Modeling Microblogs using Topic Models

Kriti Puniyani
School of Computer Science
Carnegie Mellon University
kpuniyan@cs.cmu.edu

ABSTRACT

As the popularity of micro-blogging increases, managing friends and followers and their tweets is becoming increasingly complex. In this project, we explore the usage of topic models in understanding both text and links in micro-blogs. On a data set of 21306 users, we find that LDA can find good topics that seem to capture meaningful topics of discussion in twitter. We also find that knowing whether two users tweet about similar topics is more useful in predicting links between them than standard network analysis metrics that ignore text.

1. INTRODUCTION

Language use is overlaid on a network of social connections, which exerts an influence on both the topics of discussion and the ways that these topics can be expressed [3]. In the past, efforts to understand this relationship were stymied by a lack of data, but social media offers exciting new opportunities. By combining large linguistic corpora with explicit representations of social network structures, social media provides a new window into the interaction between language and society. Our long term goal is to develop joint sociolinguistic models that explain the social basis of linguistic variation.

In this paper we focus on *microblogs*: internet journals in which each entry is constrained to a few words in length. While this platform receives high-profile attention when used in connection with major news events such as natural disasters or political turmoil, less is known about the themes that characterize microblogging on a day-to-day basis. We perform an exploratory analysis of the content of a well-known microblogging platform (Twitter), using topic models to uncover latent semantic themes [1]. We then show that these latent topics are predictive of the network structure; without any supervision, they predict which other microblogs a user is likely to follow, and to whom microbloggers will address messages. Indeed, our topical link predictor outperforms a competitive supervised alternative from traditional social network analysis. We explore the application of supervision to our topical link predictor, using regression to learn weights that emphasize topics of particular relevance to the social network structure. Finally, we propose two topics models that should better capture properties of Twitter than LDA.

2. DATA

We acquired data from Twitter's streaming "Gardenhose" API, which returned roughly 15% of all messages sent over a period of two weeks in January 2010. This comprised 15GB of compressed data; we aimed to extract a representative subset by first sampling 500 people who posted at least sixteen messages over this period, and then "crawled" at most 500 randomly-selected followers of each of these original authors. The resulting data includes 21,306 users, 837,879 messages, and 10,578,934 word tokens.

2.1 Text

Twitter contains highly non-standard orthography that poses challenges for early-stage text processing.[1] We took a conservative approach to tokenization, splitting only on whitespaces and apostrophes, and eliminating only token-initial and token-final punctuation characters. Two markers are used to indicate special tokens: `#`, indicating a topic (e.g. `#curling`); and `@`, indicating that the message is addressed to another user. Topic tokens were included, but address tokens were removed. All terms occurring less than 50 times were removed, yielding a vocabulary of 11,425 terms. Out-of-vocabulary items were classified as either words, URLs, or numbers. To ensure a fair evaluation, we removed "retweets" – when a user reposts verbatim the message of another user – if the original message author is also part of the dataset.

2.2 Links

We experiment with two social graphs extracted from the data: a **follower graph** and a **communication graph** (also called a message graph). The follower graph places directed edges between users who have chosen to follow each other's updates; the message graph places a directed edge between users who have addressed messages to each other (using the `@` symbol). [4] argue

[1]For example, some tweets use punctuation for tokenization (`You look like a retired pornstar!lmao`) while others use punctuation inside the token (`lOv!n d!s th!ng call3d l!f3`).

that the communication graph captures direct interactions and is thus a more accurate representation of the true underlying social structure, while the follower graph contains more connections than could possibly be maintained in a realistic social network.

3. MODELING TEXT

3.1 Latent Dirichlet Allocation (LDA)

We constructed a topic model over twitter messages, identifying the latent themes that characterize the corpus. In standard topic modeling methodology, topics define distributions over vocabulary items, and each document contains a set of latent topic proportions [1]. However, the average message on Twitter is only sixteen word tokens, which is too sparse for traditional topic modeling; instead, we gathered together all of the messages from a given user into a single document. Thus our model learns the latent topics that characterize *authors*, rather than messages.

3.2 Twitter-LDA

While LDA assigns topics to each word in all tweets of an author, the topic assignments of all words in a tweet are independent of each other. However given the small number of words in a single tweet, it is expected that the topic of a tweet will be coherent, and a single tweet will usually not talk about multiple things. Hence a simple extension of LDA would constrain LDA topic assignments to all words in a single tweet to have the same topic.

