Bayesian Methods for Frequent Terms in Text: Models of Contagion and the Δ^2 Statistic

Edoardo M. Airoldi[1], William W. Cohen[1] and Stephen E. Fienberg[2,1]
[1] School of Computer Science, and [2] Department of Statistics
Carnegie Mellon University, Pittsburgh, PA USA 15213

Abstract

Most statistical approaches to modeling text implicitly assume that informative words are rare. This assumption is usually appropriate for topical retrieval and classification tasks; however, in non-topical classification and soft-clustering problems where classes and latent variables relate to sentiment or author, informative words can be frequent. In this paper we present a comprehensive set of statistical learning tools which treat words with higher frequencies of occurrence in a sensible manner. We introduce probabilistic models of contagion for classification and soft-clustering based on the Poisson and Negative-Binomial distributions, which share with the Multinomial the desirable properties of simplicity and analytic tractability. We then introduce the Δ^2 statistic to select features and avoid over-fitting.

1 Introduction

Fifty years ago, Herbert Simon (1955) argued that, "as a text progresses, it creates a meaningful context within which words that have been used already are more likely to appear than others." A simple but important notion of "context" for a particular word w is how often w has appeared previously. Unfortunately, even this type of context is not captured by the usual multivariate Bernoulli and multinomial models; however, it is captured by using *contagious distributions*, such as the Poisson or Negative-Binomial, to model word frequencies in documents [36].

Contagious distributions for language modeling are not new. They were used by Mosteller and Wallace (1964, 1984) to model the frequency of function words (as indicators of personal writing styles) for the authors of the Federalist Papers; later, Church and Gale (1995) showed that Poisson mixtures often fit the observed word-frequency data better than standard Poissons [30, 12]. Nonetheless, most modern language models–even very sophisticated ones–are based on multinomial

models of frequency, largely, because of (1) the mathematical convenience of the very simple Dirichlet conjugate prior, (2) the tendency of more complex word-frequency models to overfit, and (3) the good performance of multinomial distribution on some classical problems, e.g., topic classification.

In this paper we derive practically useful contagious distributions that naturally fit into modern language models. To this end, we first introduce a new hierarchical Bayesian model, which naturally extends that of Mosteller and Wallace (1964, 1984) and makes use of re-parameterization of the Poisson and Negative-Binomial distributions in order to take advantage of the Dirichlet as a natural non-informative prior for the new parameters while maintaining analytical tractability. Second, in order to avoid overfitting, we present a novel statistic for selecting features according to their importance, the Δ^2 statistic, which helps avoid over-fitting by using sound assumptions about the particular contagious distribution for the occurrence of words along with False Discovery Rate arguments in order to control the overall probability of selecting irrelevant words. Third, we demonstrate that our distributions improve on the cross-validated classification accuracy achieved by multivariate Bernoulli and multinomial models.

Further, in order to boost the speed of feature selection strategies based upon the Δ^2 statistic, we derive its asymptotic distributions, with different degrees of precision, assuming both Poisson and Negative-Binomial word counts; this allows one to *compute* p-values for Δ^2, instead of *sampling*.

Most importantly, the analytic tractability of these contagious distributions enables fast inference mechanisms for more complex language models, such as latent Dirichlet allocation (Blei, Ng, and Jordan, 2003) or author-topic models (Erosheva, Fienberg, and Lafferty 2004), by simply plugging in these more realistic distributions and then updating the formulas—with some necessary approximations. For example, we use simple approximations to obtain a lower bound for variational inference in closed form for a soft-clustering version of our models [7, 16, 3].

1.1 Background and Related Work

The naïve Bayes approach is usually associated with multivariate Bernoulli and multinomial models. It consists of a simple application of Bayes' theorem to solve a classification problem. Its "naïvety" is in the fact that different words are considered to be pairwise independent, and the model is not specific as to the position of the words in the text [14]. Domingos and Pazzani (1997) give a complete characterization of naïve Bayes models, and study conditions for their optimality from a decision theoretic perspective [14].

Several works focused on the analysis of the limitations of naïve Bayes Bernoulli

and multinomial models [25, 34]; in particular the assumption that occurrences of a same word happen independently of one another has seriously been challenged and strong evidence, both theoretical and empirical, has been produced against it in extensive studies of textual data [41, 40, 26, 22, 17]. Ad hoc models have also been proposed to go beyond the independence assumption [11, 6, 12, 38]. Our work investigates a principled approach to relaxing the independent-occurence assumption.

