On the Naive Bayes Model for Text Categorization

Susana Eyheramendy
Department of Statistics
Rutgers University
Piscataway, NJ 0855
susanae@stat.rutgers.edu

David D. Lewis
Independent Consultant
858 W. Armitage Ave., #296
Chicago, IL 60614
ddlewis2@worldnet.att.net

David Madigan
Department of Statistics
Rutgers University
Piscataway, NJ 0855
madigan@stat.rutgers.edu

Abstract

This paper empirically compares the performance of four probabilistic models for text classification - Poisson, Bernoulli, Multinomial and Negative Binomial. We examine the "naive Bayes" assumption in the four models and show that the multinomial model is a modified naive Bayes Poisson model that assumes independence of document length and document class. Despite the fact that this last assumption might not be correct in many situations, we find that, in general, relaxing it does not change the performance of the classifier. Finally we propose and evaluate an ad-hoc method for incorporating document length.

1 Introduction

The text classification literature describes many applications of the so-called naive Bayes classifier. Lewis (1988), McCallum and Nigam (1998), and Yang and Liu (1999), for example, present analyses and extensive references. Two different versions of the model exist - the binary independence model and the multinomial model. A number of authors have provided precise descriptions of the binary independence model[1]. In contrast, standard references have obscured the core independence assumption implied by the multinomial model.

The naive Bayes classifier makes the strong assumption that the predictor variables ("features" or "words") are conditionally independent given the class. Besides this assumption, probabilistic classifiers adopt some assumed form for the conditional distribution of each feature given the class. The most popular of these probabilistic models are the ones mentioned above - the multinomial model and the binary independence model, but the literature also discusses Poisson models, Poisson mixture models and negative binomial models.

In this paper we focus on three different aspects of text classification. First, we provide precise descriptions of the two most popular so-called naive Bayes classifiers - the binary independence model and the multinomial model. Second, we empirically compare the classification performance of three different models that take into account the frequency of appearance of a word - negative binomial, multinomial and Poisson, along with the binary independence model. Third, we present an exploratory analysis that seeks to incorporate document length in the classification process.

2 The Models

Here we describe the different binary classification models and how we estimate their parameters. We represent documents by a set of random variables $X_0, X_1, \ldots, X_d$. X_0 takes values in $\{1, \ldots, C\}$ and represents the class of the document. $X_1, \ldots, X_d$ take values in $\{0, 1\}$ and represent the presence or absence of a particular term or feature ("word") in the document. Later we will consider count-valued features.

We consider probabilistic models that compute the probability of class membership for each test document and we assign the document to the class with the highest probability. We only consider binary classification ($C = 2$) in this paper.

2.1 The Binary Independence Model

The classical naive Bayes model (see, for example, Spiegelhalter and Knill-Jones, 1984 and Hand and Yu, 2002) imposes a conditional independence constraint on the joint probability distribution of these $d+1$ variables, namely that $X_1, \ldots, X_d$ are conditionally inde-

[1] Some authors refer to binary independence model as the multivariate Bernoulli model.

pendent given X_0. Figure 1 presents this model as a graphical Markov model (or Bayesian network).

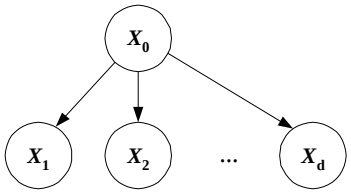

Figure 1: *The binary independence naive Bayes model.*

The probability of a document given its class is then the product of the probabilities of the words given the class:

$$\begin{aligned} & p(X_1 = x_1, \ldots, X_d = x_d | X_0 = i) \\ = & \prod_{j=1}^{d} p(X_j = x_j | X_0 = i) \\ = & \prod_{j=1}^{d} p_{ij}^{x_j} (1 - p_{ij})^{1 - x_j} \end{aligned}$$

We estimate the probabilities by $\hat{p}_{ij} = \frac{c_1 + N_{ij}}{c_2 + N_i}$ where N_{ij} is the number of documents in class i with word j and N_i is the number of documents in class i. c_1 and c_2 are constants that, in a Bayesian setting, correspond to the parameters of a beta prior distribution for p_{ij}. In the experiments reported below, we set $c_1 = 0.1$ and $c_2 = 0.04$.

