Microblog Sentiment Topic Model

Aman Ahuja
BITS Pilani K.K. Birla Goa Campus
Zuarinagar, Goa, India
ahujagwl@gmail.com

Wei Wei
Carnegie Mellon University
Pittsburgh, PA, USA
weiwei@cs.cmu.edu

Kathleen M. Carley
Carnegie Mellon University
Pittsburgh, PA, USA
kathleen.carley@cs.cmu.edu

***Abstract*—With the prevalence of social media, such as Twitter, short-length text like microblogs have become an important mode of text on the Internet. In contrast to other forms of media, such as newspaper, the text in these social media posts usually contains fewer words, and is concentrated on a much narrower selection of topics. For these reasons, traditional LDA-based sentiment and topic modeling techniques generally do not work well in case of social media data. Another characteristic feature of this data is the use of special meta tokens, such as hashtags, which contain unique semantic meanings that are not captured by other ordinary words. In the recent years, many topic modeling techniques have been proposed for social media data, but the majority of this work does not take into account the specialty of tokens, such as hashtags, and treats them as ordinary words. In this paper, we propose probabilistic graphical models to address the problem of discovering latent topics and their sentiment from social media data, mainly microblogs like Twitter. We first propose MTM (*Microblog Topic Model*), a generative model that assumes each social media post generates from a single topic, and models both words and hashtags separately. We then propose MSTM (*Microblog Sentiment Topic Model*), an extension of MTM, which also embodies the sentiment associated with the topics. We evaluated our models using Twitter dataset, and experimental results show that our models outperform the existing techniques.**

I. Introduction

The rapid growth of Internet in the recent years has led to the growth of social media websites like Twitter. People use these micro-blogging platforms to post about different aspects of their life and about the things happening in their surroundings. Many companies also use these platforms to promote their products. Because of their widespread use, the vast quantity of data available on these social media websites can be used in a variety of ways. For example, it can be used to discover trending topics, estimate the public support toward different candidates in a presidential election, or gauge the performance of a product in the market.

However, mining topics and their sentiments from social media posts, such as microblogs, differs from sentiment analysis in longer and more detailed text, such as in newspaper articles and blogs, in a variety of ways. This is mainly because in such posts, sentiment is usually conveyed using short and concise messages. Also, the text used in these posts has many abbreviations and misspelled words, since the information needs to be conveyed using only few words/characters. Since these posts (or *tweets in Twitter*) are generally short in length, mainly because of the constraint on the number of characters allowed (eg., *140 in Twitter*), it is highly likely that each post is associated with one topic and one sentiment. This is in contrast to other forms of text, such as newspaper articles and product reviews, where the text is more elaborate and deals with the subject matter in more detail. A final feature of social media data is the use of special tokens like *hashtags*. A hashtag is a token in which a hash (#) character precedes a word, which can be used to link the content in a post to a specific topic, such as a major event or crisis. For example, all the tweets on Twitter with *#Halloween* suggest that the content of the tweets will be highly related to Halloween, while *#Steelers* are likely to link to football events. It is also observed that posts related to a major topic/event or advertising campaigns generally contain more hashtags than posts that are related to daily life of the users. All these characteristics of social media text require unique methods to discover topics and sentiments.

Although many techniques, such as JST model [1] and ASUM [2], have been proposed to discover topics and sentiments from text data, their capability to model short-length text, like tweets, with special tokens such as hashtags remains unexplored. In this work, we propose two generative models: Microblog Topic Model (*MTM*) and Microblog Sentiment Topic Model (*MSTM*), which take into account these characteristics of microblogs by modeling words and hashtags separately, and assign a single topic to each post. This is in contrast to other previously proposed models that model each document/post as a mixture of topics and ignore the semantic differences between words and hashtags. MSTM is an extension of MTM, which also incorporates the sentiment associated with social media posts at the document level, and also gives a set of words and hashtags with different sentiment polarities for each topic. The user-based modeling of topics in MTM and MSTM also allows to discover the interest-distribution of social media users. We evaluated these models both qualitatively and quantitatively against several baseline models using a Twitter dataset and found that these models outperform the existing techniques.

The rest of this paper is organized as follows: Section 2 gives an overview of the existing techniques related to the proposed models. Section 3 describes the two proposed

models and outlines the solutions to the Bayesian inference. Sections 4 and 5 describe the experimental setup and the results obtained in the experimental evaluation, followed by Section 6, which concludes the paper.

II. Related Work

A. Topic Modeling

The success of topic modeling and its potentiality in mining latent representations of text has gained much attention in recent years. One of the earliest works in the field of topic modeling was the probabilistic Latent Semantic Indexing (pLSI), proposed by Hoffman [3], which models a document as a mixture of topics. Since pLSI is based on the likelihood principle, there is no generative process for determining the document-topic distribution, which leads to problems while assigning probabilities to documents outside the training set. Most of the recent work in the field of topic modeling is based on the Latent Dirichlet Allocation (LDA) [4], which overcomes the shortcomings of pLSI by assuming a hierarchical Bayesian dependency between the documents, topics and words. In LDA, each document in the corpus can be regarded as a mixture over topics, and topics as a mixture over words. A variant of this model is the SLDA model [2], which constrained the words in a single sentence to belong to a single topic.

