

A probabilistic neighbourhood translation approach for non-standard text categorisation

Ata Kabán

School of Computer Science, The University of Birmingham,
Birmingham, B15 2TT, UK
A.Kaban@cs.bham.ac.uk

Abstract. The need for non-standard text categorisation, i.e. based on some subtle criterion other than topics, may arise in various circumstances. In this study, we consider written responses to a standardised psychometric test for determining the personality trait of human subjects. A number of state-of-the-art text classifiers that having been very successful in standard topic-based classification problems turn out to perform poorly in this task. Here we propose a very simple probabilistic approach, which is able to achieve accurate predictions, and demonstrates this peculiar problem is still solvable by simple statistical text representation means. We then extend this approach to include a latent variable, in order to obtain additional explanatory information beyond a black-box prediction.

1 Introduction

Automatic text classification has been a highly researched topic over the past decade and many successful methods have been devised. However, in benchmark test beds, the notion of class is most often associated with that of a topic. Vector-space representations [10] of text collections tend to be well separable w.r.t. topics, and so methods like linear SVM and Naive Bayes [14, 1] typically obtain high accuracy.

However, there are cases when text categorisation needs to be based on some non-standard and more subtle criteria, other than topics. For example, in this work, we deal with responses of human subjects to a standardised psychometric test, for determining their personality traits. All subjects take the same test, and the data consists of their written English prose responses to the same set of questions. The topics are therefore common to all documents. Instead, in this case, the classification criterion of interest is the personality trait of each subject. The question is this: Is it possible to automatically predict the personality trait of human subjects based on their written responses and a set of hand-labelled examples?

Several methods that have been highly successful on topic based text classification turn out to perform poorly on this task. One may even worry that perhaps a bag-of-words representation is inappropriate and perhaps a more sophisticated representation, involving syntactic and semantic characteristics, and / or natural

language processing approaches might be required. Naturally, searching for another feature representation in an infinite space of possibilities is not a practical prospect. Therefore, in this work we set out to investigate whether a statistical approach would still bear any fruit.

We devise a fairly simple probabilistic approach that works on standard bag-of-words features, and has a natural and clear probabilistic formulation. On some inspection, our model has some resemblance with both tf-idf and nearest-neighbour classification, and for the latter, we have chosen to call it nearest neighbour translation. Despite its simplicity, our method is able to deal with peculiar non-linear separation boundaries and, based on the data set tested, it appears to be highly suited to the non-standard text classification task under study. We then extend this approach to include a latent variable in order to obtain some explanatory information in addition to black-box prediction.

In the remainder of the paper, Section 2 introduces our method. Section 3 presents its latent variable modelling extension, including an iterative procedure for parameter estimation. Section 4 presents comparative prediction results, as well as interpretability results. Finally, Section 5 concludes the paper.

2 Probabilistic neighbourhood translation

Consider a training set of N text documents, together with their associated labels. Each document will be represented as a discrete distribution over all the terms of the dictionary. For document n , this is denoted by $P(.|n)$. Similarly, each term is a distribution over all the documents of the corpus. For term t , this is denoted by $P(.|t)$.

Using the above representation, we define the probabilistic neighbourhood of document n by marginalising over the terms, as the following:

$$P(n'|n) = \sum_{t \in \text{Dictionary}} P(n'|t)P(t|n) \quad (1)$$

Document n' is a neighbour of document n with probability $P(n'|n)$. In general, all documents are neighbours of all other documents with some (possibly zero) probability.

Further, let us denote the label of document n by $P(z|n)$. This is a vector of length C , where C is the number of classes. For the training set instances, the labels are known, so $P(z_{true}|n)$ is just a 1-of- C label encoding — i.e. $P(z = c|n) = 1$ if document n belongs to class c and zero otherwise.

Now, we can write the label probability distribution of a previously unseen document, as the following:

$$P(z|n_{new}) = \sum_{n \in \text{TrainingSet}} P(z|n)P(n|n_{new}) \quad (2)$$

$$= \sum_{n=1}^N P(z|n) \sum_{t=1}^T P(n|t)P(t|n_{new}) \quad (3)$$

$$= \sum_{t=1}^T P(z|t)P(t|n_{new}) \quad (4)$$

where $P(t|n_{new})$ is the term distribution representation of the new document, and all other quantities are pre-computed from the training set.

