APPLICATION OF LOGISTIC REGRESSION MODELS TO AUTOMATED DOCUMENT CLASSIFICATION

IVETA STANKOVIČOVÁ
Comenius University in Bratislava, Faculty of Management
Department of Information Systems
Odbojárov 10, Bratislava, Slovakia
e-mail: iveta.stankovicova@fm.uniba.sk

ANDREJ MIHÁLIK
GPK Event- und Kommunikationsmanagement GmbH
Gußhausstraße 14/2, Vienna, Austria
e-mail: andrej.mihalik@gpk.at

Abstract

The paper presents an application of logistic regression to the problem of automated text classification and detection of presence of relevant data in a text originating from web sources. The presented example is based on a corpus of manually labelled data collected for the largest recruiting study in a German-speaking region and known as "BEST RECRUITERS". The paper focuses on key areas of extraction, normalization and tokenization of text from web sources; conversion of text into a matrix of independent variables and is then concerned with development of a logistic regression model. The paper concludes with a suggestion of deployment of the model in a business context.

Key words: *document classification, natural language processing, logistic regression, automatization.*

1. Introduction

In the following sections an approach to automatized document classification will be presented. The methods demonstrated in this paper rest on the foundation of text mining and machine learning. The former enables us to process unstructured data, the latter helps us to make sense of it once transformed into a more structured format and to make inferences about new data inputs based on insights gained from an existing dataset.

Apart from the technicalities of data mining and statistic modelling a bearing theme of this paper is automatization of information processing which we see as a major target of these research areas. The approach presented in this paper is an attempt to hint at the larger general movement towards automation of a growing array of professions seen in most branches of industry today. Authors Frey and Osborne (2013, p. 44) examined the automation potential of 702 occupations in the United States and estimated that a total of 47% of jobs are at a high risk, 19% at a medium risk and 33% at a low risk of being automated.

Text mining is a scientific discipline with intersections with statistics, data mining, machine learning, managerial sciences and artificial intelligence and which presents technologies for the analysis and processing of unstructured data. (Miner et al., 2012, pp. 30-32) Similarly to data mining in general, text mining is a collection of techniques for processing of textual data. Every one of them is at a different stage of maturity, ranging from experimental approaches to well-established best practices. Every one of them focuses on textual data at a different level (word, sentence, document) and of various structures (plain text or documents linked with each other as is often the case for web documents). The goal may be a general explorative analysis or the creation of a predictive model to solve a specific problem.

Text mining is a rapidly growing research field and the scientific literature on this topic is rich with examples of successful applications in the industry. These include: Maresse-Taylor et al. (2013) who proposed a way of identifying customer preference of tourism products based on online reviews, Phawattanakul and Luenam (2013) developed a suggestion mining system to detect customer desires from television program reviews and brand association maps by Akiva et al. (2008) represent a way to automatically create a visual concise representation of ideas the customers connect to a particular brand using online web content only.

This paper looks at text at the document level and the aim is the creation of a classification model. In the field of text mining, this task is often described as document classification. Document classification according to Miner et al. (2012, pp. 30-32) "aims to assign a category from a set of known categories to the document based on its features by utilizing a training set of documents which already are assigned a category."

2. Data

Data used for the demonstration of this approach was collected as a part of the BEST RECRUITERS study and was used here with the kind permission of the director of CAREER Verlag Mag. Julia Hauska. „Best Recruiters surveys annually the recruiting quality of the top employers in Switzerland and Liechtenstein, Germany and Austria based on more than 100 scientific criteria grouped into 4 pillars. Using this extensive dataset and resulting optimization strategies for employers, the study provides an important contribution toward the continual improvement of recruiting quality." (Best Recruiters, 2016)

The study evaluates the online recruiting presence of the company, quality of job postings offered, the recruiting process itself including the handling of job applicants and their feedback. The second pillar which entails analysis of job postings has been judged as the most suitable for the presentation of the document classification methodology. In this category, companies are awarded points if their job postings contain information about the character of the job, a responsible contact person, the starting date, a list of tasks assigned to the job as well as skills required etc. For this paper we have chosen a relatively uncontroversial example of the inclusion of the starting date information in the text of a job posting. In our classification model, this will be the target variable. Its value has been recorded in the study database.

For the independent variables we had to use the opportunistic data of job posting screenshots used as a proof for point allocation. The character of this data brought with it two major obstacles. Firstly, some of the files were stored in the PDF format enabling us text extraction with relative ease, however, for some, screenshots in image formats were created, where a trivial extraction of text is not possible. Secondly, due to the presence of multinational companies and the multiplicity of national languages in the Switzerland there was a need to filter postings written in German. Every language requires a separate classification model, German has been chosen here as it is the most prevalent language with the most observations.

