

Model-based estimation of word saliency in text

Xin Wang and Ata Kabán

School of Computer Science, The University of Birmingham,
Birmingham, B15 2TT, UK

`{X.C.Wang,A.Kaban}@cs.bham.ac.uk`

Abstract. We investigate a generative latent variable model for model-based word saliency estimation for text modelling and classification. The estimation algorithm derived is able to infer the saliency of words with respect to the mixture modelling objective. We demonstrate experimental results showing that common stop-words as well as other corpus-specific common words are automatically down-weighted and this enhances our ability to capture the essential structure in the data, ignoring irrelevant details. As a classifier, our approach improves over the class prediction accuracy of the Naive Bayes classifier in all our experiments. Compared with a recent state of the art text classification method (Dirichlet Compound Multinomial model) we obtained improved results in two out of three benchmark text collections tested, and comparable results on one other data set.

1 Introduction

Information discovery from textual data has attracted numerous research efforts over the last few years. Indeed, a large part of communications now-days is computer-mediated and is being held in a textual form. Examples include email communication, digital repositories of literature of various kinds and much of the online resources. Text analysis techniques are therefore of interest for a number of diverse subjects.

Interestingly, it is apparent from the literature that relatively simple statistical approaches, ignoring much of the syntactical and grammatical complexity of the language are found to be surprisingly effective for text analysis, and are able to perform apparently difficult tasks such as topic discovery, text categorisation, information retrieval and machine translation, to name just a few. It appears as though the natural language contains so much redundancy that simplifying statistical models still capture useful information and are able to work effectively in principle.

It is common-sense however that word occurrences do not carry equal importance and the importance of words is context-dependent. Yet, many current approaches to text modelling make no attempt to take this into consideration. Although the problem of selection/weighting of salient features for supervised text categorisation problems has been studied extensively in the literature [10], a model based approach that could be used for structure discovery problems is

still somewhat lacking. Model-based approaches exist for continuous valued data [3], however these are not directly applicable to discrete data such as text.

In this paper we overcome the deficiency of other text modeling approaches by considering word saliency as a context-dependent notion. We formulate and investigate a generative latent variable model, which includes word saliency estimation as integral part of a multinomial mixture-based model for text modelling. The obtained model can be used either for supervised classification or unsupervised class discovery (clustering). The unsupervised version of the problem is particularly challenging because there is no known target to guide the search. In this case we need to assess the relevance of the words without class labels, but with reference to our model of the structure inherent in the data.

In previous work [9], we have studied model-based feature weighting in a Bernoulli mixture framework. We found that, for documents, the absences of words are less salient than their presences. This has brought the suitability of the binary representation of text into question. Based on these and earlier results [4], a frequency based representation is arguably more appropriate and therefore in this paper we build on multinomial mixtures. We recognise a close relationship of our model with the Cluster Abstraction Model (CAM) [2], which however has not been previously used for word saliency estimation and its capabilities for down-weighting non-salient words have not been made explicit. The question of how salient is a particular word in comparison to other words has not been asked previously in a model-based manner and indeed in its basic form, CAM cannot answer this question. Making this capability of the model explicit is therefore a useful addition made in this paper, which could be used e.g. in text summarisation problems.

We demonstrate experimental results showing that a multinomial mixture-based model equipped with a feature saliency estimator is able to automatically down-weight common stop-words as well as other corpus-specific common words and this enhances our ability to capture the essential structure in the data, ignoring irrelevant details. As a classifier, our approach improves over the class prediction accuracy of the Naive Bayes classifier in all our experiments. Compared with a more recent state of the art text classifier, the Dirichlet Compound Multinomial (DCM) [2], we obtained improved results in two out of three benchmark text collections tested, and comparable results on one other data set.

2 The model

Our method is based on the multinomial mixture model and a common, cluster-independent multinomial. In addition, a binary latent variable ϕ is introduced for each draw of a word from the dictionary, to indicate whether the cluster-specific or the cluster-independent multinomial will generate the next word. The cluster-specific multinomials generate salient words whereas the common multinomial generates common words. The process can be formulated as a generative model.

