ARTIFICIAL INTELLIGENCE
TO SUPPORT THE DEVELOPMENT OF HEALTH TECHNOLOGY ASSESSMENT REPORTS

ISPOR 19th Annual European Congress
Vienna, Austria, October, 2016

Sekiewicz B, Rodak W, Rutkowski J, Plisko R, Miazga P
HTA Consulting, Krakow, Poland

The analysis was conducted by **HTA Consulting**

Our aim was to train iStefan, a computer program with the elements of the artificial intelligence (AI), to identify scientific publications potentially useful in the Health Technology Assessment (HTA). The program in its learning process relied on the abstracts from previous systematic reviews along with the information on publications included based on 'human' decisions. This abstract reports on methods and results of iStefan learning process.

Background and objective

The systematic review of medical publications and data constitutes an important and influential step in the development of HTA reports. In the initial phase, following the execution of search strategy, two analysts need to assess independently the relevancy of publications based on found abstracts. This task is most often very time-consuming due to the load of abstracts to sift through. It also shapes the further steps of the HTA process, thus its accuracy and speed is of great interest to be improved.

From perspective of machine learning (the subfield of AI), the problem of abstracts analysis boils down to text analysis and assignment of following tags: the text contains the pertinent information or the text is irrelevant. This issue is very similar to the common problem known to e-mail users, namely the identification of spam messages. There is many anti-spam filters developed based on AI method, which segregate incoming e-mails as either spam or not spam. This type of activity is called the classification.

Our goal was to examine the possibilities of applying AI methods to optimize the process of abstracts analysis.

Methods

Input data

The already evaluated abstracts from completed analyses served to train and test a classifier. For that purpose, the seven sets of quality of life (QoL) analysis and five sets of clinical analysis abstracts were collected and utilized.

Text representation

The representation of abstract texts was achieved with the application of Bag-of-Words (Bag-of-Features) model. Firstly, all documents were cleaned from non - alphanumeric characters, letters were formatted to lower case, stop words deleted and every term/phrase converted to the root word. Subsequently, each document was transformed to set of key–value elements coding for a word and its frequency of occurrence in a text. Grammar and words order were neglected. To illustrate this approach, the selected line from the Beatles song will be accordingly processed:

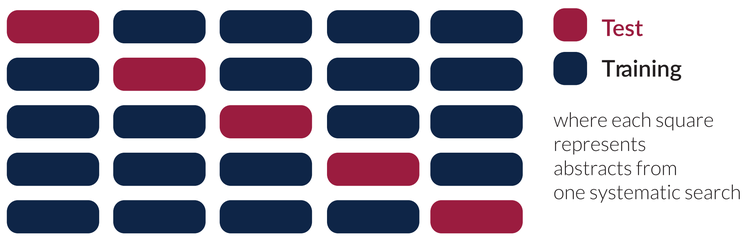
The original text:

"Let it be, let it be. Let it be, let it be"

'Be' and 'it' are considered to be stop-words, 'let' is a stemmed version, thus the Bag-of-Words representation looks as follows:

let : 4

Figure 1. Leave-one-out cross-validation - 5 clinical analyses



Model of classification

The Naïve Bayes Classifier (NBC), a Machine Learning generative model, is one of the basic tools in Natural Language Processing (NLP). Using Bayes Law, the classifier assigns a class that can maximize the right site of the following equation for a given document:

$$P(C|D) = P(C) * P(D|C)/ P(D)$$

Where:

P(C|D) is the probability of a class given document
P(C) is the probability of a class
P(D|C) is the probability of a document given class
P(D) is the probability of a document

Since the P(D) is a constant for a given document, thus:

$$P(C|D) \approx P(C) * P(D|C)$$

And assuming the independence of words in a given document:

$$P(D|C) = P(W_1|C) * P(W_2|C) *...* P(W_N| C)$$

Where:

$P(W_i|C)$ is the probability of i-th word in the document given a class

Imbalanced Class Problem

The strong disproportion was recognized in number between rejected and accepted abstracts in the process of HTA analyses (namely approximately 7:1 in dataset from quality of life analyses and 12:1 in dataset from clinical analyses) and this may have a negative influence on the classifier's performance. Thus, it was decided to train few classifiers with the use of equinumerous classes. For that purpose, all accepted abstracts and consecutive parts of the rejected were utilized.

