What's Worthy of Comment? Content and Comment Volume in Political Blogs*

Tae Yano and **Noah A. Smith**
{taey,nasmith}@cs.cmu.edu
School of Computer Science
Carnegie Mellon University

Abstract

In this paper we aim to model the relationship between the text of a political blog post and the comment volume—that is, the total amount of response—that a post will receive. We seek to accurately identify which posts will attract a high-volume response, and also to gain insight about the community of readers and their interests. We design and evaluate variations on a latent-variable topic model that links text to comment volume.

Introduction

What makes a blog post noteworthy? One measure of the popularity or breadth of interest of a blog post is the extent to which readers of the blog are inspired to leave *comments* on the post. In this paper, we study the relationship between the text contents of a blog post and the volume of response it will receive from blog readers. Modeling this relationship has the potential to reveal the interests of a blog's readership community to its authors, readers, advertisers, and scientists studying the blogosphere, but it may also be useful in improving technologies for blog search, recommendation, summarization, and so on.

There are many ways to define "popularity" in blogging. In this study, we focus exclusively on the aggregate volume of comments. Commenting is an important activity in the political blogosphere, giving a blog site the potential to become a discussion forum. For a given blog post, we treat comment volume as a target output variable, and use generative probabilistic models to learn from past data the relationship between a blog post's text contents and its comment volume. While many clues might be useful in predicting comment volume (e.g., the post's author, the time the post appears, the length of the post, etc.) here we focus solely on the text contents of the post.

We first describe the data and experimental framework, including a simple baseline. We then explore how latent-variable topic models can be used to make better predictions about comment volume. These models reveal that part of the variation in comment volume can be explained by the topic of the blog post, and elucidate the relative degrees to which readers find each topic comment-worthy.

Predicting Comment Volume

Our goal is to predict some measure of the volume of comments on a new blog post.[1] Volume might be measured as the number of words in the comment section, the number of comments, the number of distinct users who leave comments, or a variety of other ways. Any of these can be affected by uninteresting factors—the time of day the post appears, a side conversation, a surge in spammer activity—but these quantities are easily measured.

In research on blog data, comments are often ignored, and it is easy to see why: comments are very noisy, full of non-standard grammar and spelling, usually unedited, often cryptic and uninformative, at least to those outside the blog's community. A few studies have focused on information in comments. Mishe and Glance (2006) showed the value of comments in characterizing the social repercussions of a post, including popularity and controversy. Their large-scale user study correlated popularity and comment activity. Yano et al. (2009) sought to predict *which* members of blog's community would leave comments, and in some cases used the text contents of the comments themselves to discover topics related to both words and user comment behavior. This work is similar, but we seek to predict the *aggregate* behavior of the blog post's readers: given a new blog post, how much will the community comment on it?

Tasks and Dataset

We first consider predictions of the volume measured in word tokens, and in comments. The task is to predict merely whether a blog post will have higher volume than the average seen in training data. More fine-grained predictions are possible as well (e.g., predicting the absolute number of words in the comments). Other future possibilities include the prediction of the rate of positive or negative polarity words, or words referring to a certain named entity.

Our experiments use data from two blogs, Matthew Yglesias (http://www.matthewyglesias.theatlantic.com, denoted MY) and RedState (http://www.redstate.com, denoted RS); these

*The authors acknowledge research support from HP Labs and helpful comments from the reviewers and Jacob Eisenstein.

[1] Most commenting ends within a few hours or days of a post. The posts are downloaded at least six days after the original post. It is possible, though unlikely, that comments left long after the original post would make a difference in our results.

corpora are a subset of those used by Yano et al. (2009). The data are drawn from 2007–2008; in each case we use the same temporal training-test splits as Yano et al. (i.e., the test posts come strictly later than the training posts). All posts are represented as text only (images, hyperlinks, and other non-text elements were ignored). Words occuring two or fewer times in the training data and stop words were removed. No stemming was performed. Posts with fewer than five words were discarded.