The training procedure then only requires us to prepare the probability matrix $P(n|t)$ by appropriately normalising the documents-by-terms matrix, and $P(z|n)$ are the given label assignments. Moreover, the marginalisation over training documents can also be pre-computed, as in (4), which essentially results in each term being assigned a label probability. Observe in addition, the obtained label probabilities $P(z|n_{new})$ are guaranteed to be properly normalised to sum to one w.r.t. z , without any further effort. We may threshold this label probability to obtain a hard partitioning of the data into classes, but at the same time the actual value of the predicted label probability tells us about the confidence of the prediction. A probability close to 0 or 1 indicates a highly confident prediction, whereas, in two-class problems, a value close to 0.5 predicts a low confidence of the class prediction. Thus we have set up a non-parametric and fully probabilistic predictor, capable of making class predictions as well as uncertainty estimates.

To make some connection with related methods, let us inspect the formulations (1)-(4) and note a resemblance with the popular 'term frequency inverse document frequency' (tf-idf) method [10]. In (3), we have $P(t|n)$ exactly the term-frequencies (tf), and an analogy between the term $P(n|t)$ and inverse document frequencies may also be seen, since $P(n|t) = \#(t, n) / \sum_n \#(t, n)$ is inversely proportional to the frequency of term t throughout the entire corpus. Contrarily to tf-idf however, our formulation has a very clear probabilistic foundation.

Secondly, from the form (2), we can see an analogy to a nearest neighbour model, with a neighbourhood kernel defined by (1). Because of this analogy, and since the neighbourhood kernel is a stochastic translation matrix, and the entire model is formulated in probability terms, we call this approach 'probabilistic nearest neighbour translation'. The probabilistic nature of our formulation makes it fairly straightforward to incorporate extensions as appropriate and the next section demonstrates this with the purpose of allowing us to gain, besides prediction, some additional explanatory information from the data.

3 An extension for uncovering term-associations

Though the simple approach presented in the previous section can be readily used to perform classification, i.e. to predict class labels for previously unseen data, one would often prefer to be able to map text documents from the space of words into a more conceptual space. In the present case, this conceptual space would be necessarily other than a topical space, for it being defined by some non-standard labelling that we try to accommodate. Technically, this would be useful e.g. in the cases when documents happen to have no overlapping words, despite they share content w.r.t. to the given labelling — and in addition it may provide additional insights in terms of interpretability. To achieve this, we extend our model by introducing a latent 'bottleneck' variable having K different discrete values, $k = 1, \dots, K$, and aim for a version of eq. (2) where the space of terms is replaced with the space of our new latent variable:

$$P(z|n_{new}) = \sum_{n=1}^N P(z|n) \sum_{k=1}^K P(n|k)P(k|n_{new}) \quad (5)$$

To make the connection to the actual words space, we write:

$$P(n|k) = \sum_{t=1}^T P(n|t)P(t|k); \quad \text{and} \quad P(k|n_{new}) = \sum_{t'=1}^T P(k|t')P(t'|n_{new}) \quad (6)$$

We can think of this extension as having added a probabilistic translation of both terms and documents into the latent space. The use of a 'bottleneck' latent variable is quite common in generative latent variable modelling of text, especially in unsupervised learning [6, 11]. Here in turn, we employ this technique as part of our conditional model, in the supervised learning context.

Now, replacing into (5) we have:

$$P(z|n_{new}) = \sum_{n=1}^N P(z|n) \sum_{t=1}^T P(n|t) \sum_{k=1}^K P(t|k) \sum_{t'=1}^T P(k|t')P(t'|n_{new}) \quad (7)$$

and we may additionally also interpret the part $\sum_k P(t|k)P(k|t')$ as a compressed (aggregated) term association probability matrix $P(t|t')$ — somewhat in the spirit of [11] — which we are going to estimate.

