Extracting Topics Based on Authors, Recipients and Content in Microblogs

Nazneen Fatema Rajani
Department of Computer Science
University of Texas at Austin
Austin, TX 78712
nrajani@cs.utexas.edu

Kate McArdle
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712
kate.mca@utexas.edu

Jason Baldridge
Department of Computer Science
University of Texas at Austin
Austin, TX 78712
jbaldrid@mail.utexas.edu

ABSTRACT

Microblogs such as Twitter are important sources for spreading vital information at high speed. They also reflect the general people's reaction and opinion towards major events or stories. With information traveling so quickly, it is helpful to be able to apply unsupervised learning techniques to discover topics for information extraction and analysis. Although graphical models have been traditionally used for topic discovery in microblogs and text streams, previous work may not be as efficient because of the diverse and noisy nature of microblogs.

In this paper, we demonstrate the application of the Author-Topic and the Author-Recipient-Topic model to microblogs. We extensively compare these models under different settings to an LDA baseline. Our results show that the Author-Recipient-Topic model extracts the most coherent topics establishing that joint modeling on author-recipient pairs and on the content of tweet leads to quantitatively better topic discovery. This paper also addresses the problem of topic modeling on short text by using clustering techniques. This technique helps in boosting the performance of our models. Our study reveals interesting traits about Twitter messages, users and their interactions.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

1. INTRODUCTION

Online social media systems like Twitter and Facebook are used as sources of information especially during events related to natural disasters, political turmoil or other such crises. Tweets and Facebook status messages are short and may not carry many contextual clues about the content's subject matter. Hence, applying traditional natural language processing algorithms on such data is challenging. This is because of following reasons: tweets are very short in length, with a 140-character limit; tweets are very informally written and often consist of ungrammatical text; and tweets may contain implied references to locations, [7] thus making named entity recognition difficult.

We believe that clustering of tweets using topic models will help to easily categorize them based on their properties. Using such clusters, we seek to identify the topics or particular event about which the tweet is written. In this paper, we present different approaches that classify an incoming tweet as a mixture of the topic clusters learned by the topic model. Topic models do not make any assumptions on the ordering of the words in a document and also disregard the grammatical structure. Such a model is also known as the bag-of-words model. This approach is particularly suited to handling irregularities in microblog messages.

Although LDA is a well-known tool for clustering documents based on topics, it does not perform well on microblogs due to the reasons discussed above. Thus, we experimented with two directed graphical models, the *Author-Topic* (AT) model and the *Author-Recipient-Topic* (ART) model. The AT model [6] learns topics conditioned on the mixture of authors that composed a document, this has been discussed further in section 2.2. Experimental results show that the state-of-the-art Author-Topic model fails to model hierarchical relationships between entities in social media settings [2]. The ART model [3] is similar to the AT model, but with the crucial enhancement that it conditions the per-message topic distribution jointly on both the authors and recipients, rather than on individual authors. Thus the discovery of topics in the ART model is influenced by the social structure in which messages are sent and received. This setting has been used previously for role discovery in social networks [3]. In this paper we present the ART model for microblogs and analyze its performance with other models. To the best of our knowledge, our work is the first time the ART model has been implemented for topic discovery in mciroblogs. Our results and analysis have enabled us to make important inferences about Twitter messages, users and their interactions.

2. METHODOLOGY

In this section, we describe the Bayesian network approaches we used to perform topic modeling in the context of tweets: LDA, AT and ART. We use the following terminology: a set of documents forms a *corpus*. The set of unique words that are used in the corpus forms the corpus's *vocabulary*, while

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609537.

