from the questionnaire datasets, so that all different data types can be handled in a uniform manner.

Answers to open-ended questions often contain valuable information. The main problem associated with the analysis of survey data is that the manual handling is both cumbersome and very costly, especially when it exists in large volume. However, the analysis method for the open-ended answers has not been established well enough, and classification based on the content of the answers often needs manual operations. The costs of such operations are high and the result of human judgment is a lack of objectivity. In general, processing of answers in natural language is difficult because of the enormous variation in linguistic expression. This problem might be solved by applying language processing techniques, such as information extraction or automatic classification.

Our aim was to find the best computational approaches, using machine learning methods for the automatic classification of collected open-ended questionnaires, in order to speed up and reduce costs of a questionnaire's analysis. The presented approach is based on segmentation of open answers into words and conducting an analysis of the word, as well as in phrase levels. We have developed a survey analysis system that works on these principles. The proposed text mining methods provides a new way of analyzing natural-language responses to questionnaires. Using multi-label categorization techniques, we are able to extract semantic information about the open-ended questions, which is complex and multi-dimensional. This paper reports the results of our preliminary experiments, using svm, naive bayes for questionnaire classification.

[6] Y. Chen, C. Weng, *Mining fuzzy association rules from questionnaire data*, Knowledge--Based Systems Journal 2009.

1. Methods

1.1. Questionnaire of reviewers and applicants

Questionnaire foundations. Information Processing Institute supports many processes of grant funding in Poland by providing information systems. The first information system have been developed for science funding streams (OSF) managed by Ministry of Science and Higher Education. It has been launched on-line in 2004, and after this success more science funding processes have been computerized, for instance: Polish-Norwegian Research Fund (PN FBN), Polish-Swiss Research Programme (PSPB), Innovative Economy (PO IG). All of them are managed by Information Processing Institute. These systems usually contain the following modules: tools for on-line proposals preparing; tools for proposals processing used by an agency; a database and algorithms for selecting of reviewers; on-line tool for reviews.

Almost 19k reviewers have been asked since July 2011, whether they can prepare reviews using these systems. As a result, 132.5k requests for reviews were sent but 20.5k of them were returned by reviewers. The vast majority of reviews was prepared for grant programs managed by Ministry of Science and Higher Education. The reviewer's distribution was: 44% professors, 30% associate professors[7], 20% assistant professors[8] and 7% others. Most of them were employed at universities (67,1%), and 14,2% in research institutes, and 18,7% in other places[9].

Peer review process assumes that experts assessors are qualified and able to perform reasonable review about any scholarly work and research project, but in fact, the peer review is widely criticized. Neff and Olden[10] maintain that this process is open to misuse and influences on the editor and reviewer

[7] In Polish: dr hab.

[8] In Polish: doktor.

[9] *Procedures for review and selection of reviewers*, ed. J. Protasiewicz, Vol. 1 (in Polish), Information Processing Institute 2012.

[10] B.D. Neff, J.D. Olden, *Is peer review a game of chance?*, BioScience 2006, 56 (4), pp. 333-340.

integrity. For only 47% of scientists an article published in peer-reviewed journal proves its high quality[11].

Information obtained from foreign literature and desk research, were the inspiration for conducting an anonymous online survey. The aim of this study was to verify researchers perception of problems with peer review process in Poland. The survey was conducted on a group which included research staff, both reviewers (almost 20%) and applicants (45%). 35% of respondents had experience in both areas. Most respondents were assistants professors (43%), 28% were professors, 24% were associate professors and 5% with unreported degree. Respondents came from different disciplines, such as medicine, biology, economy, chemistry, physics, history, philology or computer science. 95% of the respondents had experience in Ministry of Science and Higher Education grant programs. 18% of scientists took part in the Innovative Economy and 17% in the National Centre for Research and Development programs. Polish-Swiss Research Programme applies 14% of respondents and Polish-Norwegian Research Fund 4%[12].

Answer categories and subcategories. The survey contained 14 closed-ended questions about researcher's perception of the peer review process in Poland, and one open-ended question which was a request for any further comments or suggestions about the experience of the peer review process. The questionnaire was completed by 8190 people, but the open-ended question was commented out only by 2615 of them (about 32%). According to the OPI experts, 301 answers were incomplete or irrelevant. The analysis of the answers would be time consuming and expensive. Therefore, our aim was to carry out an automatic classification using machine learning methods. The answers have been categorized in five categories of problems which consisted of sixteen subcategories[13] (Table 1).

