Mixtures of Dirichlet-Multinomial distributions for supervised and unsupervised classification of short text data

Advances in Data Analysis and Classification manuscript No.
(will be inserted by the editor)

Mixtures of Dirichlet-Multinomial distributions for supervised and unsupervised classification of short text data

Laura Anderlucci · **Cinzia Viroli**

Received: date / Accepted: date

Abstract Topic detection in short textual data is a challenging task due to its representation as high-dimensional and extremely sparse document-term matrix. In this paper we focus on the problem of classifying textual data on the base of their (unique) topic. For unsupervised classification, a popular approach called Mixture of Unigrams consists in considering a mixture of multinomial distributions over the word counts, each component corresponding to a different topic. The multinomial distribution can be easily extended by a Dirichlet prior to the compound mixtures of Dirichlet-Multinomial distributions, which is preferable for sparse data. We propose a gradient descent estimation method for fitting the model, and investigate supervised and unsupervised classification performance on real empirical problems.

Keywords Clustering · Gradient descent algorithm · Mixture models · Text Data Analysis

1 Introduction

Text classification has become an increasingly important task with the availability of internet sources and new technologies. Commonly, textual information is transformed into numerical representation via the so-called *bag-of-word* representation (Harris, 1954): under this framework, each text is represented as a high-dimensional vector storing the counts of each word in the document. In so doing, a document-term matrix

L. Anderlucci
Department of Statistical Sciences - University of Bologna
via Belle Arti, 41 - 40126 Bologna (Italy)
Tel.: +39 0512098267
Fax.: +39 0512086242
E-mail: laura.anderlucci@unibo.it

C. Viroli
Department of Statistical Sciences - University of Bologna
via Belle Arti, 41 - 40126 Bologna (Italy)
E-mail: cinzia.viroli@unibo.it

of frequencies is obtained and many statistical methods can be used for classification on the original frequencies or their transformations (Ko, 2012). For text clustering tasks, Mixture of Unigrams (MoU) (Nigam et al., 2000; Rigouste et al., 2007) can be used, under the assumptions that (i) each document corresponds to a single theme/topic and (ii), conditionally on the groups, the word frequencies are modelled as multinomial distributions.

For very short textual data, whose availability is increasing due to the popularity of social media like Twitter and Facebook, a very challenging problem is related to large-volume characteristics with elevate sparsity. Our motivating example lies in a set of data collected by an important Italian mobile carrier company; specifically, from the subject matter of the phone calls received by the customer service. Basically, when a customer calls the assistance service, a so-called *ticket* is created; the company operator labels it as *e.g.* a complaint, a request of clarification, a request of information for specific services, deals or promotions. Our dataset contains tickets related to five classes of services, previously classified from independent operators. The aim is to derive an effective clustering strategy that allows to automatically assign the tickets to such classes without the human judgment of an operator. The data are textual and information are collected in a document-term matrix with raw frequencies at each cell. They present a very complex and a high-dimensional structure, due to the large number of tickets and terms used by people that call the company for a specific request and by an elevate degree of sparsity (after a pre-processing step, the tickets exhibit indeed an average length of 5 words only and, thus, the document-term matrix contains zero almost everywhere).

To mitigate the effect of sparsity a useful strategy consists in putting a Dirichlet prior on the parameters of the multinomial distribution; in such a way it is possible to obtain a compound distribution, called Dirichlet-Multinomial, which proves to be more suitable to model large amount of zeros in the data. Mixtures of Dirichlet-Multinomial distributions have been successfully applied to the study of microbial diversity (see Holmes et al., 2012) for data coming from next generation sequencing, which have similar characteristics to the short texts: they are discrete, high-dimensional and really sparse. In the context of textual data the probabilistic model has been used by Yin and Wang (2014), who proposed a collapsed Gibbs Sampling algorithm for short text clustering. This Bayesian estimation procedure, together with the large number of terms, results in a heavy computational burden. In order to speed up and simplify the estimation problem, in this paper we propose a gradient descent algorithm to estimate the mixture of Dirichlet-Multinomials. We also show how the approach can be extended in a supervised classification; the classification performances are evaluated in a comparative study.