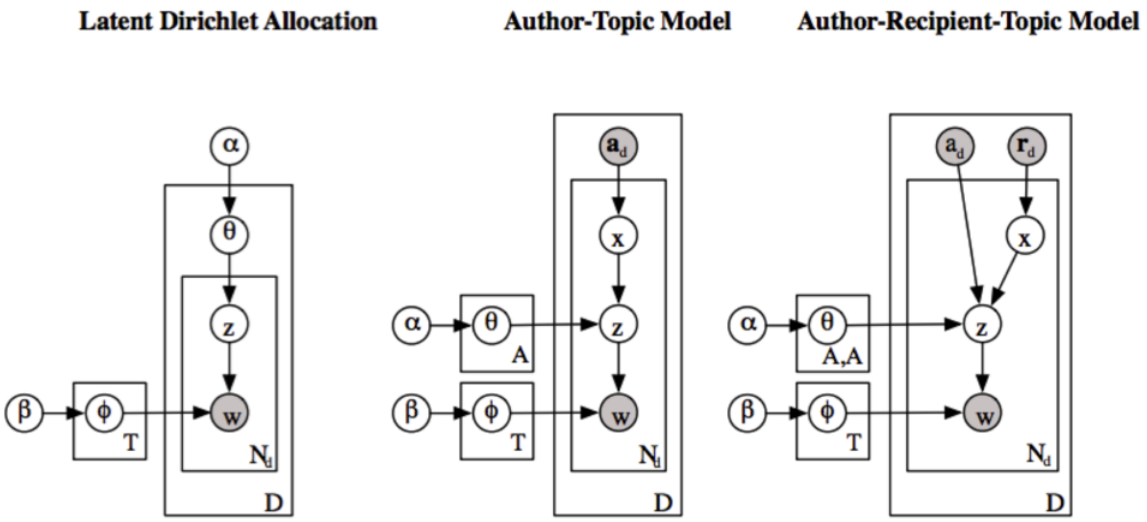

Figure 1: LDA, AT, and ART models. Modified from [3]

we refer to the collection of words that appear in a given document as *word tokens*. The word tokens found in a document are not necessarily unique words from the vocabulary. For example, a tweet that appears as "twinkle twinkle little star" uses the following words from the vocabulary: twinkle, little, star. The word tokens in this tweet are: twinkle, twinkle, little, star.

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation was first introduced by Blei et al. [1]. LDA models each text document in a corpus as a mixture of an underlying set of topics. Figure 1 displays a graphical representation of LDA. Each document d has a multinomial distribution θ_d of topics, and each topic z has a multinomial distribution ϕ_z of words. A document's topic distribution is randomly sampled from a Dirichlet distribution with hyperparameter α, and each topic's word distribution is randomly sampled from a Dirichlet distribution with hyperparameter β. Thus, topic assignment in LDA is modeled solely on the document's word token content.

2.2 Author-Topic Model

The Author-Topic model [6] builds on LDA, by modeling a document's topics based on the document's content, as in LDA, and by conditioning on the document's authors. Figure 1 displays a graphical representation of the AT model. Each document d has a set of observed authors a_d. A document's topic distribution is influenced by this set of authors. To generate each word token in the document, an author x is randomly and uniformly sampled from a_d, and then a topic z is sampled from the author's topic distribution θ_x, which comes from a Dirichlet distribution with hyperparameter α. From this topic, the word token is sampled from the topic's word distribution ϕ_z, which comes from a Dirichlet distribution with hyperparameter β. Thus, topic assignment in the AT model is based on the document's authors and word token content.

2.3 Author-Recipient-Topic Model

The Author-Recipient-Topic model [3] builds on LDA and AT, by modeling a document's topics based on the document's content, as in LDA, the document's authors, as in AT, and the document's recipients. Thus, ART is only appropriate for documents with specific recipients (e.g., emails) and is not appropriate for documents without recipients (e.g., scholarly articles). Figure 1 displays a graphical representation of ART. Each document d has a set of authors a_d and a set of recipients r_d. A document's topic distribution is influenced by the set of observed author-recipient pairs. To generate each word token in the document, an author-recipient pair ar is randomly and uniformly sampled from this set, and then a topic z is sampled from the author-recipient pair's topic distribution θ_{ar}, which comes from a Dirichlet distribution with hyperparameter α. From this topic, the word token is sampled from the topic's word distribution ϕ_z, which comes from a Dirichlet distribution with hyperparameter β. Thus, topic assignment in the ART model is based on the document's authors, recipients, and word token content.

3. DATASET

We perform topic modeling on a set of Twitter tweets from August to October 2008. We crawled tweets starting from an initial node and then recursively iterating over the followers of the node into consideration. In case we encounter a private blog, we backtrack. The initial set contained $160,000$ tweets from this time period. In this section, we describe the filtering we performed on the set of tweets and the filtering we performed on the word tokens within each tweet.