[11] N. Macnab, G. Thomas, *Quality in research and the significance of community assessment and peer review: education's idiosyncrasy*, International Journal of Research & Method in Education 2007, 30(3), pp. 39-352.

[12] *Procedures for review and selection…*

[13] Categories and subcategories were identified by the OPI experts, but mainly by Agata Kopacz.

Table 1

The categories and the subcategories of answers to the open-ended question

Category	Reviewing	Evaluation	Work quality	Anonymity	Formalism
subcategory	away of reviewers choice	range	review quality	disclosure	formalism
	recall	criteria	reviewer's knowledge	anonymity	
	guidelines	aggregation	honesty		
	dialogue	ratings discrepancy	subjectivism		
	control of reviewer				

Source: own.

Problem definition. Lets consider a set of answers to the open-ended question in the questionnaire and denote it as

$$D = [d_1, d_2, ..., d_n]^T \tag{1}$$

Each answer $d_i, i = 1,2,...n$ may contain many statements

$$d_i = [s_{i,1}, s_{i,2}, ...s_{i,m}]^T \tag{2}$$

and they can refer to various problems mentioned by responders. These problems we defined in Table 1. Let denote a category as c_a and corresponding subcategory as sc_{ab}. An answer d_i can belong to many categories or subcategories. The task is to build a classifier which will be able to automatically assign categories and subcategories to each answer d_i. We have divided the set D into the training set D_{Train} and the testing set D_{Test}. The experts have manually prepared the training set in a special way: all answers d_i in the training set were split into statements $s_{i,j}$ and next subcategories *sc* were assigned to them. One subcategory was assigned to one statement. A statement is treated as a set of sentences or one sentence which should contain a consistent message in the same category.

2. Classifiers

Selected classification algorithms. Among many classification algorithms, there are some especially important, such as Support Vector Machines (SVM) and classifiers based on Bayes theorem: Naive Bayes (NB) and Multinominal Naive Bayes (MNB).

Naive Bayesian classifiers are based on two assumptions. Firstly, they consider documents as a bag of words where word position in a document does not affect the result of classification. Secondly, they assume that probability of word's occurrence in a document d_i is independent from probability of other word's occurrences for the given class. Therefore, we can easily calculate conditional probability that a sentence d_i combined form a bunch of words $x_{i,1}, x_{i,2}, ...x_{i,K}$ belongs to a class $c_l \in C$.

$$P(c_l \mid x_{i,1}, x_{i,2}, ...x_{i,K}) \approx P(c_l) \prod_{k=1}^{K} P(x_{i,k} \mid c_l) \tag{3}$$

and finally determines to which class belongs the document

$$c_{winner} = \arg\max c_l P(c_l \mid x_{i,1}, x_{i,2}, ...x_{i,K}) \tag{4}$$

Although an assumption of features independence is rather untrue, a Naive Bayes classifier works surprisingly well in practice. In this case a distribution of each feature $P(x_{i,k} \mid c_l)$ is not defined. If we assume that each feature has multinomial distribution, then we have Multinomial Naive Bayes. This assumption works well, for instance, in case of text classification where can be used in the word counts model. A bayesian classifier is learned from a set D_{Train} and this process involves: extracting vocabulary; computing a prior $P(c_l)$; calculating a likelihood $P(x_{i,k} \mid c_l)$ of belonging each word $x_{i,K}$ to each decision class c_l. These values are calculated as ratio between a number of documents or words representing a particular class and a total number of documents or words in class. There is possibility that a particular word in the test set D_{Test}, does not occur in the training set D_{Train}, so its likelihood will be equal to zero. Thus, due to the multiplication of the probabilities, an entire reviewer's answer will not be properly classified. There are several ways to solve this problem. The most frequent solution is to use

Laplace smoothing or determination the likelihood of low value correlated to all other probabilities[14].

Support Vector Machines was firstly presented in 1995 by Valdimir Vapnik. SVM uses a principle of structural risk minimization. The main idea of algorithm is to find such decision boundary which can separate classes - usually a positive one and a negative one. Regarding the classification problem there are distinguished linear and nonlinear cases. The SVM classifiers consider a document or a sentence as a bag of words **x** similarly to Naive Bayes. In the linear case the classes are separated by a hyperplane:

$$w * x - b = 0 \tag{5}$$

where the weights **w** are selected during teaching process using the train set D_{Train} and quadratic programming. Nonlinear cases are solved by using soft margin methods which allows some errors or by using a kernel function such as multinomial, gaussian or hyperbolic tangent[15].