B. Sentiment Analysis

Sentiment analysis of social media data remains a key area of research. Many techniques have been proposed to detect the sentiment of Twitter messages. The earlier work in the field of sentiment analysis of Twitter messages used a naive Bayes classifier, as proposed in [5]. Later, an analysis of the usefulness of different lexical features in sentiment classification of Twitter messages was presented in [6]. These techniques are supervised in nature and can be used for sentiment classification of individual tweets. The potentiality of sentiment analysis of Twitter data is also studied in [7], [8], where the authors use Twitter messages to analyze public opinion in political elections and use keyword-based features to determine the public sentiment in political elections. Unlike most of these techniques, we aim to determine the sentiment of documents as well as topics.

Two sentiment models are closely related to our work. The JST model [1] is one of the earliest attempts that models sentiments and topics using a probabilistic graphical model. It models each document as a mixture over topics and sentiments. A sentiment label, in addition to the topic label, is used to jointly select the words. Since sentiments can be positive or negative, this model gives positive and negative word distributions for each topic. However, since this model also has a LDA-like hierarchy, it does not fit well in case of short-length social media text. The ASUM model [2] overcomes this problem by assigning a single topic and sentiment to all the words in a sentence. However, this assumption does not work well in case of microblog data, where a tweet has multiple short sentences that usually belong to a single topic. In tweets, a single document, rather than a sentence, is likely to belong to a single topic. Another problem with JST and ASUM is that both these models ignore the unique semantic property of hashtags and treat them as words, when used for social media text.

C. Modeling Social Media Data

A number of techniques based on LDA have been proposed for social media data. Because of the limitations of LDA in modeling short-length text, techniques such as tweet pooling [9] have been proposed to improve topic modeling in tweets. [10] discussed the application of the author-topic model [11], which models each author as a distribution over topics or interests. [12] proposed a variant of the author-topic model, which aims to discover interest distribution for different users. Unlike these models, the Twitter-LDA model [13] incorporates the small-length property of tweets and assigns a single topic to all the words in a tweet. Additionally, topic models can be used to discover events, which can be defined as clusters of documents with similar content, time, and spatial coordinates, from social media data, as proposed in [14]. The models proposed in this paper are largely inspired from Twitter-LDA and ASUM, based on the assumption that topic assignment at the document level can give a more realistic modeling of text in microblogs.

III. Models

A. Microblog Topic Model (MTM)

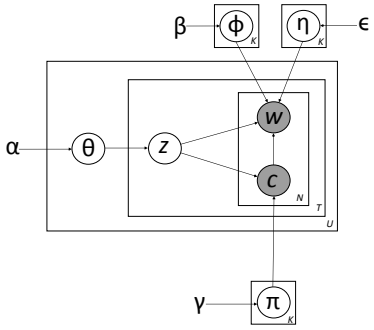

Figure 1. Plate notation of MTM

1) Model Definition: MTM is a generative model to discover latent topics in short-length social media posts that contain both words and hashtags. The plate notation of this model is illustrated in Figure 1. Unlike LDA, which models each document as a mixture of topics, MTM models each user *u* as a mixture over topics. In addition, each document *t* by user *u* in MTM can only have one topic z_{ut}, which is selected based on the user-topic distribution θ_u. This coarse granularity assumption generally holds well in case of social media text, such as Twitter, where the length of the

messages is restricted to 140 characters. Also, since words and hashtags are separate entities with different semantic meanings, we use a binary category variable c_{utn} for each token in the post (*u, t*). The model uses a dependency between the topic variable *z* and the category *c* of the tokens used, since it is generally observed that some topics that are related to advertising campaigns, or a major event, have a higher proportion of hashtags than other common topics. The value of this category variable indicates whether the the corresponding token is a word or hashtag. More specifically, if $c_{utn} = 1$, the token w_{utn} is considered to be a word, and is drawn from the topic-word distribution ϕ_z. Otherwise, if $c_{utn} = 0$, w_{utn} is considered to be a hashtag and is drawn from the topic-hashtag distribution η_z.