The text extraction has been performed in the JAVA programming language. For PDF files Apache PDFBox® library has been applied. Optical character recognition from image files was performed using Tess4J Java bindings of the popular Tesseract Open Source OCR Engine. The next step was the isolation of German texts from texts written in English, French or Italian. For this task a very simple scoring model has been devised. For each of these four languages a file of stop words has been downloaded from the Google's stop-words project at https://code.google.com/archive/p/stop-words/. Stop words are words which have the highest frequency in a given language. Then, for every text file, an algorithm has tested the presence of every word in any of the four lists of stop words. If a word belonged to a particular list, that language has been awarded a point. At the end, the language which scored most points for a given file has been used as the estimated language. This approach has worked very well in this case. The language of all texts identified as German has indeed been German.

The text files resulting from the transformation of screenshots provided us with independent variables for modelling. To match them with database entries, screenshot file names have been matched with customer IDs in the survey database. The final dataset contains 1597 records originating in surveys performed in the period from 2013 to 2016 for chosen survey groups.

3. Features

Once we have matched the texts with corresponding classes identifying whether a criterion is fulfilled or not, we have to perform the text mining step of transforming unstructured textual data into a structured format, a set of independent variables.

Before we can decide which words or word groups to use for predictive modelling, the textual data has to be cleaned and normalized. During the process of image conversion to text, noise was introduced to our data when the OCR algorithm tried to translate non-textual elements into characters. Therefore, at the beginning, all non-alphabetic characters are removed from the text and it is converted to lowercase. This is important to assign all variations of letter capitalization to the same feature. Capitalization can potentially have an effect on the meaning of the word, however, this depends on the application. For highly inflected languages such as Slavic languages, bringing words to their root forms may be considered.

In the next step a decision has to be made as to what is the maximal number of words contained within one feature. Including more words enables us to detect phrases – groups of words which when combined acquire a specific meaning. However, this comes at the cost of being more computationally intensive. For our purposes we limited the maximum number of words in a feature to three. An algorithm has then examined each of the texts, creating a set of all possible features containing individual words, bigrams and trigrams. During this process, the total count of texts in which a word has occurred has been recorded as well as the proportion of texts containing information on the starting date from this total. The process described above has been performed on the training dataset only to not interfere with estimates of model performance.

To measure the ability to distinguish between the target classes, the proportion of texts containing information about the starting date and including a given feature has been used. A commonly used measure of the degree of class heterogeneity is the entropy in the notation of Hastie et al. (2009, p. 309) defined as:

$$\sum_{k=1}^{K} \hat{p}_{m,k} \log \hat{p}_{m,k} \,. \tag{1}$$

For our purposes we will use a slightly adapted version taking into account the prior class distribution. We adjusted the estimated probabilities according to the formula:

$$p_{adjusted} = p^{\log_{p_+} 0.5}, \tag{2}$$

where p_+ denotes the proportion of cases belonging to a given class.

Analysis of our training dataset consisting of 1121 records has detected 525 355 potential features. This extensive set has to be filtered to find most suitable candidates. Suitability in this case is constituted by two major factors: First, the feature has to occur in a great enough number of texts and it has to be able to distinguish target variable classes. Therefore, we have arbitrarily created a filter according to which only features occurring at least in 2% of texts were chosen having entropy of at most 0.8. This has narrowed down the set of candidate features to just 53. Ten of them most indicative of texts containing information on starting date can be seen in Table 1.

Table 1: Top ten features indicating the presence of starting date information in terms of entropy

Feature	Translation	Count	Occurrence in positive postings	Entropy
oder nach vereinbarung	or as agreed upon	28	96.43%	0.1760
sofort oder	now or	23	95.65%	0.2052
oder nach	or as	31	93.55%	0.2773
nach vereinbarung	agreed upon	51	92.16%	0.3208
vereinbarung	agreed	55	89.09%	0.4074
sofort eine n	now a	23	86.96%	0.4617
eintrittsdatum	starting date	27	85.19%	0.5038
suchen wir zum	we are looking for to	65	84.62%	0.5168
möglichen	possible	32	84.38%	0.5222
sofort	now	169	84.02%	0.5300

Source: the authors.

The texts are then represented as an indicator matrix, whose cells constitute the occurrence or non-occurrence of a given feature in a given text. An alternative would be a matrix with cells containing the counts of occurrences of features in texts or a transformation of these (such as the logarithm).