Multi-class and multi-label classification. Typically a bayesian classifier assigns only one class with the highest probability while testing a particular answer (eq. 4). But as we mentioned previously, an answer $d_i, i = 1,2,...n$ to an open-ended question can belong to many subcategories, which we denote as the classes $c_l, l = 1,...L$. Therefore, this case contains either multi-class and multi-label problems, because the data set contains many classes (categories and subcategories - see Table 1). and the answers are assigned to many classes (labels). We can solve this issue in two ways. The first approach assumes that it is possible to use only one classifier in the manner of multi-label classification. The classifier e.g. Multinomial Naive Bayes produces as an output a vector of probabilities - one value for each class (eq. 3). The classes with the highest

[14] D. Fragoudis, D. Meretakis, S. Likothanassis, *Best terms: An efficient feature-selection algorithm for text categorization*, Knowledge and Information Systems 2005, 8 (1), pp. 16-33; T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York 2009; Z. Hoare, *Landscapes of naive bayes classifiers*, Pattern Analysis and Application 2008, 11 (1), pp. 59-72.

[15] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data,* Springer, New York, 2010; W. Noble, *What is a support vector machine?*, "Nature Biotechnology" 2006, No. 24, pp. 1565-1567; C. Silva, B. Ribeiro, *On text-based mining with active learning and background knowledge using svm*, Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications 2007, 11(6), pp. 519-530.

probability are taken as an outcome, but someone must decide how many classes should be taken into account. The second approach is using the procedure called one vs others. This procedure implies the use *L - 1* classifiers to solve the multi-class problem. Each classifier e.g Multinomial Naive Bayes is trained in a binary manner to recognize one class and all others. In classification stage all classifiers verify a new example and finally many classes can be assigned to it. There can be a situation when all classifiers choose class "others" and the tested example will be unclassified, or on the other hand too many classes will be assigned. In order to avoid the over classification someone has to experimentally choose a probability threshold of belonging to the class[16].

Model improvements. Before classification, the texts are pre-processed what involves: lemmatization, removing stopwords, determination the validity of the words, using TF-IDF (term frequency - inverse document frequency). The classifiers are trained using TF-IDF values of words from pre-processed sentences. We call it a basic form of our classification model. It is easy to notice that the quality of classifiers depends on the quality of texts pre-processing. We propose three improvements of the basic classification model. Firstly, the answers to open-ended question contain many misspellings what can interfere the lemmatization process. They can be corrected by an electronic vocabulary set. In case of the questionnaire it could be the Polish dictionary, for instance http://www.sjp.pl. Secondly, we deal with the texts in Polish, and we know that the Polish language has different grammar than English, so it needs special algorithms in order to properly extract keywords. We have developed the algorithm - Polish Keyword Extractor[17], which is based on Rapid Automatic Keyword Extraction (RAKE) and KEA. Finally, we should note that effectiveness of classification models depend on the quality of a training set and especially often on its size. The experts have agreed that the answers containing up to 220 words (about one or two sentences) should be classified in only one subcategory.

[16] G. Tsoumakas, I. Katakis, *Multi-label classifcation: An overview*, Int J Data Warehousing and Mining 2007, 1-13.

[17] *Procedures for review and selection of reviewers*, ed. J. Protasiewicz, Vol. 2 (in Polish), Information Processing Institute 2012.

SVM classifier parameter optimization. The parameters choice for SVM classifier is a nontrivial and laborious task, because there is no automatic and deterministic method which would allow selection of the best parameters to a specific issue. It is a nonlinear problem, and additionally involves many computations in case of classification of the questionnaire. Therefore, we propose applying a differential evolution (DE) algorithm[18] to optimize the parameters of SVM classifier. DE as a one of the evolutionary algorithms uses a population containing the vectors, which represent potential solutions. Finding the best vector means finding the best classifier parameters. It involves the following steps: initialization - a population of vectors is randomly created while keeping constraints for each parameter; mutation - for each vector is created a mutated vector, assuming that they differ from each other; recombination - a new vector is created in order to increase diversity of the population, provided that at least one parameter is derived from a mutated vector; selection - a vector formed during recombination is tested by an objective function, and the better one (new or old) is added to the new population. The algorithm stops when achieves fixed number of generations and the best matched vector is returned as an outcome.