The paper is organized as follows. In the next section the Mixture of Unigrams model is described. In Section 3 the mixture of Dirichlet-Multinomials is presented. Section 4 is devoted to the estimation algorithm for fitting the model, in the traditional unsupervised framework and in the supervised perspective. Experimental results on real data are presented in Sections 5. We conclude the paper with some final remarks (Section 6).

2 Mixtures of Unigrams

Given a document-term matrix $\mathbf{X}$ of dimension $n \times T$ containing the word frequencies of each document, we denote with $\mathbf{x}_d$ the single document of length T, with $d = 1,\ldots,n$. Suppose that each document refers to a single theme, among a total of k topics. In a clustering perspective k is the number of homogeneous sets in which documents could be grouped.

In the Mixture of Unigrams model, each document is modeled as a multinomial distribution function conditionally on the values of a discrete latent allocation variable z_d describing the membership to each topic:

$$p(\mathbf{x}_d) = \sum_{i=1}^{k} \pi_i p(\mathbf{x}_d|z_d = i), \tag{1}$$

with $p(z_d = i) = \pi_i$, where π_i are positive weights, with (i) $\pi_i > 0$ for $i = 1,\ldots,k$ and (ii) $\sum_{i=1}^{k} \pi_i = 1$.

In equation (1) $p(\mathbf{x}_d|z_d = i)$ is the multinomial distribution with cluster-specific parameter vector $\boldsymbol{\omega}_i$:

$$p(\mathbf{x}_d|z_d = i) = \frac{N_d!}{\prod_{t=1}^{T} x_{dt}!} \prod_{t=1}^{T} \omega_{ti}^{x_{dt}}, \tag{2}$$

with $N_d = \sum_{t=1}^{T} x_{dt}$. The multinomial distribution assumes that, conditionally to the cluster membership, all the terms can be regarded as independently distributed.

The model is indeed a simple mixture of multinomial distributions that can be easily estimated by the EM algorithm (see Nigam et al. (2000) for further details) under the assumption that a document belongs to a single topic and the number of groups coincides with the number of topics.

3 Mixtures of Dirichlet-Multinomial distributions

Mixtures of Unigrams can be extended by allowing a further layer in the probabilistic generative model, through a Dirichlet prior on the Multinomial parameter. In so doing we obtain a hierarchical architecture able to describe the data structure with larger flexibility.

The proportions $\boldsymbol{\omega}$ in (2) are now regarded as realizations of latent variables with a Dirichlet distribution of positive parameters $\boldsymbol{\theta}_i$, which are vectors of length T:

$$p(\boldsymbol{\omega}|z = i) = \frac{\Gamma\left(\sum_{t=1}^{T} \theta_{it}\right)}{\prod_{t=1}^{T} \Gamma(\theta_{it})} \prod_{t=1}^{T} \omega_{ti}^{\theta_{it}-1}, \tag{3}$$

where Γ denotes the Gamma function.

By combining equations (2) and (3), the latent variable $\boldsymbol{\omega}$ can be integrated out from the model estimation, thus leading to the Dirichlet-Multinomial compound model. For the sake of a simple notation, in the following we will refer to a generic document with $\mathbf{x}$.

The Dirichlet-Multinomial model is

$$
\begin{aligned}
p(\mathbf{x}|z=i) &= \int p(\mathbf{x}|z=i,\boldsymbol{\omega})p(\boldsymbol{\omega}|z=i)d\boldsymbol{\omega} \\
&= \frac{\left(\sum_{t=1}^{T} x_t\right)!}{\prod_{t=1}^{T} x_t!}\frac{\Gamma\left(\sum_{t=1}^{T}\theta_{it}\right)}{\prod_{t=1}^{T}\Gamma(\theta_{it})}\int\prod_{t=1}^{T}\omega_{tij}^{x_t+\theta_{it}-1}d\boldsymbol{\omega}_{ij} \\
&= \frac{\left(\sum_{t=1}^{T} x_t\right)!}{\prod_{t=1}^{T} x_t!}\frac{\Gamma\left(\sum_{t=1}^{T}\theta_{it}\right)}{\prod_{t=1}^{T}\Gamma(\theta_{it})}\frac{\prod_{t=1}^{T} x_t+\Gamma(\theta_{it})}{\Gamma\left(\sum_{t=1}^{T} x_t+\theta_{it}\right)} \\
&= \frac{\sum_{t=1}^{T} x_t}{\prod_{t=1}^{T} x_t}\frac{B\left(\sum_{t=1}^{T} x_t,\sum_{t=1}^{T}\theta_{it}\right)}{\prod_{t=1}^{T} B(x_t,\theta_{it})},
\end{aligned}
\tag{4}
$$

where B denotes the Beta function.