The mean volume is approximately 1424 words (35 comments) for MY and 819 words (29 comments) for RS. The distribution is skewed, with roughly one third of the posts having below-average volume. The MY data shows a strange effect: the test set has a much greater rate of high-volume posts (66%) compared to the training data (35%), potentially making the prediction task much harder.

Bag of Words Naïve Bayes Model

Naïve Bayes is a simple model for classification that can be used for the binary prediction task. Let $\bar{v}$ be the mean value of the volume variable we seek to predict, calculated on the training data. Let V be the (unknown volume) for a blog post that is represented as a word sequence $\boldsymbol{w} = \langle w_1, \ldots, w_N \rangle$.

$$\begin{aligned} p(V > \bar{v}, \boldsymbol{w}) &= p(V > \bar{v}) \times \prod_{i=1}^{N} p(w_i \mid V > \bar{v}) \\ p(V < \bar{v}, \boldsymbol{w}) &= p(V < \bar{v}) \times \prod_{i=1}^{N} p(w_i \mid V < \bar{v}) \end{aligned}$$

The generative model assumes that, first, the class ("high volume" or "low volume") is chosen according to a binomial distribution, then the words are generated IID conditioned on the class. Maximum likelihood estimates for the parameters are obtained straightforwardly from training data. Unobserved words are ignored at test time.

The results are shown in Table 1 for both blogs. We report precision and recall for the "high volume" class, using both word counts and comment counts to measure volume. The Naïve Bayes model tends to err on the side of precision. Note that comment volume on RS is harder to predict from words.

Beyond the performance of the predictor on this task, we may ask what the model tells us about the blog and its readers. The Naïve Bayes model does not provide much insight. Ranked by likelihood ratio, $p(w \mid V > \bar{v})/p(w \mid V < \bar{v})$, the strongest features for "high word volume" from MY are *alleged*, *instability*, *current*, *crumbling*, *canaries*, *imaginations*, *craft*, *cars*, *imagine*, *funnier*.

Regression

Regression is another approach suitable for predicting numerical values. We tested linear regression with elastic net regularization (Friedman, Hastie, and Tibshirani 2010).[2] This approach permits easy tuning of the regularization constant. We trained 1,000 regression models (at different regularization settings) with the same word features as above. We report here the binary prediction performance of the best model, selected using a 10% held-out set.

		# words					# comments				
		prec.		rec.		F_1	prec.		rec.		F_1
MY	NB	72.5		41.7		52.9	42.6		38.8		40.6
	Reg.	81.5		44.1		57.2	60.8		55.2		57.8
	T-Pois.	70.1	(±1.8)	63.2	(±2.5)	66.4	41.3	(±2.1)	53.1	(±3.5)	46.4
	k=30	71.8	(±2.0)	60.1	(±3.4)	65.4	45.3	(±2.1)	54.2	(±5.3)	49.3
	k=40	71.0	(±1.9)	63.4	(±2.7)	66.9	44.0	(±2.1)	58.8	(±3.3)	50.3
	T-NBin.	69.7	(±2.3)	62.5	(±2.5)	65.9	38.4	(±2.2)	45.7	(±3.3)	41.7
	C-LDA	70.2	(±2.3)	68.8	(±2.5)	69.4	37.2	(±1.5)	50.4	(±3.3)	42.8
RS	NB	64.1		25.7		36.6	37.8		34.1		35.0
	Reg.	52.0		26.8		35.5	20.5		19.5		20.0
	T-Pois.	52.4	(±2.8)	33.5	(±2.0)	40.8	25.4	(±2.6)	27.9	(±2.9)	26.7

Table 1: Experiments: precision and recall for "high volume" posts. NB= Naïve Bayes, Reg. = regression, T-Pois. = Topic-Poisson, T-NBin. = Topic-Negative Binomial, C-LDA = CommentLDA. Topic models are "ave. (±s.d.)" across 10 runs.

Topic Models for Prediction

We seek a model that will not only perform well on the predictive task, but actually provide insight as to *why* some blog posts inspire people to leave comments in greater volume. A natural generalization is to consider how the *topic* (or topics) of a post influence commenting behavior.