To classify a new document with a given feature vector $x_1, \ldots, x_d$, this model uses Bayes rule to compute class-specific probabilities:

$$\begin{aligned} & p(X_0 = i | X_1 = x_1, \ldots, X_d = x_d) \propto \\ & p(X_0 = i) \prod_{j=1}^{d} p(X_j = x_j | X_0 = i), i = 1, ..., C. \end{aligned}$$

We estimate the class probabilities $p(X_0 = i)$ with the MLE:

$$p(X_0 = i) = \frac{\#\, of\ documents\ in\ class\ i}{total\ \#\, of\ documents}$$

Typically, authors assign the document to the class with highest probability, although more generally, the classification could account for varying misclassification costs.

2.2 The Multinomial Model

For the multinomial model, we now represent each document by a set of random variables $X_0, X_1^c, \ldots, X_d^c$. As before X_0 takes values in $\{1, \ldots, C\}$ and represents the class of the document. $X_1^c, \ldots, X_d^c$ take values in $\{0, 1, \ldots\}$ and represent the number of occurences of particular words in the document. To classify a new document with a given feature vector $x_1, \ldots, x_d$, this model computes class-specific probabilities as:

$$\begin{aligned} & p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d) \\ & \propto p(X_0 = i) p(X_1^c = x_1, \ldots, X_d^c = x_d | X_0 = i, \textstyle\sum_j x_j) \end{aligned}$$

and assumes that $p(\sum_j x_j | X_0 = i) = p(\sum_j x_j | X_0 = k)$ for all $i, k \in \{1, \ldots, C\}$ (i.e., that document length and document class are marginally independent). By conditioning on $\sum_j x_j$, and by further assuming that within each class the individual words in each document are independent and identically distributed, the model expresses $p(X_1^c = x_1, \ldots, X_d^c = x_d | X_0 = i, \sum_j x_j)$ using the standard multinomial formula.

Note that we cannot represent this model as a graphical Markov model involving $X_0, X_1^c, \ldots, X_d^c$ since the core independence assumptions do not involve these random variables. Figure 2 shows a graphical Markov model representation where a single node represents a document's vector of counts. A square represents document length to indicate that the model treats it as fixed.

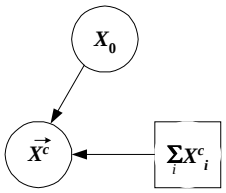

Figure 2: *The multinomial model.*

The probability of a document given its class is:

$$\begin{aligned} & p(X_1^c = x_1, \ldots, X_d^c = x_d | X_0 = i) = \\ & (\textstyle\sum_j x_j)! \prod_{j=1}^{d} \frac{p_{ij}^{x_j}}{x_j!} \end{aligned}$$

We estimate p_{ij} by:

$$\hat{p}_{ij} = \frac{c_1 + B_{ij}}{c_2 + B_i}$$

where B_{ij} is the number of time word j appears among documents in class i, B_i is the total number of words in class i, and, as before c_1 and c_2 are constants that, in a Bayesian setting correspond to a Dirichlet prior distribution. In the experiments reported below we set $c_1 = \bar{L}/d$ and $c_2 = \bar{L}$ where $\bar{L}$ is the average document length.

2.3 The Poisson Naive Bayes Model

A natural way to incorporate term frequency information into the binary independence model of Section 2.1

is to represent the document features with Poisson-distributed count-valued random variables (Lewis, 1998, §5.1). Denote by λ_{ij} the Poisson parameter for the conditional distribution of X_j^c given $X_0 = i$. The model assumes that $X_1^c, \ldots, X_d^c$ are conditionally independent given X_0. Figure 3 shows the corresponding graphical Markov model.

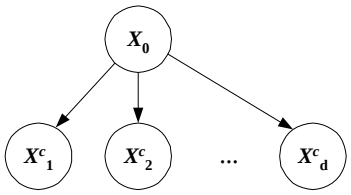

Figure 3: *The Poisson naive Bayes model.*

For a new document with a feature vector $x_1, \ldots, x_d$ the class-specific probabilities for $i = 1, \ldots, C$ are:

$$p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d)$$
$$\propto p(X_0 = i) \prod_{j=1}^{d} \exp(-\lambda_{ij}) \lambda_{ij}^{x_j}.$$

We estimate the parameter λ_{ij} as:

$$\hat{\lambda}_{ij} = \frac{c_1 + B_{ij}}{c_2 + B_i}$$

where B_{ij} is the number of times word j appears among documents in class i, B_i is the total number of documents in class i and c_1 and c_2 are constants, corresponding in the Bayesian setting to the parameters of a gamma prior distribution. In the experiments reported below we set $c_1 = 0.001$ and $c_2 = 1$.