2.1. Classification and assessment

To analyse open-ended questions using supervised methods we need to build a training set at first. Therefore, we divide our evaluation process into two stages (Preliminary model selection and Final classification procedure). In the first stage we build training set and provide classifier models which best match this problem. In the second one, we classify all open-ended questions by the classifier models selected in the first stage.

Preliminary model selection. We propose a preliminary classification stage in order to select the appropriate models and pre-processing procedures. This stage involves four experiments - we denote them as experiment 1, 2, 3, 4 in the section Results. Each experiment contains the following steps:

[18] R. Storn, K. Price, *Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces*, Journal of Global Optimization 1997, 11, pp. 341-359.

1. The experts create an initial training set D_{Train} with the same size (a number of answers d_i) for each subcategory sc_l.

2. Various classification models are tested using cross-validation procedure and the best classification model is chosen for further experiments.

3. Classification of the answers which have not been yet assigned to subcategory (usually 100 answers) using the model selected in the previous step.

4. The experts verify the experiment outcomes.

5. Based on classification errors the classification models are adjusted.

6. The training set D_{Train} is increased by classified answers (label assigned by experts), and a new experiment starts from point 2.

The training set sizes for consecutive experiments were as follows: 14 for experiment 1, 24 for experiment 2, 34 for experiment 3, 43 for experiment 4.

Using the above algorithm we have tested two approaches to classification problem: using one classifier in comparison to using many classifiers; model improvements, which were discussed above. There are experiments 1-4 for which details can be found in the section *Results*.

Final classification procedure. After selection of an adequate classification model, we conduct classification experiments of all answers to the open--ended question by repeating the following steps:

1. A classification program randomly selects 100 new answers from d_i, which have been unclassified yet.

2. The best classifier among tested in the previous experiment iteration classifies the answers.

3. The experts (people) verify the classification results.

4. All classifier types carry out experiments and the best one is chosen according to the selection criteria.

5. The classification program adds the verified answers by experts to the train set, and the next iteration is performed starting from point 1.

Using the above algorithm we perform the final classification and also optimize the SVM classifier parameters. There are experiments 5-17, which details can be found in the section *Results*.

Assessment measures. Classifiers need to be assessed on the basis of their outcomes. There are several measures that would be useful, but we should be aware of their meaning and use only the most suitable for our single

label and multi-label problem. Really simple and useful are measures based on comparison of a real subcategory and classifier decisions - as a result, the following values are received: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). A combination of these values gives three measures:

– precision

$$\mathrm{Pr}ec = \frac{TP}{TP + FP} \tag{6}$$

– recall, called also sensitivity

$$\mathrm{Re}c = \frac{TP}{TP + FN} \tag{7}$$

– F-measure (or F-score), which is harmonic mean of precision and recall

$$F = 2\frac{\mathrm{Pr}ec * \mathrm{Re}c}{\mathrm{Pr}ec + \mathrm{Re}c} \tag{8}$$

The evaluation of the multi-label data is difficult because it can be partially correct, we use Exact Match Ratio (EM)[19]. This measure indicates the percentage of examples that have all their labels correctly classified.

$$Exact\ Match,\ EM(D) = \frac{1}{k}\sum_{i=1}^{k} I(cs_i = l_i) \tag{9}$$

where, k is test example, I is the indicator function, l_i is a label subcategory vector of the i-th example, sc_i is predicted subcategory vector.

Another important issue is measuring of multi-label data, that can be represent (just like a single label data) by number of examples (n) and the number of subcategories (cs). We select three measures specific to the multi-label

[19] M.S. Sorower, *A Literature Survey on Algorithms for Multi-label Learning*, 2010.

problem, introduced in[20]. Label Cardinality (LC_{ARD}) is standard measure, that simply take the average number of labels associated witch each example:

$$LC_{ARD}(D) = \frac{\sum_{i=1}^{N} |l_i|}{N} \tag{10}$$

where $|l_i|$ is a number of subcategories in i-th example.

The second one is a Label Density (LD_{ENS}), relates to (LC_{ARD}) and includes the size of the label space. These measure gives good idea how frequently label occurs:

$$LD_{ENS}(D) = \frac{1}{L} LC_{ARD}(D) \tag{11}$$

Very often we use average values counted from many experiments. Therefore, measures presented above are denoted with prefix Avg in the section *Results*.