Topic-Poisson Model

Latent Dirichlet allocation (Blei, Ng, and Jordan 2003) is a generative probabilistic model of text that has been widely used in recent research. LDA goes beyond bag-of-words models by positing a hidden topic distribution, drawn distinctly for each document, that defines a document-specific mixture of bag-of-words models. The topics are unknown in advance, and are defined only by their separate word distributions, which are discovered through probabilistic inference from data. Like many other techniques that infer topics as measures over the vocabulary, LDA often finds very intuitive topics. A key advantage of LDA is that it can be extended to model other variables as well (Steyvers and Griffiths 2007, *inter alia*).

Our generative model, which we call the Topic-Poisson model, proceeds as follows. The number of topics, K, is fixed in advance.

1. (Once for the text collection:) For k from 1 to K, choose a distribution $\boldsymbol{\phi}_k$ over words according to a symmetric Dirichlet distribution parameterized by β.
2. For each blog post d from 1 to D:
   (a) Choose a distribution $\boldsymbol{\theta}_d$ over topics according to a symmetric Dirichlet distribution parameterized by α.
   (b) For i from 1 to N_d (the length of the dth post):
      i. Choose a topic $z_{d,i}$ from the distribution $\boldsymbol{\theta}_d$.
      ii. Choose a word $w_{d,i}$ from $\boldsymbol{\phi}_{z_{d,i}}$.
   (c) For k from 1 to K, let

$$m_{d,k} \leftarrow \frac{\text{freq}(k; \boldsymbol{z}_d) + \alpha_k}{\sum_{k'=1}^{K} \text{freq}(k'; \boldsymbol{z}_d) + \alpha_{k'}} \qquad (1)$$

[2] http://cran.r-project.org/web/packages/glmnet/

Then choose a comment volume v_d from the mixture distribution $\sum_{k=1}^{K} m_{d,k} p(\cdot; \lambda_k)$.

Note that the model is identical to LDA until step 2c, where we define document-specific mixture coefficients and generate the volume from a mixture of distributions over volume. Here, volume is always an integer value, and the mixture components are Poissons. This model is essentially a type of "supervised" or "annotated" LDA (Blei and McAuliffe 2008; Ramage et al. 2009), where a side target variable is generated based on topics and therefore influences what topics are learned.

We use the training portion of a blog's posts to estimate the model parameters to maximize likelihood. For a new (test) blog post, we infer its topic distribution $\boldsymbol{\theta}$, then compute the expected value for v, the comment volume. As with the Naïve Bayes model, we can calculate the accuracy of the "higher or lower than average" prediction, but this model is more powerful as it gives a distribution over values for v, permitting more fine-grained prediction and analysis.

Parameter Estimation

In this study we seek a maximum *a posteriori* estimate of $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$, marginalizing out $\boldsymbol{\theta}$, each word's topic z, and fixing $\alpha = 0.1$ and $\beta = 0.1$. During training, the words and volumes are, of course, observed. We use collapsed Gibbs sampling for inference (Heinrich 2008; Griffiths and Steyvers 2004). The only detail to point out is that each topic depends in part on the volume, so that the Gibbs sampler draws topic $z_{d,i}$ for word $w_{d,i}$ according to:

$$p(z \mid \boldsymbol{z}_{d,-i}, \boldsymbol{w}_d, v_d, \alpha, \beta, \boldsymbol{\theta}_d, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \frac{\text{freq}(z; \boldsymbol{z}_{d,-i}) + \alpha}{\sum_{k=1}^{K} \text{freq}(k; \boldsymbol{z}_{d,-i}) + \alpha}$$
$$\times \frac{\text{freq}(w_{d,i}, z; \boldsymbol{z}_{d,-i}) + \beta}{\sum_{w'} \text{freq}(w', z; \boldsymbol{z}_{d,-i}) + \beta} \times \left(\sum_{k=1}^{K} m_{d,k} p(v_d | \lambda_k) \right)$$