3.1 Filtering on Tweets

The set of crawled tweets was then filtered in two ways, which we describe here. The first filtering we performed was for "@mention". Our goal in this paper is to compare the relative performances of the topic models described in Section 2, one of which is the Author-Recipient-Topic Model. This model requires that every document have at least one recipient, so we filtered our original dataset to only keep tweets that include @mention. We then consider the Twitter handle mentioned in @mention to be the recipient of the tweet. In the case of multiple @mention inclusions in a single

tweet, we consider each Twitter handle listed as separate recipients. Thus, each document consists of three attributes: the tweet's content, the tweet's author, and a set of one or more recipients. We call this set of tweets the Recipient Dataset.

The second filtering we performed was for hashtags. Our motivation for this filtering comes from [5], which suggests that the performance of topic modeling on tweets is generally poor, due to the inherently short nature of each document. In [4], the authors show that one approach to overcoming this pitfall is to cluster together tweets that contain the same hashtag. Each cluster constitutes one document, and the topic model is trained on this set of documents. For completeness, we compare the performance of the topic models described in Section 2 when trained on an unclustered dataset to when trained on a clustered dataset. In order to make such comparisons, we are required to remove any tweets from the Recipient Dataset that do not have at least one hashtag in the tweet's content. We call the resulting dataset the Single-Tweet Dataset, as each document consists of a single tweet (whose contents contain at least one hashtag), a single author, and a set of one or more recipients. The Single-Tweet Dataset consists of 7288 tweets, 1176 unique authors, and 7830 unique author-recipient pairs.

We create a second dataset such that the tweets in the Single-Tweet Dataset are clustered into documents by hashtag. In the case of a single tweet with multiple hashtag inclusions, the tweet is included in the document corresponding to each hashtag. We call this dataset the Clustered Dataset. In this dataset, each document consists of a set of one or more tweets (each tweet of which contains the same hashtag), one or more authors (such that the number of authors is less than or equal to the number of tweets), and one or more recipients (such that the number of recipients is greater than or equal to the number of authors). The Clustered Dataset consists of 2563 documents. The numbers of unique authors and unique author-recipient pairs are the same as for the Single-Tweet Dataset, since the underlying set of tweets is the same in both datasets.

4. EXPERIMENTAL RESULTS

We present our results performing topic modeling on the Single-Tweet Dataset and the Clustered Dataset. To train the LDA and Author-Topic models, we use Gibbs sampling to approximate the inference step of extracting topics, since it cannot be done exactly for LDA and similar models [3]. To train the Author-Recipient-Topic model, we note that the approach is identical to the Author-Topic model, if one considers a document's author-recipient pair to be its author.

4.1 Model Settings

For all models, we set the model hyper parameters α and β to $\frac{50}{|topics|}$ and $\frac{200}{|vocabulary|}$, respectively. In different experiments that we ran on training the models, we used either 500 or 1000 iterations in Gibbs sampling, and we extracted one of the following numbers of topics: 10, 20, 30, 40, 50, 75, 150, 300, 500. We trained our models, LDA, AT, and ART, on both the Single-Tweet Dataset and the Clustered Dataset.

4.2 Model Evaluation

To evaluate our results, we implemented a metric called the Pointwise Mutual Information (PMI) score for a trained topic model. PMI measures the coherence of the topics that are created by a trained topic model, by determining the statistical independence of two words from the same topic appearing together in the same document [4]. The PMI for a pair of words is

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

In our case, for both the Single-Tweet Dataset and the Clustered Dataset, when calculating PMI we consider each tweet to be a document, so we calculate PMI using empirical probabilities of the Single-Tweet Dataset. The probability of a single word, $p(w_i)$, is the ratio of the number of tweets that contain word w_i to the total number of tweets. The probability of a pair of words, $p(w_i, w_j)$, is the ratio of the number of tweets that contain both words w_i and w_j to the total number of tweets.

To calculate PMI for a given model, we used the approach outlined in [4]: for each topic, calculate the PMI of each of the possible word pairs among the top ten words with the highest probabilities. The PMI for the given topic is the average of the PMI scores for the word pairs, and the PMI for the given model is the average of the PMI scores for the topics. A higher PMI score indicates better topic coherence, and thus we compare each of the trained models based on their PMI scores.