The final allocation of the documents to the clusters is thus given by the posterior probability $p(z|\mathbf{x})$ that can be obtained as follows:

$$
p(z=i|\mathbf{x}) = \frac{p(\mathbf{x}|z=i)p(z=i)}{\sum_{i=1}^{k} p(\mathbf{x}|z=i)p(z=i),} \qquad i=1,\ldots,k. \tag{5}
$$

4 Model Estimation

4.1 Unsupervised classification framework

Maximum Likelihood Estimation for the model parameters can be efficiently obtained through a gradient descent algorithm. Let $\boldsymbol{\Theta}=\{\boldsymbol{\theta},\boldsymbol{\pi}\}$ be the whole set of model parameters.

Given the log-likelihood of the model defined as

$$
\ell(\boldsymbol{\Theta}) = \sum_{d=1}^{n}\log\sum_{i=1}^{k}\pi_i p\left(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i\right), \tag{6}
$$

where $p\left(\mathbf{x}|z=i;\boldsymbol{\theta}_i\right)$ is the Multinomial-Dirichlet distribution derived in (4), the gradient descent method requires the computation of the first derivatives with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. The score with respect to the Multimnomial-Dirichlet parameters is

$$
\begin{aligned}
\frac{\partial\ell(\boldsymbol{\Theta})}{\partial\boldsymbol{\theta}} &= \sum_{d=1}^{n}\frac{1}{p\left(\mathbf{x}_d;\boldsymbol{\Theta}\right)}\sum_{i=1}^{k}\pi_i\frac{\partial p\left(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i\right)}{\partial\boldsymbol{\theta}_i} \\
&= \sum_{d=1}^{n}\sum_{i=1}^{k}\frac{\pi_i}{p\left(\mathbf{x}_d;\boldsymbol{\Theta}\right)}p\left(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i\right)\frac{\partial\log p\left(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i\right)}{\partial\boldsymbol{\theta}_i} \\
&= \sum_{d=1}^{n}\sum_{i=1}^{k} p\left(z_d=i|\mathbf{x}_d;\boldsymbol{\Theta}\right)\frac{\partial\log p\left(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i\right)}{\partial\boldsymbol{\theta}_i}.
\end{aligned}
\tag{7}
$$

In this simplified form the gradient is the posterior weighted sum of the single log-likelihood gradients.

Therefore, maximization of the positive vectors $\boldsymbol{\theta}_i$ involves the derivative of $\log p\left(\mathbf{x}|z=i;\boldsymbol{\theta}_i\right)$ that can be rewritten as

$$\begin{aligned}\log p(\mathbf{x}_d|z_d=i;\boldsymbol{\theta}_i) \propto \log\Gamma\left(\sum_{t=1}^{T}\theta_{it}\right) - \log\Gamma\left(\sum_{t=1}^{T}x_{dt}+\theta_{it}\right)\\ -\sum_{t=1}^{T}\log\Gamma\left(\theta_{it}\right)+\sum_{t=1}^{T}\log\Gamma\left(x_{dt}+\theta_{it}\right)\end{aligned}$$

The gradient of the previous terms with respect to the vector $\boldsymbol{\theta}_i$ can be easily computed. To this aim, we need the definition of the digamma function $\psi(x) = \frac{d}{dx}\log\Gamma(x)$. The first derivative is thus $\psi\left(\boldsymbol{\theta}_i^\top\mathbf{1}\right)(\mathbf{1}^\top)$, where $\mathbf{1}$ is a column vector of ones of length T. Similarly, the second score is $\psi\left(\sum_{t=1}^{T}x_{dt}+\theta_{it}\right)(\mathbf{1}^\top)$. The gradients of the third and forth terms are row vectors of length T with elements $\psi(\theta_{it})$ and $\psi(x_{dt}+\theta_{it})$, respectively.