The Poisson distribution has a rich history in information retrieval. The Poisson naive Bayes model has generally not outperformed the binary independence model. Lewis (1998), however, points out that most evaluations have taken place in the context of text retrieval, using little or no training data, rather than text classification. Further, ad hoc text retrieval formulas inspired by the Poisson model *have* outperformed binary independence models (Robertson and Walker, 1994).

2.4 The Connection Between the Poisson and Multinomial Models

Here we note that the multinomial model is equivalent to the Poisson naive Bayes model with an extra assumption concerning document length. We proceed as follows. First augment the Poisson model of Section 2.3 with a deterministic variable that is the document length - see Figure 4. Next consider classifying a new document with a given feature vector $x_1^c, \ldots, x_d^c$:

$$\begin{aligned} & p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d) \\ = \; & p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d, \sum_j x_j) \\ \propto \; & p(X_0 = i) p(\sum_j x_j | X_0 = i) \times \\ & p(X_1^c = x_1, \ldots, X_d^c = x_d | X_0 = i, \sum_j x_j) \end{aligned}$$

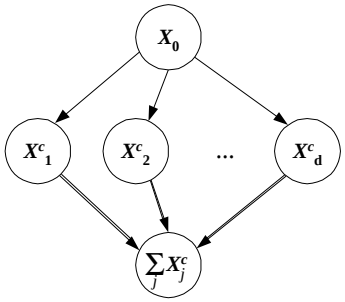

Figure 4: *The Augmented Poisson naive Bayes model.*

The first equality follows from Figure 4 since $\sum_j X_j^c$ is conditionally independent of X_0 given $X_1^c, \ldots, X_d^c$. Note that conditional on X_0, X_1 through X_d are independent Poisson random variables. A standard result in mathematical statistics (e.g., Santner and Duffy, 1989, p.17) states that conditional on their sum, independent Poisson random variables have a multinomial distribution. Hence, if in addition to the naive Bayes assumption, we further assume that $p(\sum_j x_j | X_0 = i) = p(\sum_j x_j | X_0 = k)$ for all $i, k \in \{1, \ldots, C\}$ (i.e., that document length and document class are marginally independent), the class-specific probability becomes:

$$p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d) \propto$$
$$p(X_0 = i) \prod_{j=1}^{d} \left(\frac{\lambda_{ij}}{\sum_j \lambda_{ij}} \right)^{x_j}$$

and the Poisson model reduces to the multinomial model.

So the multinomial model *is* a naive Bayes model insofar as it assumes that document word frequencies are conditionally independent given document class, but it also imposes the further assumption that document length and document class are independent. McCallum and Nigam (1998) explicitly mention the independence of document length and document class, but do not explicate the connection with the Poisson model.

Adopting a Bayesian perspective, we note that the Poisson model using conjugate gamma prior distributions for the λ_{ij} will have more hyperparameters

$(2 \times d)$ than the corresponding multinomial model with a Dirichlet distribution for p_{ij} and a gamma prior distribution for document length $(d+2)$. Consequently, the two models can lead to different Bayesian predictions. It is straightforward to derive hyperparameter constraints that do lead to identical Bayesian predictions and we will include these in the final version of the paper if space permits.

2.5 The Multinomial Word Model

An alternative but equivalent description of the multinomial model suggests that it is a generative model for words rather than documents. The multinomial word model represents each word as a pair of random variables (Y_0, Y_1), where Y_0 takes values in $\{1, \ldots, C\}$ and represents the class of the word (which, in actuality, derives from the class of the document in which the word resides), and Y_1 takes values in $1, \ldots, d$, where d, as before is the total number of words. A vector of p probabilities summing to one describes the distribution of Y_1 with a possibly different vector for each value of Y_0. Figure 5 shows the corresponding graphical Markov model.