Our results for 500 iterations are displayed in Table 1. For each number of topics, the model and dataset combination with the highest PMI is displayed in bold. Table 2 displays the top 10 words belonging to topic related to "Austin" for each of the LDA, AT and ART topic models.

Our results indicate that, as expected, the Clustered Dataset results in better-trained topic models than the Single-Tweet Dataset, regardless of the number of topics. We compared the results with 500 iterations and with 1000 iterations but did not see a big difference, indicating that our models are converging by 500 iterations. Our results suggest that LDA performs better than the other models on clustered tweets for a small number of topics, while ART performs better than the other models on a mid-range number of topics, and AT performs better than the other models on a higher number of topics. Table 2 indicates that the ART model extracts most coherent topics, followed by the AT and the LDA models respectively.

5. DISCUSSION

In this paper, we presented the performance of the ART topic model for microblogs, which addresses the issues of short text modeling. As expected, our experimental results demonstrate that all three types of model perform better on clustered documents than unclustered documents. Tweets belonging to one cluster tend to represent more coherent topics as shown by [4]. Thus, models trained on longer text yield better results than those trained on short text.

Our experiments demonstrate that, on average, for fewer than 300 topics, the performance of the ART model is the best, followed by LDA, and finally the AT model. The poor performance of AT model on smaller number of topics enabled us to make important inferences. Firstly we claim that an average Twitter user tweets about a wide range of topics and these topics have a high distance when compared using

Model	Dataset	PMI score for the following number of topics:								
		10	**20**	**30**	**40**	**50**	**75**	**150**	**300**	**500**
LDA	Single-Tweet	0.565	1.002	1.317	1.528	1.715	1.921	2.093	2.152	2.244
LDA	Clustered	**0.770**	**1.168**	**1.479**	**1.615**	1.778	2.066	**2.596**	3.269	3.169
AT	Single-Tweet	0.634	0.932	1.315	1.372	1.609	1.954	2.232	2.723	2.982
AT	Clustered	0.712	0.994	1.215	1.514	1.607	1.973	2.377	**3.298**	**3.336**
ART	Single-Tweet	0.523	1.047	1.291	1.555	1.724	1.981	2.412	2.272	2.412
ART	Clustered	0.639	0.953	1.243	1.555	**1.790**	**2.103**	2.538	2.769	2.867

Table 1: PMI scores for LDA, AT and ART models trained on the Single-Tweet and Clustered Datasets, with 500 iterations.

LDA	AT	ART
hit	time	sxsw
stuff	sxsw	love
life	apple	panel
night	real	**austin**
key	store	row
takes	app	party
austin	talk	rocks
uh	current	things
disappointed	click	lots
start	**austin**	student

Table 2: Top 10 words belonging to topic related to "Austin" for each of the LDA, AT and ART topic models on the Single-Tweet Dataset.

a similarity metric, thus implying that they have very little or no overlap. Secondly it it very difficult to distinguish between any two average Twitter users; this inference follows from our first claim. Our results provide evidence to these claims, the performance of AT model keeps improving as we increase the number of topics. This means that allowing more topics in the model gives room to cluster authors into more topics and enables distinguishing between their messages. Another point of contention as discussed in [2] is that the the reason may be the "OR" nature of the AT model: a message is either *generated* by the message or by an author.

Increasing the number of iterations from 500 to 1000 has very little effect on the performance of the models. Thus, we are assured that all our models converge. Also we verified that our models are robust to different random initializations to the Gibbs chains.

6. CONCLUSION

We demonstrated the application of three types of graphical models, the LDA, the AT and the ART to microblogs. We also addressed the issue of topic modeling in a microblogging environment. More specifically, through our experiments we showed that for short and unstructured text, it is more meaningful to cluster the documents before modeling them. Our results show that discovering topics by conditioning on the author-recipient relationships in a corpus of tweets works best. To the best of our knowledge, this paper is the first to demonstrate the effectiveness of such a model to microblogs.

We compared the models based on a number of aspects including how the topics learned by these models differ qualitatively. We also demonstrated that clustering tweets using *hashtags* leads to superior performance in classification. We believe that our research would lay groundwork for future work in story or event detection in microblogs by implementing topic discovery using joint modeling over author-recipient and tweet content.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[3] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, page 3, 2005.

[4] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. 2013.

[5] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[6] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[7] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.