The estimates for the mixing proportions $\boldsymbol{\pi}$ of the mixture model have to be computed via $\frac{\partial\ell(\boldsymbol{\Theta})}{\partial\boldsymbol{\pi}}$ under the constraints that they are positive and sum to one. This can be obtained by the softmax transform $\pi_i = \frac{\exp(q_i)}{\sum_{i'=1}^{k}\exp(q_h)}$. The constraint optimization problem leads to the following estimates

$$\hat{\pi}_i = \frac{\sum_{d=1}^{n}p(z_d=i|\mathbf{x}_d)}{n}. \tag{8}$$

The ingredients needed for the model estimation are therefore all available. The algorithm consists of an initialization step and an estimation step, which increases the likelihood in each step and is repeated until convergence. The scheme of the algorithm is the following:

1. *Initialization*: Set $h=0$. For each component $i=1,\ldots,k$, choose values for the vectors $\boldsymbol{\theta}_i^{(h)}$ and fix equispaced probabilities for $\pi_i^{(h)}$.
2. *Estimation step*: Repeat the following until $\ell(\boldsymbol{\Theta})$ stops changing:
   (a) Compute the posteriors using (5);
   (b) For $i=1,\ldots,k$ compute new values for $\boldsymbol{\theta}_i$ using the gradients in (7) according to $\boldsymbol{\theta}_i^{(h+1)} = \boldsymbol{\theta}_i^{(h)} + \alpha\frac{\partial\ell(\boldsymbol{\Theta})}{\partial\boldsymbol{\theta}_i}$.
   (c) For $i=1,\ldots,k$ compute new values for $\boldsymbol{\pi}_i$ using (8).
   (d) h=h+1.

Usually, the presented algorithm needs few iterations to reach convergence (not more than 50 with a very low tolerance level). This represents a high advantage with respect to other estimation procedures: for example a Gibbs-sampling algorithm (Yin and Wang, 2014) with 10000 runs is about 25 times more time-consuming than the proposed gradient descent procedure.

4.2 Supervised Classification framework

The previous model can be implemented in a supervised perspective where for a subset of observations (training set) the labels denoted by z_d are known, and the remaining part (test set) is unlabeled (Ambroise and Govaert, 2000; Zhu and Goldberg, 2009). In this manner, the mixture model is an automatic prediction tool for the unlabelled units in a discriminant framework. In a fully unsupervised mixture model, usually the allocation variable z is unknown, and its posterior distribution is estimated and used for clustering. In a semisupervised context, the labelled data enter into the estimation problem with their allocation. Suppose the dataset is composed by n_1 labelled data with known allocations v_{di} and by $n - n_1$ not labelled units. The allocations v_{di} are one in case of membership and zero viceversa.

The log-likelihood in (6) can be written as

$$\ell(\boldsymbol{\Theta}) = \sum_{d=1}^{n_1} \log \sum_{i=1}^{k} v_{di} p\left(\mathbf{x}_d | v_{di} = 1; \boldsymbol{\theta}_i\right) + \sum_{d=n_1+1}^{n} \log \sum_{i=1}^{k} \pi_i p\left(\mathbf{x}_d | z_d = i; \boldsymbol{\theta}_i\right). \quad (9)$$

In the gradient descent algorithm, the derivative with respect to the Multinomial-Dirichlet parameters is rephrased as

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\Theta})}{\partial \boldsymbol{\theta}} &= \sum_{d=1}^{n_1} \sum_{i=1}^{k} v_{di} \frac{\partial \log p\left(\mathbf{x}_d | v_{di}; \boldsymbol{\theta}_i\right)}{\partial \boldsymbol{\theta}_i} \\ &+ \sum_{d=n_1+1}^{n} \sum_{i=1}^{k} p\left(z_d = i | \mathbf{x}_d; \boldsymbol{\Theta}\right) \frac{\partial \log p\left(\mathbf{x}_d | z_d = i; \boldsymbol{\theta}_i\right)}{\partial \boldsymbol{\theta}_i} \end{aligned} \quad (10)$$

The estimates for the weights become

$$\hat{\pi}_i = \frac{\sum_{d=1}^{n_1} v_{di} + \sum_{d=n_1+1}^{n} p(z_d = i | \mathbf{x}_d)}{n}.$$

The algorithm is iterated until convergence. Compared with the unsupervised setting, less iterations are usually required to reach the convergence; specifically, the larger the ratio n_1/n the faster the convergence.