Results

Model selection. Initial experiments focused on assessing two classification models. We have tested MNB single multi-label classifier in comparison to using many classifiers by the procedure one vs others. In four experiments (we denote them as 1-4) 994 answers (36,1% of all answers) were classified using preliminary model selection procedure (see section Classification and assessment). Basing on the results which are presented in Table 2 we can conclude that the individual MNB classifier gives better or similar results as the procedure one vs others. Moreover this classification model is also less complicated and easier to implement.

[20] G. Tsoumakas, I. Katakis, *op. cit*.

Table 2

Comparison standard MNB classier with One vs others MNB classier

	Assigned number of class	**Experiment 1**				**Experiment 2**				**Experiment 3**				**Experiment 4**			
MNB classifier		EM	Prec	Rec	F	EM	Prec	Rec	F	EM	Prec	Rec	F	EM	Prec	Rec	F
	one	35,7	41,1	33,8	37,1	42,5	55,8	40,4	46,8	24,7	54,9	29,1	38,1	25	48,1	30	37,2
	two	0	27,7	45,6	34,5	39,8	51,1	44,9	47,8	16,1	48,5	45,7	47,1	21,2	43	46,1	44,5
	three	0	20,7	51,5	29,5	39,8	48,3	46,8	47,6	14	40,5	54,9	46,6	-	-	-	-
One vs others MNB classifier (threshold - 0,9)	-	28,6	41	36,2	38,5	36,3	51,1	43,1	46,8	22,1	60,9	37,4	46,4	26,7	46,1	35,8	40,3

Source: own.

After selecting the classification model we have tested three improvements of the basic classification model: misspelling correction using Polish dictionary (SJP); finding the most important words using Polish Keyword Extractor (PKE); enlarging data set (see section *Classifiers. Model improvements*). In the experiments involving model improvements we used the same data set as in the previous experiments therefore we denote them also as 1-4. The results presented in the Table 3 indicate that models containing improvements could be more efficient than basic Multinomial Naive Bayes classifier. Especially spelling correction using Polish language dictionary and large data set significantly improve the quality of classification.

Table 3

Classification experiments 1-4 with model improvements

	Assignet number of class	Experiment 1				Experiment 2				Experiment 3				Experiment 4			
		EM	Prec	Rec	F	EM	Prec	Rec	F	EM	Prec	Rec	F	EM	Prec	Rec	F
MNB classifier	one	35,7	41,1	33,8	37,1	42,5	55,8	40,4	46,8	24,7	54,9	29,1	38,1	25	48,1	30	37,2
	two	0	27,7	45,6	34,5	39,8	51,1	44,9	47,8	16,1	48,5	45,7	47,1	21,2	43	46,1	44,5
	three	0	20,7	51,5	29,5	39,8	48,3	46,8	47,6	14	40,5	54,9	46,6	-	-	-	-
Model using SJP	one	28,7	35,7	29,4	32,3	36,3	48,7	35,3	40,9	25,8	59,1	32	41,5	25	51	32,7	39,9
	two	0	21,4	41,2	28,2	34,5	45,6	39,7	42,5	15,1	48,5	47,2	47,8	20,2	41,8	48,1	44,7
	three	0	20,8	57,4	30,6	34,5	44,4	42,3	43,3	12,9	40,9	56	47,3	-	-	-	-
Model using PKE	one	25,4	35,7	29,4	32,3	28,3	40,7	29,5	34,2	21,5	58,1	30,9	40,3	21,2	41,4	26,1	32
	two	1,8	25,9	42,7	32,2	27,4	40,2	35,3	37,6	14	46,7	44	45,3	17,3	37,9	40,6	39,2
	three	0	20,2	50	28,8	26,6	38,4	37,2	37,8	11,8	40,9	55,4	47,1	-	-	-	-
Model using larger DS	one	33,9	44,6	33,8	38,5	49,6	62,8	46,8	53,6	29	63,4	33,1	43,5	25	51	31,5	39
	two	0	30,4	54,4	39	46,9	59,1	51,9	55,3	18,3	52,7	49,1	50,9	22,1	46,3	49,1	47,7
	three	0	20,5	64,7	31,2	46,9	55,6	53,2	54,4	15,1	42,2	57,7	48,7	-	-	-	-

Source: own.