The overall generative process of MTM is as follows:

- For each topic *k*,
  - Draw a category distribution $\pi_k \sim Beta(\gamma)$
  - Draw a word distribution $\phi_k \sim Dir(\beta)$
  - Draw a hashtag distribution $\eta_k \sim Dir(\epsilon)$
- For each user *u*,
  - Draw user-topic distribution $\theta_u \sim Dir(\alpha)$
  - For each post *t* by the user *u*,
    - Choose a topic $z_{ut} \sim Multi(\theta_u)$
    - For each token *n* in the post (*u, t*),
      - Choose a category $c_{utn} \sim Ber(\pi_{z_{ut}})$
      - Draw a word/hashtag w_{utn} as follows:

$$w_{utn} \sim \begin{cases} Multi(\phi_{z_{ut}}), & \text{if } c_{utn} = 1 \\ Multi(\eta_{z_{ut}}), & \text{if } c_{utn} = 0 \end{cases}$$

2) Inference: To infer the latent parameter *z* for each post (*u, t*), we use collapsed Gibbs sampling technique [15]. We first integrate out the model parameters: θ, π, ϕ, and η. We then sample z_{ut} using the conditional probability distribution of *z* according to the following equation:.

$$P(z_{ut} = k | \boldsymbol{Z_{-ut}}, \boldsymbol{C}, \boldsymbol{W}, \alpha, \beta, \epsilon, \gamma) \propto \frac{N_{u,(.)}^{k,-ut} + \alpha_k}{\sum_{i=1}^{K} N_{u,(.)}^{i,-ut} + \alpha_i} \frac{\prod_{r \in W_{ut}} \prod_{j=0}^{n_{ut}^{w,r}-1} (M_{w_r}^{k,-ut} + \beta_r + j)}{\prod_{j=0}^{n_{ut}^{w,(.)}-1} ((\sum_{r=1}^{W} M_{w_r}^{k,-ut} + \beta_r) + j)} \frac{\prod_{r \in H_{ut}} \prod_{j=0}^{n_{ut}^{h,r}-1} (M_{h_r}^{k,-ut} + \epsilon_r + j)}{\prod_{j=0}^{n_{ut}^{h,(.)}-1} ((\sum_{r=1}^{H} M_{h_r}^{k,-ut} + \epsilon_r) + j)} \frac{\prod_{r=0}^{1} \prod_{j=0}^{n_{ut}^{r,(.)}-1} (C_r^{k,-ut} + \gamma_r + j)}{\prod_{j=0}^{n_{ut}^{(.),(.)}-1} ((\sum_{r=0}^{1} C_r^{k,-ut} + \gamma_r) + j)} \quad (1)$$

After sampling *z*, the model parameters that were integrated out can be recovered using the equations 3, 5, 7, and 9.

B. Microblog Sentiment Topic Model (MSTM)

1) Model Definition: Although MTM accommodates reasonable assumptions to model microblog posts and discover latent topics in these posts, it cannot model the sentiment associated with these latent topics.

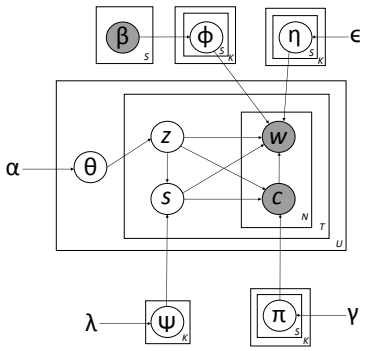

Figure 2. Plate notation of MSTM

MSTM is an extension of MTM, which also incorporates the sentiment. The plate notation of MSTM is illustrated in Figure 2. To model sentiments, MSTM has a sentiment variable *s* at the document level, which represents the sentiment polarity of the post (*u, t*). This is drawn from the sentiment distribution ψ_z of the topic *z* associated with the post. For each token *n* in the post (*u, t*), after determining the category c_{utn} (word or hashtag) of the token, it is drawn from the respective topic-sentiment-word distribution $\phi_{k,s}$ or topic-sentiment-tag distribution $\eta_{k,s}$, based on the value of the c_{utn}. The prior sentiment polarity of words can be incorporated into MSTM, in the values of the hyper-parameter β, based on the assumption that a word with positive polarity is more likely to appear in a positive sentiment topic, whereas a word with negative polarity is more likely to appear in a negative sentiment topic.

Intuitively, the model can be described as: when a user *u* decides to write a post *t*, he (i) decides the post's topic z_{ut} based on his interest distribution θ_u, (ii) sentiment s_{ut} and the type (word or hashtag) of tokens in the post, and (iii) chooses the tokens w_{utn} based on the topic z_{ut} and sentiment s_{ut} of the post and the category c_{utn} of the tokens.

The generative process of MSTM is as follows:

- For each topic *k*,
  - Draw a sentiment distribution $\psi_k \sim Dir(\lambda)$
  - For each sentiment label *s*,
    - Draw a word distribution $\phi_{k,s} \sim Dir(\beta_s)$
    - Draw a hashtag distribution $\eta_{k,s} \sim Dir(\epsilon)$
    - Draw a category distribution $\pi_{k,s} \sim Beta(\gamma)$
- For each user *u*,
  - Draw user-topic distribution $\theta_u \sim Dir(\alpha)$
  - For each post *t* by the the user,
    - Choose a topic $z_{ut} \sim Multi(\theta_u)$
    - Choose a sentiment label $s_{ut} \sim Multi(\psi_{z_{ut}})$
    - For each token *n* in the post (*u, t*),
      - Choose a category $c_{utn} \sim Ber(\pi_{z_{ut}, s_{ut}})$

