2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)

Tag-Latent Dirichlet Allocation: Understanding Hashtags and Their Relationships

Zhiqiang Ma, Wenwen Dou, Xiaoyu Wang, Srinivas Akella
Department of Computer Science
The University of North Carolina at Charlotte
Charlotte, NC, USA
Email: zma5@uncc.edu, wdou1@uncc.edu, xwang25@uncc.edu, sakella@uncc.edu

***Abstract*—A hashtag is defined to be a word or phrase prefixed with the symbol "#". It is widely used in current social media sites including Twitter and Google+, and serves as a significant meta tag to categorize users' messages, to propagate ideas and topic trends. The use of hashtags has become an integral part of the social media culture. However, the free-form nature and the varied contexts of hashtags bring challenges: how to understand hashtags and discover their relationships? In this paper, we propose Tag-Latent Dirichlet Allocation (TLDA), a new topic modeling approach to bridge hashtags and topics. TLDA extends Latent Dirichlet Allocation by incorporating the observed hashtags in the generative process. In TLDA, a hashtag is mapped into the form of a mixture of shared topics. This representation further enables the analysis of the relationships between the hashtags. Applying our model to tweet data, we first illustrate the ability of our approach to explain hard-to-understand hashtags with topics. We also demonstrate that our approach enables users to further analyze the relationships between the hashtags.**

***Index Terms*—topic model; hashtag; Twitter analysis**

I. Motivation, Challenges, and Research Goal

Hashtags, commonly used on Twitter and Google+, have become a unique tagging convention to organize social media content and associate events, trends, or topic information. Over the years, the number of hashtags utilized has been on the rise. On Twitter, for example, one out of eight tweets contains at least one hashtag[1]. Hashtags establish a bi-directional interaction between users and the online information. On the one hand, hashtags as categorizations enable users to follow and acquire news, people's opinions, and status updates [1]. On the other hand, hashtags further embody user participation in the process of creating hashtags, for the purpose of initiating and propagating content throughout the social networks.

There are no restrictions on how a hashtag can be constructed, resulting in various lengths, forms, and structures of hashtags. Specifically, a hashtag is a brief phrase that could consist of a single word, numbers, abbreviation or a combination of these. According to Twitter, a hashtag is comprised of the symbol "#" followed by a sequence of keywords or phrases (without spaces) and is used to "mark keywords or topics in tweets"[2]. Such characterization enables the retrieval of all tweets including a certain hashtag. It empowers users to follow conversations of interest and facilitates their search for similarly tagged tweets.

Although the use of hashtags has become a convention, *how well users understand and use the hashtag information* is still unclear. Sifting through trending hashtags on social media has become a popular way to learn what events have occurred, as shown in the up-to-the-minute trending topics listed on Twitter. But the intrinsically polylingual, fragmented, and dynamic nature of hashtags is also a disadvantage in eliciting valuable information. Users can be overwhelmed with the noise of unrelated messages and conflicting information.

Therefore, it is necessary for us to identify a solution that can help users effectively make sense of hashtags. The goal of our research is to enable users to understand the meaning of hashtags as well as the correlations between hashtags.

A. Challenge 1: How to Understand the Meaning and Context of a Hashtag?

Some of the existing hashtags are constructed in an intuitive manner, serving as metadata to categorize what that post is about. For instance, "#grammys" and "#ImmigrationStory" denote the Grammy music awards and President Obama's pledge to share stories of immigrant families to support his immigration reformation, respectively.

However, other hashtags are not as easy to make sense of. For instance, hashtag "#NatGat" (national gathering) and "#tcot" (Top Conservatives on Twitter) are difficult to decode by reading just the hashtags themselves. Also, hashtags such as "#Jan25"[3] are challenging to comprehend since they are too general to derive the significance without any context.

Although the loose format of hashtags encourages users' creativity, without help from proper computational methods, users can hardly make sense of the brief phrases. While there is work on crowd-sourcing solutions[4], our goal is to identify a computational approach that will automate the process, making it more effective for users to make sense of hashtags.

B. Challenge 2: How to Semantically Correlate Relevant Hashtags?

Ideally, one would like to maintain a clear structure of hashtags with a one-to-one relationship to the corresponding

[1] http://www.nytimes.com/2012/11/04/magazine/in-praise-of-the-hashtag.html

[2] https://support.twitter.com/articles/49309-what-are-hashtags-symbols

[3] "#Jan25" is used to indicate Egyptian revolution began on Jan 25, 2011.

[4] http://www.livescience.com/26151-twitter-mechanical-turk-workers.html

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.38

topics or events. However, in practice, this is not an option due to the creativity of the users. Users could create multiple hashtags for the same event or topic. For example, hashtags related to "#MichaelJackson" can also be seen in the form of "#KingOfPop" or "#MJ". "#occupywallstreet" and "#OWS" are both used to characterize the same event, but with different expressions.

Sometimes, multiple hashtags were created to denote different aspects of the same event. For instance, the discussions of the "Occupy movement" on Twitter contain various hashtags to denote the who, what, when, and where information. Specifically, such hashtags include "#usdor"—the organizing party of OWS, "#sep17"— information about occupy date, "#occupywallstreet" or "#occupyChicago" denoting where the protests occur, and "#pepperspray" that describes a sub event in the Occupy movement.