4. Model

We have arrived at a sparse indicator matrix representing independent variables and a class variable to be predicted. Although, the aim of this paper is to present modelling based on logistic regression, there is a variety of classification algorithms that have been applied to document classification in the scientific literature. Decision trees can be easily applied in the context of an indicator matrix. The problem with this approach is that in a case of many variables in the process of growing the decision tree we very soon arrive at nodes with insufficient number of cases if we do not have a sufficient number of data in our training dataset. This approach is also prone to overfitting. A remedy to this problem has been proposed in the form of random forests. A very popular approach based on Rosenblatt's perceptron algorithm are the separating hyperplanes and their extended version based on support vector machines, which tend to be intensive on computational resources. However, they often achieve very accurate results. On the other end of the spectrum is a relatively light-weight yet often successful Naïve Bayes algorithm. The basis of this method lies in the Bayes theorem and the assumption that all features used in the model are independent. Looking at Table 1, we see that some of the features co-occur together which would undermine the power of this algorithm. A crucial consideration for Naïve Bayes algorithm is ensuring that there are no zero conditional probabilities as this would result in the prediction of zero regardless of any other dependent variables. This could happen if a certain feature does not occur in the training dataset. This is often solved by introducing smoothing to the calculation of probabilities. In our previous research, however, logistic regression has proven superior in terms of accuracy, interpretability and the ease of use and therefore it is the method we have chosen for this demonstration (Mihálik, 2015, p. 63-73).

"Logistic regression tries to model posterior probabilities of K classes via linear functions in x and ensuring that they sum to one and remain in the range between zero and one." (Hastie et al., 2009, p. 119). In its simplest and perhaps most popular form of binary logistic regression, the model consists of a single linear function to model a zero or one outcome. The model is specified in terms of a logit transformation of the odds ratio of probability of one class to the probability of the other class:

$$\log \frac{P(G=1 \mid X=x)}{P(G=0 \mid X=x)} = \beta_0 + \beta_1^T x \,. \tag{3}$$

The regression is fit by maximizing the likelihood function L, in this case computing the log-likelihood is more convenient:

$$\log L = l(\theta) = \sum_{k=1}^{K} \log(p_{g_i}(x_i, \theta)) \,. \tag{4}$$

As we can see probability function is nonlinear in β, thus we will have to find the vector of coefficients that maximizes the likelihood iteratively using the Newton-Raphson algorithm until sufficient accuracy is achieved:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta \partial\beta^T} \right) - \frac{\partial l(\beta)}{\partial\beta} \,. \tag{5}$$

Generalization of this approach for more than two classes is referred to as multinomial logistic regression or maximum entropy classifier.

As is the case with all regression models, multicollinearity in explanatory variables decreases the predictive power of a model. To address this concern, we have expressed the 53 independent variables in terms of 27 principal components which together capture more than 90% of variability in our data. For this the prcomp function of stats package in R has been used, the data has been centered.

The model has been created using R's glm function with the maximum number of iterations set to 4000. The resulting model probability is the probability of starting date information not occurring in a given job posting. Following coefficient estimates were obtained (Table 2):

Table 2: Coefficient estimates

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.47822	0.09505	5.031	4.87e-07
PC1	0.92266	0.14141	6.525	6.81e-11
PC2	2.72766	0.19960	13.666	< 2e-16
PC3	3.31337	0.66608	4.974	6.54e-07
PC4	0.75242	0.29558	2.546	0.010910
PC5	1.63320	0.30088	5.428	5.69e-08
PC6	-1.19962	0.34632	-3.464	0.000532
PC7	-1.79774	0.44188	-4.068	4.73e-05
PC8	-0.58788	0.49686	-1.183	0.236734
PC9	-0.10996	0.43736	-0.251	0.801484
PC10	0.48972	0.60231	0.813	0.416178
PC11	-0.73796	0.43594	-1.693	0.090494
PC12	0.81090	0.45023	1.801	0.071692
PC13	-0.80447	0.61545	-1.307	0.191171
PC14	-3.86994	0.58982	-6.561	5.34e-11
PC15	-2.83096	0.56132	-5.043	4.57e-07
PC16	-0.80800	0.63373	-1.275	0.202309
PC17	1.50669	0.57832	2.605	0.009180
PC18	1.00270	0.72306	1.387	0.165520
PC19	0.39533	0.67512	0.586	0.558160
PC20	0.65671	0.63243	1.038	0.299088
PC21	2.19710	0.65289	3.365	0.000765
PC22	-0.89628	0.70158	-1.278	0.201422
PC23	-1.45235	0.65308	-2.224	0.026158
PC24	-1.26413	0.74512	-1.697	0.089783
PC25	-0.89235	0.67856	-1.315	0.188487
PC26	-0.91596	0.70265	-1.304	0.192378
PC27	0.43387	0.62863	0.690	0.490079

Source: the authors.

The interpretability of coefficients in this case is complicated by the fact that they are related to principal components rather than original textual features.