· Draw a word/hashtag w_{utn} as follows:

$$w_{utn} \sim \begin{cases} Multi(\phi_{z_{ut},s_{ut}}), & \text{if } c_{utn} = 1 \\ Multi(\eta_{z_{ut},s_{ut}}), & \text{if } c_{utn} = 0 \end{cases}$$

2) Inference: Similar to MTM, we use collapsed Gibbs sampling for inference in MSTM. The model parameters θ, ψ, π, ϕ, and η can be integrated out easily using the Dirichlet-Multinomial conjugacy. After this, the only latent parameters left in the model are z and s, which can be sampled as follows:

$$\begin{aligned} P(z_{ut} = k, s_{ut} = p | \{\boldsymbol{Z}, \boldsymbol{S}\}_{-\boldsymbol{ut}}, \boldsymbol{C}, \boldsymbol{W}, \alpha, \lambda, \beta, \epsilon, \gamma) \\ \propto \frac{N_{u,(.)}^{k,-ut} + \alpha_k}{\sum_{i=1}^{K} N_{u,(.)}^{i,-ut} + \alpha_i} \frac{L^{k,p,-ut} + \lambda_p}{\sum_{s=1}^{S} L^{k,s,-ut} + \lambda_s} \\ \frac{\prod_{r \in W_{ut}} \prod_{j=0}^{n_{ut}^{w,r}-1} (M_{w_r}^{k,p,-ut} + \beta_{p,r} + j)}{\prod_{j=0}^{n_{ut}^{w,(.)}-1} ((\sum_{r=1}^{W} M_{w_r}^{k,p,-ut} + \beta_{p,r}) + j)} \\ \frac{\prod_{r \in H_{ut}} \prod_{j=0}^{n_{ut}^{h,r}-1} (M_{h_r}^{k,p,-ut} + \epsilon_r + j)}{\prod_{j=0}^{n_{ut}^{h,(.)}-1} ((\sum_{r=1}^{H} M_{h_r}^{k,p,-ut} + \epsilon_r) + j)} \\ \frac{\prod_{r=0}^{1} \prod_{j=0}^{n_{ut}^{r,(.)}-1} (C_r^{k,p,-ut} + \gamma_r + j)}{\prod_{j=0}^{n_{ut}^{(.),(.)}-1} ((\sum_{r=0}^{1} C_r^{k,p,-ut} + \gamma_r) + j)} \end{aligned} \quad (2)$$

Similar to MTM, the model parameters can be calculated using the equations 3, 4, 6, 8, and 10.

$$\theta_u^k = \frac{N_{u,(.)}^k + \alpha_k}{\sum_{i=1}^{K} N_{u,(.)}^i + \alpha_i} \quad (3) \qquad \psi_k^p = \frac{L^{k,p} + \lambda_p}{\sum_{s=1}^{S} L^{k,s} + \lambda_s} \quad (4)$$

$$\pi_k^c = \frac{C_c^k + \gamma_c}{\sum_{r=0}^{1} C_r^k + \gamma_r} \quad (5) \qquad \pi_{k,p}^c = \frac{C_c^{k,p} + \gamma_c}{\sum_{r=0}^{1} C_r^{k,p} + \gamma_r} \quad (6)$$

$$\phi_k^v = \frac{M_{w_v}^k + \beta_v}{\sum_{r=1}^{W} M_{w_r}^k + \beta_r} \quad (7) \qquad \phi_{k,p}^v = \frac{M_{w_v}^{k,p} + \beta_{p,v}}{\sum_{r=1}^{W} M_{w_r}^{k,p} + \beta_{p,r}} \quad (8)$$

$$\eta_k^v = \frac{M_{h_v}^k + \epsilon_v}{\sum_{r=1}^{H} M_{h_r}^k + \epsilon_r} \quad (9) \qquad \eta_{k,p}^v = \frac{M_{h_v}^{k,p} + \epsilon_v}{\sum_{r=1}^{H} M_{h_r}^{k,p} + \epsilon_r} \quad (10)$$

IV. Dataset

A. Text Data

Since the models proposed in this paper are specifically designed for short length social media, such as *tweets*, we use Twitter dataset containing nearly 2.4 million tweets by 11,509 users, collected using the Twitter Decahose API, within a one-month time period, from May 1, 2011 to May 31, 2011, to evaluate our models. The dataset we have is a 10% random sample of all the geo-tagged tweets that have spatial coordinates and fall into the spatial boundaries of United States, since we use an English sentiment lexicon as prior sentiment knowledge in MSTM, and most of the tweets from USA are in English.

These tweets were preprocessed using basic preprocessing techniques like removal of URLs, co-mentions, common stop words, and infrequent words (words that occurred less than two times in the entire corpus). Since emoticons are an essential part of sentiment classification, we removed punctuation marks from the tweets in such a way that all the valid emoticons and hashtags were preserved. Finally, we converted all the words and hashtags in the tweets to lowercase, followed by tokenization. After this, the dataset had 557,318 unique words and 100,445 unique hashtags.