Recently, a number of extensions have been proposed to the naïve Bayes approach, which prescribe hierarchical (graphical) models, with both observed and hidden variables, in order to describe, cluster, and classify documents [7, 8, 9, 18, 16]. In order to perform inferences in these models, the constants underlying the distributions of the variables in the top layer of the hierarchy have to be fixed. The empirical Bayes approach [10] is used here, often in combination with methods to approximate certain intractable (marginal) distributions, for example, MCMC [35], variational methods [21], and expectation propagation [27]. As the main ingredients of these extensions are the multinomial model and its conjugate Dirichlet prior, many of them could be adapted to handle other frequency models with tractable conjugate priors. We discuss to what extent our models can be combined with these more sophisticated language models in Section 3.3.

1.2 Notation

Our data consists of the number of times words appear in the texts. For each category ($c \in C$) we have a collection of D_c documents, and we represent each as a random vector $\boldsymbol{X}_{dc} := [X_1, X_2, ..., X_V]_{dc}$, that is, a bag of word counts, where the words indexed by $w = 1, ..., V$ belong to a pre-specified vocabulary. We denote the observed word counts, instances of the corresponding random numbers, with lowercase x's.

2 Contagious Distributions for Words and Context

Contagious distributions provide a better fit for frequent terms by relaxing the assumption of independence of successive occurrences of the same word across the text. Intuitively, *contagion* means that the occurrence of a word makes its subsequent occurrences more likely. We argue that this notion of contagion introduces a natural notion of context, i.e., the more a word is used, the more likely it is that it will occur again, thus defining the writing style of an author, or the prevalent sentiment in a sentence. In our experiments these distributions also led to lower cross-validated classification errors.

In this section, we reparametrize widely used contagious distributions such as the Poisson and the Negative-Binomial for word frequency. Our goal is to make their connection explicit, and introduce quantities that will help correcting the estimates of the relevant parameters by taking into account the different lengths of the texts.

2.1 The Poisson Model Revisited

For text data, using the Poisson model implicitly assumes that words or terms occur randomly and independently, but with some mean frequency. Stated differently, suppose the usage of word each word w is modeled as a random variable T denoting the expected "time till usage" of w. The Poisson distribution gives a particular form for the density of T, since one may interpret a Poisson distribution with parameter $\frac{\tau}{E(T)} =: \theta$ as the probability of w being used x times in a time interval of length τ. If X_{wd} encodes the number of times w appears in document d, then

$$Poisson\,(X_{wd} = x \,|\, \Theta_w = \theta) \;=\; \frac{\theta^x e^{-\theta}}{x!}, \quad x \geq 0.$$

We rewrite $\theta = \omega_d \mu$, where ω_d is the observable size of document d in thousands of words and μ is the rate of occurrence of a word per thousand words, so that

$$Poisson\,(X_{wd} = x \,|\, \omega_d,\, M_w = \mu) \;=\; \frac{(\omega_d \mu)^x\, e^{-\omega_d \mu}}{x!}, \quad x \geq 0.$$

The maximum likelihood estimator for M_w is $\frac{\sum_d X_{wd}}{\sum_d \omega_d}$, which takes into account the variable length of the texts, ω_d. Note that $\omega_d > 0$.

2.2 The Negative-Binomial Model Revisited

The Negative-Binomial distribution can be obtained from expansion of $(Q-P)^{-\kappa}$, where $Q = (1+P)$, $P > 0$, and and κ is positive real. Note that P need not be in $(0,1)$. If X_{wd} encodes the number of times w appears in document d,

$$Neg\text{-}Bin\,(X_{wd} = x \,|\, P_w = p,\, Q_w = q) \;=\; \binom{\kappa + x - 1}{\kappa - 1} \left(1 - \frac{p}{q}\right)^{\kappa} \left(\frac{p}{q}\right)^{x},$$

for any $x \geq 0$. In this parameterization, the mean equals κP_w and the variance equals $\kappa P_w(1 + P_w)$. The standard parameterization is obtained by introducing a single parameter $P'_w = \left(1 - \frac{P_w}{Q_w}\right) \in (0,1)$.

	Poisson Model		Negative-Binomial Model	
Pool of words	Reagan− (38 texts)	Reagan+ (75 texts)	Reagan− (38 texts)	Reagan+ (75 texts)
50 highest frequency words	12 (50)	3 (50)	31 (50)	49 (50)
21 semantic features	3 (21)	1 (21)	21 (21)	20 (21)
27 words by information gain	0 (7)	0 (8)	7 (7)	8 (8)

Table 1: Goodness of fit of Poisson and Negative-Binomial models for various pools of words. The pools are selected from positive (written by Reagan) and negative (written by Hannaford) examples of Reagan's radio addresses. Unbracketed counts are predicted number of words; in brackets we give the actual number of words. Predictions were made using p-values from a two-sample Kolmogorov Smirnov test. Source [1].