2.1 Notation

– Data

Let $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ denote N documents. Each document

$\mathbf{x}_n = (x_1, \dots, x_{L_n})$ contains L_n words from a T -size dictionary.

$\mathcal{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ is the terms \times documents matrix of \mathcal{X} where each \mathbf{y}_n is a T -dimensional histogram over the dictionary.

– Hidden variables

Let $\mathcal{Z} = (z_1, \dots, z_N)$ denote the class labels of the N documents. With each document, $\Phi_n = (\phi_1, \dots, \phi_{L_n})$ is the sequence of the binary indicators of the saliency for each word in $\mathbf{x}_n, n = 1, \dots, N$.

– Parameters

θ_k is the k -th multinomial of the K components of the model, λ denotes the common multinomial component. The prior probability that a word is picked from cluster k of the K clusters is denoted by ρ_k , as opposed to $1 - \rho_k$, which is the probability with which the common component activates and generates words. $\alpha_1, \dots, \alpha_K$ will denote the individual prior probabilities of these clusters. Let $\Theta \equiv \{\alpha, \theta_1, \dots, \theta_K, \lambda, \rho\}$ denote the full parameter set.

2.2 The generative process

Assume a document (data sequence) $\mathbf{x} = (x_1, \dots, x_L)$ is to be generated.

- As in a standard finite mixture, a component label $z = k$ is selected by sampling from a multinomial distribution with parameters $(\alpha_1, \dots, \alpha_K)$; Then for each word $l = 1, \dots, L$:
- Generate ϕ_l from flipping a biased coin, whose probability of getting a head is ρ_k .
- If $\phi_l = 1$, then use the cluster-specific multinomial distribution θ_k to generate the word $x_l = t$, with probability θ_{tk} ;
- Else ($\phi_l = 0$), use the common multinomial distribution λ to generate the word $x_l = t$, with probability λ_t .

Model formulation Following the above generative process, the joint probability of \mathcal{X} and Φ , given the model parameters and under the i.i.d assumption of document instances is:

$$\begin{aligned}
 P(\mathcal{X}, \Phi | \Theta) &= \prod_{n=1}^N \sum_{k=1}^K \alpha_k \prod_{l=1}^{L_n} \left[\rho_k P(x_{ln} | \theta_k) \right]^{\phi_{ln}} \left[(1 - \rho_k) P(x_{ln} | \lambda) \right]^{1 - \phi_{ln}} \\
 &= \prod_{n=1}^N \sum_{k=1}^K \alpha_k \prod_{l=1}^{L_n} \prod_{t=1}^T \left[\left[\rho_k P(x_{ln} = t | \theta_k) \right]^{\phi_{ln}} \left[(1 - \rho_k) P(x_{ln} = t | \lambda) \right]^{1 - \phi_{ln}} \right]^{\delta(x_l, t)} \\
 &= \prod_{n=1}^N \sum_{k=1}^K \alpha_k \prod_{l=1}^{L_n} \prod_{t=1}^T \left[\left[\rho_k \theta_{tk} \right]^{\phi_{ln}} \left[(1 - \rho_k) \lambda_t \right]^{1 - \phi_{ln}} \right]^{y_{ln}}
 \end{aligned}$$

Therefore the marginal probability (likelihood function) by summing out the hidden variable ϕ_l is the following.

$$\begin{aligned}
P(\mathcal{Y}|\Theta) &= \sum_{\Phi} P(\mathcal{X}, \Phi|\Theta) \\
&= \prod_{n=1}^N \sum_{k=1}^K \alpha_k \prod_{l=1}^{L_n} \prod_{t=1}^T \sum_{\phi_l=0}^1 \left[\left[\rho_k \theta_{tk} \right]^{\phi_{ln}} \left[(1 - \rho_k) \lambda_t \right]^{1 - \phi_{ln}} \right]^{y_{tn}} \\
&= \prod_{n=1}^N \sum_{k=1}^K \alpha_k \prod_{t=1}^T \left[\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t \right]^{y_t} \tag{1}
\end{aligned}$$

Model estimation *E-step*: the class posteriors, that is the expected value of the latent variables (z_n) associated with each observation given the current parameter estimates are calculated

$$\gamma_{kn} \equiv P(z_{kn} = 1 | \mathbf{y}_n) \propto \alpha_k \prod_{t=1}^T \left[\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t \right]^{y_{tn}} \tag{2}$$

We note that the parameter estimation can be conveniently carried out without computing the posterior probabilities of the saliency variables, and so those will be computed after the parameter estimation is complete.