B. Text Data with Sentiment Labeling

We use a set of 390 labeled tweets to serve as text level sentiment ground truth. Those labeled tweets are a subset of the dataset prepared by Go [16], which originally contains 500 human labeled sentiments. We deleted all the tweets

For all the terms shown in equations,

- for any dimension *d*, *(.)* denotes that the term is not limited to the specific value of *d*
- *-ut* denotes that the term excludes the current post (*u, t*)

Table I
Notations

U	number of users	T	number of posts/tweets
N	number of tokens in post	K	number of topics
S	number of sentiments	W	the size of word vocabulary
H	the size of hashtag vocabulary	z	topic
w	word	c	category (word or hashtag)
s	sentiment polarity	θ	user-topic distribution
π	topic-category distribution	ϕ	topic-word distribution
η	topic-hashtag distribution	ψ	topic-sentiment distribution
α	Dirichlet prior vector for θ	γ	Beta prior vector for π, $\pi_{k,s}$
β	Dirichlet prior vector for ϕ	β_s	Dirichlet prior vector for $\phi_{k,s}$
ϵ	Dirichlet prior vector for η, $\eta_{k,s}$	λ	Dirichlet prior vector for ψ

Table II
Auxillary Notations

$N_{u,t}^k$	number of times post *t* by user *u* is assigned topic *k*
$M_{w_r}^k$	number of occurrences of the r^{th} word from the word vocabulary in topic *k*
$M_{h_r}^k$	number of occurrences of the r^{th} hashtag from the hashtag vocabulary in topic *k*
C_r^k	total number of occurrences of tokens from category *r* in topic *k*
$n_{ut}^{c,r}$	number of occurrences of r^{th} token from category *c* in post (*u, t*)
W_{ut}	set of words present in post (*u, t*)
H_{ut}	set of hashtags present in post (*u, t*)
$L^{k,p}$	total number of posts in topic *k* with sentiment label *p*
$M_{w_r}^{k,p}$	number of occurrences of the r^{th} word from the word vocabulary in topic *k* with sentiment *p*
$M_{h_r}^{k,p}$	number of occurrences of the r^{th} hashtag from the hashtag vocabulary in topic *k* with sentiment *p*
$C_r^{k,p}$	total number of occurrences of tokens from category *r* in topic *k* with sentiment *p*

with neutral sentiment and 390 tweets with positive and negative labels were used in the experiment.

C. Sentiment Lexicon

To incorporate the prior sentiment polarity of words in MSTM, we used Vader [17] sentiment lexicon. This choice was based on the fact that Vader is specifically designed for words that frequently appear in social media posts, particularly Twitter, and is highly optimized for such datasets. Also, many of these commonly occurring words in tweets are present only in Vader, and cannot be found in any other sentiment lexicons, such as MPQA subjectivity corpus [18] or SentiWordnet [19]. Since we consider only positive and negative sentiments in our experiments, we separate out the positive and negative words from Vader based on the score assigned to them. After this, the sentiment lexicon had 3,300 positive words and 4,100 negative words.

V. Experimental Results

A. MTM

In order to evaluate MTM, we first need to determine the values of the hyper-parameters α, β, γ, and ϵ, which serve as prior information for the model. We use symmetric values for all the hyper-parameters, which were derived empirically. Specifically, we set $\alpha = 1$, $\beta = 0.05$, $\epsilon = 0.05$, and $\gamma = 5$. The model was run for 800 Gibbs sampling iterations, using different values for the number of topics K.

1) Qualitative Results: To demonstrate the qualitative results, we present the top 10 words and hashtags from two different topics ranked by their topic-word-distribution ϕ and topic-hashtag-distribution η. We can see from Table III that the first topic is related to the death of Osama Bin Laden, which happened on May 2, 2011, as it includes words such as "bin", "laden", "osama", and hashtags such as "#osama", "#obama", "#pakistan". The second topic is related to food since it contains words like "eat", "food", "chicken" and hashtags such as "#fattweet", "#yummy", "#delicious".

Table III
Top 10 words and #tags for two different topics obtained using MTM

T1:Words	T1:#tags	T2:Words	T2:#tags
bin	#binladen	eat	#fattweet
laden	#osama	good	#win
obama	#syria	food	#yum
osama	#news	chicken	#yummy
news	#obama	:)	#hungrytweet
dead	#pakistan	icecream	#hungry
death	#cnn	eating	#munchies
world	#usa	breakfast	#love
killed	#osamabinladen	cheese	#delicious
man	#dead	drink	#ny

2) Detecting Abnormal Topics Using Topic-Hashtag Distribution: Although the proportion of hashtags and words differs from topic to topic, extreme use of either may indicate spammers or abnormal users [20]. Figure 3 illustrates the ratio of hashtags and words for each topic (i.e., $\pi_{k,0}/\pi_{k,1}$) in the experiment where the number of topics $K = 50$. We see that most topics have a ratio of *0.25*, i.e., four words associated with each hashtag in a tweet. However, some topics have an abnormally high value of this ratio. We have listed the top words and hashtags, along with a sample tweet for these topics, in Table IV. Due to Twitter's policy on user privacy, we processed the example tweets so that no personal identifiers can be accessed. Topic 7, for example, has a ratio of 0.78; its top words include "jobs", "manager", and "sales", and hashtags like "#jobs", "#retail", and "#tweetmyjobs". This indicates that it is an advertising tweet, which tries to attract people by making them click on links that might get them jobs. The abnormal use of hashtags triggers Twitter's automatic matching algorithms so that more people can see this tweet. After examining the tweets sent by this user, we found that he only sends tweets