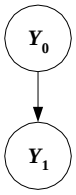

Figure 5: *The multinomial word model.*

So, the multinomial word model assumes that the words that belong to a particular class are independent and identically distributed multinoulli random variables. This model has no representation for documents *per se*.

To classify a new document comprising n words $y_1^1, \ldots, y_1^n$, the multinomial word model treats n as fixed and, for $i = 1, \ldots, C$, calculates the class-specific probability associated with n independent and identically distributed multinoulli variables:

$$p(Y_0^1 = i, \ldots, Y_0^n = i | Y_1^1 = y_1^1, \ldots, Y_1^n = y_1^n) \propto \prod_{j=1}^{n} p(Y_0^j = i) p(Y_1^j = y_1^j | Y_0^j = i).$$

The multinomial and Poisson models are richer than the binary independence model in the sense that they use word frequencies; a word that appears many times in a particular class will have more influence on future classifications than a word that makes few appearances. Note however that the models make no distinction between a word that appears ten times in one document and a word that appears once in each of ten documents.

2.6 The Negative Binomial Naive Bayes Model

The negative binomial distribution represents an alternative to the Poisson distribution for word frequencies, with two parameters per word instead of one. Katz (1996) and others have observed over-dispersion of word-count distributions in several document collections. Katz (1996) presents evidence that the negative binomial provides a better fit than the Poisson. It is interesting to note that Mosteller and Wallace (1984) modeled word counts with the negative binomial in their celebrated analysis of the Federalist Papers. The negative binomial distribution generalizes the Poisson distribution, insofar as the negative binomial is an infinite gamma-mixture of Poisson distributions. Denote by r_{ij} and p_{ij} the negative binomial parameters for the conditional distribution of X_j^c given $X_0 = i$. The graphical Markov model is identical to that of Figure 3. The negative binomial model has almost twice as many parameters as the Poisson model.

For a new document with a feature vector $x_1, \ldots, x_d$ the class-specific probabilities for $i = 1, \ldots, C$ are:

$$p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d) \propto p(X_0 = i) \prod_{j=1}^{d} \frac{(r_{ij} + x_j - 1)!}{r_{ij}!} p_{ij}^{r_{ij}} (1 - p_{ij})^{x_j}$$

and we estimate the parameters r_{ij} and p_{ij} using a modified method of moments that replaces all negative values of $\hat{p}_{ij}$ with zero (Mosteller and Wallace, 1984, p.97).

3 Experimental Results

We evaluated the multinomial model, the negative binomial model, the Poisson model, and multiviariate Bernoulli model in the context of three publicly available datasets. The multinomial model generally outperformed the other models. The negative binomial model, in particular, performed poorly.

We also show that the assumption, in the multinomial model, that the length of the document is independent of its class, even though in many cases might not be true, does not harm the classifier. We use three datasets in our analysis: MDR, Newsgroup and the ModApte version of the Reuters−21578 which we now describe.

3.1 Datasets

The MDR dataset contains information from CDRH's (Center for Devices and Radiological Health) device experience reports on devices which may have malfuntioned or caused a death or serious injury. The reports were received under both the mandatory Medical Device Reporting Program (MDR) from 1984−1996, and the voluntary reports up to June 1993. The database currently contains 620179 reports that are divided into three disjoint classes: malfunction, death and serious injury. We randomly split the dataset into 75% for training and 25% for testing. The data are available at: `http://www.fda.gov/cdrh/mdrfile.html`.

The Newsgroups dataset contains 18828 articles divided into 20 disjoint categories. Again we randomly split the dataset into 75% for training and 25% for testing. We took this version of the dataset from http://www.ai.mit.edu/people/jrennie/20Newsgroups/. It differs from the original in that this one has duplicates and most headers removed.

We use the ModApte version of Reuters−21578 dataset. It contains 7769 documents in the training set and 3019 in the testing set. The collection defines 135 categories corresponding to newswire article topics. A document may belong to 0, 1, or many categories, but we treat each category as a separate binary classification problem. We provide results on two subsets of the categories: (a) the 90 categories for which there is at least one class member in both the training and test set, and (b) the 10 categories with the highest number of class members in the corpus. For experiments with the latter set, we used only the 6775 training documents and 2258 test documents that belong to at least one of the 10 categories. The Reuters data are available at: `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

We remove stopwords and punctuation marks before we do the analyses.