Intuitively, the Negative-Binomial distribution can be though of as a Poisson distribution with extra variability[1]. In order to make this connection explicit and obtain parameters easy to interpret when we specify our model for the word counts, we introduce the extra-variability parameter D_w, and we set $P_w = \omega_d D_w$ and $Q_w = (1 + P_w) = (1 + \omega_d D_w)$ to get

$$Neg\text{-}Bin\,(X_{wd} = x \,|\, \omega_d,\, M_w = \mu, D_w = \delta) \;=\; \frac{\Gamma(x+\kappa)}{x!\Gamma(\kappa)}\,(\omega_d\delta)^x(1+\omega_d\delta)^{-(x+\kappa)}$$

for any $x \geq 0$, where $\mu > 0$, $\delta > 0$, $\omega_d > 0$, and κ is a redundant parameter such that $\kappa \cdot \delta = \mu$. As in the Poisson case, ω_d is the observable size of document d in thousands of words and μ is the rate of occurrence of a word per thousand words. The parameter δ is the *non-Poissonness* parameter, that is, a parameter that controls how far the Negative-Binomial distribution is from its corresponding Poisson limit. More formally, as $D \to 0$, and $\kappa \to \infty$, the Negative-Binomial converges in distribution to its Poisson limit, for a fixed rate μ. The Negative-Binomial with parameters (ω_d, M, D) has mean equal to $\omega_d M$, and variance equal to $\omega_d M(1 + \omega_d D)$, that is, the same mean as its corresponding Poisson limit with an extra variability factor, $(1 + \omega_d D)$. Thus the extra-Poissonness parameter, D, allows us to model heavy tails, or extra variability, relative to the Poisson.

For example, in a prior studies of authorship attribution Airoldi et al. (2005) used this parameterization for the Negative-Binomial, in terms of (ω_d, M, D), and observed that estimates of δ were relatively stable across words and authors. For most words, $\hat{\delta} \in [0,\, 0.75]$. The Negative-Binomial model often captures the variability in observed word-frequency data better than the Poisson model. In Table 1 we present some data demonstrating that the flexibility gained by introducing two

[1] A thorough treatment of these two distributions is given in Johnson et al. (1992) [20].

Dataset	# of classes	Selection	Naïve Bayes	Poisson	Neg-Bin
Newsgroups	5	Info. Gain	*4.34%	**3.86%**	4.04%
Reuters	3	Info. Gain	*9.97%	**6.42%**	8.12%
Spam-Assassin	3	Info. Gain	6.03%	**1.79%**	2.42%
Fraud Detection	3	Info. Gain	**0.78%**	0.91%	1.09%

Table 2: The prediction errors on popular data sets. Errors refer to words selected by information gain; we used internal five-fold cross-validation to select how many. Whenever the best accuracy was obtained using all words there is no selection involved. The baseline is naïve Bayes with $p(x_{wd}|\theta_w) \propto (\theta_w)^{x_{wd}}$, where the estimates of θ_w were corrected for the different lengths of documents. Further, to provide a stronger baseline, we give the best accuracy between TFIDF-scaled (marked with a *) and unscaled naïve Bayes.

numeric parameters for each word allows one to better capture the way frequent and/or semantically important words are used [1].

Mosteller and Wallace (1964, 1984) gave non-Bayesian method-of-moment estimators for the Negative-Binomial parameters [30, 31]. Their estimators account for the different word-length of the documents, ω_d for $d = 1, \ldots, N$, and are "optimal" at the Poisson limit:

$$\left\{ \begin{array}{rcl} \hat{\mu}_w & = & m_w, \\ \hat{\delta}_w & = & d_w = \max\left\{0, \frac{v_w - m_w}{m_w r}\right\}, \end{array} \right. \qquad \begin{array}{rcl} m_w & = & \frac{\sum_j x_{wd}}{\sum_d \omega_d}, \\ v_w & = & \frac{1}{N-1} \sum_d \omega_d \left(\frac{x_{wd}}{\omega_d} - m_w\right)^2, \\ r & = & \frac{1}{N-1}\left(\sum_d \omega_d - \frac{\sum_d \omega_d^2}{\sum_d \omega_d}\right). \end{array}$$

2.3 Conditioning on the Word-Length of a Document

Word count models based on the Multinomial distribution fail to account for some variability in the length of the documents, in various ways. Sampling a document from our models, instead, guarantees the desired word length on average, rather than exactly, thus accounting for some variability. Specifically, the parameter ω is used to condition on the size of the documents, in the (ω, M) and (ω, M, D) parameterizations for the Poisson and Negative-Binomial, respectively.