5 Empirical applications

5.1 Binary clustering: Reuters-21578

We used a subset of the Reuters-21578 text categorization collection (Sebastiani, 2002) to demonstrate the usefulness of the proposed model and its improved performance over the mixture of Unigrams and other classical competitors, such as the k-means with cosine distance and Euclidean distance (on data transformed according to Semantic Analysis), Partition Around Medoids (PAM), Mixture of Gaussians (on semantic-based transformed data), hierarchical clustering according to different criteria, Latent Dirichlet Allocation and Mixtures of Unigrams estimated via the EM algorithm.

We considered the subset containing the two topics acq and crude, which are contained in the R package tm (Feinerer and Hornik, 2018; Feinerer et al., 2008). The number of observations is 70: the first fifty documents refers to acq and the latter twenty to crude. After a preprocessing step aimed to remove digits, punctuation marks and stopwords, the final document-term matrix has dimension 70 × 1517, with an average number of words per document of about 50. Therefore, this dataset is not composed by extremely short documents; however it is extremely sparse and high-dimensional, with highly skewed variables. The performance of several clustering methods on these data is illustrated in Table 1. For each method we considered $k = 2$ groups and reported both the obtained Adjusted Rand Index and accuracy rate (the number of correctly classified documents over the total number of documents).

The ARI of traditional clustering methods are about zero or even negative due to the fact that the units tend to be allocated to a single group as a consequence of the high sparsity; indeed, classical clustering procedures have not been designed to account for it. However, as the two classes are imbalanced, the theoretical minimum accuracy is generally high; if the group structure is not detected, i.e. units are all assigned to a single cluster, the returned accuracy is 71.43%.

Table 1: Real data Reuters-21578. Adjusted Rand Index (ARI) and Accuracy for different methods, multiplied by 100.

Method	ARI	Accuracy
k-means with cosine dissimilarity	88.39	97.14
k-means with Euclidean distance on Semantic	-5.61	65.71
PAM with cosine dissimilarity	49.86	85.71
Mixture of Gaussians on Semantic	-3.19	68.57
Hierarchical - Ward's method with cosine dissimilarity	-1.43	50.00
Hierarchical - Centroid method with cosine dissimilarity	-1.69	70.00
Hierarchical - Single linkage with cosine dissimilarity	-1.69	70.00
Hierarchical - Complete linkage with cosine dissimilarity	-3.37	57.14
Hierarchical - Average linkage with cosine dissimilarity	-1.69	70.00
Latent Dirichlet Allocation	24.89	75.71
Mixtures of Unigrams	51.30	87.14
Mixtures of Dirichlet-Multinomials	88.39	97.14

The Mixture of Dirichlet-Multinomials works very well in identifying the underlying group structure, yielding the best performance in terms of accuracy, together with k-means. However, the challenge of the clustering task here is limited, as the problem at hand deals with binary classification of documents that are not characterized by a very few words.

5.2 Multiclass clustering: Ticket data

The final dataset that motivates this work contains $n = 2129$ documents (= *tickets*) and $T = 489$ terms. The pre-processing phase included stemming, so as to reduce inflected or derived words to their unique word stem, and the removal of some terms that represents very common non-informative words in the Italian language (stopwords).

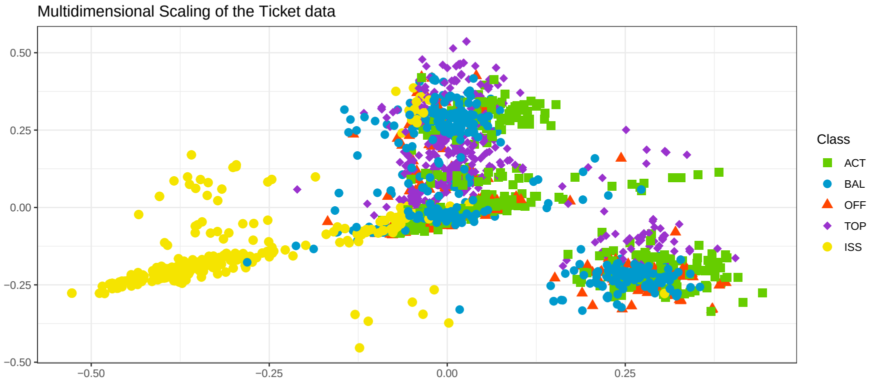

Fig. 1: Graphical representation of the ticket data via Multi-dimensional Scaling on the cosine dissimilarity.