Figure 1 shows the Twitter-LDA model, using the same notation as LDA. Each user has multiple tweets, and each tweet is constrained to have a single topic (z) that generates the words (w) and hash tags (#) of the tweets. Since the words and hashes are both generated from z, and have multinomial distributions, we do not treat them separately, and generate them from the same β vector, whose length is a sum of the vocabulary size of the words and the hash. Note that the β vector will have to be normalized separately for the words and the hash tags, but we do not consider this issue in this paper, since it is easily solvable. Hence, for notational convenience, we will refer to the words and hash tags of a tweet together as w_n for $n = 1...N$, with N being the total number of words and hash tags.

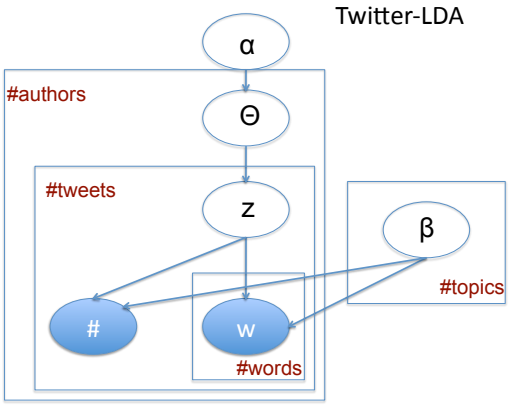

Figure 1: Plate representation of Twitter-LDA

Once we have defined the model, we develop a variational inference procedure for approximate inference, and use it in a variational expectation maximization algorithm for parameter estimation.

3.2.1 *Variational Inference*

In posterior inference, we seek to compute the posterior distribution of the latent variables conditioned on the observations. Exact posterior inference is intractable, hence we use variational inference. We use a fully factorized family for the variational inference,

$$q(\Theta, Z|\gamma, \Phi) = \prod_u \{q_\theta(\theta_d|\gamma_d) \prod_{tw} q_z(z_{u,tw}|\phi_{u,tw})\} \quad (1)$$

where u represents the users and tw the tweets of user u. γ is the set of Dirichlet parameters, one for each user, and Φ is the multinomial parameters, one for each tweet, for each user.

Minimizing the relative entropy is equivalent to maximizing the Jensen's lower bound on the marginal probability of the observations.

$$L = \sum_u \sum_{tw} \sum_w E_q(\log p(w_{u,tw}|\beta_{1...K}, z_{u,tw})) +$$
$$\sum_u \sum_{tw} E_q(\log p(z_{u,tw}|\theta_u) + \sum_u E_q(\log p(\theta_u|\alpha) + H(q)$$

As in LDA, minimizing this bound leads to the following updates, that can be used in a coordinate descent algorithm to find the best variational parameters γ and Φ. The derivation is very similar to the LDA, and the final updates are given below

$$\phi_{ni} \propto (\prod_{v:v \in tw_n} \beta_{iv}) \; exp\{\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_i)\} \quad (2)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \quad (3)$$

3.2.2 *Variational EM*

We fit the model by finding maximum likelihood estimates for the parameters α and β. The α update is the same as in LDA, while to find the MLE of β, we observe that now there is an additional level in the hierarchy to sum over, hence the update for β has an additional sum.

For topic i, and word j

$$\beta_{ij} \propto \sum_{u:users} \sum_{tw:tweets} \sum_{w:words} \phi_{u,tw,i} w^j_{u,tw,w} \quad (4)$$

3.3 Modeling Links

Authors with similar topic proportions are likely to share interests or dialect, suggesting potential social connections. Author similarity can be quantified without supervision by taking the dot product

of the topic proportions. If labeled data is available (partially observed network), then regression can be applied to learn weights for each topic. [2] describe such a regression-based predictor, which takes the form $\exp\left(-\eta^T(\bar{z}_i - \bar{z}_j) \circ (\bar{z}_i - \bar{z}_j) - \nu\right)$, denoting the predicted strength of connection between authors i and j. Here $\bar{z}_i$ ($\bar{z}_j$) refers to the expected topic proportions for user i (j), η is a vector of learned regression weights, and ν is an intercept term which is only necessary if a the link prediction function must return a probability. We used the updates from Chang and Blei to learn η in a post hoc fashion, after training the topic model.

3.3.1 Global Link Modeling

One limitation of the above models is that the topic inference and link prediction are done separately, thus the existence of a link cannot influence the topics of the corresponding users. However, since homophily has been observed in the Twitter social network (with up to 70% of links being reciprocal), users that link to each other can be expected to talk about similar things. The Relational Topic Model takes one step in this direction, however, it only takes into account similarity of topics when predicting links. Global measures of the network itself like presence of hubs etc. cannot be captured by the RTM.