Let us consider the Poisson case, where the rate $\Theta = \omega_d M$, and let us assume that our observations are number of times a certain word occurs in a set of documents with possibly different word-lengths. The new parameterization $\omega_d M$ breaks the rate into two parts: M, which is the rate of occurrence of the word under study, say, in a thousand consecutive words of text, or ℓ consecutive words in general, i.e., the rate as measured on a document with a reference length, in terms

of number of words; and ω_d, which is the length of a document expressed as a pure number, multiple of the word-length of the reference document, e.g., $\omega_d = 1.67$ for a document 1670 word long if the reference text is a thousand words, i.e., if $\ell = 1000$. This allows us to express the rate θ as the rate of occurrence of a word in a document ℓ word long, M, conditionally on the desired, or observed, length of the text, ω_d, expressed as a multiple of the word-length of the reference text. Similar considerations can be made in the Neagative-Binomial case.

3 Bayesian Models of Contagion for Words and Context

The models presented in sections 2.1 and 2.2 depend on a large number of parameters, one or two for each word in the vocabulary. In this section we introduce more parsimonious models by assuming that the population of parameters can be described compactly, in terms of distributions with simple functional forms that ultimately depend on a set of at most five underlying constants.

The Bayesian models we introduce here are hierarchical generative models. In this class of models the focus is on the hierarchy of probabilistic assumptions about the parameters and the data. Classification and soft-clustering tasks are then two sides of the same coin, differing mostly in the amount of labeled documents available for initializing the inference, that is, for training parameters or initializing latent categories [5]. The parameterizations in terms of (ω, M) and (ω, M, D) account for the natural variability in the length of the texts, but it is hard to posit a set of natural prior distributions for then, in cases where we have little or no information about the parameters. Below we introduce an novel idea for parametrizing contagious distributions and we discuss the properties it entails.

3.1 Sum/Ratio Parameterizations

The idea behind what we term *sum/ratio parameterizations* is very intuitive; we map a parameter vector in $\mathbb{R}^C$ to a new parameter vector in $\mathbb{R} \times [0,1]^{C-1}$. We do this by introducing a *sum* parameter, sum of the components in the original vector, and additional *ratio* parameters, obtained dividing components of the original vector by the sum parameter[2]. In the models of contagion we introduced above, for each word w:

$$\begin{array}{rclcrcl} \sigma_w & = & \sum_{c=1}^{C} \mu_{wc}, & & \tau_{wc} & = & \frac{\mu_{wc}}{\sum_{c=1}^{C} \mu_{wc}}, \\ \zeta & = & \log(1+\delta), & & & & \\ \xi_w & = & \sum_{c=1}^{C} \zeta_{wc}, & & \eta_{wc} & = & \frac{\zeta_{wc}}{\sum_{c=1}^{C} \zeta_{wc}}. \end{array}$$

[2]Note that the ratio parameters sum to one, so we only need $C-1$ of them.

for $c = 1, ..., C-1$, where the log transformation $\zeta = \log(1 + a\delta)$ serves the purpose of dampening the heavy tails of the distribution of $\hat{\delta}$ we explored in separate studies [1]. This transformation is one possibility among many; more generally we could use $\log(1 + a\delta)$, where a is depends on the document length. The parameters $(\tau_{w,C}, \eta_{w,C})$ are redundant. Further, we make the following assumptions:

(A1) the vectors $(\sigma_w, \tau_{w,1}, ..., \tau_{w,C-1}, \xi_w, \eta_{w,1}, ..., \eta_{w,C-1})$ are independent across words,

(A2) ξ_w, the vector $(\eta_{w,1}, ..., \eta_{w,C-1})$ and the vector $(\sigma_w, \tau_{w,1}, ..., \tau_{w,C-1})$ are independent from each other for each word w.

This class of parameterizations has the major advantage of separating the overall rate of occurrence from the way it's allocated to the various categories, independently of whether they are observed or latent. This simplifies, at times, inference calculations in complex language models. Further, it naturally[3] leads to simple analytic forms for the non-informative priors. In most cases is possible to derive an expression for the maximum likelihood estimators of the sum parameter, on which we can condition on in the inference process, both in the classification and soft-clustering versions of our models, in order to improve fit to the data.