C. Our Approach to Depicting and Correlating Hashtags

The examples in Section I-A and I-B highlight the challenges in understanding the meaning and relationships of the hashtags. While prior research has analyzed linguistic features of hashtags [2] and the spreading of hashtags as ideas [3], to the best of our knowledge, no prior work has studied utilizing tweet content to denote, correlate, and make sense of hashtags.

To this aim, we develop a new topic model that extends the latent Dirichlet allocation (LDA) [4] by incorporating the observed hashtags. Our model can address the following two questions:

1) How can one understand and interpret hashtags and the context in which they are used?
2) How can one discover the relationships and correlations between the hashtags?

To highlight the differences between our approach and previous research, our model not only provides topic results for hashtags, but also enables the analysis of the relationships between the hashtags.

More specifically, we propose a Tag-Latent Dirichlet Allocation (TLDA) model that accounts for hashtag information together with tweet content. We assume topics are shared among hashtags and a hashtag therefore can be seen as a mixture of topics. A message is viewed as a mixture of observed hashtags. The design of the model could uncover the statistical relationship between hashtags and their associated messages, the relationship between topics and hashtags, as well as the relationship between topics and terms. Our model particularly suits the case when tweets have multiple hashtags. To effectively represent the modeling results, we apply visualizations to help users understand the complex probabilistic output. To empirically evaluate our approach, we apply TLDA on two tweet datasets. In the first experiment, we demonstrate that the results of the TLDA model can semantically explain the Twitter hashtags, and the context in which they are used, with the help of visualization tools. In the second experiment, we demonstrate that TLDA supports the exploration of the relationships between the hashtags.

II. Related Work

There has been a lot of research on social network structure, identifying influential users, etc. [5][6][7][8]. Although there is some work focusing on hashtags, but very little focuses on analyzing the meaning of hashtags. Most work considers hashtags as ideas, opinions, or information that flow and propagate over the social media network.

Cunha et al. [2] studied how the hashtags are created and used from the perspective of linguistic theory. Romero et al. [3] studied the widely used hashtags and found different hashtags exhibit different spreading patterns. Kamath et al. [9] combined two hypotheses of information spread and developed a probabilistic model to understand the global spread of social media. They found hashtags have local characteristics and therefore distance is the most significant factor influencing the spread. Tsur and Rappoport [10] developed a linear regression based approach to predict the spread of the hashtags, where content features combined with temporal and topological features would deliver the best prediction performance.

Lin et al. [11] assumed hashtags are indicators of topics of interest and they tracked the topics in continuous streams of Twitter by integrating a "foreground" model and "background" model. One function of hashtags is to track "trending topics", which is most often mentioned [1]. To identify the trendsetters, Saez-Trumper et al. [12] proposed a ranking algorithm in an information network with the temporal factor integrated. Sentiment analysis on tweets can also be performed with hashtags [13][14]. To the best of our knowledge, very little work has discussed the meaning of the hashtag. In this paper, we leverage topic modeling to analyze the hashtags.

Topic models aim to discover "latent topics" that pervade the document collections, discovering statistical relationships of words and latent topics. Popular topic modeling algorithms include probabilistic latent semantic indexing (pLSI) [15] and its extension LDA model. LDA [4] is a powerful statistical model that can be used to perform dimension reduction, classification, and other tasks on discrete data. LDA and its variations have been applied to short messages from social media [16][17]. However, LDA is an unsupervised algorithm and does not take into account the metadata.

To integrate metadata such as tags or labels into the unsupervised algorithm, researchers have proposed a few new approaches. Blei and McAuliffe [18] introduced a supervised topic model. They took into account a response variable, which could be a rating or category associated with a document. Their goal is to predict the response variables for new documents. The work most similar to ours is the author-topic model introduced by Rosen-Zvi et al. [19]. The author-topic model assumes a uniform distribution of authors' contributions in a given document, but our model assumes the distribution is multinomial with a Dirichlet prior. Ramage et al. [20] developed a labeled-LDA model. Labeled-LDA links the latent topics to the labels in a one-to-one mapping, given a user labeled document collection. Therefore, the relationship between labels and words can be established since one topic corresponds to one label. This one-to-one constraint is relaxed in their later work on the partially labeled topic model [21], where one label contains multiple topics. However, in partially labeled LDA, one topic can only be assigned to one label exclusively. One advantage of PLDA is that the connection between a label and its topic is clear because of the exclusive attribution of the topic to the label. But the disadvantage is that the relationships of the labels cannot be directly computed based on the topic proportions since labels are not described in the same topic space. In contrast, our work does not have such restrictions; the tags are expressed as a distribution over all the topics and the topics are shared among all tags. One advantage of our model is that the observed tags can be presented with the contributions of topics such that the relationship between tags themselves is measurable.

III. Tag-Latent Dirichlet Allocation

A. Model Description

Latent Dirichlet Allocation is a probabilistic generative model, that can be used to extract information, "latent topics", in a collection of documents. The latent topics consist of vocabulary terms, and the terms follow a multinomial distribution in each topic. Typically, these multinomial distributions are sampled from Dirichlet distributions.