All cases with model probabilities of over 50% were assigned the class 0 (starting date information not present), the rest was assigned to class 1. Predicting the target variable for the test dataset using this model and comparing the results with test dataset actual target variable values we could construct the following confusion matrix (Table 3):

Table 3: Model confusion matrix based on test dataset

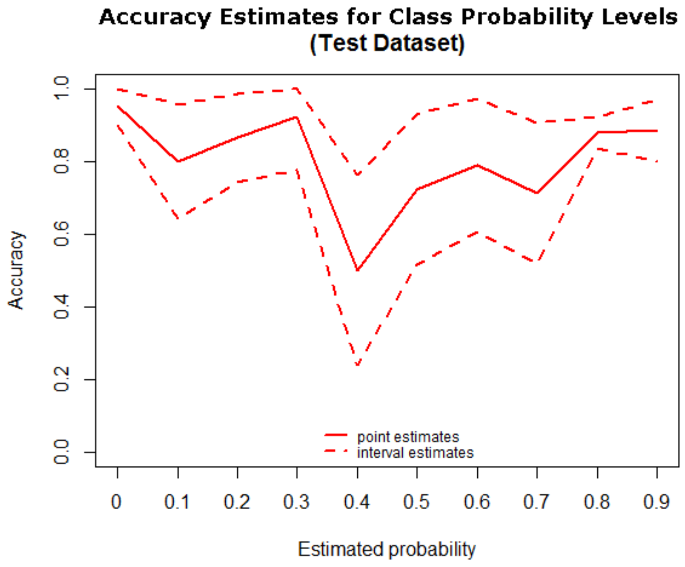
	Predicted negative	Predicted positive	Actual totals
Actual negative	126	20	146
Actual positive	47	283	330
Predicted totals	173	303	476

Source: the authors.

Analysis of this table yields estimates of model accuracy at 0.8412, sensitivity at 0.73 and specificity at 0.9123. Let us now present graphically model accuracies at different class probability levels (Figure 1).

Figure 1 seems to indicate that the model provides more accurate estimates at both ends of probability estimates.

Figure 2: Model accuracy estimates at various class probability levels



Source: the authors.

5. Summary

The paper has presented an approach of obtaining a document classification model from a set of manually labelled textual data. The data used in this demonstration consisted of 1597 job postings, the target variable being the inclusion of information on the starting date for a potential job applicant. Explanatory variables for target class prediction have been obtained from the total set of 525 355 of all individual words, bigrams and trigrams occurring in the training dataset by using only those which occur at least in 2% of texts and have an entropy of at most 0.8. This has narrowed down the set of independent variables to only 53 features which have then been expressed in terms of 27 principal components explaining 90% of data variability to counter multicollinearity. Logistic regression has been applied to class prediction, yielding a regression model with an accuracy of 0.8412, sensitivity of 0.73 and specificity of 0.9123 for the test dataset consisting of 476 randomly chosen texts. A graphical representation of accuracy estimates at different levels of estimated class probability has indicated that probabilities closer to both extremes of 0 and 1 seemed to correspond to better quality predictions than probability values in the middle.

For the application of this model in the practice of automatized document classification the character of the textual data and the target variable have to be considered. The greater the variety of expressions used in association with the target classes, the more difficult it will be to come to a satisfying set of model features and the more data will be necessary. Also, each application is associated with a different level of a tolerable error rate. The deployment of a prediction model similar to the one presented in this paper might not initially lead to the full automatization of the task of document classification. One approach would be to automatically classify only texts with estimated class probabilities near to 0 or 1. For the rest, the class would be still assigned manually. These manual assignments would then be used to further refine the classification model until the desired level of accuracy has been achieved.

References

[1] AKIVA, N. et al. 2008. Mining and visualizing online web content using BAM : Brand Association Map™. [cit. 10-09-2016]. http://aaaipress.org/Papers/ICWSM/2008/ICWSM08-028.pdf.

[2] BEST RECRUITERS 2016. Die repräsentative Recruiting-Studie. [cit. 10-09-2016] http://www.bestrecruiters.de/die-studie/die-studie/.

[3] FREY, B., OSBORNE, M. 2013. The future of employment : How susceptible are jobs to computerisation? [cit. 10-09-2016] http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

[4] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. 2009. The elements of statistical learning : Data mining, inference, and prediction. 2nd edition. New York : Springer, 2009. ISBN 978-0-387-84857-0.

[5] MARRESE-TAYLOR, E. et al. 2013. Identifying customer preferences about tourism products using an aspect-based opinion mining approach. In Procedia Computer Science, 2013, vol. 22, p. 182 – 191.

[6] MIHÁLIK, A. 2015. Text mining and sentiment analysis in marketing research. Dissertation thesis. Bratislava : Comenius University in Bratislava, 2015.

[7] MINER, G. et al. 2012. Practical text mining and statistical analysis for non-structured text data applications. Watham, MA : Academic Press, 2012. ISBN 978-0123869791.

[8] PHAWATTANAKUL, K., LUENAM, P. 2013. Suggestion mining and knowledge construction from Thai television program reviews. In Proceedings of the International MultiConference of Engineers and Computer Scientists. 2013, vol. 1. ISSN 2078-0958.