3.1.1 Natural Non-Informative Priors: Full Specification

The parameter vector $\{\tau_{wc}, c \in C\}$ and $\{\eta_{wc}, c \in C\}$ both have support in $[0, 1]^C$, and we assume their values follow a symmetric Dirichlet distribution; this entails the same expected rate of occurrence, $\frac{1}{C}$, for the parameters $\{\mu_{wc},\ c \in C\}$. The residual parameter σ_w is greater than zero. We rely on prior studies in order to pick the functional form of the non-informative priors for this parameter [30, 12, 1]. In summary, for frequent terms we propose:

(A3) $(\tau_{w,1}, ..., \tau_{w,C-1})|\sigma_c$ is symmetric Dirichlet with parameter $(\beta_1 + \beta_2\, \sigma_w)$,

(A4) σ_w has an improper [4] constant density,

(A5) $(\eta_{w,1}, ..., \eta_{w,C-1})$ is symmetric Dirichlet with parameter (β_3),

[3] Briefly, we model $\boldsymbol{\tau}$ according to a Dirichlet distribution. We argue this is a *natural* choice since alternative sampling schemes are equivalent, exactly or asymptotically. For example, if we model the components of the vector of rates, $\boldsymbol{\mu}_w$, with independent Gamma distributions, the *sum* parameter is Gamma and the *ratio* parameter vector follows a Dirichlet distribution. See Kotz et al. (2000) for details and similar results [23].

[4] An improper constant density is constant density over an infinite support; it is *improper* as it does not integrate to 1.

(A6) ξ_w is Gamma with parameters $(\beta_5, \frac{\beta_4}{\beta_5})$.

When $\beta_2 > 0$ in (A3) the model encodes the notion that words that occur often are a-priori less likely to be useful in discriminating categories. The higher the overall occurrence of word w (i.e., the higher σ_w) the higher the Dirichlet parameter, and the lower the a-priori variability of the elements of τ_w.

3.2 Inference and Parameter Estimation

The Bayesian models of contagion for frequent words and context can be used for classification and soft-clustering tasks; here we present some calculations that relate to the classification task. To that extent, we assume there are C categories and we predict the category of a new document by evaluating the log-odds of each class c, i.e., $f(x_{new}, \theta, c) = \log \frac{p(c|x_{new},\theta_c)}{p(1|x_{new},\theta_1)}$. The log-odds are function of the parameters $\{\theta_{wc}\}$ that need be learned from training documents. Our models posit a hierarchy of probabilistic assumptions on the parameters, and Bayesian inference is required to learn their values. Note that we posit a separate model for each category, thus a new index appears, c, which runs from 1 to C.

3.2.1 MAP Estimation

We first evaluated the log-odds at the mode of the posterior distribution of the parameters given the data. We derived closed form expressions for the first and second derivatives of a quantity proportional to the posterior, for both our hierarchical Bayesian models. We then used Newton-Raphson to perform the maximization. Note that the maximization may fail for fairly rare words, as the matrix of second derivatives corresponding to the Negative-Binomial model may vanish.

3.2.2 MCMC

As an alternative we evaluated the log-odds at the mean of the posterior distribution of the parameters given the data; this is theoretically more sound, but computationally more expensive. For both models we used a Metropolis in Gibbs sampler with Gaussian proposals. Briefly, an outer loop iteratively samples one-dimensional full conditionals (Gibbs) and an inner loop is called upon to sample from those conditionals that are known up to a proportionality constant (Metropolis) [35]. In the Dirichlet-Poisson model, for example, the posterior distributions of $(\tau_{w,1}, ..., \tau_{w,C-1}, \sigma)|x_{wdc}$, entails the following full conditionals.

$$\log P(\tau_c|\tau_{(-c)}, \sigma, x_{wdc}) \;\propto\; -\sum_d w_{dc}\,\tau_c\,\sigma - \sum_d w_{dC}\,\tau_C\,\sigma + (\beta_1+\beta_2\,\sigma)\,\log(\tau_c\,\tau_C)$$

for all $c = 1, \ldots, C-1$, and $\tau_C = 1 - \sum_{c=1}^{C-1} \tau_c$.

$$\begin{aligned} \log P(\sigma|\tau_1, ..., \tau_{C-1}, x_{wdc}) \quad \propto \quad & -\sum_c \sum_d w_{dc}\, \tau_c\, \sigma + \sum_c \sum_d x_{dc}\, \log(\sum_d w_{dc}\, \tau_c\, \sigma) \\ & + (\beta_1 + \beta_2\, \sigma)\, \log(\prod_c \tau_c) + \log \left[\frac{\Gamma[C\,(\beta_1 + \beta_2\, \sigma)]}{\Gamma[(\beta_1 + \beta_2\, \sigma)]^C} \right]. \end{aligned}$$

Similar derivations give the set of full conditionals to perform inference in the Dirichlet-Negative-Binomial model.