The tickets have then been classified by independent operators into five main classes. Table 2 report a brief description of the clusters and their size (in terms of number of documents).

Table 2: Real data `Tickets`. Brief description of each class and corresponding size.

Class	Description Frequencies	
ACT	Activation of SIM, ADSL, new contracts	407
BAL	General information about current balance, consumption, etc.	471
OFF	Request of information about new offers and promotions	376
TOP	Top-up	435
ISS	Problems with password, top-up, internet connection, etc.	440

A graphical representation can be obtained via multi-dimensional scaling, computed on the cosine dissimilarity. Figure 1 shows clearly that, a part from class ISS, all the clusters widely overlap.

The crucial (and challenging) aspect of this dataset is the limited number of words included in each text, on average. Indeed, after the pre-processing phase, the tickets exhibit an average length of five terms. As previously pointed out, this translates into a sparse document-term matrix with many zeros: as a consequence, most ‘traditional’ clustering strategies fail. Table 3 shows the Adjusted Rand Index and the accuracy of different methods: k-means with cosine distance and Euclidean distance on data transformed according to Semantic Analysis, Partition Around Medoids (PAM), Mixture of Gaussians on semantic-based transformed data, hierarchical clustering according to different criteria, Latent Dirichlet Allocation and Mixtures of Unigrams estimated via the EM algorithm. For comparative reasons, for all the methods we considered the true number of clusters $k = 5$ as known.

Table 3 suggests that this classification task is particularly difficult. As previously noticed, the ARI of most traditional methods are about zero or negative. Mixtures

Table 3: Real data `Tickets`. Adjusted Rand Index (ARI) and Accuracy for different methods, multiplied by 100.

Method	ARI	Accuracy
k-means with cosine dissimilarity	23.30	51.57
k-means with Euclidean distance on Semantic	13.53	46.41
PAM with cosine dissimilarity	0.00	22.22
Mixture of Gaussians on Semantic	13.45	40.25
Hierarchical - Ward's method with cosine dissimilarity	-0.07	22.45
Hierarchical - Centroid method with cosine dissimilarity	-0.01	22.26
Hierarchical - Single linkage with cosine dissimilarity	-0.01	22.22
Hierarchical - Complete linkage with cosine dissimilarity	-0.12	22.40
Hierarchical - Average linkage with cosine dissimilarity	-0.01	22.26
Latent Dirichlet Allocation	4.93	31.75
Mixtures of Unigrams	55.68	76.30
Mixtures of Dirichlet-Multinomials	72.24	87.08

of Dirichlet-Multinomials improve upon the simple unigrams model and prove to be the most effective method for classifying the tickets, probably because they are less affected by the large number of zeros, like the other methods. The Latent Dirichlet Allocation model, despite its flexibility, is not able to improve the classification on these short documents; the association between multiple topics and tickets is not likely in this empirical context.

5.3 Supervised classification: Ticket data

The semi-supervised strategy presented in Section 4.2 has been applied on the ticket data in a 10-fold cross-validation study (Kohavi et al., 1995). At each step, each algorithm is evaluated in terms of the accuracy rate. We compared the accuracy of the supervised Dirichlet-Multinomial mixtures with several other methods: Naïve-Bayes with Multinomial distributions and weighted probability of each class (John and Langley, 1995; Hand and Yu, 2001), *k*-Nearest Neighbor (Kumbhar and Mali, 2016; Cover and Hart, 1967) with cosine distance, classic Linear Discriminant Analysis, Centroid Classifier (Tibshirani et al., 2003), classification trees Breiman et al. (1984), Random Forests Breiman (2001), Support Vector Machines with linear kernel (Cortes and Vapnik, 1995), Neural Networks and Deep Neural Networks (Lai et al., 2015; Khan et al., 2010). All these methods have been previously tuned in a dedicated training phase. The accuracy results of the selected methods are reported in Table 4.

Results from Table 4 clearly show that the supervised Dirichlet-Multinomial mixtures outperform all the other methods; only the deep neural networks perform slightly better thanks to their flexibility.