M-step: the parameters are re-estimated as follows,

$$\begin{aligned}
\hat{\alpha}_k &= \frac{\sum_n \gamma_{kn}}{\sum_{nk} \gamma_{kn}} = \frac{\sum_n \gamma_{kn}}{N} \propto \sum_n \gamma_{kn} \\
\hat{\theta}_{tk} &\propto \frac{\rho_k \theta_{tk}}{\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t} \sum_{n=1}^N \gamma_{kn} y_{tn} \\
\hat{\lambda}_t &\propto \sum_{k=1}^K \frac{(1 - \rho_k) \lambda_t}{\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t} \sum_{n=1}^N \gamma_{kn} y_{tn} \\
\hat{\rho}_k &= \frac{1}{\sum_{t=1}^T \sum_{n=1}^N \gamma_{kn} y_{tn}} \sum_{t=1}^T \frac{\rho_k \theta_{tk}}{\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t} \sum_{n=1}^N \gamma_{kn} y_{tn}
\end{aligned}$$

Scaling It is to be noted that this algorithm can be efficiently implemented using sparse matrix manipulation routines. In particular, the data is typically very sparse – many entries y_{tn} are zero, since in most text documents the majority of the words of the overall dictionary are not used. Therefore the log of the E-step expression can be rewritten as a multiplication between a sparse and a dense matrix. Also each one of the M-step equations, where the data appears, can be re-arranged as matrix multiplications where one of the matrices (the data matrix) is sparse. Therefore the scaling per iteration is linear in the number of non-zero entries in the terms \times documents matrix.

Inferring the probability that a word is salient In this section we show that word saliency estimates can be computed from the presented model. This interpretation adds a new and useful functionality to the model and also helps us to better understand the working of the model. Despite the calculations are straightforward, this issue has not been addressed or noticed in any of the related previous works.

After the parameter estimation is completed, the expected saliency of a word t can be inferred as the following. The probability that an arbitrary occurrence of the term t is salient will be denoted $P(\phi = 1|t)$ and this probability is evaluated as

$$\begin{aligned} P(\phi = 1|t) &= \frac{\sum_{n=1}^N \sum_{l=1}^{n_l} P(\phi_l = 1|x_{ln} = t)}{\sum_{n=1}^N y_{tn}} = \frac{\sum_{n=1}^N y_{tn} P(\phi_l = 1|x_{ln} = t)}{\sum_{n=1}^N y_{tn}} \\ &= \frac{1}{\sum_{n=1}^N y_{tn}} \sum_{n=1}^N y_{tn} \frac{\sum_{k=1}^K \gamma_{kn} \rho_k \theta_{tk}}{\sum_{k=1}^K \gamma_{kn} [\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t]} \\ &= \frac{1}{\sum_{n=1}^N y_{tn}} \sum_{n=1}^N y_{tn} \sum_{k=1}^K \gamma_{kn} \frac{\rho_k \theta_{tk}}{\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t} \end{aligned}$$

where, according to Bayes rule,

$$P(\phi_l = 1|x_{ln} = t) = \frac{\sum_{k=1}^K \gamma_{kn} \rho_k \theta_{tk}}{\sum_{k=1}^K \gamma_{kn} \rho_k \theta_{tk} + \sum_{k=1}^K (1 - \rho_k) \gamma_{kn} \lambda_t}$$

The latter computes the probability that a specific occurrence of a word t is salient. Note that all words may have both salient and non-salient occurrences.

Note also that such saliency based ranking of the dictionary-words would not be possible to obtain from the model parameters alone. From the parameters θ_k and λ we can obtain a list of most probable words for each cluster and the list of the most probable common words respectively, similarly to CAM [2]. However these do not answer the question of how salient a word is and how it compares to other words in terms of saliency, simply because θ_{tk} is not comparable to λ_t . This is a very important issue in problems where one needs to know which handful of words are the most responsible for the inherent topical structure of a document collection, e.g. text summarisation problems. Therefore we regard this as a notable contribution of our approach, and it allows us to have a principled model-based estimation of word saliency as an integral part of a model-based cluster analysis.