3.2.3 Full Bayes

The sets of constants (β_1, β_2) and $(\beta_1, \ldots, \beta_5)$, underlying Poisson and Negative-Binomial models respectively, need to be fixed. Following a fully Bayesian approach, we did not estimate the underlying constants using our data. Instead, we relied on results from a prior study, on a separate data set, and selected 20 sets of constants that lead to reasonable tails for the priors [1]. We then evaluated the cross-validated error rates corresponding to each set of underlying constants, $\boldsymbol{\beta}$, in order to get a sense for how sensitive our predictions may be. The errors reported in Table 3 were obtained with $\beta = (2, 1)$

3.3 A Note on Soft-Clustering

The reparameterization in section 3.1 partially maps the parameters of Poisson and Negative-Binomial models to the simplex, thus allowing one to combine into a hierarchy of probabilistic assumptions the Dirichlet density, a natural non-informative prior for frequent terms, with powerful contagious distributions, which introduce an intutive notion of context. This is not only of interest for the understanding of the mathematical connections of our models with, for example, the latent Dirichlet allocation of Blei et al. (2003), but improves their analytical tractability as well. Specifically, separating the sum of the rates, Σ, from its split across classes allows us to estimate Σ directly from the data and condition other estimates on it—in classification—and allows us to carry out variational inference conditionally on it—in soft-clustering—leading to some closed formula variational EM updates.

Elsewhere, we posit fully generative models for soft-clustering that share the same hierarchy of probabilistic assumptions about the parameters as that of the models for classification presented here [3]. The focus of the Bayesian paradigm on the set of probabilistic assumptions enables us to fit practically useful contagious distributions into complex language models. Briefly, the soft-clustering version of our models allow for a variational lower bound in closed form. We devise M-step updates (in a variational EM algorithm) conditionally on parameters that can be

reliably estimated from the data, i.e., Σ, as hinted above. Ultimately, our models extract a richer set of categories than competing latent allocation models.

4 Feature Selection with Δ^2

Using more expressive classes of distributions to represent word frequency can cause overfitting. Here we propose a distribution-based feature selection strategy, which tests for feature relevance according to a specific word-frequency model, e.g., Poisson. The test also produces a well-defined p-value, so that feature selection over many features can be performed in a principled way by using standard methods for combining multiple statistical tests, such as the False Discovery Rate [37].

Let X_{wdc} denote the number of times the wth word in the dictionary appears in the dth document belonging to the cth class, and let $\{x_{wd1} : d = 1, ..., D_1\}$ and $\{x_{wd2} : d = 1, ..., D_2\}$ denote the observed counts in the texts. We define Δ^2 for word w as follows.

$$\Delta_w^2 \quad = \quad \frac{\left(\sum_{d=1}^{D_1} x_{wd1} - \sum_{d=1}^{D_2} x_{wd2}\right)^2}{\sum_{d=1}^{D_1} x_{wd1} + \sum_{d=1}^{D_2} x_{wd2}}. \tag{1}$$

We use the Δ^2 statistic to test the null hypothesis: "word w is irrelevant to the extent of discriminating between documents in categories one and two." Specifically, we assume a contagious frequency model for word w, $P(x_w|\theta_w)$, and test whether $\theta_{w1} = \theta_{w2}$. The p-value will provide a probabilistic assessments on whether word w occurred in the two categories *differently enough* to discard the hypothesis that such differences are the outcome of pure chance, i.e., that word w is irrelevant for discrimination.

In order to perform the test of irrelevance for a word, we (i) compute the observed value of the statistic, Δ_{obs}^2, (ii) use the estimators in sections 2.1 and 2.2 to estimate the parameters[5] underlying the word-frequency model, $\hat{\theta}_w$, and (iii) compute the p-value, i.e., we evaluate the following integral: $P(\Delta_w^2 > \Delta_{obs}^2 | \theta_{w1} = \theta_{w2} = \hat{\theta}_w)$.

The naïve solution is that of sampling the distribution of $P(\Delta_w^2|\hat{\theta}_w)$ in step (iii). This may be expensive, especially for rare words. Alternatively, we approximate analytically the distribution of Δ^2 under the Poisson and Negative-Binomial

[5]We estimate one set of parameters corresponding to the collection of documents. It is possible to use document labels to weight the parameter estimates corresponding to different classes.

models and compute the p-value using the approximate density. Tedious calculations lead to the following normal approximations, corresponding to expansions at different orders:

$$\begin{aligned} \Delta^2 &\sim N\left(1, 2 + \tfrac{1}{2\mu\omega}\right) && \text{2-nd} \\ \Delta^2 &\sim N\left(1 - \tfrac{1}{2\mu\omega}, \tfrac{1+\mu\omega(25+2\mu\omega(11+8\mu\omega))}{8\mu^3\omega^3}\right) && \text{3-rd} \end{aligned}$$

for the case of $X \sim Pois(\omega\mu)$ and $Y \sim Pois(\omega\mu)$, X and Y independent.