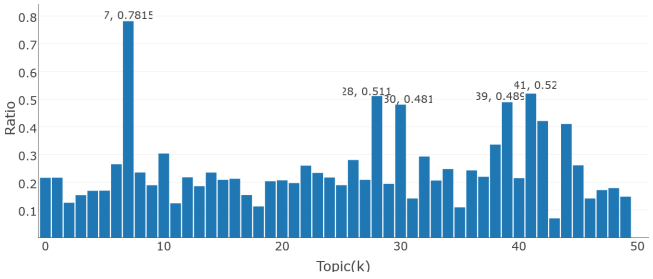

Figure 3. $\pi_{k,0}/\pi_{k,1}$ for different topics when K=50

Table IV
Top words and #tags for topics marked in Figure 3

k	Words	#tags
	Example Tweet	
7	jobs, manager, ca, health, medical, sales, assistant	#jobs, #getalljobs, #nursing, #healthcare, #finance, #retail, #tweetmyjobs
	Travel Occupational Therapist - Skilled... - **[HT]** , **[Location]**,**[URL]**. Get Nursing Jobs **#Nursing #jobs #job #GetAllJobs**	
28	local, play, game, map, news, restaurant, weather	#texas, #california, #iowa, #virginia, #kansas,, #florida, #west, #ohio
	Choose a hotel in **[Location]**, **#California** here! **[URL]**	
30	copy, love, vote, tweet, fact1, fact2, votes, voting	#slavenames, #freedownload, #oldpplnames, #facebooknames, #oldpplenames
	[HT] [HT] Presents **[book name]** - **#FreeDownload** -KP **[URL]**	
39	899fm, video, subscribe good, watch, link, remix	#free, #livehere, #teamheat, #soundcloud, #listen, #nowwatching, #teamfollowback
	please support my bro **[userid]** by watching his vid **[URL]**	
41	follow, back, followers, follower, newest, promo	#teamfollowback, #ff, #shoutout, #tfb, #follow, #mustfollow, #followback
	BREAKING DAWN Books - as low as $9.80 **[URL] [HT] #breakingdawn**	

that are exactly the same as the example tweet shown here, meaning that this user is a spammer. Similarly, topic 30 contains an abnormally high proportion of hashtags in its tweets, with content associated with free ebooks, illustrated by the hashtag "#freedownload". This is also a typical spam message that we see repeatedly in the dataset. One can validate similar conclusions on other topics. We omit the details for topic 44 because it contains offensive words.

3) Quantitative Results: To evaluate MTM quantitatively, we choose LDA and SLDA as the baseline models, and compare the numerical results of these models with that of MTM. This choice of baseline models was made to quantitatively examine the performance of MTM, which assigns a single topic to the document, against LDA, which models a document as a mixture of topics, in case of short-length text like tweets. To examine the assumption of treating words and hashtags separately, we select SLDA, which is structurally similar to MTM in terms of topic hierarchy, but does not differentiate between words and hashtags. We evaluate perplexity, a commonly used metric to evaluate topic models. The perplexity of a model for a test set of *M* documents is:

$$Perp(\mathcal{D}_{test}) = exp\left\{ \frac{-\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{11}$$

The perplexity of MTM can be calculated as:

$$Perp(\mathcal{D}_{test}) = \frac{1}{\sum_{u=1}^{U}\sum_{t=1}^{T} N_{ut}} \sum_{u=1}^{U}\sum_{t=1}^{T} log \Big(\sum_{k=1}^{K} \theta_{u,k} \Big(\sum_{n=1}^{N_w^{ut}} \pi_{k,1}\phi_{k,n} + \sum_{n=1}^{N_h^{ut}} \pi_{k,0}\eta_{k,n} \Big) \Big) \tag{12}$$

As per the definition, a lower perplexity score indicates a better predictive accuracy of the model. The perplexity of MTM was compared with that of LDA and SLDA using different values of *K*, ranging from 5 to 100. As shown in Figure 4, both SLDA and MTM, which assign topic at the document-level, have a consistently lower perplexity than LDA. In case of short-length social media data, the assumption of assigning a single topic to each sentence/document results in better fitting of the model and thus lower perplexity, which is why SLDA and MTM have better perplexity than LDA. Although theoretically, LDA should generalize SLDA, but in reality, a simpler model with fewer parameters such as SLDA, converges faster than more complicated models and often yields better results with fewer sampling iterations. As long as the simplified sentence-topic assumption agrees with the dataset, as in case of microblogs, it is expected that SLDA will outperform LDA. MTM has the advantage of SLDA, along with the ability to distinguish between text and hashtags. Hence, MTM outperforms both the models in the perplexity comparison.