Learning model

Our method of classifier learning estimates for every feature in a training dataset, the probability of its occurrence in each class. The outcomes can be affected by following parameters:

» **Sparsity** –the lowest number of documents which must have a given word in order for a word to be considered as a feature; [0, 5, 10, 15, 20, 25]

» **K** – which defines how many of the most informative words (counted with the use of the chi – squared statistics) will be incorporated in the model; [25, 50, 75, 100, 200, all]

» **Alpha** – is a smoothing parameter, which assigns an implicit value to unknown words (Laplace smoothing); [0.01, 0.1, 1, 10, 100]

» **T** – is a threshold value corresponding to the lowest number of positive classifications assigned to the abstract to obtain the final positive classification recommendations; a threshold value depends on class imbalance: [QoL analyses: 3, 4, 5, 6], [clinical analyses: 3, 4, 5, 6, 7, 8, 9, 10]

For each type of analysis (QoL or clinical), the leave-one-out cross-validation was performed. This type of cross-validation is often utilized in a case of small datasets. For each set of parameters (the combinations of above--mentioned parameters), the models were trained on n-1 systematic re--views (where n is the number of available systematic reviews) and tested on abstracts from the remaining systematic review. The implementation of this method can generate suspiciously optimistic results, therefore much appropriate solution would be to use the nested-cross-validation. Due to the data scarcity, this approach was infeasible. Due to the randomness in abstract selection during the single model development, the entire process was repeated fifty times. The model which achieved the highest scores was used for further testing.

Results

Five most optimal models were selected for each type of HTA analysis (QoL and clinical). Two scenarios were investigated:

» „Artificial analyst" model, which outcomes correspond to the results of an analyst's workings: the maximization of precision; this model to be selected based on F-measure;

» „Filter tool" model, which identifies and removes from the further analysis the abstracts failing the inclusion criteria; the maximization of recall.

Table 1. Artificial analyst - the best results of five models (Qol analyses)

No.	Sparsity	K	Alpha	T	Accuracy	Precission	Recall	F-measure
1	15	75	1	6	67%	45%	75%	0,562
2	10	100	1	6	68%	45%	74%	0,558
3	20	100	1	6	67%	43%	75%	0,550
4	15	100	1	6	67%	44%	72%	0,550
5	15	75	1	6	65%	44%	72%	0,549

Table 2. Artificial analyst - the best results of five models (Clinical analyses)

No.	Sparsity	K	Alpha	T	Accuracy	Precission	Recall	F-measure
1	0	all	1	5	74%	29%	59%	0,385
2	0	all	1	6	75%	29%	56%	0,379
3	0	all	1	4	72%	28%	60%	0,379
4	20	all	1	2	73%	28%	58%	0,378
5	5	all	1	3	73%	28%	57%	0,377

Table 3. Filter tool - the best results of five models (Qol analyses)

No.	Sparsity	K	Alpha	T	Accuracy	Precission	Recall	F-measure
1	0	all	100	2	26%	16%	100%	0,272
2	0	all	100	3	28%	16%	100%	0,277
3	0	all	100	2	25%	16%	100%	0,270
4	0	all	100	2	26%	16%	100%	0,272
5	0	all	10	2	31%	17%	100%	0,288

Table 4. Filter tool - the best results of five models (Clinical analyses)

No.	Sparsity	K	Alpha	T	Accuracy	Precission	Recall	F-measure
1	0	all	100	2	51%	18%	84%	0,302
2	0	all	100	2	51%	19%	84%	0,306
3	0	all	100	3	53%	20%	83%	0,321
4	0	all	100	3	53%	19%	83%	0,308
5	5	all	100	2	54%	19%	82%	0,313

Conclusions

The developed model does not substitute the analysts in the process of abstracts selection, but it may significantly expedite their work. Its current low precision may lead to the inclusion of excessive number of publications in the phase of the full-text analysis (false positive observations). However, the model's very high recall despite its low precision allows to narrow down the abstract number by even several dozens percent.

The model returns promising results but it requires further refinement. In our process we have implemented the most primary method of machine learning. In future, we plan to use the modified versions of NBC method and introduce more advanced tools such as the support vector machine. Additionally, it is necessary to extend the training input dataset by new search results as well as supplementary information including the search strategy or medical indications.

Abbreviations

AI	*Artificial Intelligence*	**NLP**	*Natural Language Processing*
HTA	*Health Technology Assessment*	**QoL**	*Quality of life*
NBC	*The Naïve Bayes Classifier*		

HTA consulting