In the sequel, we provide experimental evidence of the working of our approach, and its advantages over a mixture of multinomials that does not have an in-built word saliency estimator.

3 Experiments

3.1 Synthetic data

In a first experiment, 300 data points have been generated over a 100-symbol dictionary, of which 60 are uninformative. The data is shown on the leftmost plot of Fig.1. To avoid local optima, the model estimation was repeated 40 times and the estimated model with the best in-sample log likelihood was selected. The average length of each sequence is 75 (Poisson parameter). There are four clusters mixed with one common component. The parameter ρ_k is set to 0.53, that is, when generating each word, the chance of using the cluster component is 0.53. The middle plot of Fig.1 shows the estimated feature saliences — indeed the 40

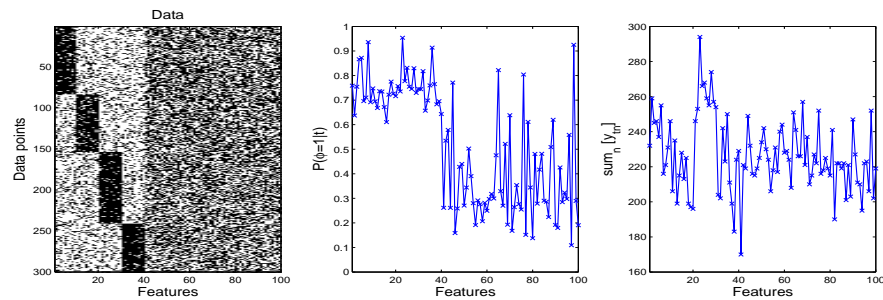


Fig. 1. Left: The data, with 4 clusters defined by 40 salient symbols (features) and having another 60 common features. Darker means higher frequency of occurrence, white stands for zero. Middle: Estimated saliency for each symbol. Right: Frequency of occurrences for each feature. We see, the frequency would have been misleading in this case, whereas the model-based approach identifies the salient features much more accurately. It is also obvious, that other possible feature weighting schemes that are unrelated to the essential structure of the data would also be fooled in some cases. E.g. if the common features are distributed more sparsely across the data set, then the tf-idf weighting would be misleading as well.

informative features, which define the four true clusters, are accurately identified. The rightmost plot depicts the frequency of occurrence for each symbol, to show that frequency counts would have been misleading for determining which are the important features in this case. Clearly, other possible feature weighting schemes that are unrelated to the essential structure of the data would also be fooled in some cases. E.g. if the common features are distributed more sparsely across the data set, then the tf-idf weighting would be misleading as well. Thus the principle behind the advantage of a model-based approach is now evident.

3.2 Real data

Finding common terms in 10 Newsgroups data We apply the model to text document data from 10 newsgroups of the 20 Newsgroups corpus¹: alt.atheism, comp.graphics, comp.sys.ibm.pc.hardware, misc.forsale, rec.autos, rec.sport.baseball, sci.electronics, sci.med, sci.space and talk.politics.mideast. The data was processed using the Rainbow toolbox², without stop-word removal but word stemming only. Rare words, with less than 5 occurrences were also removed in the preprocessing phase. The resulting data matrix is $22,945 \times 10,000$. Each class contains 1000 documents. We observed that the algorithm is sensitive to initialisation and to alleviate this problem of local optima, we initialise the common component to the sample mean of the data.

Fragments from the lower end of the word saliency ranking obtained are shown in Table 1 together with their actual estimated saliency probabilities. Clearly, most of them are common words, as checked against a standard stop-list [1]. However, notably, there are some corpus-specific common terms identified too, which don't fall into the scope of general common stop words, such as 'subject', 'question', 'article', 'write', 'information', 'people', 'world', etc. Therefore using a pre-defined stop-word list would not be able to eliminate these.