3.1 Parameter estimation by maximising the leave-one-out error

For the ease of the subsequent manipulations, let us introduce the following notations for elements precomputed from the training set: $\mathbf{U}_{-n} \in \mathbb{R}^{C \times T}$ with elements $P(z|t) = \sum_{n' \neq n} P(z|n')P(n'|t)$, $\mathbf{v}_n \in \mathbb{R}^{T \times 1}$ with elements $P(t|n)$, and $\mathbf{y}_n \in \mathbb{R}^{K \times 1}$ with elements $P(z_{true}|n), \forall n \in \text{TrainingSet}$. The unknown parameters will be denoted by $\mathbf{A}_1 \in \mathbb{R}^{T \times K}$ with elements $P(t|k)$ and $\mathbf{A}_2 \in \mathbb{R}^{K \times T}$ with elements $P(k|t')$ respectively. Using these notations, the r.h.s. of the model (7) may now be written compactly as $\mathbf{U}_{-n_{new}} \mathbf{A}_1 \mathbf{A}_2 \mathbf{v}_{n_{new}}$.

Now, to estimate the parameters \mathbf{A}_1 and \mathbf{A}_2 , we maximise the *leave-one-out error*, using the training set. The idea of maximising the leave-one-out error was previously proposed in [5] for a predictive k-nearest-neighbour model termed the 'neighbourhood component analysis'. Our approach in this section may be seen to have some analogies with that model, though an obvious difference is that our parameters are all probabilities, which leads to different (and simpler) estimation equations and allows straightforward interpretation in the context of our text analysis application.

Thus, we formulate the objective to minimise the sum of Kullback-Leibler divergences between the given labels \mathbf{y} and their predictions¹,

$$\{\mathbf{A}_1, \mathbf{A}_2\} = \underset{\mathbf{A}_1, \mathbf{A}_2}{\operatorname{argmin}} \sum_{n=1}^N KL(\mathbf{y}_n || \mathbf{U}_{-n} \mathbf{A}_1 \mathbf{A}_2 \mathbf{v}_n) \quad (8)$$

which, up to an additive constant (i.e. the sum of entropies of \mathbf{y}_n), is equivalent to maximising the following objective:

$$\operatorname{Obj}(\mathbf{A}_1, \mathbf{A}_2) = \sum_n \mathbf{y}_n \log \{ \mathbf{U}_{-n} \mathbf{A}_1 \mathbf{A}_2 \mathbf{v}_n \} \quad (9)$$

We add Lagrange multipliers to ensure that all columns of both \mathbf{A}_1 and \mathbf{A}_2 are proper probabilities i.e. sum up to one.

After straightforward algebra, the stationary equations will have the form of fixed point non-linear equations that can be solved iteratively. These iterative updates are the following:

$$\mathbf{A}_1^{new} \propto \mathbf{A}_1^{old} \odot \sum_n \mathbf{U}_{-n}^T \frac{\mathbf{y}_n}{\mathbf{U}_{-n} \mathbf{A}_1^{old} \mathbf{A}_2 \mathbf{v}_n} (\mathbf{A}_2 \mathbf{v}_n)^T \quad (10)$$

$$\mathbf{A}_2^{new} \propto \mathbf{A}_2^{old} \odot \sum_n (\mathbf{U}_{-n} \mathbf{A}_1)^T \frac{\mathbf{y}_n}{\mathbf{U}_{-n} \mathbf{A}_1 \mathbf{A}_2^{old} \mathbf{v}_n} \mathbf{v}_n^T \quad (11)$$

where \propto denotes proportionality, \odot stands for element-wise multiplication and the division above also operates element-wise.

The algorithm is then to alternate the above updates to convergence. Convergence to a local optimum of the objective (9) is guaranteed, for similar arguments as in E-M algorithms.

A remaining issue is to estimate the optimal dimensionality of the latent space, i.e. K . This may be done e.g. by cross-validation. However, in our experiments a wide range of K yielded very similar results, therefore for the next section we decided to use a $K = 3$, for best serving interpretability. The reason is, we are then able to use $p(k|n)$ for visualising the text corpus. In addition, we may also use $P(t|k)$ to inspect the latent concepts inferred, as a result of the word associations captured, in the form of ordered lists of representative words. Finally, since the neighbourhood kernel is now parametrised, visualising those is also likely to be meaningful.

¹ The KL-divergence is a natural choice for measuring the dissimilarity between two probability distributions. Other divergence functions may also be employed alternatively.