B. MSTM

Similar to MTM, MSTM also has a set of hyper-parameters, which serve as prior knowledge for the model. The values of the hyper-parameters α, γ, and ϵ were kept the same as in case of MTM. MSTM has an additional prior λ on the topic-sentiment distribution ψ. The value of λ was determined experimentally as 5.

As described earlier, the prior sentiment knowledge about word polarity can be incorporated into MSTM by using unsymmetrical value for β. We fix the number of sentiments S as 2, using only positive and negative polarities, considering the fact that MSTM can identify non-polar neutral topics also, based on the value of ψ, so we do not need a third sentiment label S for neutral topics. For each word r that was present in the sentiment lexicon, the value of β was assigned as follows: if the polarity value $polarity(r)$ (pos or neg) agrees with the sentiment variable of the topic s, the β is chosen to be 0.09. Otherwise, β will be chosen as 0.01. For every other word r, whose prior sentiment knowledge was not known, a symmetric β_r = 0.05 was assigned.

$$\beta_{r_s} = \begin{cases} 0.09, & \text{if polarity(r) = s} \\ 0.01, & \text{if polarity(r)} \neq s \end{cases}$$

During the initialization step for each post (*u, t*), we determine the number of positive (*pos*) and negative (*neg*) words by comparing each word in the post against the sentiment lexicon. After this, the sentiment s_{ut} was assigned as follows:

$$s_{ut} = \begin{cases} 1, & \text{if } pos > neg \\ 0, & \text{if } pos < neg \\ random\{0,1\}, & \text{otherwise} \end{cases}$$

We evaluated the results for MSTM after running 800 Gibbs sampling iterations using different values of K ranging from 5 to 100.

1) Qualitative Results: We now present a qualitative analysis of the inferred topics, and explain how to determine the topic polarity using MSTM. To determine the polarity of a topic *k*, we use the value of the topic-sentiment distribution parameter ψ_k. This can be verified by examining the sentiment words and hashtags obtained for each topic.

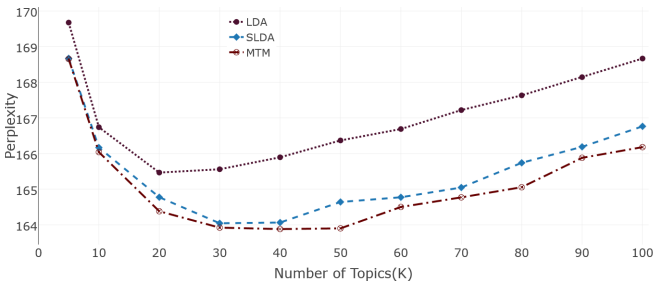

Figure 4. Perplexity comparison for MTM

Table V shows the top 10 positive and negative words and hashtags for a sample topic obtained using MSTM.

Table V
POSITIVE AND NEGATIVE WORDS AND HASHTAGS FOR A TOPIC OBTAINED USING MSTM

+ve Words	+ve #tags	-ve Words	-ve #tags
lol	#billboardawards	lol	#lmao
love	#thevoice	whoa	#billboards
song	#americanidol	lil	#loud
beyonce	#idol	shit	#garbage
video	#nowplaying	voice	#np
gaga	#1	video	#co
music	#beyonce	online	#boaw
sing	#oprah	internet	#justsaying
good	#winning	song	#fb
performance	#teamminaj	watch	#bored
$\psi_{k,1} = 0.972$		$\psi_{k,0} = 0.0278$	

As is evident from the results shown in Table V, the topic shown here is about music and awards, since it contains words such as "music", "video", "billboardawards", and hashtags such as "#billboardawards", "#thevoice", "#americanidol". The positive sentiment tokens for this topic contain words such as "lol", "love", "good", and hashtags such as "#winning", which clearly highlight the positive aspect of this topic. At the same time, this topic also has a set of negative words such as "shit", and hashtags such as "#garbage". The value of the positive sentiment distribution $\psi_{k,1}$ for this topic is much greater than the value of the negative sentiment distribution $\psi_{k,0}$, indicating that this topic is more likely to be a positive topic.

2) Quantitative Results: For the quantitative evaluation, we compare the numerical results of MSTM with the JST model and ASUM. In addition to perplexity, we also compare the sentiment accuracy of MSTM with the two baseline on a test set of 390 sentiment-labeled tweets. This sentiment accuracy indicates how well the prediction by the model aligns with the human judgment. A high sentiment accuracy of a model indicates how well the model can incorporate the sentiment in the generative process.