Like LDA, our model is also a probabilistic generative model. It simulates the procedure of generating documents and further discovers the relationship between tags and topics. Tags are observed and associated with documents. We define a tag as a multinomial distribution over topics. To describe the model, we introduce some notation. We have a corpus of documents $\mathcal{C} = \{d_1, d_2, \ldots, d_M\}$. Each document d consists of a set of words $\boldsymbol{w}_d = \{w_1, w_2, \ldots, w_{N_d}\}$, which we assume meets the "bag of words" assumption, and a set of tags $\boldsymbol{\delta}_d = \{p_1, p_2, \ldots, p_{L_d}\}$. We also define a set $\Delta = \{p_1, p_2, \ldots, p_L\}$ containing all tags without replication in the corpus. $\boldsymbol{\delta}_d$ therefore is a subset of Δ. In addition, we assume all the elements of $\boldsymbol{w}_d$ come from a corpus-wide vocabulary $\mathcal{V}$.

More formally, when generating document d, a subset of tags $\boldsymbol{\delta}_d$ are first selected from Δ. For the ith word in $\boldsymbol{w}_d$, a tag e is chosen from $\boldsymbol{\delta}_d$ based on the multinomial distribution $\boldsymbol{\theta}_d$ of tags in this document. $\boldsymbol{\theta}_d$ is sampled from a symmetric Dirichlet distribution with hyperparameter $\boldsymbol{\alpha}$. Then, under the chosen tag, a topic is sampled from a multinomial distribution $\boldsymbol{\gamma}_e$, where $\boldsymbol{\gamma}_e$ is also assumed to be generated from a symmetric Dirichlet distribution. Topics in TLDA are also described as multinomial distributions $\boldsymbol{\beta}$ over the vocabulary, and the distributions are independently drawn from a Dirichlet($\boldsymbol{\phi}$). To discover the tag-topic relationship, $\boldsymbol{w}_d$ can be explicitly thought of as a mixture of observed tags and implicitly a mixture of topics, because words are in fact generated from the topics under the tags. Let T be the total number of topics predefined. So with the notation given in Table I, the generative steps of our model are listed below:

1) For each tag $p \in \Delta$, sample $\boldsymbol{\gamma}_p$ over all topics from $\gamma_p \sim$ Dirichlet($\boldsymbol{\rho}$).

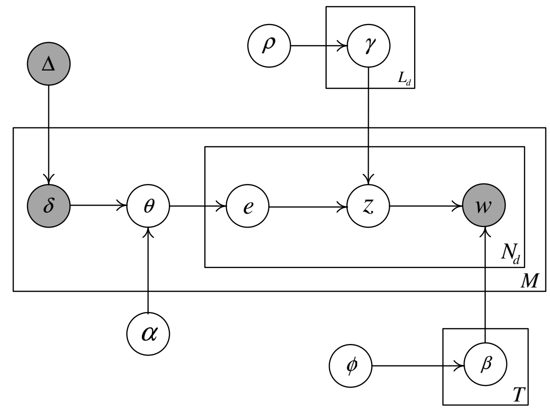

Figure 1. Graphical model of TLDA. Word w and tags δ are observed. Latent variables e and z are the tag and topic assignment to the word. Variables θ, γ, and β are latent variables. Tag set Δ is included so as to keep the completeness of the generative process.

symbol	size	description
M	scalar	number of documents
L	scalar	number of distinct tags
Δ	$1 \times L$	tags
$\mathcal{C}$	$1 \times M$	corpus
$\boldsymbol{w}_d$	$1 \times N_d$	words of document d
$\boldsymbol{\delta}_d$	$1 \times L_d$	tags of document d
$\boldsymbol{\gamma}_p$	$1 \times T$	tag-topic multinomial dist.
$\boldsymbol{\beta}_t$	$1 \times \lvert\mathcal{V}\rvert$	topic-term multinomial dist.
$\boldsymbol{\theta}_d$	$1 \times L_d$	doc-tags multinomial dist.
$\boldsymbol{\rho}$	$1 \times T$	Dirichlet hyperparameters
$\boldsymbol{\phi}$	$1 \times \lvert\mathcal{V}\rvert$	Dirichlet hyperparameters
$\boldsymbol{\alpha}$	$1 \times L_d$	Dirichlet hyperparameters

Table I
NOTATION TABLE.

2) For each topic $t \in \Gamma = \{t_1, t_2, \ldots, t_T\}$, sample $\boldsymbol{\beta}_t$ over $\mathcal{V}$ from $\boldsymbol{\beta}_t \sim$ Dirichlet($\boldsymbol{\phi}$).
3) For each document d with $\boldsymbol{w}_d$ from the corpus:
   a) Sample a distribution over observed tags from $\boldsymbol{\theta}_d \sim$Dirichlet($\boldsymbol{\alpha}$).
   b) For ith word in document d
      i) Sample a tag $e \sim$ Multinomial($\boldsymbol{\theta}_d$).
      ii) Sample a topic $z \sim$ Multinomial($\boldsymbol{\gamma}_e$), a Multinomial probability conditioned on current tag assignment e.
      iii) Sample a term $w \sim$ Multinomial($\boldsymbol{\beta}_z$), a Multinomial probability conditioned on current topic assignment z.

The graphical model in Figure 1 demonstrates our model. Since $\boldsymbol{\delta}_d$ is already observed, the selection of it from Δ is not mathematically modeled, but for completeness, it is kept in the graphical model.