Table 1. Fragments from the lower end of the estimated word saliency ranking list in 10 Newsgroups, together with each words estimated saliency estimates. Some of the down-weighted (low saliency) words are indeed in the stop-words list, others (in the leftmost column) are common words that are specific to this corpus and would therefore not be detected, removed or down-weighted without a model-based approach.

0.0068	the	0.0741	you	0.2655	article	
0.0147	to	0.0761	as	0.2660	ask	
0.0313	in	0.0778	that	0.2691	read	
0.0489	it	0.0790	have	0.2694	once	
0.0500	this	0.0845	do	...	0.2699	post
0.0563	be	0.0853	all	0.2707	why	
0.0579	and	0.0853	so	0.2710	news	
0.0623	some	0.0855	of	0.2771	writes	
0.0644	are	0.0893	about	0.2797	wrote	
0.0667	is	0.0894	an	0.2999	information	

To further illustrate our method at work, we pick an article at random from the talk.politics.mideast group, which was not used for model training. The following paragraph shows the text with the words made with different fonts, according to their saliences as follows. The underlined words are those which are estimated to be the most salient, with saliency ≥ 0.8 and ≤ 1.0 . (Note that, the

¹ Available from <http://www.cs.cmu.edu/~textlearning>

² <http://www.cs.cmu.edu/~mccallum/bow/rainbow/>

words "prerequisites", "co-operation", "reaffirm" were removed in preprocessing due to rareness, therefore they don't have any saliency specified.) The words with saliency between 0.41 and 0.8 are in normal font, and the least salient words, with saliency ≤ 0.4 are in grey.

"(The participating States) recognize that pluralistic democracy and the rule of law are essential for ensuring respect for all human rights and fundamental freedoms. . . They therefore welcome the commitment expressed by all participating States to the ideals of democracy and political pluralism. . . The participating States express their conviction that full respect for human rights and fundamental freedoms and the development of societies based on pluralistic democracy. . . are prerequisites for progress in setting up the lasting order of peace, security, justice, and co-operation. . . They therefore reaffirm their commitment to implement fully all provisions of the Final Act and of the other CSCE documents relating to the human dimension. . . In order to strengthen respect for, and enjoyment of, human rights and fundamental freedoms, to develop human contacts and to resolve issues of a related humanitarian character, the participating States agree on the following".

As a final illustrative experiment, we look at the induced geometry of the word features. To this end, we train the model on 2 newsgroups: `sci.space` and `talk.politics.mideast`. For showing corpus-specific common terms, the words on the stop-words list are removed in this experiment. Then each word t is visualised in 3D by its coordinates $\lambda_t, \theta_{t,1}$ and $\theta_{t,2}$. So the number of points on the plot will equal the dictionary size (8,824 words in this case). This plot is seen on Fig. 2 as follows. Each point has a colour between red and green. The proportion of the red and the green component, for each word, is given by its word saliency estimate $P(\phi = 1|t)$. Pure red would stand for a saliency value of 1. Pure green would stand for a saliency value of 0. Intermediate colors signify the probability of saliency. For some of the points further away from the centre, we also show the actual word content. As we can see, the salient words for the two classes are distributed along two of the axes and are well separated from each other: 'space', 'nasa', 'earth', 'launch' clearly are salient, coming from the newsgroup on `sci.space`. Similarly, 'israel', 'armenian', 'jews', 'turkish' are salient, coming from the group `talk.politics.mideast`. In turn, the common, unsalient words lie along the third axis: 'people', 'write', 'article', 'time' are all common words that are specific to newsgroup messages. In other corpora and other contexts they may well be salient — therefore they are not on a general-purpose stop-words list and would be difficult to appropriately deal with in a non-model-based approach.