where $m_{d,k}$ comes from Eq. 1, with $z_i = z$. Note that the sampling distribution depends on the mixture coeficients, which are calculated directly from the document's topics $\boldsymbol{z}_d$ in the current sample according to Eq. 1. We use a mixture of Poissons, so that for all $v \in \mathbb{N}$,

$$p(v \mid \lambda_k) = e^{-\lambda_k} \lambda_k^v \big/ v! \qquad (2)$$

To estimate the λ_k, we embed the Gibbs sampler in a stochastic EM algorithm (Casella and Robert 2004) that reestimates the λ_k after resampling the $\boldsymbol{z}$ for each document in turn, according to the maximum likelihood formula:

$$\lambda_k \leftarrow \left(\sum_{d=1}^{D} \theta_{d,k} v_d \right) \Big/ \left(\sum_{d=1}^{D} \theta_{d,k} \right) \qquad (3)$$

Experimental Results

We first consider predictions of the topic model with $K = 15$, $\alpha = 0.1$, and $\beta = 0.1$. Table 1 shows all precision-recall results discussed here, for MY and RS, respectively.

On the MY data, the Topic-Poisson model improves recall substantially over Naïve-Bayes, on both measures, with a slight loss in precision. Its precision lags behind the regression model, gaining in recall on word volume prediction but not on comment volume prediction.

The effect is similar on RS data when predicting *word* volume, but the loss in precision is much greater, and the model is ineffective for *comment* volume. The regression model is much less effective on the RS data set, falling behind Naïve-Bayes on both tasks.

Topics

An attractive property of topic models is that they discover topics. In Topic-Poisson, we can characterize each topic by λ_k (the mean for its Poisson distribution over volume values) and, more traditionally, by the words associated with each topic. Table 2 shows the topics discovered in MY by the word-volume Topic-Poisson model. Topics are ranked by λ_k; words are selected as by Blei and Lafferty (in press).

The most comment-worthy topics on the liberal blog MY appear to deal with the gender/race issue and the Democratic presidential primary race. On both blogs, discussion of internal races for party nominations is clearly the most likely to incite readers of these partisan blogs to comment.

Some clear issues arise as topics. On MY, the Middle East and energy are the seventh and eighth topics; healthcare is slightly below the overall average. RS (not shown) rates religion very high (fourth topic), with the economy just above average and Iraq/Afghanistan well below.

Note that the least comment-worthy topics on MY have to do with sports and reading material, which interest Matthew Yglesias but perhaps not his readers.

Table 2 also shows the binary accuracy on posts associated with each topic. We assign a post to each topic k that has $\theta_{d,k} \geq 0.25$ (a post can go to zero, one, or more topics), and measure binary prediction accuracy within the topic. These accuracies are based mostly on very small numbers of posts, so our analysis is tentative. On MY, the most comment-worthy topics are also the ones that our model is most accurate at classifying. Part of the variation across topics may be due to temporal changes in topics and reader interest between training and (later) testing periods.

Model Variations

We test variations of the Topic-Poisson model on MY data.

More topics We tested the Topic-Poisson with more topics. These are shown as "$K = 30$" and "$K = 40$" in Table 1. More topics had a negligible effect on word-volume task but improved the comment volume task substantially. On inspection, the topics discovered by these models were more difficult to understand.

Negative binomial We tested the model with a mixture of negative binomials instead of Poissons:

$$p(v; \rho_k, r_k) = \binom{v + r_k - 1}{r_k - 1} \rho_k^{r_k} (1 - \rho_k^v) \qquad (4)$$

In the M step of the training algorithm, we estimate each ρ_k and r_k using moment matching (Casella and Berger 2001). The experimental results of this change (Table 1) were negative. While the negative binomial offers more expressive power and degrees of freedom than the Poisson, it tends toward the Poisson as $\rho \to 0$; estimated ρ_k values were, indeed, close to 0.