B. Learning Parameters

Given a document with words $\boldsymbol{w}_d$, the associated tags $\boldsymbol{\delta}_d$, and all the hyperparameters, we would like to compute the posterior distribution of the latent variables

$$P(\boldsymbol{e}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho}) = \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z} | \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{\sum_{\boldsymbol{e}} \sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}, | \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}.$$

However, this posterior distribution is not computable, because the denominator is intractable to compute. In the LDA literature, researchers mainly use two types of approaches to indirectly tackle this problem: 1) variational inference [4], and 2) Monte Carlo Markov chain (MCMC) sampling [22]. Our work adopts Gibbs sampling, a MCMC sampling method.

To build the Gibbs sampler, we require the joint distribution of observed words and their tag and topic assignments $P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})$. This joint distribution can be factorized as below

$$P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho}) = P(\boldsymbol{w}|\boldsymbol{\phi}, \boldsymbol{z}) \cdot P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\rho}), \quad (1)$$

based on the independence of variables. We analyze the two terms of the right part of Equation 1 separately. The first term is the same as the LDA model. We simply write down the derivations as follows and refer interested readers to references [23] and [22] for detailed explanations.

$$P(\boldsymbol{w}|\boldsymbol{\phi}, \boldsymbol{z}) = \prod_{t=1}^{T} \frac{B(\boldsymbol{n}_t + \boldsymbol{\phi})}{B(\boldsymbol{\phi})},$$

where $\boldsymbol{n}_t = (n_t^1, n_t^2, \cdots, n_t^{|V|})$ is a vector of length $|\mathcal{V}|$ consisting of the number of each term assigned to topic t, and $B(\cdot)$ is a multinomial beta function.

Now we turn to computing the second term in Equation 1. The second term can be further factorized to $P(\boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\rho}) = P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})P(\boldsymbol{e}|\boldsymbol{\alpha})$ applying Bayes rule and independence assumption. Let us first look at $P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})$. We notice that $P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})$ can be obtained after integrating out $\boldsymbol{\gamma}$:

$$P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho}) = \int_{\boldsymbol{\gamma}} P(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{e})P(\boldsymbol{\gamma}|\boldsymbol{\rho})d\boldsymbol{\gamma}. \quad (2)$$

For word i in document d, given its tag assignment e_{di}, $P(z_{di} = t|e_{di})$ is a multinomial distribution with parameter $\gamma_{e_{di}t}$. So we can obtain

$$P(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{e}) = \prod_{p=1}^{L_d} \prod_{t=1}^{T} \gamma_{pt}^{n_p^t}.$$

We already assume $P(\boldsymbol{\gamma}|\boldsymbol{\rho})$ follows a Dirichlet distribution. Substitute $P(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{e})$ and $P(\boldsymbol{\gamma}|\boldsymbol{\rho})$ in Equation 2 and apply Dirichlet integrals:

$$P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho}) = \prod_{p=1}^{L_d} \frac{B(\boldsymbol{n}_p + \boldsymbol{\rho})}{B(\boldsymbol{\rho})},$$

where $\boldsymbol{n}_p = (n_p^1, n_p^2, \cdots, n_p^T)$ contains the occurrence of each topic assigned to tag p. The derivation of the tag distribution $P(\boldsymbol{e}|\boldsymbol{\alpha})$ is quite similar to that of $P(\boldsymbol{z}|\boldsymbol{e}, \boldsymbol{\rho})$ via integrating out $\boldsymbol{\theta}$. Therefore, the derivation similarly yields

$$P(\boldsymbol{e}|\boldsymbol{\alpha}) = \prod_{d=1}^{M} \frac{B(\boldsymbol{n}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

and $\boldsymbol{n}_d = (n_d^1, n_d^2, \cdots, n_d^{L_d})$ contains the occurrences of each tag presenting in document d. Note that n_d^p will always be zero, if tag p is not associated with document d, i.e., the tag distribution per document must be over the associated observed tags only. Finally, the joint distribution of Equation 1 can be written down:

$$P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho}) = \prod_{t=1}^{T} \frac{B(\boldsymbol{n}_t + \boldsymbol{\phi})}{B(\boldsymbol{\phi})} \cdot \prod_{p=1}^{L_d} \frac{B(\boldsymbol{n}_p + \boldsymbol{\rho})}{B(\boldsymbol{\rho})} \cdot \prod_{d=1}^{M} \frac{B(\boldsymbol{n}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}. \quad (3)$$

In Gibbs sampling, the value of each variable is sampled sequentially conditioned on the current values of all other variables. Therefore the update equation that the sampler uses to update the topic and tag assignment for the ith word in document d is a conditional distribution:

$$P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) = \frac{P(\boldsymbol{w}, \boldsymbol{e}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}{P(\boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-|\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\rho})}, \quad (4)$$

where the subscript $-$ indicates the current updating assignments of topic and tag of word w_{di} are excluded. After we obtain the joint distribution as Equation 3, the update equation is given by

$$P(e_{di} = p, z_{di} = t|\boldsymbol{w}, \boldsymbol{e}_-, \boldsymbol{z}_-, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\phi}) \propto \frac{n_{t-}^{w_{di}} + \phi}{\sum_w^{\mathcal{V}} (n_{t-}^w + \phi)} \cdot \frac{n_{p-}^t + \rho}{\sum_k^{\Gamma} (n_{p-}^k + \rho)} \cdot \frac{n_{d-}^p + \alpha}{\sum_l^{\boldsymbol{\delta}_d} (n_{d-}^l + \alpha)}. \quad (5)$$

In Equation 5, n_{t-}^w is the count of term w under topic t excluding the current topic assignment of this term. Very similarly, n_{p-}^k denotes the count of tag p assigned to topic k excluding the current tag assignment to the topic. n_{d-}^l is the count of words in document d assigned to tag l excluding the current word. The explanation of the update equation is straightforward. The first term in the right part of Equation 5 represents the probability of term w_{di} under topic t with prior ϕ, the second term is the probability of topic t under tag p with prior ρ, while the last term is the probability of tag p in document d with prior α. Thus, the current topic and tag assignment to a word is proportional to the tag proportion in the document, the topic proportion under the tag, and the term proportion under the topic.