Final classification. According to above findings, for final classification we decided to use the MNB classifier but enriched by Polish language dictionary and larger data set (we call them MNB classifier with improvements). Moreover, in the new experiments we have evaluated questionnaires using SVM classifier with default parameters and also parameters selected manually in an intuitive way. Before proceeding to the final classification, the experts have improved the training set. They examined the shortest texts that may adversely affect the quality of the classifier result by adding more relevant data from original reviewer's response. We carried out five experiments (we denote them as 5-9) by classifying in each one 100 new answers. The results presented in the Table 4 shows the best results are achieved when the classifier assigns two classes like in the case of experiments 1-4. In all cases the average recall and precision are between 49-52%, the F-score is about 50%. The average exactly match (AvgEM) is better when classifier returns only one class and is 29,08% for SVM with gaussian kernel. There is a small difference between performance of SVM and MNB classifier in this case.

Table 4

Classification experiments 5-9

	Assignet number of class	**AvgEM**	**AvgPrec**	**AvgRec**	**AvgF**
MNB classifier with improvemnts	one	28,78	63,51	34,2	44,46
	two	19,98	51,72	48,28	49,94
SVM classifier (Polynomial kernel, eksponent = 1; C = 1;)	one	28,61	61,26	34,64	44,26
	two	18,61	51,03	49,6	50,3
SVM classifier (RBF kernel, gamma = 0.01; C = 21)	one	29,08	58,53	33	42,2
	two	19,76	49,51	47,88	48,68
One vs others MNB classifier (threshold - 0,9)	-	21,21	55,68	36,87	44,81

Source: own.

After analysing the results of experiments 5-9 the experts decided to join two subcategories (disclosure and anonymity) into one because these subcategories were difficult to differentiate. Moreover, they suggested to increase number of answers in one experiment to 150 in order to obtain a more repre-

sentative sample of data. Therefore, the next eight experiments (we denote them as 10-17) we carried out by classifying 150 new answers in each one. The other parameters were the same like in the previous experiments. The results are presented in the Table 5. There is no significant improvements but on the other hand SVM classifier with parameters selected manually achieved slightly better results. When we used SVM algorithm we achieved F-score about 1-1,5 percentage points better than MNB classifier and 6,82 percentage points better than model one vs others using MNB classifier.

Table 5

Classification experiments 10-17

	Assignet number of class	**AvgEM**	**AvgPrec**	**AvgRec**	**AvgF**
MNB classifier	one	27,14	72,46	35,57	47,72
	two	21,34	59,74	51,14	55,11
SVM classifier (Polynomial kernel, eksponent = 1; C = 1)	one	27,32	72,34	35,77	47,84
	two	22,57	60,47	52,43	56,13
SVM classifier (RBF kernel, gamma = 0.01; C = 21)	one	**28,92**	**75,06**	37,12	49,64
	two	23,59	60,99	**52,9**	**56,63**
One vs others MNB classifier (threshold - 0,9)	-	26,62	65,67	40	49,81

Source: own.

Optimization of SVM parameters. In the previous experiments we have used default parameters or manually selected for the SVM classifier. We believe that it is possible to find optimal parameters, which can improve classification quality, and it can be done by using differential evolution (DE) algorithm (see section *Classifiers. SVM classifier parameter optimization*).

In order to find the optimal parameters for SVM classifier we used again data from experiments 10-17. The half of experiments was carried out using the training set, and the half using the test set. The cost function can be presented as:

$$cost\ function = 100 - F_{avg10-13} \tag{12}$$

where $F_{avg10-13}$ is an average F-score from experiments 10-13.

Given the fact, there was a huge number of iterations of training set's evaluations we decided to set small population size equal to 20 and maximum iteration equal to 100. Other parameters of the DE algorithm were chosen intuitively: standard deviation (0.1), scale factor (0.9) and recombination probability (0.9). Vectors created from SVM parameters were the input data for DE algorithm. Before evaluation, it was necessary to set minimum and maximum values for all included parameters. Experiments involved comparing optimization on polynomial and RBF kernel (The best results shown on Table 6) to the primary performance.