4 Results

4.1 The data

The input to our two algorithms consist of responses of 669 human subjects to a psychometric test that encompasses four standardised questions, the same for all subjects tested. The actual questions are not part of the data. Further, the training data was hand-labelled by domain experts and comprised 281 dominant and 388 submissive examples².

For this study, we assemble the four pieces of texts produced by each subject into a subject-specific text document, so each document will have four paragraphs. Topically, all of these are necessarily strongly overlapping, since the topical subject has been set and fixed through the use of the same set of four questions within the standardised test. The aim is then to investigate the possibility of automatically deciding the 'dominant' versus 'submissive' personality traits of the subjects on the basis of their written answers.

First, we perform a standard preprocessing the raw text to produce a term-document frequency based bag-of-words representation, using the Rainbow toolkit [8]. In this process, we kept all words that occurred more than once, and we switched off both stemming and stop-word removal, because unlike in topic-based discrimination, the use of these language features may well contribute to expressing one or the other of the personality traits.

This preprocessing resulted in a dictionary size of 4030 distinct terms. We did not attempt to correct for spelling mistakes, so misspelled versions of the same word occasionally coexist as distinct words in the dictionary.

4.2 Illustration

As already mentioned in Sec. 2, the basic version of our method is a lazy-learner, i.e. it does not require any effort for training. The latent variable extension (Sec. 3) in turn, requires an iterative algorithm for parameter estimation at the training stage. The evolution of the objective function Eq (9) through the iterations is shown in Figure 1. As expected, we observe a monotonic increase to convergence.

Fig. 2 further illustrates the working of the method, showing the probabilistic label predictions for each document, in a leave-one-out experiment, in comparison to the true labels. The accurate matching is quite apparent.

4.3 Personality trait prediction results

To measure the performance of the automated categorisation methods that we investigate, we create 100 independent random splits of the data into 320 training examples and 349 examples set aside for testing. We measure the classification

² The data was anonymised and kindly provided for research purposes by Dr Marco de Boni, Unilever Ltd.

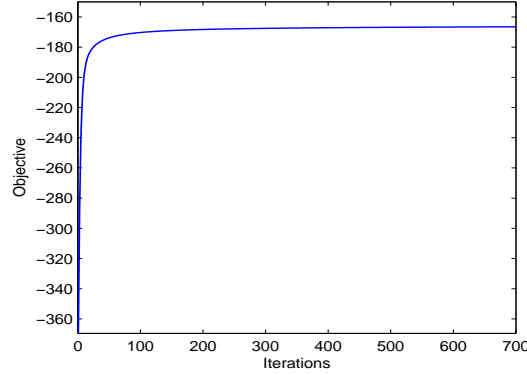


Fig. 1. Evolution of the objective function Eq (9) through iterations.

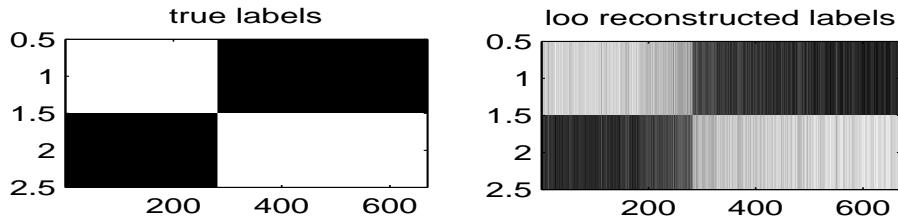


Fig. 2. Left: the true labels. Right: Leave-one-out predictions.

accuracy in terms of the percentage of correct predictions, as well as in terms of area under the Receiver Operating Characteristic curve (AUC) [4]. The results are presented in Table 1.

In the table, pNN refers to the approach presented in Section 2 and pNN-aggr3 is the extension detailed in Section 3, where $K = 3$ was set. We see, both variants of our method produced highly accurate predictions. The results of these two methods are comparable to each other, there is a very slight (and statistically insignificant) performance sacrificed for interpretability in the latter method. We see the AUC values are high and even closer, which reflects the methods' ability to also produce good uncertainty estimates.