We propose a very simple extension of the RTM, where the link between two users a and b, is dependent on three things:

- The similarity of the topics the two users talk about, i.e. $(\bar{z_a} \circ \bar{z_b})$.
- The popularity of user b, when user a links to b, approximated by the number of users following b, denoted by E_b.
- The number of common neighbors between a and b, denoted by $N_{a,b}$.

The probability of observing a link $y_{a,b}$ can then be expressed as a function of these 3 factors, with a weight for each. The link function used here is the logistic function ($logit$), though it may be replaced by any other suitable function that maps our features to $(0, 1)$.

$$P(y_{a,b}|z_a, z_b, N_{a,b}, E_b) = logit(\eta^T(\bar{z}_a \circ \bar{z}_b) + \delta E_b + \omega N_{a,b} + \nu) \quad (5)$$

where $x \circ y$ represents the Hadamad or element-wise product between x and y.

The model can be visualized in figure 2.

This model extends RTM by adding two observed nodes to the model. Since no new hidden nodes are added, the inference step does not change. In the M-step of the variational E-M, two new parameters need to be estimated, δ, and ω. However, since they appear in a logistic function, in the same way as η, they are learned by using the same updates as η by doing gradient descent. Since only positive links are observed and modeled by GLM, a regularization penalty is added to the logistic function, parametrized by ρ, which assumes ρ negative links in the network (without specifying which ones), and incorporates them into the estimates.

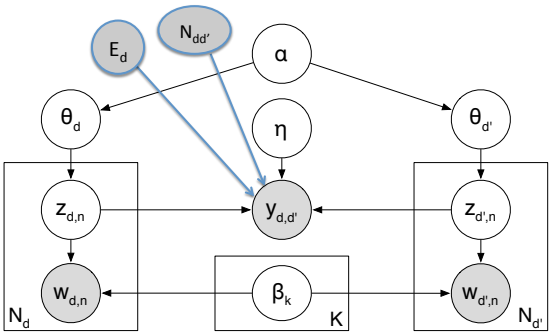

Figure 2: Global link modeling is a simple extension of RTM

Thus, the objective function to estimate these parameters is

$$\min_{\eta,\delta,\omega,\nu} \sum_{(a,b)\in Edges} \log p(y_{a,b}|z_a, z_b, \eta, \delta, \omega, \nu) + \rho \; logit(\eta^T(\bar{\Pi} \circ \bar{\Pi}) + \delta\bar{E} + \omega\bar{N} + \nu)$$

where $\bar{\Pi} = \frac{1}{|U|}\sum_u \bar{z}_u$ is the mean topic vector over all users, $\bar{E} = \frac{1}{|U|}\sum_u E_u$, and $\bar{N} = \frac{1}{|U|}\sum_a \sum_b N_{a,b}$ are the mean influence and mean common neighbors across all nodes in the network (and not just the ones with links). Since this objective is differentiable, given ρ, the gradient can be computed and conjugate gradient descent can be used to find the parameter values.

4. RESULTS

We constructed topic models using Latent Dirichlet Allocation (LDA)[2]. We implemented Twitter-LDA and GLM in Java, however, we believe that GLM has a bug, hence its results are not reported here.

4.1 Text Analysis

We ran LDA with 50 topics, detailed results of the run can be found at `http://sailing.cs.cmu.edu/socialmedia/naacl10ws/`. There were 8 topics from different languages (French, Dutch, Indonesian, Italian, Portuguese, Spanish, German, mixture of English and Indoneisan). Two topics seemed to capture stopwords in different communities, and one topic was a mixture of travel and spam. The remaining 39 topics were manually classified into the six broad categories shown in Table 1. As can be seen, many of the topics focus on content (for example, electronics and sports), others capture distinct languages and even dialect variation. Such dialects are particularly evident in stopwords (`you` versus `u`).

We repeated the same experiment with Twitter-LDA, and a brief analysis is shown in Figure 3. Over 50 topics, the number of conversation topics seem to increase in Twitter-LDA versus LDA, at the expense of content topics(news, markets etc.) - there were 13 conversation topics in Twitter-LDA as opposed to 10 in LDA. The one exception was the emergence of a topic about Haiti in Twitter-LDA; in LDA, tweets about Haiti got diffused into different categories like food (`Send food, supplies to Haiti!`),

[2] `http://www.cs.princeton.edu/~blei/lda-c`

travel (`Doctors without Borders travel to Haiti`), prayer etc. The number of foreign language topics remained fixed.