3.2 Results

To evaluate the performance of the different distribution models we report micro and macro averaged recall, precision, and F values (see, Lewis, 1991, and Lewis, 1995, for definitions). In Table 1, Table 2, Table 3 and Table 4 we show these values for the 4 different models. In all the datasets the multinomial model has at least one highest value among the micro and macro average of the F measures, and in two of the datasets both of them are the highest. The Poisson model behaves similarly to the multinomial in the Newsgroups dataset and a little weaker than this but comparable with the Bernoulli model in the Reuters-21578 with 10 categories. The Bernoulli performs similarly to the multinomial in two of the datasets - MDR and Reuters−21578. The negative binomial is by far the weakest of the models in all datasets.

We also consider an augmented multinomial model that includes a class dependent model for document length - see Figure 6. Specifying a Poisson distribution for document length, i.e., for $p(\sum_j x_j | X_0 = i)$, yields a model equivalent to the Poisson naive Bayes model of Section 2.3 and suffers from the same overdispersion problem. In Tables 1-4 we present results using a non-parametric Gaussian kernel density estimate for $p(\sum_j x_j | X_0 = i)$.

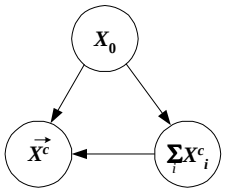

Figure 6: *An augmented multinomial model that incorporates a class dependent model for document length.*

The outcomes barely differ from the multinomial model. Table 5 and 6 show basic statistics of the *woe* ("weight of evidence") and the part of the *woe* that comes from adding a distribution in the length ("length factor") of the document. Note that *woe* is on a log scale i.e. we have a positive prediction if $woe > 0$.

$$
\begin{aligned}
woe &= log\left(\frac{p(X_0 = i | X_1^c = x_1, \ldots, X_d^c = x_d)}{p(X_0 \neq i | X_1^c = x_1, \ldots, X_d^c = x_d)}\right) \\
&= log\left(\frac{p(X_0 = i)}{p(X_0 \neq i)}\right) + \sum_j log\left(\frac{p(X_j^c = x_j | X_0 = i)}{p(X_j^c = x_j | X_0 \neq i)}\right) \\
&\quad \underbrace{+ log\left(\frac{p(\sum_j x_j | X_0 = i)}{p(\sum_j x_j | X_0 \neq i)}\right)}_{length factor}
\end{aligned}
$$

These tables show that the scale of the "length factor" is much smaller than the one for the *woe*, hence, in general, the *woe* is little affected by the length factor. A negative binomial model for document length behaves similarly. It is well known that the naive Bayes model tends to produce probability estimates for X_0 that are either close to 1 or close to 0 and are badly calibrated (Rennie, 2001). In a sense, the length factor cannot compete with this intrinsic bias that is characteristic of naive Bayes-type models.

Table 1. Summary of classifiers performances for MDR dataset

	Multinomial	Poisson	Bernoulli	Negative Binomial	Density Estimation
number of words	89077	89077	89077	89077	89077
micro recall	0.85436	0.83502	0.82874	0.60202	0.85562
micro precision	0.80225	0.59039	0.77874	0.75909	0.80761
micro F	0.82749	0.69171	0.80297	0.67149	0.83092
macro recall	0.88407	0.84539	0.86316	0.46785	0.88088
macro precision	0.64433	0.80215	0.63381	0.63102	0.64980
macro F	0.7454	0.8232	0.73091	0.53732	0.74790

Table 2. Summary of classifiers performances for the Newsgroups dataset

	Multinomial	Poisson	Bernoulli	Negative Binomial	Density Estimation
number of words	137782	137782	137782	137782	137782
micro recall	0.86807	0.85129	0.84194	0.23752	0.87232
micro precision	0.85877	0.85877	0.62825	0.99643	0.84503
micro F	0.86239	0.85501	0.71956	0.3836	0.85846
macro recall	0.86239	0.84651	0.83758	0.22768	0.8665
macro precision	0.85813	0.86200	0.73129	0.99279	0.84781
macro F	0.86025	0.84687	0.85419	0.37042	0.85705