In turn, in the remainder of the table we see results obtained on the same data, by some of the most successful existing text classification methods, which all turned out to perform quite poorly on this non-standard classification task. We found it useless to give the the AUC values for these — SVM does not give probabilistic predictions or uncertainty estimates, and for the other three methods the probabilistic predictions were close to 0 or 1 in all cases. It is clear already from the 0-1 errors that these methods are barely above a random

Table 1. Classification results over 100 random splits into 320 subjects for training and 349 subjects for testing (mean \pm standard deviation).

Method	% classification accuracy	AUC
pNN	0.9822 \pm 0.0120	0.9996 \pm 0.0004
pNN-aggr3	0.9722 \pm 0.0119	0.9977 \pm 0.0015
Multinomial Naïve-Bayes	0.6002 \pm 0.0152	
Dirichlet Compound Multinomial	0.6098 \pm 0.0082	
linear-SVM	0.6183 \pm 0.0131	
sparse (L1) Logistic Regression	0.5977 \pm 0.0682	

guessing (though the difference from random guessing was found statistically significant in all cases).

In these comparisons, the multinomial Naive Bayes (implemented in Rainbow) may be considered as a baseline for its simplicity, nevertheless its previous good performance in topic-based classification has been quite remarkable (see e.g. [9]). The Dirichlet Compound Multinomial is a fairly recent enhancement on Naive Bayes [2], endowed with an ability to model word burstiness in the language, which offers a more realistic word distribution than the multinomial model. Despite its remarkable previous success, it does not excel in the non-standard categorisation task investigated here. The linear SVM has been among the best performing and most popular text classifiers [7]. It has an in-built capability of avoiding overfitting in the presence of large numbers of relevant features (words). However, from the results obtained, its limitation in this peculiar problem is most apparent³ Next, we tested the possibility that perhaps the poor performance in our task is because too many words may be irrelevant to the target. We employed the sparse logistic regression, which again has been previously found to have a state-of-the-art performance for text categorisation [3]. We used the efficient implementation available in [13], which can deal with high dimensional data. However, the results obtained in this problem have been again not much better than random.

4.4 Discussion

It is not trivial to generalise and trace ultimate conclusions from these results, nevertheless, we are able to answer the most concerning of the questions. Namely, it is certainly the case, for this data set, that a statistical approach using just word frequency information (without any more sophisticated NLP technique) is still capable to predict the classes of interest. Thus, the automated prediction of personality traits from psychometric tests seems feasible to a reasonably high degree of accuracy.

³ The result given in the table is with the C parameter optimised using an internal leave-one-out validation. However, we have not experimented with other kernels so far.

As for the large difference in performance between our simple approach versus several of the previously most successful methods, we conjecture at this point this may be because — unlike in topic-based classification, where topic classes tend to be well separated — here we may encounter peculiar non-linear boundaries between the representatives of the two personality traits in the data space. Our rather simple approach, with its kNN flavour is able to deal with this successfully. Further investigations using other, non-linear classifiers may shed more light on this issue. However, one advantage of our approach, besides of being extremely simple and computationally inexpensive, is its probabilistically clear foundation. This enables extensions in a straightforward manner, as we have already seen in Section 3. Rather than employing a universal black-box predictor and having to search for the appropriate non-linearity to suit the problem at hand, we have a natural way of defining similarities and optimising the leave-one-out prediction error directly, in order to gain further explanatory information in support of a subsequent interpretation of the results.

4.5 Interpretability

Naturally, the extended (parametrised) version of our method is computationally more demanding than the basic variant. Let us inspect therefore the additional information that it provides. Figure 3 shows the visualisation of the text collection in the coordinate basis defined by the estimated parameters $P(k|n)$. Contrary to any unsupervised or topic-based visualisation method, by our model construction, the groupings are necessarily w.r.t. the particular label specification. The colours and markers reflect the true labels for the ease of visual evaluation. Indeed the two fairly distinct clouds of points have a good correspondence to the true personality trait labels. Beyond just label predictions, such a visualisation may have the benefit of revealing the more detailed topological proximity / similarity relationships between the subjects w.r.t. the categorisation studied, beyond the hard partitioning into disjoint classes.