One question we could ask using Twitter-LDA was whether the hash-tags captured meaningful associations with topics. We observed that for content topics, the top hash tags accurately describe the topic, while for conversation and language topics, this is not true. For example, the top hash tags associated with the politics topic shown in figure 3 are `#sarahpalin`, `#news`, `#fox`. Conversation topics cannot have "topical" tags, but instead capture common tags in the dataset like `#WhatSheSaid`, `#ItsSoStupid` etc.

Structured topic models that explicitly handle these two orthogonal axes of linguistic variation versus content are an intriguing possibility for future work.

4.2 Link Analysis

We evaluate our topic-based approach for link prediction on both the **message** and **follower** graphs, comparing against an approach that only considers the network structure. [7] perform a quantitative comparison of such approaches, finding that the relatively simple technique of counting the number of shared neighbors between two nodes is a surprisingly competitive predictor of whether they are linked; we call this approach common-neighbors. We evaluate this method and our own supervised LDA+regression approach by hiding half of the edges in the graph, and predicting them from the other half. As mentioned earlier, due to the possibility of a bug in our code, we do not present RTM and GLM results here, and leave it for future work. (The current results I observed seem to imply that RTM does worse than LDA+regression, which contradicts the results reported in the RTM paper.)

For each author in the dataset, we apply each method to rank all possible links; the evaluation computes the average rank of the true links that were held out (for our data, a random baseline would score 10653 – half the number of authors in the network). As shown in Figure 4, topic-based link prediction outperforms the alternative that considers only the graph structure. Interestingly, post hoc regression on the topic proportions did not consistently improve performance, though joint learning may do better (e.g., Chang and Blei, 2009). The text-based approach is especially strong on the message graph, while the link-based approach is more competitive on the followers graph; a model that captures both features seems a useful direction for future work.

5. DISCUSSION

This project explored the use of topic models in understanding both text and links in Twitter. On a dataset of 21306 users, we found that LDA finds meaningful topics, that seem to capture the kinds of conversations observed in Twitter, with topics being equally divided into conversations between users, spam, and content topics (like news, travel, gaming, etc). Twitter-LDA proposed a simple extension that allowed us to find representative tweets for each topic, by assigning a single tweet to a single topic. This had the effect of finding more conversation topics at the expense of content topics. Interestingly enough, the conversation topics seemed to capture more linguistic variations in Twitter-LDA, with separate topics for positive words and a separate one for negative words. One of the topics found has top topic words "niggas, hoes, neva" etc. which was not observed in LDA. Clearly, a more detailed study is needed.

We also predicted links in Twitter using topic similarity alone. We observed that topic-similarity is more helpful in predicting links on the message graph than on the follower graph. This confirms to our intuition that people who talk to each other (message graph) are more similar to each other in the content of their tweets than people who follow each other (follower graph), where following a user may be explained by their fame or the number of common neighbors etc, instead of topic-similarity. On the basis of this conclusion, we derived a new model that captures global network properties, specifically influence of a user and number of common neighbors. Future evaluation will study important of each topic in predicting links. For example, are spammers more likely to link to each other than regular people? Is content similarity more important than linguistic similarity when predicting links?

This work has studied text and links separately, to enable us to derive better intuition of the contribution of the two ideas in understanding twitter. A joint model, that predicts a single topic per tweet, and also predicts links, much like GLM, using topic similarity and global network properties as features, seems to be a useful direction for future work.

6. ADDITIONAL AUTHORS

Jacob Eisenstein and Shay Cohen. Jacob and Shay contributed equally with Kriti in collecting data, running LDA, and running the LDA+regression link prediction work. Kriti derived and implemented Tiwtter-LDA, and GLM, and did the topic analysis for both models.

7. REFERENCES

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

[2] J. Chang and D. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 2009.

[3] M. Halliday. *Language as social semiotic: The social interpretation of language and meaning*. University Park Press, 1978.

[4] B. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1–5), January 2009.

[5] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSN '08: Proceedings of the first workshop on Online social networks*, pages 19–24, 2008.

[6] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 577–584, New York, NY, USA, 2006. ACM.

[7] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. of CIKM*, 2003.

[8] D. Mimno, W. Li, and A. Mccallum. Mixtures of hierarchical topics with pachinko allocation. In *In Proceedings of the 21st International Conference on Machine Learning*, 2007.

[9] D. Shen, J. Sun, Q. Yang, and Z. Chen. Latent friend mining from blog data. In *Proc. of ICDM*, 2006.