To estimate these three multinomial parameters after sampling, compute their expectations in the Dirichlet distributions. The estimate of topic-term distribution $\boldsymbol{\beta}$ is

$$\hat{\beta}_{tw} = \frac{n_t^w + \phi}{\sum_w^{\mathcal{V}} (n_t^w + \phi)},$$

where the estimation is identical to the LDA model. The estimates of document-tag distribution $\boldsymbol{\theta}$ and tag-topic distribution $\boldsymbol{\gamma}$ can be derived similarly as follows:

$$\hat{\theta}_{dp} = \frac{n_d^p + \alpha}{\sum_l^{\boldsymbol{\delta}_d} (n_d^l + \alpha)},$$

$$\hat{\gamma}_{pt} = \frac{n_p^t + \rho}{\sum_k^{\Gamma} (n_p^k + \rho)}.$$

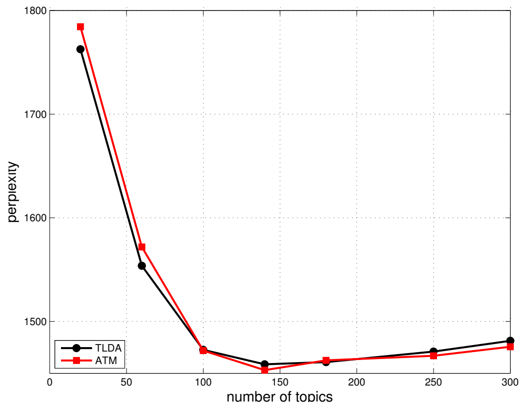

Figure 2. Perplexity of the test data for subset I with one hashtag per tweet.

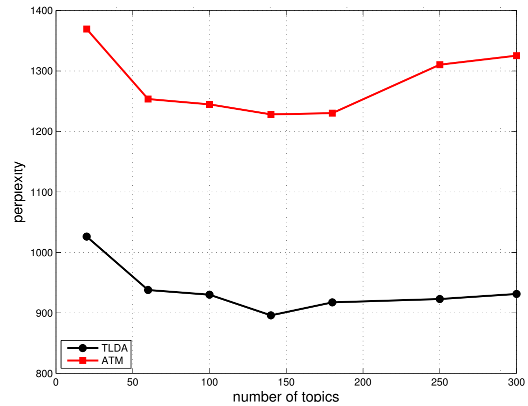

Figure 3. Perplexity of the test data for subset II with two or more hashtags per tweet.

C. Model Evaluation

We use the testing perplexity to evaluate the model and to profile the performance variation of the model as the number of topics varies. We also compare our model with the author-topic model [19]. The perplexity of the test data is expressed as:

$$Perplexity(\mathcal{D}_{test}) = \exp\left[-\frac{\log(P(\boldsymbol{w}|\mathcal{D}_{test}))}{N_{\mathcal{D}_{test}}}\right],$$

where $P(\boldsymbol{w}|\mathcal{D}_{test})$ is the likelihood and given by

$$P(\boldsymbol{w}|\mathcal{D}_{test}) = \prod_{w}\sum_{z\in\Gamma}\sum_{e\in\Delta} P(w|z)P(z|e)P(e|d).$$

$N_{\mathcal{D}_{test}}$ is the total number of words in the test data. After obtaining the trained model, it is employed to perform inference on the test data and obtain the likelihood. Perplexity measures the generalization performance of the model, with lower scores indicating better generalization.

The data used here is the TREC 2011 microblog dataset, details of which are introduced in Section IV. We note that, when there is only one tag in the document, the third term in Equation 5 has no influence on the sampling and makes our model the same as the author-topic model. To better investigate the difference between these two models, we split the data into two subsets: 1) subset I including tweets with exactly one hashtag; 2) subset II containing tweets with more than one hashtag. The experiment setup is that 10% of each subset is drawn randomly as the corresponding testing data and the remainder is used to train. The number of topics is varied from 20 to 300. As shown in Figure 2, the perplexity of TLDA and author-topic model is very close for tweets with one hashtag. However for tweets with multiple hashtags (Figure 3), TLDA demonstrates a better generalization with a lower perplexity.

IV. Experiments and Empirical Studies

To demonstrate the capability of our proposed method, we conducted experiments on two sets of tweet samples. In our experiments, we extracted hashtags from individual tweets as their corresponding tags for TLDA.

The first dataset contains retail store related tweets. To compile the dataset, we first collected tweets through Twitter's public "Garden-hose" API, which yields one percent samples. To select hashtags of similar nature, we filtered tweets containing hashtags of 12 well-known retail stores in the US: #macys, #belk, #nordstrom, #dillards, #neimanmarcus, #target, #walmart, #sears, #kohls, #jcpenney, #bestbuy, and #staples. The resulting dataset contains 3500 tweets posted from May 2012 to January 2013, with the retail store as the corresponding tag.