Further, we may inspect the three latent variables through their associated lists of probable words. As discussed earlier, these are unlikely to be topical aspects, but instead, aspects that define the imposed non-topical grouping of the documents. In the present case, these will be correlated lists of words whose usage characterises the two personality traits. The top end of these lists are shown in Table 2.

To find out how these factors relate to the known classes, we look at the probabilities $P(z|k) = \sum_n P(z|n)P(n|k)$. In this case, we find a probability very close to 1 for the 'submissive' class in the case of the aspects $k = 1$ and $k = 2$, and very close to 1 for the 'dominant' class in the case of $k = 3$. So the interpretation of the above lists of words is clear. We should note though that, in general, these estimated 'conceptual aspects' need not be exclusively tied with one of the classes but they may be shared by all classes with a certain probability $P(z|k)$ that we can calculate.

Finally, to conclude the discussion on parameter interpretability, Figure 4 shows the neighbourhood kernels as obtained by our kNN and aggregated kNN

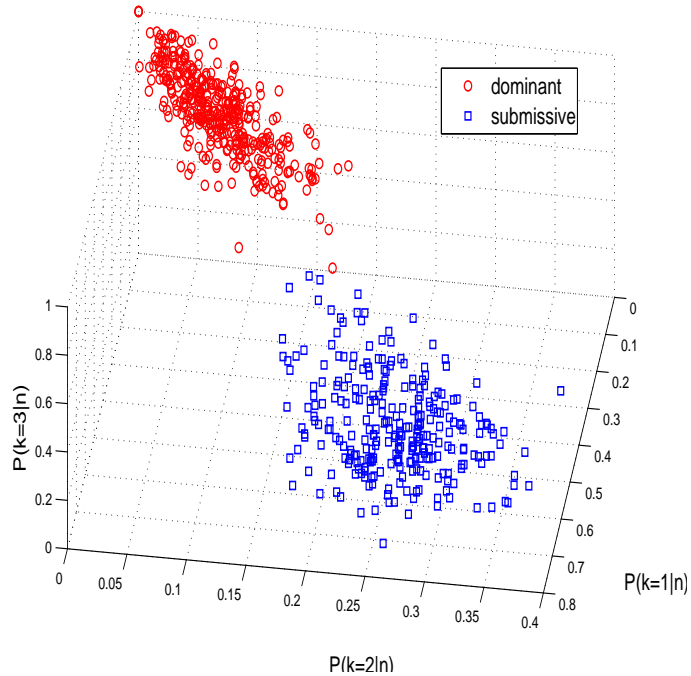


Fig. 3. Visualisation of the text collection in the coordinate basis defined by $P(k|n)$. Each points represents one subject. The superimposed true labels represented by the markers have a good correspondence with the shape of the density.

respectively, highlighting the advantage of the latter in terms of bringing out the hidden structure from the data. Since in the aggregated version the neighbourhood probabilities become a function of the additional latent variable and its associated parameters, this allows us essentially to learn the similarities implied by the given labelling. This is most apparent from the right-hand plot, where we clearly see the 2-class structure in the aggregated kNN neighbourhood. By contrary, this global structure is not readily seen in the simpler kNN neighbourhood (left-hand plot), as this model extracts the predictive information in a 'local', i.e. instance-specific manner without distilling any global structural information.

5 Conclusions

We presented a novel probabilistic approach for text categorisation and text analysis for automating the prediction of personality traits on the basis of standardised psychometric tests taken by human subjects. This is a non-topical, non-standard text categorisation problem, which presents difficulties to a number of state of the art text classifiers. Through our approach, we demonstrated

Table 2. The ordered lists of the most probable words associated with the three possible values of the latent variable in our model.

ran 0.020; peoples 0.019; ends 0.018; excersise 0.017; discretion 0.016;
diseased 0.016; encourages 0.015; motivate 0.014; becasue 0.013;
appeal 0.012; achoice 0.012; questioning 0.012; steadied 0.011; epecially 0.011;
turns 0.011; puffing 0.011; energy 0.011; baulk 0.010; expression 0.010; ...

throats 0.030; ran 0.029; tying 0.028; lung 0.026; speak 0.026;
encourages 0.023; ready 0.023; wanna 0.021; atmospheres 0.018; wee 0.018;
pregnant 0.018; previous 0.017; attitude 0.016; greed 0.016; circumstances 0.016;
achoice 0.015; garden 0.015; epecially 0.015; closely 0.015; ...

countries 0.146; home 0.109; easily 0.103; restrictions 0.050; burger 0.050;
hours 0.045; things 0.044; health 0.036; social 0.031; longer 0.029;
clubs 0.027; matter 0.024; unlike 0.023; arrival 0.021; media 0.019;
bombarded 0.017; boys 0.016; public 0.013; sec 0.012; ...