Similar approximations for the Negative-Binomial are available. We extend the Δ^2 statistics for word w to multiple categories, e.g., by iteratively computing the p-values for a class versus all the others and keeping the smallest p-value.

5 Experiments

We compared the cross-validated accuracies of naïve Poisson and Negative-Binomial and that of the Bayesian Dirichlet-Poisson to the baselines (multinomial and multivariate Bernoulli) on eleven data sets.

5.1 Data Sets

In the *Newsgroups* problem we want to classify newsgroups' posts according to their topic [29]. In the *Reuters* problem we abandon the typical breakdown into very narrow categories, a scenario where low frequency keywords drive the classification, and create our own high level categories—*Money*, *Crops*, and *Natural Resources*—in order for medium frequency, weakly topical words to drive the classification [24]. In the *Fraud detection* problem we want to find messages that contain fraudulent intent [4]. In the three *Opinion Extraction* problems we want to categorize the overall opinion expressed in online news articles (courtesy of Infonic.com) as being *Positive*, *Neutral* or *Negative* [2]. In the *Spam* problem we want to classify emails as *Easy Ham*, *Hard Ham*, and *Spam*, where ham is the term that indicates legitimate emails[6]. In the *Web-Master* problem the task is to classify web site update requests as *Add*, *Change*, or *Delete* [13]. In the *Reagan's Data* the problem is that of attributing authorship to text of Ronald Reagan's radio addresses broadcasted over the years 1975-1979 [1]. In the *Movie Reviews* problem we want to associate a positive or negative sentiment with each movie review [32]. In the *Medical Data:* the task is to classify whether a patient has a certain disease given outcomes of different tests.

[6]The SpamAssassin corpus is available online at http://www.spamassassin.org/.

Dataset	Class	Selection	Naïve Bayes	Poisson	Neg-Bin	Dir-Pois
Reagan's Data	2	IG	8.97%	8.29%	7.72%	7.50%
		Δ^2	8.50%	7.81%	6.95%	**6.50%**
Movie Reviews	2	IG	30.64%	28.71%	28.86%	25.75%
		Δ^2	28.07%	26.50%	26.14%	**21.93%**
Medical Data	2	IG	13.34%	7.67%	7.52%	6.13%
		Δ^2	11.28%	7.01%	6.13%	**5.95%**
Opinions: Finance	3	IG	*35.66%	29.83%	31.00%	**27.61%**
		Δ^2	*35.66%	29.83%	31.00%	**27.61%**
Opinions: Mixed	3	IG	*29.00%	28.17%	28.33%	**24.83%**
		Δ^2	*29.00%	28.17%	28.33%	**24.83%**
Opinions: M & A	3	IG	*30.33%	26.33%	27.33%	**24.83%**
		Δ^2		24.33%		
Web-Master	3	IG	*11.17%	9.97%	8.93%	7.56%
		Δ^2				**7.46%**

Table 3: The prediction errors refer to words selected by information gain; we used internal five-fold cross-validation to select how many. Whenever the best accuracy was obtained using all words there is no selection involved. The baseline is naïve Bayes with $p(x_{wd}|\theta_w) \propto (\theta_w)^{x_{wd}}$. The estimates of θ_w were corrected for the different lengths of documents. Further, naïve Bayes is sometimes improved by scaling the counts with TFIDF weights [33]. To provide a stronger baseline, we give the best accuracy between TFIDF-scaled and unscaled naïve Bayes. (Accuracies for scaled naïve Bayes are marked with a *). The errors in the central columns refer to our parameterizations for Poisson and Negative-Binomial models, as given in Sections 2.1 and 2.2. The errors for the DiP model were obtained with $\beta = (2, 1)$.

5.2 Results

To allow for a fair comparison we corrected the parameter estimates for the baseline models to account for different length of documents and transformed the word counts with TFIDF. In fact, naïve Bayes is sometimes improved by scaling the counts with TFIDF weights [33]. The tables report the best accuracy between TFIDF-scaled and unscaled naïve Bayes[7]. We compared the accuracies on sets of words selected by information gain and Δ^2 for different values of α and different number of words to make results comparable.