The second dataset is drawn from the TREC 2011[5] microblog data, which contains 16 million tweets sampled between January 23rd and February 8th, 2011. In this dataset, there are around 1.78 million tweets containing at least one hashtag. Among tweets containing hashtags, 19.9% uses two or more hashtags. We ranked and selected the 200 most frequent hashtags. We then filtered out the hashtags that mainly appeared in non-English tweets, which left 161 hashtags and 150K tweets; nearly 22.0% of these tweets have more than one hashtag. For the purpose of cleaning the data, the words in the tweets are stemmed by utilizing the NLTK package [24], and the resulting vocabulary size is 21139.

To facilitate the interpretation of the topic results of tag-topic distributions from the TLDA model, we further apply visualizations so that our modeling results are clear to users. In the following subsections, we describe our experimental methods, visualization tools, and findings.

A. Addressing Challenge 1: Understanding Hashtags and Their Contexts

Interpreting Hashtags through Topics: To understand the hashtags in the TREC2011 dataset, 140 topics were extracted from the corpus. The number of topics was determined based on the perplexity.

To illustrate our results, we selected a few hashtags belonging to different categories (sports, politics, world) and list the topic with the highest probability for each hashtag

[5]http://trec.nist.gov/data/tweets/

Categories	Hashtag	Topic with the highest probability
Politics	#tcot	teaparty gop obama ocra tlot sgp obamacare bill palin twisters tpp vote repeal reagan repeal vote tpp conservative gore republicans constitution
Sport	#steeler	game yellow afc black championship beat fan lose nfl nyjets steelernation jersey blackandyellow sunday twitpic pittsburgh picks
World	#mubarak	egypt jan mubarak people pro square thugs protesters internet protests egyptian cnn tahrir watch government police news anti egyptians feb violence
	#tahrir	square thugs pro jan liberation clashes blessed cairo live protesters aje egyptian breaking mubarak armed twitpic

Table II
THE LIST OF HIGH PROBABILITY TERMS OF THE HIGHEST PROBABILITY TOPIC FOR EACH HASHTAG.

in Table II. At a first glance, one can see that some hashtags are difficult to interpret. The hashtag #tcot (Top Conservatives on Twitter) for example, it is a coalition of conservatives on the Internet. However, without reading the topic, the meaning of the hashtag is hard to infer. The keywords of the highest probability topic for hashtag #tcot capture terms related to conservative American political parties. Another interesting example is the hashtag #tahrir, which is an arabic word meaning liberation. The topic results for the hashtag clearly indicate that the hashtag refers to a specific location – Tahrir Square in Cairo, Egypt. In addition, the topic also contains information regarding possible protests in Tahrir Square that might warrant further investigation. The examples demonstrate that topic terms could greatly help with the interpretation of the hashtags, which could otherwise be difficult to decode through reading the hashtags alone or merely examining a few tweets containing the hashtag.

Comparing Hashtags through Topic Distribution: Through demonstrating tag-topic results from the TREC2011 data, we have showcased that the highest probability topic for each hashtag can be used to understand and distinguish the meaning of the hashtags. However, just like a document could exhibit multiple topics in the LDA model, a tag in the TLDA model could also be described by multiple topics. To depict the tag distribution over topics and distinguish single-topic vs. multi-topic tags, we apply a visualization tool to facilitate the analysis. For this experiment, we used the retail store tweet collection with the store names as tags for each tweet. 25 topics were extracted to depict the 12 retail store hashtags. In the visualization (Figure 4), each (colored) vertical axis denotes a topic, and the highlighted black polyline running across all topics represents a tag. Therefore, the distribution of one tag over all topics is intuitively depicted in the visualization. Furthermore, the visualization is interactive in the sense that users could brush on any topic to select tags with high probability on the topic. As we will illustrate later, some tags have only one topic with high probability while others may have multiple topics with equally high probabilities.

Figure 4 illustrates two retail stores that seem to exhibit

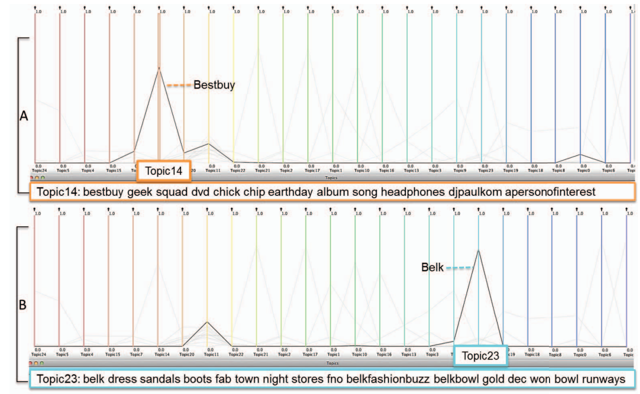

Figure 4. Visualization of two retail stores with a single dominant topic. Each vertical (colored) axis represents one topic, while the black polyline denotes the distribution of a certain store over all topics. The top visualization shows the topic distribution of the electronics retail store Bestbuy and the topic terms for the topic with the highest probability. The bottom visualization illustrates the topic distribution of the department store Belk.

focused discussions (evolving around a single topic) on their products and promotions. In particular, the tweets related to the electronics retail store *BestBuy* (Figure 4 A) center around the orange topic which includes keywords (in the orange rectangle) such as “geeksquad”, “dvd”, “song”, “headphones”, and popular on-sale albums such as “apersonofinterest” by “djpaulkom”. Similarly, tweets related to the fashion retail store *Belk* (Figure 4 B) focuses more on clothing and fashion including “dress”, “sandals”, “boots”, “belkfashionbuzz”, and a Belk sponsored event “belkbow”. The example illustrates that the visualization enables users to easily identify retail stores with discussions on Twitter mainly focusing on a single dominant topic.