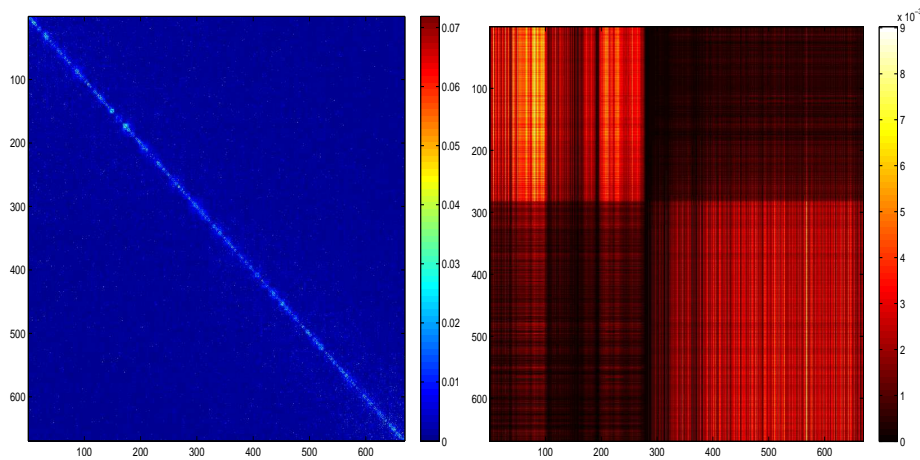


Fig. 4. The obtained neighbourhood probabilities (kernels) $p(n|n')$ in pNN and the 3D aggregated version of pNN.

that such a peculiar task can still be successfully automated within a simple statistical approach using a standard word frequency representation of documents. Further work may consider more data sets of this kind to test the so far well-performing method further. In addition, other existing or novel classifiers could be included in the investigation to further pin down the reasons that make some methods more suitable than others for this problem.

Acknowledgement

Thanks to Dr Marco de Boni from Unilever Ltd. for sharing the data and the problem.

References

1. F. Colas and P. Brazdil. Comparison of SVM and Some Other Classification Algorithms in Text Classification Tasks. *Artificial Intelligence in Theory and Practice*. Vol. 217/2006, pp. 169–178.
2. R.E. Madsen, D. Kauchak and Ch. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, 2005
3. S. Eyheramendy, A. Genkin, W.-H. Ju, D.D. Lewis, and D. Madigan. Sparse Bayesian Classifiers for Text Categorization. Technical Report, Department of Statistics, Rutgers University, 2003.
4. T Fawcett. ROC graphs : Notes and practical considerations for researchers, Technical report, HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto CA 94304, USA, April 2004.
5. J Goldberger, S Roweis, G Hinton, R Salakhutdinov. Neighbourhood Component Analysis. *Neural Information Processing Systems 17 (NIPS'04)*. pp. 513-520.
6. Th. Hofmann. Probabilistic Latent Semantic Analysis. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 1999.
7. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning*, 1998.
8. A.K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. 1996. www.cs.cmu.edu/~mccallum/bow.
9. T. Mitchell. *Machine Learning*. McGraw Hill, 1997. (ch.6)
10. G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18, 613–620, 1975.
11. L. Saul and F. Pereira, Aggregate Markov Models for statistical language processing, *Proc of the Second Conference on Empirical Methods in Natural Language Processing*, pp.81–89, 1997.
12. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1–47, 2002.
13. S.K. Shevade and S.S. Keerthi. A Simple and Efficient Algorithm for Gene Selection using Sparse Logistic Regression, Technical Report No. CD-02-22, Control Division, Department of Mechanical Engineering, National University of Singapore, Singapore - 117 576, 2002.
14. Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, vol. 1, nr. 1/2, pp. 69–90, 1999.