Conversations (out of 10)	Chit-chat	geting ready go to dinner and movie with some friends!!!!!!!!!	-OOV-, my, to, and, in, with, at, for, the, a, new, going, day, work, so, am, night, today, time, is
	Slang	@MsGumdrop lmao das ok now u 1 step smarter den da next person onlything i remember is math n science smh	-OOV-, u, lol, i, me, my, a, 2, im, ur, on, up, lmao, like, in, shit, its, get, got, n
	Weather	@JPRennquist concerned on the amount of ice that may form over Lake Superior, that would play a bigger role on our weather by March thru May	the, a, -OOV-, in, it, of, on, is, to, 's, and, cold, for, this, snow, just, here, good, morning, today, weather
	Optimism	Feeling the hope that a new day brings.	-OOV-, for, the, of, in, thanks, 's, to, great, at, a, it, out, -URL-, today, new, so, 2, you, see
SPAM (9)	Social media	say hi to some new #followfriday frenz: @Kdark74 @MAKEAMILLION1 @janettefuller @The_Scallywags @therealamaru @mslegalhelp	you, follow, thanks, ff, for, thank, rt, the, to, back, your, me, are, followfriday, a, great, u, -OOV-, is, hi
	Weight loss	Start Losing Weight NOW. Shed holiday pounds safely and effectively. Free Trial Offers. Act today! spon http://tinyurl.com/yd9m8po	-URL-, -OOV-, free, weight, to, for, the, on, loss, and, your, get, save, off, a, with, fitness, diet, at, lose, fat
News (6)	Financial	Market breadth is very negative; a bearish indication see http://bit.ly/lPIyW #News #Trading #Investing #Finance #SPX #Mkt #Business #Stocks	-URL-, for, -OOV-, business, to, in, home, real, the, estate, credit, a, mortgage, on, finance, economy, money, market, and
	Gadgets	The Switch From iPhone To Android, And Why Your First Impression Is Wrong #gadgets http://bit.ly/4GTPmo	-URL-, -OOV-, the, to, iphone, google, for, and, of, a, on, apple, ces, app, one, 's, in, with, nexus, new
Market (4)	Jobs	#jobs Oracle Project Manager 12i in Denver, Colorado http://jobshouts.com/job/10128/	-OOV-, -URL-, job, jobs, for, a, to, me, and, follow, the, in, need, looking, we, php, at, little, rock, i
	Gadgets	don't miss : Canon PowerShot SD960 IS Digital ELPH Camera (Pink) + 4GB SD Card + Case + NB-4L + Accessory Kithttp://bit.ly/60JFM2	digital, card, 2.0, new, usb, reader, in, one, memory, multi, camera, sdhc, lcd, zoom, mp, -URL-, 2.4, 4x, coffee, pink
Entertain. (3)	Television	RT @joshmmartin: How about nearly everyone on the Atlanta episode of #americanidol tonight w/ a country accent hasn't been from GA	ha, -OOV-, la, idol, flu, a, american, boxing, of, the, show, fight, castle, floyd, hee, on, swine, americanidol, you, dude
	Film and music	RARE Michael Jackson interview with Soulbeat 1979 http://bit.ly/RBsDk Pls ReTweet & #FollowFriday us ! MJ music 4ever!	-OOV-, -URL-, the, 's, music, of, new, in, and, a, michael, to, on, movie, for, jackson, with, guitar, by, is
Lifestyle (7)	Prayer	Lord someone is in that rubble in Haiti who needs to hang on and believe for a miraculous rescue. Breathe on them!Be their rescue!#tcot#pray	-OOV-, for, god, please, love, and, my, haiti, pray, in, to, the, you, praying, of, her, help, hugs, all, is, jesus
	Food	Good Food, Good Wine and a Bad Girl: A Slice of Heaven - Chocolate Banana Bread: Chocolate Banana Bread http://bit.ly/5LiSwp	-OOV-, -URL-, the, to, and, a, in, for, haiti, food, of, 's, with, our, we, recipe, cooking, on, at, wine

Table 1: Examples from the 50 topics found by the LDA model. In each broad category, the first column roughly summarizes the topic, the second shows a representative tweet from the same topic, and the third column shows the top 20 words in this topic.

Figure 3: Visualizing four out of 50 topics found by Twitter-LDA. The words are the top-25 words of the topic, and two tweets assigned to the topic are shown.

Figure 4: Mean rank of test links (lower is better), reported over 4-fold cross-validation. Common-neighbors is a network-based method that ignores text; the LDA (Latent Dirichlet Allocation) methods are grouped by number of latent topics.