While the aforementioned retail store hashtags present distributions that are centered around one major topic, others exhibit more diverse topic distributions. Figure 5 illustrate two retail stores with tweets covering different topics and events. Figure 5A shows that the world’s largest retail corporation *Walmart* exhibits four almost equally important topics, ranging from blackfriday shopping to Walmart workers strike to the recent Walmart bribery scandal in Mexico. In this example, the topics not only cover shopping related events, but also public-relations related events. All four topics contribute to understanding the corporation’s profile on social media, and people’s opinion towards certain events related to the corporation. Another example of the retail store (hashtag) exhibiting multiple topics is *Macy’s* department store. The two major topics for hashtag #macys cover the thanksgiving day parade in New York City, celebrity product line promotions, sweepstakes sponsored by Macy’s, etc. Digging deeper into the volume of tweets related to certain topics/events, we found that the annual Macy’s Thanksgiving Day Parade in New York City attracts the most amount of traffic containing #macys on Twitter, followed by celebrities promoting their own product lines. For example, the celebrity product lines such as the

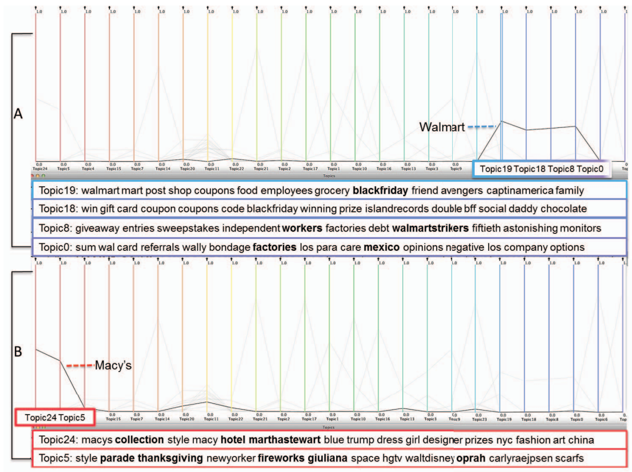

Figure 5. Visualization of two retail stores exhibiting multiple topics. Each vertical axis represents one topic, while the black polyline denotes the distribution of a certain store over all topics. The top visualization shows the topic distribution of Walmart which has 4 almost equally important topics. The keywords related to certain events described in Section IV-A are highlighted. The bottom visualization illustrates the topic distribution of the department store Macy's.

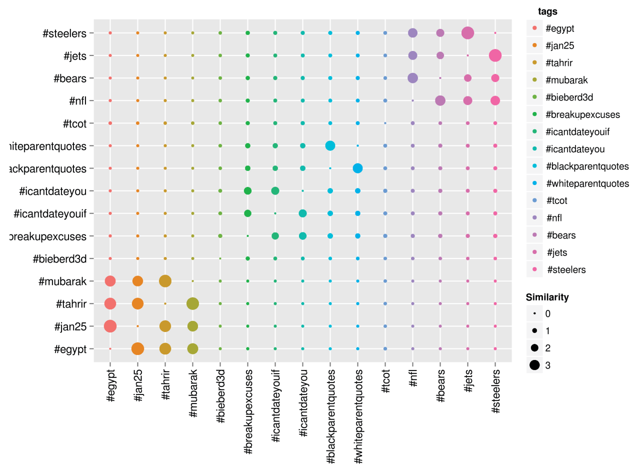

Figure 6. Visualization of the hashtag similarity matrix. Dots with larger size represents higher similarity; the larger the radius, the greater the similarity. Similarity values are scaled and self similarities on the diagonal are removed for better display.

marthastewart collection and the giuliana hotel collection get a lot of coverage on Twitter. Such findings could potentially inform the corporation's marketing strategy on social media.

In summary, the TLDA results facilitate the understanding of the meaning of the hashtags and the contexts in which they are used. Furthermore, the companion visualization enables the analysis of hashtags based on topic distributions, and supports the discovery of single vs. diverse-topic hashtags. The visualization coupled with the TLDA model not only makes the complex modeling results legible to average users, but also enables users to explore and compare the meaning of the hashtags, thus contributing to the overall understanding of the hashtag phenomenon in the social media.

B. Addressing Challenge 2: Discovering the Relationships between Hashtags

In addition to understanding the meaning and contexts of the hashtags, knowing the relationships among hashtags also contributes to proper categorization of tweets using hashtags. As mentioned in Section I-B, different hashtags are commonly used together to depict different aspects of a particular event. However, even with the topic results for each hashtag, the discovery of similar hashtags through a manual process is still a challenging and laborious task. Such a task can be adequately addressed through combining the TLDA results with proper distance measures.