Table 6

Optimization SVM parameters

Parameters settings SVM classifier	**Assignet number of class**	**Training for DE (experiment 10-13)**	**Testing for DE (experiment 14-17)**						
		AvgEM	AvgPrec	AvgRec	AvgF	AvgEM	AvgPrec	AvgRec	AvgF
Primary performance									
Polynomial kernel (eksponent =1, C = 1)	one	26,92	73,1	35,17	47,44	27,73	71,58	36,37	48,23
	two	20,98	59,44	50,6	54,61	24,15	61,49	54,25	57,64
RBF kernel (gamma = 0.01, C = 21)	one	28,39	74,38	35,74	48,23	29,46	75,74	38,49	51,04
	two	22,09	59,49	50,66	54,67	25,09	62,5	55,15	58,59
Results after optymalization									
Polynomial kernel (eksponent = 1.2149, C = 115.12282)	one	27,48	73,66	35,4	47,77	29,4	74,54	37,87	50,23
	two	20,24	59,63	50,74	54,77	24,1	61,8	54,52	57,93
RBF kernel (gamma = 0.001417, C = 70.22902)	one	29,13	77,7	37,37	50,41	30,21	76,7	38,98	51,69
	two	21,92	62,21	52,98	57,17	25,98	63	55,59	59,06

Source: own.

Dataset statistics. Increasing popularity of multi-label classification in academic literature causes the emergence of publicly available dataset[21]. In order to facilitate further analysis and evaluation of this dataset we present all multi-label specific measurements that were described in Section 2.1 (Table 7). Equally important in multi-label classification is knowing the label set frequencies (Figure 1).

Table 7

Dataset statistics

n	l	LC_{ARD}	LC_{ENS}
2314	15	1,771	0,118

Source: own.

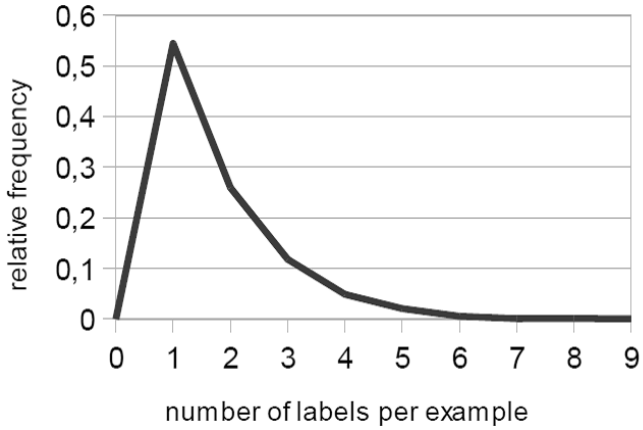

Fig. 1. The label distributions of dataset
Source: own.

[21] Our dataset can be available via email: tstanislawek@opi.org.pl.

3. Discussion

We have evaluated several machine learning methods to carry out an automatic classification of open-ended questions. There were presented the multi-label classifiers, which are responsible for labelling open-ended questions. In the classification experiments, we used the MNB and SVM methods and obtained the average precision of about 77% and the average recall of about 55%.

At first we have tested MNB a single multi-label classifier in comparison to procedure one vs others. We concluded that the individual MNB classifier gives better or similar results as the procedure one vs others and it is less complicated. Surprisingly, one vs others model has slightly higher recall than standard classifier with assigned only one class.

The experiments involving model improvements (Polish language dictionary and larger data set) achieved better results than basic Multinomial Naive Bayes classifier. In the other hand, model that using Polish Keyword Extractor algorithm is much worse in comparison to all others.

The reported factors shows clear improvement after we aggregate two most likely subcategories (experts decided to aggregate two subcategories: disclosure and anonymity into one because they were often mistaken). Compared to the previous experiments in Table 4 , F-score increased by 6%.

In order to find the best parameters in SVM classifier we used Differential Evaluation algorithm. After closer look at the Table 6, we noticed that there is no such a big difference between results achieved by SVM classifier with parameters selected manually and parameters selected by evaluation on DE algorithm (about 0,5% on AvgEM and AvgF). However, SVM classifier with default settings reaches much worse results than the two previously mentioned. This means that it is important to look for optimal parameters for SVM classifier and not necessarily use for that optimization methods like evolutionary algorithms.

Conclusion

The on-going studies on the automatic classification of open-ended texts are still in an early stage. But the desire to use the classification or analysis method of response texts of open-ended questionnaires is increasing. In this research, we conducted automatic classification of texts of an open-ended questionnaire. The results show that our best classification model (SVM classifier with parameters selected by DE algorithm) works well for multi--criteria classification and can produce questionnaire categories similar to those produced by humans.

While questionnaires are inexpensive, quick, and easy to analyze, often the questionnaire produce many problems (which influenced the achieved results by automatic classifiers). The people conducting the research may never know if the respondent understood the question that was asked. Specificity of questions causes that, the information gained can be minimal. Questionnaires conducted by mail or online produce very low return rates (only 32% of our respondents answered open-ended questions). The other problem associated with return rates is that often people that who return the questionnaire are those that have a really positive or a really negative viewpoint and want their opinion to be heard. People that are most likely to be unbiased typically don't respond because it is not worth their time.