Table 3. Summary of classifiers performances for Reuters-21578 dataset

	Multinomial	Poisson	Bernoulli	Negative Binomial	Density Estimation
number of words	24463	24463	24463	24463	24463
micro recall	0.84054	0.73024	0.76709	0.29594	0.84081
micro precision	0.60264	0.61013	0.68316	0.97278	0.60145
micro F	0.70199	0.6648	0.7227	0.45382	0.70127
macro recall	0.47614	0.31747	0.32048	0.02682	0.46545
macro precision	0.36732	0.33658	0.4136	0.18338	0.35831
macro F	0.41471	0.32674	0.36113	0.0468	0.40491

Table 4. Summary of classifiers performances for Reuters-21578 10 categories dataset

	Multinomial	Poisson	Bernoulli	Negative Binomial	Density Estimation
number of words	23080	23080	23080	23080	23080
micro recall	0.93793	0.87586	0.89037	0.4744	0.93793
micro precision	0.8243	0.80214	0.78696	0.95227	0.82197
micro F	0.87745	0.83738	0.83548	0.63331	0.87613
macro recall	0.91898	0.86254	0.83682	0.31079	0.91898
macro precision	0.72394	0.68958	0.71387	0.92817	0.72126
macro F	0.80989	0.76642	0.77047	0.46566	0.80821

Table 5. Summary of statistics for the multinomial model with density estimation on the length of the document. Reuters-21578 earn category. Training set.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
length factor	-14.27	-1.276	-0.9009	-0.9758	0.116	0.9122
woe	-3811	-282.3	-81.88	-137.9	124.1	937.8

Table 6. Summary of statistics for the multinomial model with density estimation on the length of the document. Reuters-21578 earn category. Test set.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
length factor	-13.75	-0.6366	0.3575	0.07291	0.8089	0.9271
woe	-1534	-117.4	-27.92	-55.16	53.3	402

4 An Alternative Ad-Hoc Way to Incorporate Document Length

Figure 7 plots *woe* from a multinomial classifier versus document length for the Reuters-21578 test documents. This particular binary classifier concerns the Reuters category "money-fx." Classifying the documents using *woe* alone yields 88 misclassified documents. A k-nearest neighbor classifer using both *woe* and length as features yielded just 61 misclassified documents, a 30% reduction. The Figure shows the k-NN decision boundary. Notice that the decision boundary departs from the multinomial model's decision boundary which is the line $woe = 0$. The neighborhood size used here was 15; 10-fold cross-validation of the training data selected this neighborhood size. Table 7 shows results for other Reuters categories.

Figure 7 also shows a typical behavior of the multinomial model, that is, with longer documents we get more extreme *woe* values and we have a higher rate of documents being well classified, while shorter documents get less extreme value of *woe*.

5 Discussion

Our intention in this paper is to clarify the distinction between the multinomial model and naive Bayes models. We have considered alternatives to the multinomial model that incorporate term frequencies but remain within the naive Bayes framework. Our empirical results suggest that the multinomial model often outperforms these alternatives.

We have presented variants of the multinomial model, including a Poisson naive Bayes model, a negative binomial naive Bayes model, and models explicitly incorporating document length. In general, the multinomial model, despite its poorly calibrated predictions, provides classification performance that is as good as, and in most cases better than, the performance achieved by the other models.

Directly incorporating document length into the multinomial model has little effect due to the extreme probability estimates produced by naive Bayes-type models. One possibility would be to correct for the bias before introducing length (see, for example, Spiegelhalter and Knill-Jones, 1984 or Lewis and Gale, 1994).

Acknowledgements

We thank Guido Consonni, Wen-Hua Ju, Jim Landwehr, Regina Liu, Alberto Roverato, and Yehuda Vardi for helpful discussions. The NSF supported this work.

References

Hand D.J. and Yu K. (2002). Idiot's Bayes - not so stupid after all? *International Statistical Review*, to appear.

Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, **2**, 15–59.

Lewis, D.D. (1991). Evaluating text categorization. In: *Proceedings of Speech and Natural Language Workshop*, Defense Advanced Research Projects Agency, Morgan Kaufmann, 312–318.

Lewis, D.D. (1995). Evaluating and optimizing autonomous text classification systems. In: *SIGIR 95: Proceedings of the 18th Annual Information Retrieval Conference on Research and Development in Information Retrieval*, Edward A. Fox, Peter Ingwersen, and Raya Fidel (editors), Association for Computing Machinery, 246–254.

Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In: *ECML'98, The Tenth European Conference on Machine Learning*, 4–15.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In: *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*. AAAI Press.

Mosteller, F. and Wallace, D.L. (1984). *Applied Bayesian and Classical Inference (Second Edition)*. Springer-Verlag, New York.

Rennie, J.D.M. (2001). Improving Multi-class Text Classification with Naive Bayes. *AI Technical Report, Massachusetts Institute of Technology* AITR-2001-004.

Robertson, S.E. and Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: *SIGIR 94: Proceedings of the 17th Annual IR Conference on Research and Development in IR*, W. Bruce Croft and C. J. van Rijsbergen (editors), Springer-Verlag.

Santner, T.J. and Duffy, D.E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.

Spiegelhalter, D.J. and Knill-Jones, R.P. (1984). Statistical and knowledge based approaches to clinical decision support systems, with an application in gastroenterology (with discussion). *Journal of the Royal Statistical Society (Series A)*, **147**, 35-77.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In: *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*,42–49.

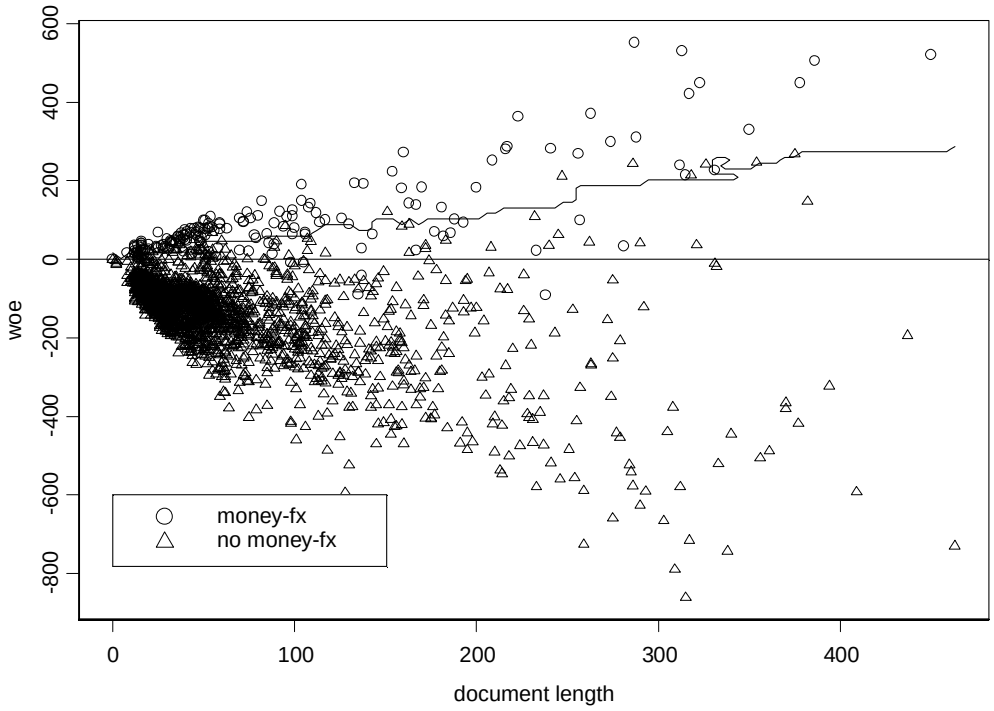

Figure 7: *k-nearest neighbour, with k=15, applied to the "money-fx" category of Reuters-21578. Each point corresponds to a test set document. The horizontal line is the decision boundary for the multinomial model. The NN boundary incorporates both the prediction of the multinomial model and document length.*

Table 7. Performance results for the K-NN model combining the multinomial model prediction with document length for the Reuters-21578 ten-category dataset.

Category	# of test errors multinomial model	# of test errors K-NN model	K	% improvement
earn	112	103	10	8%
corn	60	35	23	42%
ship	29	30	25	-3%
wheat	44	36	25	18%
interest	97	65	10	33%
trade	63	63	10	0%
crude	38	30	25	21%
grain	22	22	7	0%
moneyfx	88	61	15	31%
acq	83	103	15	-24%
Total	636	548		14%