As mentioned in Section II, the difference between TLDA and a previously proposed similar model, partially labeled LDA, lies in that all tags are modeled as a distribution over a shared topic space. Therefore, the computation of the similarity between every pair of tags becomes straightforward. Since a tag is in the form of a probabilistic distribution over topics, we utilize the Hellinger distance [25] to measure the distance between a pair of tags. Given any two discrete probability distribution $P = (p_1, p_{2,}, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, the Hellinger distance is defined as $H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i(\sqrt{p_i} - \sqrt{q_i})^2}$. Therefore a distance matrix can be constructed by measuring the distance between every pair of hashtag distributions over topics. We invert the distances between the hashtags to get the similarity measurement.

To enable users to discover similar hashtags, we visualize the similarities using a matrix metaphor. Figure 6 illustrates the similarities between pairs of hashtags in the TREC2011 data. At a glance, one can detect two major clusters in the visualization (lower left and upper right), with larger dot size denoting higher similarity. The upper right cluster in Figure 6 highlights a group of hashtags related to football teams and events. The two biggest dots in pink denote the high similarity between #steelers and #jets are high. Indeed, there are a lot of discussions in the tweets about the game between the New York Jets and Pittsburgh Steelers on that Sunday weekend.

A more interesting and intriguing example is the lower left cluster of hashtags including #egypt, #tahrir, #mubarak, #jan25. Coupling the results with the main topic for hashtag #tahrir and #mubarak (Table II), which includes keywords such as protest, protestors, anti, government, police, one can infer that there may be a big protest event occurring at Tahrir Square in Egypt on January 25. Indeed, according to Wikipedia, the event was the 2011 Egyptian Revolution against former president Hosni Mubarak, and over 50,000 protestors occupied Tahrir Square in Cairo[6]. The hashtags in the lower left cluster are related to one event and different hashtags are created to describe temporal, geospatial, and people information. This example illustrates that by coupling the hashtag similarity

[6]http://en.wikipedia.org/wiki/Tahrir_Square

results with the topics for hashtags, one can conduct deep analysis of events discussed on Twitter.

In summary, the above example illustrates that visually presenting the similarities among hashtags could help users identify groups of hashtags used to characterize similar topics or events. In addition, combining the hashtag similarity results with the hashtag-topic results supports the development of comprehensive understanding of events discussed on Twitter. Retrospective examination of the hashtags and tweet content provides an overview of what has been discussed on Twitter.

V. Conclusion

In this paper we focus on analyzing and understanding the hashtags used in social media from the content perspective. We consider two challenges: first, how to interpret hashtags, and second, how to find related hashtags? We propose a topic model, TLDA, to address these questions. The TLDA model learns the hidden topic structures for each hashtag. Since hashtags are modeled in a common topic space, we can further measure the similarities between every pair of hashtags. In the experiments, we apply our model to extract meaningful topics from tweets, and use visualization techniques to understand hashtags. We also show that our methods can help discover a group of hashtags created to describe one common event based on the topic similarity of the hashtags. Finally, while we use TLDA to study hashtags in this paper, TLDA is capable of handling other forms of discrete data with metadata serving as tags.

Acknowledgments

This research was partially supported by the National Science Foundation (NSF) under awards SBE-0915528, CCF-1048261, IIS-1019160, and the Army Research Office (ARO) under grant #A13-0154-001.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web* (WWW '10), 2010, pp. 591–600.

[2] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto, "Analyzing the dynamic evolution of hashtags on twitter: a language-based approach," in *Proceedings of the Workshop on Languages in Social Media* (LSM '11), 2011, pp. 58–65.

[3] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th International Conference on World Wide Web* (WWW '11), 2011, pp. 695–704.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (WebKDD/SNA-KDD '07), 2007, pp. 56–65.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '03), 2003, pp. 137–146.

[7] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (WSDM '10), 2010, pp. 241–250.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *International AAAI Conference on Weblogs and Social Media* (ICWSM '10), 2010, pp. 10–17.

[9] K. Y. Kamath, J. Caverlee, Z. Cheng, and D. Z. Sui, "Spatial influence vs. community influence: modeling the global spread of social media," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (CIKM '12), 2012, pp. 962–971.

[10] O. Tsur and A. Rappoport, "What's in a hashtag? content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (WSDM '12), 2012, pp. 643–652.

[11] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: topic tracking in tweet streams," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '11), 2011, pp. 422–429.

[12] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto, "Finding trendsetters in information networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '12), 2012, pp. 1014–1022.

[13] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!" in *International AAAI Conference on Weblogs and Social Media* (ICWSM '11), 2011, pp. 538–541.

[14] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING '10), 2010, pp. 241–249.

[15] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '99), 1999, pp. 50–57.

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

[17] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *International AAAI Conference on Weblogs and Social Media* (ICWSM '10), 2010, pp. 130–137.

[18] D. M. Blei and J. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems* (NIPS), 2008, pp. 121–128.

[19] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (UAI '04), Arlington, Virginia, United States, 2004, pp. 487–494.

[20] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP), Singapore, August 2009, pp. 248–256.

[21] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '11), 2011, pp. 457–465.

[22] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[23] G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Tech. Rep., 2009.

[24] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[25] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen (german)," *Journal für die Reine und Angewandte Mathematik*, vol. 136, pp. 210–271, 1909.