Text classification Although our algorithm has been derived from an unsupervised model formulation, it can also be used in supervised classification if class labels are available. To be comparable with other algorithms, we carry out experiments on three standard corpora: the 'industry sector', '20 Newsgroups'

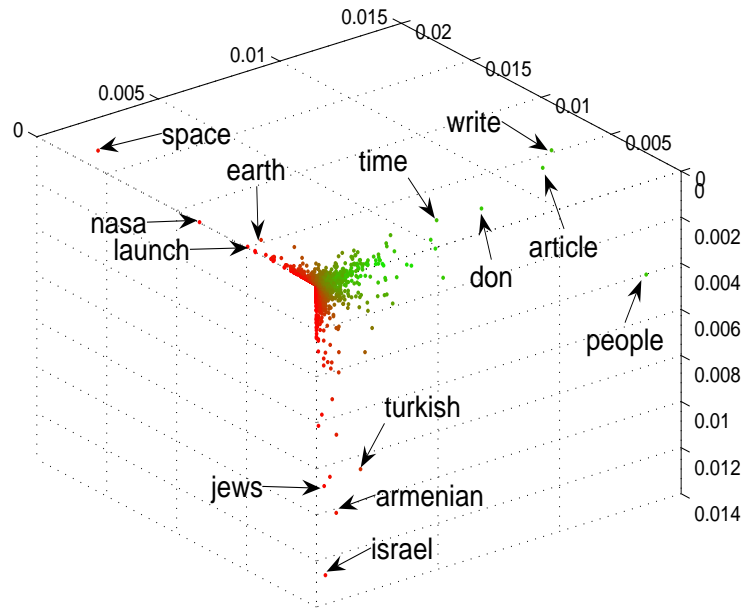


Fig. 2. The 3D view of the words in the vocabulary, as obtained from the two newsgroups `sci.space` and `talk.politics.mideast`. The colour combination visualises the estimated word saliences: redder stands for more salient, green stands for less salient. We see the salient words for the two classes are distributed along two of the main axes and are well separated from each other. The common, unsalient words lie along the third axis: ‘people’, ‘write’, ‘article’, ‘time’ are all common words specific to newsgroup messages.

and ‘Reuters-21578’ document collections³. The same experimental settings and criteria are used as in [2]. Documents are preprocessed and count vectors are extracted using the Rainbow toolbox. The 500 most common words are removed from the vocabulary in this experiment, in order to enable a direct comparison with recent results from the literature. The characteristics of the three document collections employed are given in Table 2.

Table 3 shows the classification results, that we obtained together with a result taken from [2] in the last row of the table, for comparison. For the latter, the readers can refer to [2] for more details. It should be pointed out, that for single-labelled datasets, such as the ‘Industry sector’ and ‘20 newsgroups’, the precision is used as a measure of accuracy of the classification. For multi-labelled

³ Downloaded from <http://www2.imm.dtu.dk/~rem/index.php?page=data>

Dataset	Vocabulary Size	Total nr of docs	Nr of Classes	Avg length of docs	Multiple labels	Train/Test splitting
Industry Sector	55,055	9,555	104	606	N	50/50
20 Newsgroups	61,298	18,828	20	116	N	80/20
Reuters-21578	15,996	21,578	90	70	Y	7,770/3,019

Table 2. Characteristics of the '20 newsgroups', 'Industry' and 'Reuters' data collections.

data in turn, such as the 'Reuters-21578', precision and recall are combined by computing the "break-even" point, which is defined using either micro or macro averaging [7]:

$$BE_{micro} = \frac{\sum_k TP_k}{\sum_k (TP_k + FP_k)}; \quad BE_{macro} = \frac{1}{K} \sum_k \frac{TP_k}{TP_k + FP_k}$$

K is the number of document classes. TP_k and FP_k are the number of true positives and false positives for class k respectively.

	20 Newsgroups	Industry Sector	Reuters-21578	
METHOD	PRECISION $\pm\sigma$	PRECISION $\pm\sigma$	MACRO BE	MICRO BE
Mixture model	0.866 \pm 0.005	0.8085 \pm 0.006	0.2487	0.6428
Our model	0.888 \pm 0.003	0.877\pm0.003	0.4513	0.7644
DCM	0.890\pm0.005	0.806 \pm 0.006	0.359	0.740

Table 3. Classification results for the '20 newsgroups', 'Industry sector' and 'Reuters-21578' data sets. Standard deviations are not given in the fourth column since there is a single standard training set/test set split for the Reuters-21578 corpus.

The classification results of our model are significantly superior to those obtained with a multinomial mixture, in all cases, as tested using the nonparametric Wilcoxon rank sum test at the 5% level.