The experiments suggest that the Poisson and Negative-Binomial models fit textual data better, and lead to log-odds consistently less extreme than multinomial. This need not lead to better accuracy, as in the case of the email Fraud data set. The statistic Δ^2 favors words that occur often, and leads to higher accuracies than

[7] Accuracies for scaled naïve Bayes are marked with a *.

information gain on our classification problems. An advantage of choosing words that occur often is that a small set of them may be sufficient to represent the whole collection of documents, promoting insights into the problem and interpretability of the results.

6 Conclusions

We have described a simple, principled extension to the widely-used multinomial model for text. The extension allows better modeling of frequent words by replacing the widely-used multinomial distribution with simple "contagious" distributions, that is, by relaxing the assumption of independence of different occurrences of the same word across the text. Using eleven data sets, we show that the model generally leads to better classification accuracy, sometimes to substantially better. The experiments presented here have been with simple "naïve and hierachical Bayes" models for classification; however, an important advantage of the proposed extension is that is easy to combine with more complex models of text, e.g., mixtures and hierarchical mixture models.

In the current paper we also developed tractable non-informative priors for the models, for use in settings for which a fully Bayesian or empirical Bayesian approach is appropriate. Elsewhere, we have successfully exploited the proposed hierarchy of probabilistic assumptions on the parameters to build soft-clustering counterparts of our models [3].

References

[1] E. M. Airoldi, A. G. Anderson, S. E. Fienberg, and K. K. Skinner. Who wrote Ronald Reagan radio addresses? *Bayesian Analysis*, 2005. To appear.

[2] E. M. Airoldi, X. Bai, and R. Padman. Sentiment extraction from unstructured texts: Markov blankets and meta-heuristic search. In *Lecture Notes in Computer Science*. Springer-Verlag, 2005. To appear.

[3] E. M. Airoldi and S. E. Fienberg. Serial analysis of gene expression data with the Dirichlet-Poisson model. Manuscript, September 2005.

[4] E. M. Airoldi and B. Malin. Data mining challenges for electronic safety: The case of fraudulent intent detection in e-mails. In *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining*, pages 57–66. IEEE Computer Society, 2004.

[5] E.M. Airoldi. Hierarchical mixture models: Theory and practice. Manuscript, September 2005.

[6] D. Beeferman, A. Berger, and J. Lafferty. A model of lexical attraction and repulsion. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 373–380, 1997.

[7] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *Uncertainty in Artificial Intelligence*, 2004.

[9] J. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[10] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, second edition, 2000.

[11] K. Church. One term or two? In E. Fox, P. Ingwersen, and R. Fidel, editors, *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 1995.

[12] K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering*, 1:163–190, 1995.

[13] W. W. Cohen, E. Minkov, and A. Tomasic. Learning to understand web site update requests. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005.

[14] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[15] B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators-part ii: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139, 1972.

[16] E. A. Erosheva, S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892, 2004.

[17] S. Eyheramendy, D. Lewis, and D. Madigan. On the naive bayes model for text categorization. In *Proceedings of the Workshop on Artificial Intelligence and Statistics*, 2003.

[18] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[19] M. Johnson, S. P. Khudanpur, M. Ostendorf, and R. Rosenfeld, editors. *Mathematical Foundations of Speech and Language Processing*. Springer, 2004.

[20] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. John Wiley, 1992.

[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[22] S. Katz. Distribution of content words and phrases in text and language modeling. *Natural Language Engineering*, 2(1):15–59, 1996.

[23] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions*, volume 1: Models and Applications. John Wiley, 2000.

[24] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[25] A. McCallum and K. Nigam. A comparison of event models for naïve bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, 1998.

[26] G. A. Miller, E. B. Newman, and E. A. Friedman. Length-frequency statistics for written English. *Information and Control*, 1:370–389, 1958.

[27] T. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence (UAI)*, pages 362–369, 2001.

[28] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence*, 2002.

[29] Tom M. Mitchel. *Machine Learning*. McGraw-Hill, 1997.

[30] F. Mosteller and D.L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

[31] F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. Springer-Verlag, 1984.

[32] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing*, 2002.

[33] J. Rennie and T. Jaakkola. Using term informativeness for named entity detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[34] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naïve bayes text classifiers. In *International Conference on Machine Learning*, 2003.

[35] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY, corrected second edition, 2005.

[36] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[37] J. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205, 2004.

[38] J. Teevan and D. R. Karger. Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model. In *Proceedings of the 26th Annual ACM SIGIR International Conference on Research and Development in Informaion Retrieval*, 2003.

[39] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22th Annual ACM SIGIR International Conference on Research and Development in Informaion Retrieval*, 1999.

[40] U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.

[41] G. K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, 1932.