Perplexity Comparison As discussed earlier, a lower perplexity score of a model is an indicator of better predictive performance of a model. The perplexity of MSTM for a test set can be calculated using the following equation:

$$Perp(\mathcal{D}_{test}) = \frac{1}{\sum_{u=1}^{U}\sum_{t=1}^{T} N_{ut}} \sum_{u=1}^{U}\sum_{t=1}^{T} log\Big(\sum_{k=1}^{K} \sum_{s=1}^{S} \theta_{u,k}\psi_{k,s}\Big(\sum_{n=1}^{N_w^{ut}} \pi_{s,k,1}\phi_{k,s,n} + \sum_{n=1}^{N_h^{ut}} \pi_{s,k,0}\eta_{k,s,n}\Big)\Big) \quad (13)$$

We compare the perplexity of MSTM against JST and ASUM, keeping the number of sentiments S as 2 for all the models, with different values of K ranging from 5 to 100. The numerical comparison is illustrated in Figure 5. The JST model is comparable to LDA, where each word in the document has a different topic and sentiment assignment. ASUM is a variant of SLDA, which can be considered as a special case of ASUM with a single sentiment. Consequently, ASUM can be generalized using JST. As discussed earlier, ASUM should outperform JST in most cases since it is a much simper model that leads to better convergence. The results in Figure 5 once again validate the correctness of modeling social media data using a model that assigns single topic to a document. It can be seen that MSTM outperforms the baseline models in the perplexity comparison, which justifies the assumption of modeling microblog data using document-level topic and the separate treatment of hashtags and words.

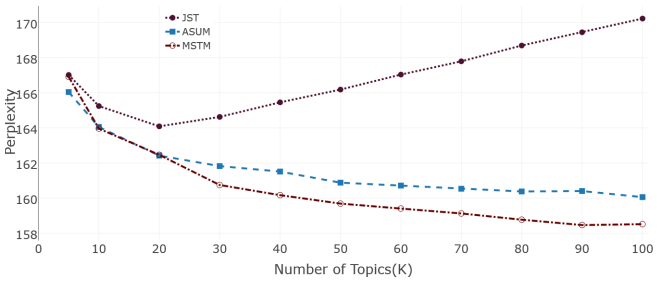

Figure 5. Perplexity comparison for MSTM

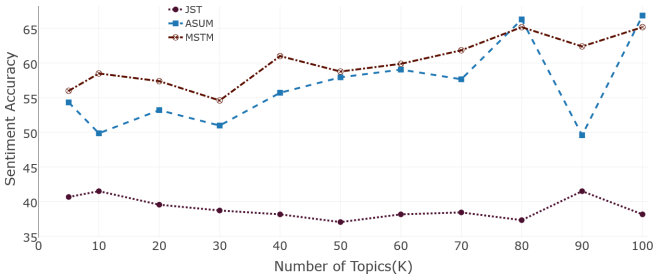

Figure 6. Sentiment accuracy comparison for MSTM

Sentiment Accuracy To quantitatively evaluate the sentiment modeling attribute of MSTM, we compare its sentiment accuracy against the baseline models. In MSTM, since the sentiment s_{ut} is a document-level parameter, the sentiment polarity of the test document can be determined using this value. After Gibbs sampling, we compare the inferred polarity s_{ut} of the 390 human-labeled tweets against their actual labels, and calculate the accuracy, which is the percentage of test tweets whose sentiment label was predicted correctly. This is illustrated in Figure 6. The sentiment accuracy of MSTM is marginally better than ASUM, but significantly better than JST model. Also, this accuracy increases as the number of topics grow, which is a result of better generalization of the dataset using a model with more parameters.

VI. Conclusion

In this work, we presented two probabilistic graphical models, namely MTM and MSTM, to discover latent topics in microblogs. In addition to topics, MSTM can also discover the sentiments associated with the topics. Both these models were based on the assumption that because of the short-length nature of microblogs, all the tokens in these social media posts belong to a single topic. Also, these models incorporate the special characteristic of these posts, i.e., the hashtags. To the best of our knowledge, no previous work incorporates these two properties of social media text. We evaluated both these models qualitatively and quantitatively, and found that these models outperformed the existing baseline techniques.

Acknowledgment

This work was supported in part by the Office of Naval Research (ONR) through MURI N000140811186 on adversarial reasoning, by the Department of Defense under the MINERVA initiative through the ONR N000141310835 on Multi-Source Assessment of State Stability, and by the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Department of Defense, or the U.S. Government.

References

[1] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 375–384.

[2] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 815–824.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

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

[5] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 1320–1326.

[6] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" *Icwsm*, vol. 11, pp. 538–541, 2011.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[8] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 115–120.

[9] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 889–892.

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

[11] 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*. AUAI Press, 2004, pp. 487–494.

[12] Z. Xu, R. Lu, L. Xiang, and Q. Yang, "Discovering user interest on twitter with a modified author-topic model," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2011, pp. 422–429.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[14] W. Wei, K. Joseph, W. Lo, and K. M. Carley, "A bayesian graphical model to discover latent events from twitter," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[15] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[16] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[17] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[18] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[19] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6. Citeseer, 2006, pp. 417–422.

[20] W. Wei, K. Joseph, H. Liu, and K. M. Carley, "The fragility of twitter social networks against suspended users," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015.