Finally, we compare our results with a recently proposed text classifier, the Dirichlet Compound Multinomial (DCM) [2]. The DCM was proposed as an alternative to the multinomial for modeling of text documents, with one additional degree of freedom, which allows it to capture the burstiness phenomenon of words. It was found to be superior to the multinomial on text modeling, and promising classification improvements have been reported [2].

Interestingly, when compared to DCM, we find our proposed model performs comparably on the full 20 Newsgroups corpus, and significantly better on the other two datasets. The latter two data sets are more sparse and we conjecture this may be a reason for the success of our method.

These results are very promising because they are obtained from different considerations than those of [2]. Therefore the possibility of combining our feature saliency modelling approach with their Dirichlet-compound-multinomial building blocks is a potentially fruitful line of further research, which may bring further improvements. It is also worth mentioning that Madsen *et. al* report even better results obtained via a heuristically modified version of multinomial, such as a log transformation of the data, complement modeling approaches. Such approaches modify the input data and distribution parameters, and therefore don't give probability distributions properly normalised [2]. All such heuristics could be included in our framework too.

3.3 Discussion

We are currently trying to analyse the reasons behind the success of our model against multinomial mixture and the implications of having included a word saliency estimator as an integral part of the mixture. Intuitively, it may appear that assuming two bags of words — a bag of topic-bearing, class-specific words and a bag of common words — for generating each document offers more flexibility than assuming one bag only. However, this is not the only reason, since a mixture of multinomials is just another multinomial. It is important, that one of these bags is constrained to be the same for all the topical clusters. The cluster-specific multinomial is then 'relieved' from having to represent both the common words and the content-bearing words, whose distributions, as observed in [2], are fundamentally different.

It is interesting to view our approach from the algorithmic point of view and follow up the effects of the parameters ρ_k and λ in terms of a 'shrinkage' of the multinomial mean parameters towards a common distribution. The probabilities ρ_k control the extent of this shrinkage — if ρ_k is small, it is more probable to have non-salient words in average, therefore the multinomial mean, $\rho_k \theta_{tk} + (1 - \rho_k) \lambda_t$ parameters are 'shrunk' closer to the common component. Shrinkage methods have widely been used for data denoising e.g. in the wavelet literature, for continuous valued data. Little work has been devoted to shrinkage estimators for discrete data, and text in particular [6] and even less to analysing shrinkage effects where these are somewhat implicit.

One may also wonder why we bother devising new classification methods that bring some improvements over Naive Bayes, when Support Vector Machines (SVM) [8] are so effective and successful for text classification [5]. The main reason for this is that SVMs are 'black-box' type methods in the sense that they do not provide explanatory information regarding the text being classified. One often wants to quickly understand or summarise a text corpus, to find out what it is about and which is the most important topical information that it contains. Tools for automating this are very useful. Ideally, we should aim for methods that exhibit both excellent class prediction accuracy as well as intuitive explanatory ability for human interpretation. However achieving this is not trivial and there is a lot more to be desired. In this work we hope to have made a small step in this direction.

4 Conclusions

We proposed a generative latent variable model for feature saliency estimation based on multinomial mixture. This provides a computationally efficient algorithm that can be used in both unsupervised and supervised classification problems. The model is able to infer the saliency of words in a model-based principled way. Experimental results have shown that, common stop-words as well as other corpus-specific common words are automatically down-weighted. As a classifier, our approach improves over the class prediction accuracy of the Naive Bayes classifier in all our experiments. Compared with DCM, we obtained improved results in two out of three benchmark text collections tested, and comparable results on one other data set.

References

1. W. Nelson Francis and Henry Kucera. Frequency analysis of English usage, 1982.
2. Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 545–552, New York, NY, USA, 2005. ACM Press.
3. Mario A. T. Figueiredo Martin H. C. Law and Fellow-Anil K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1154–1166, 2004.
4. A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.
5. Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning*, Springer, 1998.
6. Andrew McCallum, Ronald Rosenfeld, Tom M. Mitchell, and Andrew Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 359–367, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
7. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
8. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
9. Xin Wang and Ata Kabán. Finding uninformative features in binary data. In *Proceedings of IDEAL05*, pages 40–47, 2005.
10. Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.