λ_k	topic words	# posts	accuracy
1873	women black white men people liberal civil working woman rights	7	(100)
1730	obama clinton campaign hillary barack president presidential really senator democratic	13	(77)
1643	think people policy really way just good political kind going	74	(72)
1561	conservative party political democrats democratic republican republicans immigration gop right	12	(50)
1521	people city school college photo creative states license good time	19	(58)
1484	romney huckabee giuliani mitt mike rudy muslim church really republican	3	(33)
1478	iran world nuclear israel united states foreign war international iranian	16	(69)
1452	carbon oil trade emissions change climate energy human global world	6	(33)
1425	obama clinton win campaign mccain hillary primary voters vote race	22	(64)
1352	health economic plan care tax spending economy money people insurance	22	(55)
1263	iraq war military government american iraq troops forces security years	24	(58)
1246	administration bush congress torture law intelligence legal president cia government	5	(20)
1215	mccain john bush president campaign policy know george press man	20	(60)
1025	team game season defense good trade play player better best	8	(38)
1007	book times news read article post blog know media good	23	(43)
	Overall:	183	(58)

Table 2: MY Topic-Poisson model: Poisson parameter estimate and top words for each topic. See text for explanation.

User identities and comment content Following Erosheva et al. (2004) and Yano et al. (2009), we further extended LDA to model the user identities and comment words in addition to the post words. The model thus prefers the topics which explain the comment volume, as well as those additional observations. We tested the CommentLDA model of Yano et al. 2009, with "counting by comments" (see that paper for details), and achieved substantial gains in word volume prediction with similar recall to other models. This approach was harmful to comment volume prediction.

Related Work

We have discussed the most closely related work above. A few other topic model papers are relevant. Recently, various extensions to LDA with side variables are proposed. Mimno et al. (2008) and Zhu et al. (2009), also proposed LDA models with labeled (or annotated) variables in the generative story. The most widely used may be supervised LDA (Blei and McAuliffe 2008), where a generalized linear model is incorporated. Our model is different from SLDA as our volume variable is sampled from a mixture distribution. Also worth noting are Quinn et al. (2006) and Grimmer (forthcoming), which are among the first to use LDA for text analysis for social science inquiries.

Conclusion

We have considered the task of predicting which blog posts will have a greater-than-average volume of response from readers, measured in comments or words in comments. Our findings show that modeling topics can improve recall when predicting "high volume" posts, and revealed interesting patterns in two different blogging communities' notions of comment-worthiness.

References

Blei, D., and Lafferty, J. In press. Topic models. In Srivastava, A., and Sahami, M., eds., *Text Mining: Theory and Applications*. Taylor and Francis.

Blei, D., and McAuliffe, J. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems 20*.

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

Casella, G., and Berger, R. L. 2001. *Statistical Inference*. Duxbury Press, 2nd edition.

Casella, G., and Robert, C. 2004. *Monte Carlo Statistical Methods*. Springer, 2nd edition.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed membership models of scientific publications. *Proc. of the National Academy of Sciences* 5220–5227.

Friedman, J. H.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proc. of the National Academy of Sciences* 101 Suppl. 1:5228–5235.

Grimmer, J. Forthcoming. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*.

Heinrich, G. 2008. Parameter estimation for text analysis. Technical report, University of Leipzig.

Mimno, D., and McCallum, A. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proc. of UAI*.

Mishne, G., and Glance, N. 2006. Leave a reply: An analysis of weblog comments. In *Proc. of Workshop on the Weblogging Ecosystem*.

Quinn, K. M.; Monroe, B. L.; Colaresi, M.; Crespin, M. H.; and Radev, D. R. 2006. An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. Senate. Midwest Political Science Association Meeting.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of EMNLP*.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. In Landauer, T.; McNamara, D.; Dennis, S.; and Kintsch, W., eds., *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.

Yano, T.; Cohen, W. W.; and Smith, N. A. 2009. Predicting response to political blog posts with topic models. In *Proc. of NAACL-HLT*.

Zhu, J.; Amr, A.; and Xing, E. P. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *Proc. of ICML*.