Using machine learning algorithms speeded up process of questionnaires analysis. On the other hand the experts were still needed for models improvement and tuning. In future work, we plan to proceed with the analysis of characteristic expressions in texts of open-ended questionnaires based on these experimental results, and investigate other multi-label classification methods which can be applied to open-ended questions. The most critical problem is the estimation of number of classes (labels), which we will try to resolve by using prediction methods.

Acknowledgements

The authors would like to thank the consultant in the field of the classification theory Prof. Witold Pedrycz, and also Anna Plewa, Sławomir Dadas and Kinga Skolimowska for grammatical correction of the article.

References

Neff B.D., Olden J.D., *Is peer review a game of chance?*, BioScience 2006, 56 (4).

Liu B., *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*, Springer, New York 2010.

Silva C., Ribeiro B., *On text-based mining with active learning and background knowledge using svm*, "Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications" 2007, 11 (6).

Fragoudis D., Meretakis D., Likothanassis S., *Best terms: An efficient feature-selection algorithm for text categorization*, "Knowledge and Information Systems" 2005, 8 (1).

Lebart L., Salem A., Berry L., *Exploring Textual Data*, Kluwer Academic Publishers 1998.

Tsoumakas G., Katakis I., *Multi-label classifcation: An overview*, Int J Data Warehousing and Mining 2007.

Inui H., Murata M., Uchimoto K., Isahara H., *Classiffication of open-ended questionnaires based on surface information in sentence structure*, In Proceedings of the 6th NLPRS2001 2001.

Yamanishi K., Li H., *Mining open answers in questionnaire data*, "IEEE Intelligent Systems" 2002.

Sorower M.S., *A Literature Survey on Algorithms for Multi-label Learning*, Oregon State University 2010.

Macnab N., Thomas G., *Quality in research and the significance of community assessment and peer review: education's idiosyncrasy*, "International Journal of Research & Method in Education" 2007, 30 (3).

Procedures for review and selection of reviewers, Vol. 1, ed. J. Protasiewicz, Information Processing Institute, Warsaw 2012 [in Polish].

Procedures for review and selection of reviewers, Vol. 2, ed. J. Protasiewicz, Information Processing Institute, Warsaw 2012 [in Polish].

Storn R., Price K., *Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces*, "Journal of Global Optimization" 1997, 11.

Hirasawa S., Shih F., Yang W., *Student questionnaire analyses for class management by text mining both in japanese and in Chinese*, In Proc. 2007 IEEE International Conference on System, Man and Cybernetics, 2007.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer, New York 2009.

Noble W., *What is a support vector machine?*, "Nature Biotechnology" 2006, 24.

Chen Y., Weng C., *Mining fuzzy association rules from questionnaire data*, "Knowledge-Based Systems Journal" 2009.

Hoare Z., *Landscapes of naive bayes classifiers*, "Pattern Analysis and Application" 2008, 11 (1).

KLASYFIKACJA ANKIET RECENZENTÓW I APLIKANTÓW

Streszczenie

Artykuł opisuje metody wieloetykietowej klasyfikacji tekstów z pytania otwartego ankiety przy wykorzystaniu technik uczenia maszynowego. Ma to na celu zwiększenie szybkości oraz redukcję kosztów analizy otwartego pytania w ankiecie. Na początku zostały opisane różne modele klasyfikatorów wieloetykietowych, za pomocą których przyporządkowuje się kategorię do tekstów. W doświadczeniach wykorzystywane zostały klasyfikatory jednoetykietowe: Wielomianowy Naiwny Bayes (MNB) oraz Maszyna Wektorów Nośnych (SVM). Za ich pomocą uzyskaliśmy średnią precyzję na poziomie 77% oraz średnią dokładność na poziomie 55%. Eksperymenty uwzględniały wiele usprawnień (wielkość zbioru uczącego, korektę słownictwa, optymalizację parametrów klasyfikatora SVM przy użyciu metod ewolucyjnych...), dzięki którym zwiększyliśmy skuteczność klasyfikacji w porównaniu do pierwotnego modelu. Zaproponowana metoda została użyta do automatycznego przyporządkowania kategorii do tekstów z otwartego pytania w ankiecie.

Tłumaczenie Tomasz Stanisławek