6 Final remarks

In this paper we have proposed an efficient algorithm for fitting a Mixture of Dirichlet-Multinomials on a set of short texts, that allowed to extend the more 'traditional'

Table 4: Accuracy rates (multiplied by 100) of the Naïve-Bayes classifier. In brackets cross-validation standard errors (multiplied by 100) are reported.

Method	Accuracy
Bayes Classifier - Multinomial with class document frequency	97.93 (0.31)
1-nn - Cosine distance	98.03 (0.33)
Linear Discriminant Analysis	97.14 (0.30)
Centroid classifier	95.35 (0.35)
Classification trees	92.81 (0.46)
Random Forests	97.37 (0.30)
Support Vector Machines - linear	97.84 (0.28)
Neural Networks	88.03 (1.15)
Deep Neural Networks	98.92 (0.29)
Supervised Dirichlet-Multinomial Mixture	98.12 (0.38)

mixture of Unigrams, by accounting better for the high level of sparsity of the document-term matrix.

Our proposal can be employed in both unsupervised and supervised contexts: the former represent a more challenging task, as there is no explicit hint on the underlying group structure and the elevate sparsity can mislead the usual clustering methods; the latter exploits the information on class membership and finds an allocation rule with good levels of accuracy.

Results of our study show that the Mixtures of Dirichlet-Multinomials represent an excellent strategy for the clustering of textual data and that they widely improve the Mixtures of Unigrams in terms of accuracy. The hierarchical structure, indeed, provides a larger flexibility and allows to better account for the high level of sparsity, typical of textual data. The ticket dataset, that has motivated our work, presents the additional challenge of having very short texts, on average of five terms only. The scarcity of information unabled most of the clustering methods to recover the group structure. The proposed algorithm proved to be less affected by such aspect and yielded a good accuracy.

Looking at the same dataset with a supervised perspective confirms the goodness of the classifier compared with the most popular competitors, which all return low misclassification error rates. In the performance ranking, the supervised mixture of Dirichlet-Multinomial scores second-best, just after the deep neural networks. Given the limited computational time and the simpler and better interpretable structure of the mixture model, such accuracy rate highlights a promising tool also for the supervised classification of very short documents.

The presented algorithm could be extended by adding additional latent layers (instead of just a single one), so as to enrich the model of extra flexibility, in the same spirit of the deep neural networks. In addition, the procedure is developed under the implicit assumption of a single topic per ticket; further developments may account for a multi-topic perspective. Such possible directions are left for future research.

References

Ambroise, C. and G. Govaert (2000). Em algorithm for partially known labels. In H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader (Eds.), *Data Analysis, Classification, and Related Methods*, pp. 161–166. Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Belmont CA: Wadsworth.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 21–27.

Feinerer, I. and K. Hornik (2018). *tm: Text Mining Package*. R package version 0.7-6.

Feinerer, I., K. Hornik, and D. Meyer (2008). Text mining infrastructure in r. *Journal of Statistical Software 25*(5), 1–54.

Hand, D. and K. Yu (2001). Idiot's Bayes - Not so Stupid After All? *International Statistical Review 69*, 385–398.

Harris, Z. S. (1954). Distributional structure. *Word 10*(2-3), 146–162.

Holmes, I., K. Harris, and C. Quince (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One 7*(2), e30126.

John, G. and P. Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345.

Khan, A., B. Baharudin, L. H. Lee, K. Khan, and U. T. P. Tronoh (2010). A review of machine learning algorithms for text-documents classification. In *Journal of Advances In Information Technology*.

Ko, Y. (2012, 08). A study of term weighting schemes using class information for text classification. *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial intelligence*, Volume 2, pp. 1137–1145. Montreal, Canada.

Kumbhar, P. and M. Mali (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research 5*(5), 9.

Lai, S., L. Xu, K. Liu, and J. Zhao (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2267–2273. AAAI Press.

Nigam, K., A. McCallum, S. Thrun, and T. Mitchell (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning 39*, 103–134.

Rigouste, L., O. Cappé, and F. Yvon (2007). Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing & Management 43*(5), 1260 – 1280. Patent Processing.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv. 34*(1), 1–47.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 104–117.

Yin, J. and J. Wang (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on KDDM*, KDD '14, New York, pp. 233–242. ACM.

Zhu, X. and A. B. Goldberg (2009). Introduction to Semi-Supervised Learning. *Morgan & Claypool Publishers*.